Wydziałowy Zakład Systemów Informacyjnych
Wydział Informatyki i Zarządzania
Politechnika Wrocławska

MODEL UŻYTKOWNIKA W INTERNETOWYCH SYSTEMACH WYSZUKIWANIA INFORMACJI

(praca doktorska)

Agnieszka Indyka – Piasecka

Promotor:
prof. dr hab. Czesław Daniłowicz

Wrocław 2006

Pracę dedykuję mojemu Mężowi

Składam podziękowania
Jackowi Surażskiemu
i Michałowi Rolskiemu
za udostępnienie i rozszerzenie,
na potrzeby niniejszej pracy,
komercyjnej wyszukiwarki Netoskop

Dziękuję Adamowi Folmertowi
za pomoc przy implementacji
środowiska do eksperymentów

1. Wstęp

Pierwsze systemy wyszukiwania informacji powstały na potrzeby środowiska naukowego i służyły wąskiemu gronu użytkowników, którzy mieli bardzo konkretnie sprecyzowane potrzeby informacyjne. Użytkownik rozszerzał swoją wiedzę przez sięgnięcie do zewnętrznych zasobów informacji. W systemach tamtego okresu, z racji na znaczne skomplikowanie procesu wyszukiwania, dostęp do zasobów informacji był możliwy tylko poprzez człowieka–pośrednika, obsługującego zasoby i opisującego potrzebę użytkownika w sposób adekwatny do wymagań systemu. Pośrednik pomagał i ułatwiał interakcję pomiędzy użytkownikiem a zasobami informacji.

W miarę upływu czasu, zwiększała się ilość gromadzonych informacji i pojawiła się tendencja do wyeliminowania pośrednika w wyszukiwaniu informacji, a udostępnienia informacji bezpośrednio użytkownikowi końcowemu. Tendencja ta znalazła realizację w sposobie korzystania z zasobów sieci WWW. Koniecznością stało się więc uproszczenie i wprowadzenie mechanizmów ułatwiających interakcję użytkownika z systemem wyszukiwania informacji. Teraz użytkownik końcowy stał się sam odpowiedzialny za poprawne sformułowanie swojej potrzeby informacyjnej.

W klasycznych systemach wyszukiwania informacji funkcjonuje założenie, że reprezentacja potrzeby informacyjnej, formułowana przez użytkownika lub pośrednika wspomagającego użytkownika w procesie wyszukiwania informacji, jest zgodna z rzeczywistymi potrzebami informacyjnymi użytkownika. Użytkownicy traktowani są jako eksperci w swojej dziedzinie, potrafiący precyzyjnie opisać swoje zainteresowania. Założenie to zostało wprost przeniesione na grunt internetowych systemów wyszukiwania informacji. W systemach tych nie jest uwzględniane zagadnienie trudności, jakie stwarza użytkownikowi samo wyrażenie i sprecyzowanie potrzeby informacyjnej, a następnie przedstawienia jej w postaci reprezentacji akceptowalnej przez system wyszukiwania informacji. Najczęściej największe problemy w sformułowaniu poprawnej reprezentacji potrzeby informacyjnej mają użytkownicy nie będący ekspertami w dziedzinie, w której dokonują wyszukiwania. A do tej grupy zaliczyć można wielu użytkowników internetowych systemów wyszukiwania informacji. Bo czyż użytkownik nie korzysta z systemu najczęściej wtedy, gdy brakuje mu informacji na określony temat i właśnie nie jest jeszcze ekspertem w danej dziedzinie? Z powyższych faktów można wysnuć wniosek, że wiedza, doświadczenie mają znaczący wpływ na ocenę informacji dostarczonej użytkownikowi przez system (np. na ocenę relewancji). Jednocześnie, wraz z czasem poświęconym na wyszukiwanie, zmienia się rozumienie przez użytkownika problemu informacyjnego oraz informacji, które są pomocne w rozwiązaniu tego problemu.

Podstawową reprezentacją potrzeby informacyjnej użytkownika w systemach wyszukiwania informacji jest pytanie zadane przez użytkownika. Opisane powyżej

problemy użytkownika z wyrażeniem swojej potrzeby w postaci pytania skłoniły badaczy do poszukiwania rozwiązań, które wspomagałyby użytkownika w wyrażeniu tej potrzeby. Było to inspiracją powstania metod automatycznego adaptowania wyszukiwania do dziedziny zainteresowania użytkownika. Do metod tych należą m.in. metody automatycznej modyfikacji pytania użytkownika. Modyfikacja pytania ma na celu doprowadzenie pytania do takiej postaci, która będzie lepiej odzwierciedlała rzeczywiste zainteresowania użytkownika niż pytanie przed modyfikacją. W odpowiedzi na pytanie zmodyfikowane użytkownik uzyskuje więc więcej interesujących go informacji.

Wyszukiwanie informacji należy więc traktować jako dialog, w którym użytkownik jest wspomagany przez system podczas tworzenia reprezentacji swojej potrzeby informacyjnej. W wyniku procesu tworzenia reprezentacji potrzeby informacyjnej użytkownika powstaje model użytkownika, opisujący jego zainteresowania w pewnej dziedzinie. Można stwierdzić, że wyszukiwanie informacji jest samo w sobie procesem modelowania użytkownika. Problemem jest jednak jakość modelu użytkownika. Przyjmuje się zazwyczaj, że użytkownik formułując wprost potrzebę informacyjną, przez wprowadzenie pytania lub podanie terminów opisujących zainteresowania, przedstawia swoje rzeczywiste zapotrzebowanie na informacje. Zazwyczaj jednak okazuje się, że taka reprezentacja potrzeby informacyjnej odbiega od wyobrażenia użytkownika o swojej potrzebie. Efektem tej rozbieżności są niezadowalające informacje, które otrzymuje użytkownik z systemu wyszukiwania informacji. Przyczynę takiej sytuacji upatruje się w niedoskonałości systemu wyszukiwania informacji, pomijając fakt, że istotny wpływ może mieć również nieadekwatne sformułowanie potrzeby informacyjnej przez samego użytkownika, wynikające z jego niewiedzy.

Celem pracy było opracowanie modelu profilu użytkownika reprezentującego różnorodne zainteresowania użytkownika korzystającego z internetowego systemu wyszukiwania informacji oraz procedur automatycznego tworzenia, modyfikacji i wykorzystania profilu na podstawie pytań kierowanych przez użytkownika do systemu oraz dokumentów zwracanych w odpowiedzi systemu i ocenionych przez użytkownika.

W pracy przedstawiono nową koncepcję profilu, w którym różne zainteresowania użytkownika reprezentowane są w różnych *subprofilach* – częściach składowych struktury złożonego profilu. Każde pytanie użytkownika powiązane jest tylko z jednym subprofilem, który zawiera informacje o konkretnej dziedzinie zainteresowań użytkownika. Użytkownik formułując pytanie posługuje się *swoim własnym słownictwem*. Nie zawsze musi być ono prawidłowe, w sensie powszechnie stosowanych terminów, oraz może opisywać dobrze daną dziedzinę tylko i wyłącznie

z subiektywnego punktu widzenia danego użytkownika. Natomiast zadaniem subprofilu powiązanego z konkretnym pytaniem jest opis *tej samej dziedziny tematycznej*, ale z zastosowaniem słownictwa powszechnie stosowanego w tej dziedzinie w sieci. W profilu jednoznacznie łączy się *subiektywnie* sformułowane pytanie użytkownika z *obiektywnym*, automatycznie utworzonym, opisem w subprofilu (utworzonym na podstawie analizy relewantnych dokumentów odpowiedzi systemu wyszukiwawczego). Można więc powiedzieć, że zaproponowany w pracy *profil użytkownika* jest strukturą opisującą *translację* pomiędzy terminologią wykorzystywaną przez użytkownika w pytaniu, a słownictwem powszechnie stosowanym w danej dziedzinie zainteresowań użytkownika.

Praca składa się z trzech części. Rozdziały 2–3 wprowadzają w tematykę systemów wyszukiwania informacji, prezentując rozwój systemów wyszukiwania informacji od klasycznych systemów do systemów internetowych, charakteryzujących się ogromnymi kolekcjami i dużą częstotliwością zmian w kolekcjach dokumentów.

Badania w dziedzinie wyszukiwania informacji w sieci WWW są kolejnym krokiem w rozwoju metod i technologii wyszukiwania informacji w stosunku do tych stosowanych w klasycznych systemach wyszukiwania informacji. Dziedzictwem klasycznych systemów wyszukiwania informacji są dogłębnie opracowane i zweryfikowane metody wyszukiwania dokumentów tekstowych. Natomiast nowa rzeczywistość jaka pojawiła się wraz z web'owymi kolekcjami dokumentów skłania środowisko naukowe do proponowania różnorodnych modyfikacji klasycznych metod wyszukiwania. Sugeruje się uwzględnianie w procesie wyszukiwania nowych informacji, jakie mogą dostarczyć o samych dokumentach (stronach internetowych), np. źródło pochodzenia strony, częstość aktualizacji, częstość cytowania danej strony, jakość, popularność, czy użyteczność, typ i format strony (tekst, HTML, PDF, postscript, obrazy, dźwięki, wideo), oczekując poprawy wyników wyszukiwania i wzrostu zadowolenia użytkownika z odpowiedzi uzyskanej od systemu. W ośrodkach naukowych prowadzących prace badawcze w zakresie internetowych systemów wyszukiwawczych w celu wykorzystania tych nowych informacji o dokumentach proponuje się rozwiązania wykorzystujące metody sztucznej inteligencji, metody statystyczne, psychologię, czy nauki kognitywne. Autor pracy uważa jednak, że najważniejszym elementem dokumentu jest sam tekst i, świadomie skupiając się głównie na jego analizie, sądzi, że istotną poprawę wyników wyszukiwania, co pokazały wyniki pracy, można uzyskać na drodze dostosowania klasycznych metod wyszukiwania do specyfiki kolekcji internetowych oraz poprzez stosowanie bardziej precyzyjnych modeli użytkowników.

Zagadnienia związane z problematyką modelowania użytkownika zostały przeanalizowane w Rozdziale 3. W pierwszej części rozdziału, autor pracy przedstawił szerokie spektrum metod konstruowania jak i zastosowania modeli użytkowników

w systemach informacyjnych. Do najważniejszych dziedzin zastosowania modelowania użytkownika zaliczono: wyszukiwanie informacji, wspomaganie nauczania oraz wspomaganie podejmowania decyzji. Druga część Rozdziału 3 poświęcona jest problematyce modelowania użytkownika w systemach wyszukiwania informacji. Zagadnienie to zostało dogłębnie potraktowane osobno, ponieważ tezy pracy dotyczą właśnie modelowania użytkownika w internetowych systemach wyszukiwania informacji.

W drugiej części pracy, czyli w Rozdziale 4, opisano i przeanalizowano własny model profilu użytkownika oraz metody tworzenia, modyfikacji i wykorzystania profilu do personalizacji zapytania użytkownika.

Różnorodność zarówno zainteresowań, jak i stawianych pytań jest własnością wyszukiwania w systemach internetowych. Stąd pojawia się potrzeba takich modeli profilu użytkownika, które będą skutecznie reprezentowały wszystkie tematy zainteresowań ujawniane i wykorzystywane w trakcie pracy użytkownika z internetowym systemem wyszukiwania informacji. Powyższe potrzeby stały się inspiracją do zaproponowania odmiennej koncepcji profilu, w której zainteresowania użytkownika reprezentowane są w oddzielnych częściach złożonego profilu – tj. subprofilach. Zaproponowany profil użytkownika dostosowuje się do różnorodnych zainteresowań użytkownika, wyrażonych przez pytania dotyczące różnorodnej tematyki. Koncepcję nowego profilu użytkownika opisano w podrozdziale 4.1, będącym wprowadzeniem przed formalną definicją profilu użytkownika.

Szczególnego podkreślenia wymaga fakt, że koncepcja zaproponowanego w pracy profilu została tak sformułowana, że uwzględnia zarówno zmieniające się i różnorodne potrzeby użytkownika, jak również dynamikę kolekcji dokumentów (co ma miejsce dla kolekcji WWW). Dynamika kolekcji jest uwzględniona poprzez konstruowanie subprofilu na podstawie dokumentów relewantnych odpowiedzi z na bieżąco zmieniającej się web'owej kolekcji dokumentów. Subprofil jest „dostrajany" wraz z każdym wyszukiwaniem odpowiednio do dokumentów z kolekcji. Stąd zaproponowany model profilu jest szczególnie przystosowany dla internetowych systemów wyszukiwania informacji.

W ramach pracy opracowano również nowe kryterium wyboru terminów kluczowych spośród słownictwa stosowanego w dziedzinie zainteresowań użytkownika. Zastosowanie dla kolekcji WWW klasycznych metod wyznaczenia terminów kluczowych, tj. na podstawie progów wyrażanych przez raz ustalone i stałe wartości liczbowe, nie daje oczekiwanego zbioru terminów znaczących. W pracy zaproponowano dynamiczne progi, które przyjmują postać wielostopniowego kryterium, a ich wartości nie są stałe, ale wyznaczane na podstawie funkcji uwzględniających dynamikę zmian wag terminów w kolekcji web'owej.

Trzecią część pracy tworzy Rozdział 5, który zawiera opis eksperymentalnej weryfikacji zaproponowanego profilu. W koncepcji wykorzystania zaproponowanego

profilu użytkownika założono, że korzystając z internetowego systemu wyszukiwania informacji umożliwiającego personalizację wyszukiwania dzięki zastosowaniu profilu, użytkownik po pewnym czasie współpracy z systemem będzie otrzymywał w odpowiedzi na pytania z określonej dziedziny zainteresowań coraz więcej dokumentów na interesujący go temat oraz odpowiedź będzie zawierała coraz mniej dokumentów w ogóle. Przeprowadzono eksperymenty iteracyjnej modyfikacji pytania, które potwierdziły zgodność zaproponowanego profilu z intuicją wykorzystania profilu użytkownika w systemie wyszukiwania informacji.

Pewnym ograniczeniem przeprowadzonych eksperymentów było wykonywanie wyszukiwań w stałej, politematycznej kolekcji dokumentów WWW. W kolekcji znajdowały się dokumenty z sieci WWW, zebrane i poindeksowane na przestrzeni pewnego ograniczonego przedziału czasowego. Dla tak utworzonej kolekcji może pojawić się pytanie, czy na podstawie wyników przeprowadzonych eksperymentów można wysnuć wnioski dla dynamicznej kolekcji dokumentów WWW. Dyskusja przeprowadzona w podsumowaniu eksperymentów dogłębnie analizuje powyższe wątpliwości.

Rozdział 6 to podsumowanie wyników pracy oraz propozycje i możliwości dalszego rozwoju zaproponowanego profilu użytkownika.

2. Systemy wyszukiwania informacji

2.1. Tradycyjne metody wyszukiwania informacji

Inspiracją rozwoju systemów wyszukiwania informacji był od lat 40–tych szybki wzrost liczby publikowanych prac naukowych. Mająca miejsce obecnie "rewolucja informacyjna", a szczególnie bardzo szybki rozwój Internetu, wpłynęły na rozszerzenie obszarów, których dotyczy wyszukiwanie informacji. Dzisiejsze systemy obejmują nie tylko wyszukiwania literatury naukowej, ale również wyszukiwania różnych typów informacji, jak np. dane multimedialne (obrazy, dźwięki, mowa, animacje, wideo) (Daniłowicz, 2000). W dalszej części pracy, rozważania na temat systemów wyszukiwania informacji dotyczyć będą wyszukiwania dokumentów tekstowych.

W systemie wyszukiwania informacji można wyróżnić trzy podstawowe elementy (van Rijsbergen, 1979, str. 4):

- użytkownika posiadającego pewną potrzebę informacyjną,
- kolekcję dokumentów, w której prowadzone są wyszukiwania oraz
- odpowiedź systemu na potrzebę informacyjną użytkownika.

Głównym celem działania systemu wyszukiwania informacji jest znalezienie na pytanie użytkownika odpowiedzi wśród dokumentów kolekcji. Aby cel ten został osiągnięty, niezbędne jest zastosowanie takich metod i technik, które pozwolą na określenie stopnia zgodności, czyli podobieństwa tematyki dokumentu kolekcji z zadanym przez użytkownika pytaniem. Do powszechnie stosowanych metod należy indeksowanie dokumentów oraz zastosowanie reprezentacji dokumentów i pytań w postaci wektorów przestrzeni wielowymiarowej do określenia zgodności pytania i dokumentu.

2.1.1. Automatyczne indeksowanie dokumentów

W klasycznym modelu wyszukiwania informacji, każdy dokument opisany jest przez zbiór słów nazywanych *terminami indeksowymi*, *terminami dyskryminacyjnymi* lub *słowami kluczowymi*. Terminy pochodzące z dokumentu nie mają jednakowej wartości w reprezentowaniu najważniejszych pojęć występujących w dokumencie. Niezbędne jest więc rozróżnienie pomiędzy terminami istotnymi i nieistotnymi dla tematyki dokumentu. Wyznaczenie istotności danego terminu indeksowego dla opisu treści dokumentu jest procesem, w którym uwzględniana jest częstość występowania terminu w dokumencie oraz w całej kolekcji dokumentów. Znaczenie terminu dla opisu dokumentu reprezentowane jest przez wagę terminu indeksowego (Salton i McGill, 1983), (Baeza–Yates i Ribeiro–Neto, 1999). Im waga terminu jest wyższa tym jest on

bardziej istotny. Każdemu terminowi w dokumencie przypisana jest waga. Waga terminów wykorzystywana jest do wyznaczenia *terminów indeksowych*. Terminy indeksowe, dobrze opisując treść danego dokumentu, umożliwiają w procesie wyszukiwania selekcję tego dokumentu spośród innych dokumentów kolekcji, jeśli przekazane zostaną w pytaniu do systemu.

Procesy analizy dokumentów tekstowych w celu utworzenia ich reprezentacji, łatwej do przetwarzania w komputerowych systemach wyszukiwania informacji, są obszarem intensywnych badań od lat 60–tych, od momentu zaproponowania przez Luhna idei reprezentacji treści tekstu przez słowa występujące z odpowiednią częstością w tym tekście. Proces tworzenia wewnętrznej reprezentacji dokumentu nazywany jest *procesem indeksowania*.

W początkowych latach rozwoju wyszukiwania informacji proces indeksowania przeprowadzany był przez indekserów. Jednak wykonywane przez nich indeksowanie było subiektywne, obciążone znajomością zbioru terminów indeksowych, z których użytkownik będzie korzystał wyszukując określony dokument. Procedura indeksowania może być bardziej efektywna, jeśli zastosowane zostaną obiektywne kryteria wyboru terminów indeksowych podczas analizy wszystkich dokumentów kolekcji. Głównym jednak problemem indeksowania ręcznego był czas potrzebny indekserom na poindeksowanie kolekcji dokumentów. Podjęto próby automatycznego indeksowania dokumentów, a przeprowadzone doświadczenia pokazały, że automatyczne indeksowanie dokumentów daje tak samo dobre wyniki wyszukiwania jak ręczne indeksowanie dokumentów (Salton, 1971).

Techniki automatycznego indeksowania można podzielić na dwie podstawowe kategorie, które różnią się podejściem do procesu selekcji terminów indeksowych. Są to:

- *podejście statystyczne*, wykorzystujące informacje o częstości występowania słów oraz
- *podejście lingwistyczne*, wykorzystujące relacje syntaktyczne i semantyczne słów dla pewnego kontekstu.

W dalszej części pracy skupimy się na statystycznych technikach indeksowania.

Selekcja terminów indeksowych

Proces wyznaczenia reprezentacji dokumentu realizowany jest w kilku krokach (van Rijsbergen, 1979):

- usunięcie słów potocznych (*stop–lista*) z tekstu dokumentu,
- przeprowadzenie *stemmingu*,
- usunięcie powtarzających się rdzeni słów.

Usunięcie słów *stop–listy* ma na celu usunięcie słów o dużej częstości występowania w tekście, które nie wnoszą istotnych informacji o treści dokumentu i dlatego nie mają

wartości jako terminy indeksowe. Są to słowa używane potocznie (ang. *common words*) jak np.: and, the. Korzyścią płynącą z przeprowadzenia tego procesu jest zmniejszenie objętości tekstu.

Podczas wyszukiwania informacji w sieci Internet nie jest stosowana stop–lista. Jedną z przyczyn jest prowadzenie wyszukiwania w Internecie przez użytkowników dla bardzo różnorodnych zakresów tematycznych. Dla takiego szerokiego spektrum zagadnień nie jest możliwe określenie jednego zbioru terminów niedozwolonych do używania w pytaniach, czyli terminów stop–listy, dlatego też wszystkie terminy traktowane są jako terminy jednakowo znaczące dla treści dokumentu.

Ograniczenie liczby terminów reprezentujących dokument uzyskiwane może być nie przez wykorzystanie stop–listy, ale dzięki zastosowaniu odpowiedniej funkcji progowej. Funkcja ta wyznacza terminy istotne dla reprezentowania treści dokumentu. Jeśli waga terminu jest większa od progu τ, to termin uznawany jest za termin dobrze reprezentujący treść dokumentu:

$$f_{\tau}(w) = \begin{cases} w, & w \geq \tau \\ 0, & w < \tau \end{cases}$$

Za wartość w przyjmuje się wagę terminu wyznaczoną na podstawie jednego z klasycznych schematów przypisania wagi terminom w wyszukiwaniu informacji (Salton, 1988), (Rao, 1988), (Rao, 1988a).

Następnym krokiem analizy tekstu jest usunięcie końcówek dla pozostałych słów dokumentu wejściowego w celu określenia rdzeni słów. Najwięcej prac z tego zakresu powstało dla języka angielskiego i dlatego też proces ten powszechnie nazywany jest *stemmingiem.* Polega on na przedstawieniu słów powiązanych ze sobą syntaktycznie za pomocą jednego, wspólnego rdzenia, np. słowa *retrieval*, *retrieving*, *retrieves*, *retrieve* reprezentowane mogą być przez wspólny rdzeń *retriev*. Stemming, czyli metoda analizy słów w celu wyznaczenia rdzenia słowa jest dobrze rozwinięta i stosowana dla angielskojęzycznych systemów wyszukiwania informacji. Istotą tej metody jest zidentyfikowanie końcówek i przedrostków w słowie, a następnie odcięcie zidentyfikowanych przedrostka czy końcówki i przekazanie w wyniku – rdzenia analizowanego słowa. Zbiór występujących w języku końcówek i przedrostków jest zbiorem skończonym. Określenie, czy fragment słowa jest końcówką lub przedrostkiem wykonuje się przez porównanie tego elementu z ręcznie ustaloną listą końcówek i przedrostków dla danego języka. W przeciwieństwie do zastosowania analogicznej metody dla języka polskiego, dla języka angielskiego nie jest konieczne przeprowadzenie analizy lingwistycznej, gdyż rdzenie słów nie podlegają odmianie.

W polskich internetowych systemach wyszukiwania informacji stemming nie jest stosowany. Przyczyną są nierozwiązane jeszcze problemy natury lingwistycznej.

Zbiór słów będących wynikiem procesu stemmingu jest zbiorem terminów indeksowych danej kolekcji dokumentów. Dodatkowo może być zastosowany zbiór synonimów, który pozwoli na przypisanie słów będących w relacji synonimii do jednej klasy pojęć (Faloutsos i Oard,1995).

W wyniku powyższych kroków procesu tworzenia reprezentacji dokumentu, opis pojedynczego dokumentu może być przedstawiony w postaci zbioru terminów indeksowych zawartych w tym dokumencie. Tradycyjnie opis dokumentu przedstawiany jest w postaci wektora przestrzeni *n*–wymiarowej, (Salton i McGill, 1983):

$$d=(d_1, d_2, ...,d_n)$$

gdzie d_i – waga terminu t_i w dokumencie d, i=1, 2, …, n,

n – liczba dozwolonych terminów indeksowych.

Każda pozycja wektora dokumentu odpowiada danemu terminowi indeksowemu z kolekcji dokumentów. Jeżeli termin indeksowy nie występuje w dokumencie, wtedy odpowiedniej pozycji wektora dokumentu przypisywana jest wartość 0. Jeżeli termin występuje w dokumencie, to odpowiedniej pozycji wektora dokumentu przypisywana jest liczba 1 (dla binarnego wektora dokumentu) lub inna liczba dodatnia (waga terminu), która odzwierciedla ważność terminu w dokumencie.

Metody nadawania wag terminom indeksowym

Prace nad indeksowaniem dokumentów doprowadziły do opracowania kilku schematów automatycznego ważenia terminów, wykorzystujących informacje o terminach zarówno z samego dokumentu, jak i z całej kolekcji dokumentów.

Schemat ważenia terminów uwzględniający *częstość występowania terminu* w treści dokumentu (ang. *term frequency – tf*) oparty jest na założeniu, że słowa, frazy lub grupy słów, które występują w tekście dokumentu z odpowiednią częstością mają istotne znaczenie dla treści dokumentu. Waga terminu t_i w dokumencie d, w_i w tym schemacie ważenia wyznaczana jest na podstawie częstości występowania terminu t_i w dokumencie d (van Rijsbergen, 1979):

$$w_i = tf_i$$

Luhn w swoich pracach stwierdza, że najwyższa waga powinna być przypisana terminom ze średnią częstością występowania w dokumencie (Luhn, 1958). Terminy występujące bardzo często są słowami powszechnie używanymi – potocznymi, natomiast słowa o niskiej częstości występowania są słowami rzadkimi. Ani pierwsze, ani drugie nie wnoszą istotnych informacji o treści dokumentu.

Przeprowadzane eksperymenty pokazały, że również rozkład terminów indeksowych w całej kolekcji dokumentów wpływa na ważność terminu w reprezentowaniu treści dokumentu (Salton i McGill, 1983). Schemat ważenia terminów *tf* nie pozwala na

rozróżnienie pomiędzy terminami, które występują w każdym dokumencie kolekcji a takimi, które występują tylko w kilku dokumentach.

Istotą procesu selekcji terminów indeksowych jest wyróżnienie z dokumentu tych słów, które będą dobrze opisywały treść dokumentu. Jednak ze względu na zachowanie efektywności procesu wyszukiwania[1] istotne jest również, aby dzięki tym terminom indeksowym dokumenty kolekcji były dobrze rozróżniane ze względu na pytanie tzn., aby odpowiedź systemu zawierała jak najwięcej dokumentów relewantnych w stosunku do wszystkich dokumentów odpowiedzi. Salton, Yang i Yu wprowadzili pojęcie *wartości dyskryminacyjnej terminu* (ang. *term discrimination value*) (Salton i inni, 1975). Analiza wartości dyskryminacyjnej umożliwia wyznaczenia ważności terminu ze względu na to, jak dobrze dany termin wyróżnia dokument spośród wszystkich innych dokumentów kolekcji. Wartość dyskryminacyjna (ang. *discrimination value*) terminu t_i, oznaczana symbolem dv_i, jest miarą zmiany gęstości przestrzeni dokumentów[2], gdy termin t_i zostanie przypisany do dokumentów kolekcji oraz gęstości po usunięciu tego terminu z dokumentów kolekcji.

Analiza rozkładu terminów indeksowych w całej kolekcji dokumentów pozwala na rozróżnienie pomiędzy ważnością terminów występujących w każdym dokumencie, a ważnością takich, które występują tylko w kilku dokumentach. Termin t_i jest dobrym dyskryminatorem, jeśli usunięcie tego terminu z kolekcji zwiększa gęstość zbioru dokumentów, czyli zmniejsza odległość pomiędzy dokumentami (ang. *space compression*), a tym samym rośnie średnie podobieństwo dokumentów kolekcji. Termin t_i jest uznawany za zły dyskryminator, jeśli usunięcie tego terminu z kolekcji zmniejsza gęstość przestrzeni dokumentów, czyli zwiększa odległość pomiędzy dokumentami (ang. *space separation*), a tym samym maleje średnie podobieństwo dokumentów kolekcji (Dąbrowski i Laus–Mączyńska, 1978). Waga terminu t_i w dokumencie d wyznacza jest na podstawie następującego wzoru, uwzględniającego częstość występowania terminu w dokumencie oraz jego wartość dyskryminacyjną:

$$w_i = tf_i * dv_i$$

Dzięki określeniu wartości dyskryminacyjnej terminu, możliwe jest wybranie z kolekcji dokumentów tylko tych terminów, które będą dobrze wyróżniały dokumenty. Wartość dyskryminacyjna terminu promuje terminy występujące w niewielkiej liczbie dokumentów kolekcji i tym terminom nadawana jest wysoka wartość dyskryminacyjna.

Salton, Yang i Yu na podstawie przeprowadzonych eksperymentów pokazali również związek pomiędzy wartością dyskryminacyjną terminu, a *częstością*

[1] Czynnikami wpływającymi na efektywność wyszukiwania jest kompletność i dokładność. Terminy pojawiające się z dużą częstością w dokumencie zwiększają kompletności wyszukiwania natomiast, jeśli takie terminy występują z porównywalną częstością we wszystkich dokumentach kolekcji – obniżają dokładność wyszukiwania.

[2] Lub inaczej różnicą odległości elementów przestrzeni dokumentów.

dokumentową terminu (ang. *document frequency – df*) (Salton i inni, 1975). Częstość dokumentowa to liczba dokumentów kolekcji, w których wystąpił termin. Częstość dokumentowa terminu może być traktowana jako kryterium selekcji terminów indeksowych będących dobrymi dyskryminatorami. Terminy o niskiej częstości dokumentowej, czyli te, które pojawiają się w niewielu dokumentach kolekcji, dobrze wyróżniają dokument spośród pozostałych dokumentów kolekcji. Terminy te uznawane są za dobre dyskryminatory treści dokumentu.

Przy tak przyjętym kryterium, ważność terminu rośnie wraz ze wzrostem częstości występowania tego terminu indeksowego t_i w dokumencie, ale maleje wraz ze wzrostem liczby dokumentów kolekcji, w których ten termin t_i występuje. Własności te uwzględnia stosowana w literaturze miara określana nazwą odwrotnej częstości dokumentowej (ang. *inverse document frequency – idf*).

$$idf_i = \log \frac{ND}{nd_i}$$

ND – całkowita liczba dokumentów w kolekcji,

nd_i – liczba dokumentów kolekcji zawierających termin t_i.

Waga terminu w schemacie *tf–idf* wyznaczana jest na podstawie następującego wzoru:

$$w_i = tf_i * idf_i .$$

Badania Saltona i Buckleya mające na celu określenie najbardziej efektywnego schematu ważenia terminów pokazały, że najlepszą miarą ważności terminu, czyli wagą terminu, jest waga uwzględniająca zarówno częstość występowania terminu – *tf*, jak i liczbę dokumentów kolekcji, w których występuje termin t_i – nd_i, co uwzględniane jest przez odwrotną częstość dokumentową terminu (*idf*).

$$w_i = \frac{tf_i \log \frac{ND}{nd_i}}{\sqrt{\sum_j \left(tf_j \log \frac{ND}{nd_j} \right)^2}},$$

gdzie:

tf_i – częstość terminu t_i w dokumencie d,

ND – całkowita liczba dokumentów w kolekcji,

nd_i – liczba dokumentów kolekcji zawierających termin t_i,

j – kolejne terminy należące do dokumentu d.

Tak określona waga terminu indeksowego rośnie wraz ze wzrostem częstości występowania tego terminu indeksowego w dokumencie, a maleje wraz ze wzrostem liczby dokumentów kolekcji, w których ten termin występuje. Waga terminu jest często normalizowana (Salton i Buckley, 1988). Również eksperymenty przeprowadzone przez Qiu potwierdziły, że spośród zaproponowanych przez Saltona i Buckleya sposobów obliczania wagi terminu powyższy schemat dobrze wyznacza terminy

dyskryminacyjne i jest powszechnie uznany za dobrą metodę wyznaczania terminów indeksowych (Qiu, 1996).

Opierając się na analizie i weryfikacji metod ważenia terminów zaprezentowanych m.in. w pracach (Salton i Buckley, 1988), (Qui, 1996), do określania ważności terminu wykorzystano w niniejszej pracy m.in. wagę terminu, która uwzględnia częstość występowania terminu w dokumencie, liczbę dokumentów kolekcji, w których występuje ten termin, jak i normalizację wagi terminu. Wykorzystana w pracy funkcja przedstawiona jest w Rozdziale 4.4.

2.1.2. Pytanie użytkownika i odpowiedź systemu wyszukiwania informacji

Pytanie użytkownika wyraża potrzebę informacyjną użytkownika. Pytanie jest kierowane do systemu wyszukiwania informacji. W zależności od realizacji systemu, użytkownik może zadać pytanie w postaci wyrażenia boolowskiego, w którym terminy pytania połączone są operatorami boolowskimi (AND, OR, NOT), lub w języku naturalnym.

Po zadaniu pytania do systemu wyszukiwania informacji jest ono poddawane procesowi podobnemu jak proces indeksowania dokumentów, aby reprezentacja pytania mogła być przetwarzana przez system informatyczny. Pytanie może być poddane analizie tak, aby wyróżnić i otrzymać jego istotne elementy – terminy pytania. Może to być wykonane przez usunięcie słów należących do stop–listy oraz przez wykonanie stemmingu. Wynikiem jest reprezentacja pytania w postaci wektora przestrzeni n – wymiarowej.

$$q = (q_1, q_2, ..., q_n),$$

Wystąpienie terminu w pytaniu jest oznaczone na odpowiedniej pozycji q_i wektora przez wagę tego terminu.

Reprezentacja powyższa pozwala na porównanie dokumentów ze zgromadzonej kolekcji z pytaniem użytkownika, a w efekcie na określenie relewancji dokumentów z kolekcji do pytania oraz selekcję tych dokumentów. Wyszukane dokumenty tekstowe mogą być oceniane, co prowadzi do modyfikacji: pytania, potrzeby informacyjnej lub rzadziej reprezentacji dokumentu (Belkin i Croft, 1992). Na przestrzeni ostatnich 30 lat opracowane zostały trzy podstawowe modele wyszukiwania informacji – *boolowski*, *wektorowy* oraz *probabilistyczny* (van Rijsbergen, 1979), z których dwa ostatnie uwzględniają stopień relewancji dokumentu i pytania w procesie generowania odpowiedzi. Modele te różnią się procesem wyznaczania reprezentacji dokumentów, określania relewancji dokumentu w stosunku do zadanego pytania oraz procesem modyfikacji pytania.

Proces porównywania dokumentu z pytaniem, czyli określenie relewancji dokumentu do pytania, jest jednym z istotnych elementów różniących te modele.

Odpowiedź systemu, generowana na podstawie jednej z miar podobieństwa dokumentu i pytania, jest zbiorem dokumentów dokładnie pasujących do pytania (model boolowski) (Salton i McGill, 1983a) lub jest zbiorem dokumentów w postaci rankingu według malejącej miary relewancji dokumentów do pytania (model wektorowy, probabilistyczny). Wartość wspomnianej miary podobieństwa jest otrzymywana w wyniku porównywania deskryptorów z dokumentów oraz terminów pytania, uwzględniając również rozkład wszystkich terminów indeksowych w kolekcji. Opisane w literaturze eksperymenty potwierdziły, że zaproponowane modele, wykorzystane w systemach wyszukiwania informacji, spełniają wymagania dostarczenia użytkownikowi dokumentów zaspakajających jego potrzebę uzyskania informacji (Salton i Buckley, 1988). Doświadczenia przeprowadzali m.in. Salton i Buckley dla kolekcji dokumentów tworzących jednorodne, o kontrolowanym słownictwie bazy tekstów, takie jak Inspec, kolekcja CACM, MED, NLP, czy Yang, Maglaughlin, Meho i Sumner dla kolekcji dokumentów konferencji TREC (Yang i inni, 1999).

Tradycyjne systemy wyszukiwania informacji wykorzystują terminy indeksowe do indeksowania i wyszukiwania dokumentów. Wyszukiwanie oparte na tej idei jest proste i sprawdziło się w wielu zastosowaniach testowych i komercyjnych. Podstawowym założeniem wyszukiwania wykorzystującego terminy indeksowe jest wyrażenie przez ustalony zbiór terminów indeksowych zarówno semantyki dokumentu, jak i potrzeby informacyjnej użytkownika. Proces porównywania dokumentu z kolekcji i pytania odbywa się poprzez badanie podobieństwa n – wymiarowych wektorów reprezentujących dokumenty i pytanie.

2.1.3. Miary efektywności wyszukiwania

W klasycznych systemach wyszukiwania informacji do oceny efektywności wyszukiwania stosowane są dwie miary: *dokładność* (ang. *precision*) oraz *kompletność* (ang. *recall*). Dokładność określa procent dokumentów w odpowiedzi systemu wyszukiwania informacji, które są relewantne. Natomiast kompletność określa procent dokumentów relewantnych w odpowiedzi spośród wszystkich dokumentów relewantnych zgromadzonych w kolekcji (wyszukanych i niewyszukanych). Oznaczmy przez:

Rel – zbiór wszystkich dokumentów relewantnych,
$\neg Rel$ – zbiór dokumentów nierelewantnych,
$Wysz$ – zbiór dokumentów wyszukanych,
$\neg Wysz$ – zbiór dokumentów niewyszukanych,
Kol – zbiór wszystkich dokumentów w systemie wyszukiwania informacji (kolekcja dokumentów).

Zależności pomiędzy wymienionymi zbiorami można opisać następującą tabelą (van Rijsbergen, 1979):

	Relewantne	Nierelewantne	
Wyszukane	$Rel \cap Wysz$	$\neg Rel \cap Wysz$	$Wysz$
Niewyszukane	$Rel \cap \neg Wysz$	$\neg Rel \cap \neg Wysz$	$\neg Wysz$
	Rel	$\neg Rel$	Kol

Wykorzystując przyjęte oznaczenia, dokładność *Dokl* i kompletność *Kom* formalnie zapisujemy w postaci następujących wzorów:

$$Dokl = \frac{|Rel \cap Wysz|}{|Wysz|} \qquad Kom = \frac{|Rel \cap Wysz|}{|Rel|},$$

gdzie | | oznacza liczność danego zbioru dokumentów.

Dla wyszukiwań przeprowadzanych w sieci WWW nie jest możliwe określenie statycznego zbioru dokumentów, będącego kolekcją dokumentów, których dotyczy wyszukiwanie (Seo i Zhang, 2000), co jest podstawową cechą klasycznych systemów wyszukiwania informacji. Również ze względu na rozmiary zbiorów dokumentów gromadzonych w internetowych systemach wyszukiwania informacji istnieje problem z określeniem liczby wszystkich dokumentów relewantnych dla danego pytania. Skutkiem jest brak możliwości obliczenia, klasycznie rozumianej, kompletności odpowiedzi. Jako miarę efektywności wyszukiwania proponuje się więc stosowanie zmodyfikowanej dokładności i zmodyfikowanej kompletności. Jest to *dokładność* oraz *kompletność obcięta* do początkowych *m* dokumentów odpowiedzi (gdzie *m* = 10, 20, 30, ..., 100) (Rao, 1988a). W przypadku kompletności, modyfikacja z zastosowaniem obcięcia liczby dokumentów odpowiedzi do pewnej stałej liczby dla zasobów sieci WWW nie likwiduje jednak problemu związanego z koniecznością wyznaczenia wszystkich dokumentów relewantnych dla pytania w kolekcji. Z drugiej strony wiadomo, że dla statycznych kolekcji testowych, jak np. TREC, TIPSTER dokładność określa lepiej niż kompletność poprawę wyszukiwania oraz informuje o tym, czy użytkownik uzyskuje w odpowiedzi z systemu wyszukiwania informacji dokumenty relewantne (Callan i Croft, 1993). Uzasadnione jest więc, że w niniejszej pracy przyjęto dokładność odpowiedzi za zadowalającą miarę oceny efektywności wyszukiwania z wykorzystaniem profilu użytkownika w sieci WWW. Przyjęto miarę *dokładności obciętej*, obliczaną dla pierwszych 10, 20 oraz 30 dokumentów odpowiedzi. Wyznaczanie dokładności obciętej dla maksymalnie 30 dokumentów odpowiedzi jest również uzasadnione zachowaniami wyszukiwawczymi użytkownika. Przeglądając odpowiedź wyszukiwarki, użytkownicy poszukują dobrych dokumentów najczęściej wśród początkowych dokumentów odpowiedzi i wiadomo, że mało który użytkownik jest na tyle cierpliwy, aby przeglądać więcej niż 100 dokumentów odpowiedzi (Rao, 1988a).

2.2. Problemy wyszukiwania informacji w Webowych systemach wyszukiwawczych

Badania w dziedzinie wyszukiwania informacji w sieci WWW są kolejnym krokiem w rozwoju metod i technologii wyszukiwania informacji (Kobayashi i Takeda, 2000). Od początku zaistnienia Internetu w 1969 roku i powołania do istnienia przez Toma Bernersa–Lee sieci WWW w 1985 roku, ilość zasobów rośnie w sposób nieprzewidziany przez twórców sieci (Lawrence i Giles, 1998). Prowadzone systematycznie szacowania wielkości zasobów WWW oraz ilości informacji dostępnych poprzez wyszukiwarki internetowe, które są podstawowym narzędziem używanym przez 85% użytkowników sieci do zlokalizowania istotnych informacji, potwierdzają ciągłe powiększane się zasobów (Lawrence i Giles, 1999). Wyszukiwarki internetowe są indeksami, które mają pełnić rolę taką, jak tradycyjne indeksy lub katalogi biblioteczne. Podstawowa różnica w stosunku do tradycyjnych zbiorów danych to decentralizacja, duża dynamika tworzenia i modyfikacji istniejących stron w sieci WWW. Lawrence i Giles w latach 1997–1999 przeprowadzili badania i porównania ilości dostępnych informacji w najpopularniejszych wyszukiwarkach (Lawrence i Giles, 1998), (Lawrence i Giles, 1999). Badacze ci zaproponowali własne metody szacowania wielkości tych zasobów, pokazując niedociągnięcia metod stosowanych do powyższych szacunków przez producentów wyszukiwarek. Przeprowadzone przez nich analizy dla sieci WWW pozwoliły oszacować liczbę indeksowanych przez wyszukiwarki stron w sieci WWW na co najmniej 800 mln w 1999r. (gdy w 1997r. – 320 mln stron), dających zasoby danych o wielkości ok. 15 terabajtów (w tym ok. 6 terabajtów danych tekstowych) (Lawrence i Giles, 1999). W szacunkach tych nie uwzględniali oni stron niedostępnych dla wyszukiwarek z powodu przeniesienia, usunięcia lub autoryzowanego dostępu oraz stron, które nie zawierały w swojej treści słów z pytania użytkownika (analizy przeprowadzano na podstawie odpowiedzi na zbiór 1056 pytań sformułowanych przez pracowników instytutu NEC). Po określeniu wielkości tych zasobów, Lawrence i Giles sprawdzili, że pojedyncze wyszukiwarki mają dostęp jedynie do 3 – 34% z poindeksowanych stron WWW, co oznacza, że oprogramowanie do zarządzania i wyszukiwana informacji nie nadąża za rozwojem sieci WWW. Wiąże to się głównie z trudnościami w utrzymaniu aktualności indeksów wyszukiwarek internetowych.

Zasoby WWW podlegają ciągłym zmianom przeprowadzanym prawie równocześnie przez miliony użytkowników. Z tego powodu aż 2 – 14% stron poindeksowanych przez wyszukiwarki jest nieaktualnych (Lawrence i Giles, 1998). Konieczna jest więc permanentna reindeksacja sieci przez wyszukiwarki, aby wyniki wyszukiwania dokumentów przekazywane użytkownikowi były jak najbardziej aktualne. Ci sami autorzy podjęli się określenia czasu, po jakim zmodyfikowana strona zostaje zarejestrowana w indeksie wyszukiwarki. Średnia wartość wyniosła 186 dni (dla 11

badanych wyszukiwarek wartość ta wynosiła 141 – 240 dni), czyli kilka miesięcy lub dłużej. Tak więc nie tylko wielkość zasobów sieci, ale również ich zmienność jest czynnikiem wpływającym na problemy z reindeksacją danych (Lawrence i Giles, 1999).

Sieć WWW można potraktować jako bardzo dużą, nieustruktualizowaną i rozproszoną bazę danych, zawierającą dokumenty o różnorodnej tematyce. Strony WWW mogą być różnorodne ze względu na zawartość. Dokumenty mogą różnić się językiem (naturalny lub programowania), słownictwem (adresy e–mail, numery telefonów, odsyłacze, dane spakowane) oraz typem i formatem (tekst, HTML, PDF, postscript, obrazy, dźwięki, wideo) (Daniłowicz, 2000). Istotne są informacje o samych stronach, takie jak źródło pochodzenia danej strony, częstość aktualizacji, jakość, popularność lub użyteczność, czy częstość cytowania danej strony.

Dziedzictwem tradycyjnych systemów wyszukiwania informacji są dogłębnie opracowane metody wyszukiwania dokumentów tekstowych. Różnorodność zawartości dokumentów WWW sugeruje, że można rozważyć uwzględnienie w procesie wyszukiwania informacje, jakie dodatkowo niosą odsyłacze hipertekstowe, obrazy, czy dźwięki. Gdy w przypadku tych pierwszych – istnieją konkretne, wdrożone realizacje (Marchiori,1997), (Brin i Page, 1998), tak w przypadku wyszukiwania obrazów i dźwięków – rozwiązania są na etapie intensywnych badań.

Podstawową różnicą (oprócz wymienionych powyżej dotyczących wielkości, różnorodności, decentralizacji i modyfikacji danych) pomiędzy tradycyjnym systemem wyszukiwania informacji, a internetowym systemem wyszukiwania jest realizowanie procesu wyszukiwania, w tym ostatnim, tylko na podstawie posiadanego przez wyszukiwarkę indeksu, a nie na podstawie pełnego tekstu dokumentu. Tak więc systemy wyszukiwania informacji w sieci WWW nie są pełnotekstowymi systemami wyszukiwania informacji (Baeza–Yates i Ribeiro–Neto, 1999), (Tanudjaja i Mui, 2002). Jest to również przyczyną, że w odpowiedzi na zadane do wyszukiwarki pytanie, zawartych jest niewiele informacji o wyszukanej stronie. Zamieszczone w odpowiedzi jedynie informacje takie jak: adres URL strony, kilka pierwszych linijek tekstu, wielkość strony, nie są wystarczającymi informacjami, aby użytkownik mógł zdecydować o relewancji odpowiedzi. Dlatego też użytkownik zmuszany jest do otwierania kolejnych stron w odpowiedzi, z których wiele okazuje się nierelewantnymi (Daniłowicz, 1999), (Pretschner i Gauch, 1999), (Pretschner i Gauch, 2000).

W porównaniu do wzrostu wielkości sieci WWW oraz znaczenia wyszukiwarek internetowych, liczba szczegółowych publikacji na temat aktualnie działających wyszukiwarek nie jest zbyt duża (Pinkerton, 1994). Najczęściej szczegóły techniczne systemów komercyjnych nie są ujawniane. Dlatego też większość publikowanych danych opiera się na wynikach przeprowadzonych doświadczeń, a nie na danych udostępnianych przez producentów tego oprogramowania (Pinkerton, 1994), (Lawrence i Giles, 1998), (Lawrence i Giles, 1999), (Brewington i Cybenko, 2000), (Choroś, 2002).

2.3. Próby dostosowania tradycyjnych technologii na potrzeby wyszukiwania informacji w sieci WWW

Przedstawione powyżej problemy, a równocześnie ogromne znaczenie, jakie ma sieć WWW w rozpowszechnianiu i wymianie informacji stymulują do prowadzenia intensywnych badań nad znalezieniem rozwiązań efektywnego zarządzania i wyszukiwania informacji w sieci WWW (Indyka-Piasecka, 2000).

Rozwiązaniem wielu problemów z wyszukaniem relewantnych informacji w sieci WWW wydaje się zebranie wszystkich dokumentów znajdujących się w sieci w jedną kolekcję dokumentów. Mając taką kolekcję dokumentów, klasyczne metody wyszukiwania informacji mogłyby być zastosowane tak, jak dla każdej innej kolekcji dokumentów. Na takim podejściu bazują najpopularniejsze wyszukiwarki, m.in.: Google, AltaVista, HotBot, Northen Light, Excite, korzystające ze scentralizowanej architektury *crawler–indexer* (Baeza–Yates i Ribeiro–Neto, 1999), (Hu i inni, 2001). W architekturze tej zbieranie oraz przesyłanie nowych i zmodyfikowanych stron należy do programów zwanych pająkami (ang. *crawlers, spiders, wanderers, knowbots*). Programy te działają na lokalnym komputerze, na którym zainstalowana jest wyszukiwarka i wysyłają zapytania o strony do serwerów WWW oraz przechodzą do kolejnych stron po odsyłaczach istniejących na innych stronach. Aby zarejestrować zmiany wprowadzane na stronach, powracają one do odwiedzonych stron w określonych cyklach czasowych, np., co miesiąc lub dwa miesiące (Baeza–Yates i Ribeiro–Neto, 1999).

Indeksowanie stron przesłanych na serwer lokalny, wykonywane jest przez program indeksujący. Większość indeksów posiada strukturę zbiorów odwróconych (ang. *inverted file*), w których każde słowo z uporządkowanej listy słów posiada wskaźniki do stron, w których słowo to występuje. W niektórych wyszukiwarkach stosowana jest eliminacja słów należących do stop–listy w celu zredukowania wielkości zbiorów (Pinkerton, 1994). Stosowanie stop–listy jednak nie zawsze przynosi korzyści. Listy takie można utworzyć dla określonej dziedziny lub dla określonego języka. Istnieją stop–listy dla języka angielskiego (Frakes i Baeza–Yates, 1992), zawierające ok. 450 słów, jednak nie można ich zastosować dla innego języka, ponieważ nie wiadomo, czy np. ciąg liter *the* nie jest znaczącym słowem w innym języku lub skrótem, nazwą firmy. Przyjęte jest, że jeśli stop–lista wykorzystywana jest w procesie wyszukiwania w sieci WWW, to lista ta nie jest zbyt obszerna i zawiera tylko kilka popularnie używanych słów (spójniki: *and*, *or*, przyimki: *a*, *the*, przysłówki). Jednak w większości wyszukiwarek internetowych stop–lista nie jest w ogóle stosowana.

Indeks wyszukiwarki internetowej zawiera również krótki opis wyszukiwanej strony, aby zasygnalizować treść dokumentu (tytuł, kilka pierwszych linii tekstu dokumentu, data utworzenia, rozmiar). Tak budowany, scentralizowany indeks nie zawiera pełnych tekstów dokumentów, na postawie których mogłyby być udzielane odpowiedzi na

pytania użytkowników. Odpowiedź, będąca listą adresów stron WWW, generowana jest po przeszukaniu indeksu będącego najczęściej posortowaną listą słów z przypisaną do każdej pozycji listą stron WWW, w których słowo wystąpiło – inaczej indeksem odwróconym (ang. *inverted index*). Gdy pytanie jest złożone z kilku słów, odpowiedź generowana jest na podstawie złączenia wyników przeszukania indeksu dla każdego ze słów pytania. Utrzymywanie indeksów wymaga znacznych zasobów sprzętowych. W 1998 r. wyszukiwarka AltaVista potrzebowała do efektywnego działania 20 wieloprocesorowych komputerów o pamięci RAM 130 GB i dyskach 500 GB. Szacuje się, że przy dzisiejszych technikach indeksowania, zbiory odwrócone pozwalają na redukcję indeksowanego tekstu, o 30%, co dla 100 mln stron daje wielkość 150GB potrzebnej pamięci dyskowej (Baeza–Yates i Ribeiro–Neto, 1999).

Szukanie w zgromadzonym indeksie odpowiedzi na pytanie użytkownika wykonywane jest przez program szukający (ang. *search engine*), który dokonuje również rankingu znalezionych pozycji z indeksu. Określanie relewancji strony, podobnie jak szukanie odpowiedzi, odbywa się tylko na podstawie indeksu, bez dostępu do pełnego tekstu strony. W literaturze nie są publikowane szczegółowe informacje na temat stosowanych w komercyjnych wyszukiwarkach technik określania relewancji stron WWW dla pytania użytkownika. Wiadomo jednak, że w metodach tych wykorzystywany jest najczęściej klasyczny schemat ważenia terminów na podstawie częstości występowania terminu w tekście (ang. *term frequency – tf*) w stosunku do liczby dokumentów kolekcji, które zawierają analizowany termin (ang. *inverted document frequency – idf*) (Pinkerton, 1994), (Baeza–Yates i Ribeiro–Neto, 1999). Ta metoda ważenia terminów oznaczana jest skrótem *tf–idf*[1].

Uwzględnienie w algorytmach rankingu dokumentów informacji zawartych w odsyłaczach hipertekstowych jest istotną różnicą pomiędzy systemem wyszukiwania informacji w sieci WWW, a tradycyjnym systemem wyszukiwania informacji. Przyjmuje się, że miarą popularności i jakości strony WWW może być liczba odsyłaczy wskazujących na daną stronę.

Przykładem efektywnej wyszukiwarki internetowej wykorzystującej w rankingu dokumentów miarę ważności odsyłacza hipertekstowego jest opracowana na Uniwersytecie Stanford wyszukiwarka Google (Brin i Page, 1998). Na podstawie utworzonego grafu odsyłaczy sieci WWW, twórcy wyszukiwarki Google określają miarę *PageRank* – obiektywną wartość strony ze względu na posiadane odsyłacze – dla 518 mln stron (w 1998r, ponad 4,2 mld stron w styczniu 2004 r), wskazywanych przez odsyłacze hipertekstowe. Metoda PageRank symuluje zachowanie użytkownika, który wybiera losowo stronę WWW i przechodzi do kolejnych stron, klikając kolejne odsyłacze, jednak nigdy nie wraca do poprzednio odwiedzanej strony. Prawdopodobieństwo, że użytkownik odwiedzi daną stronę jest przypisaną jej wartością

[1] Schemat ważenia *tf–idf* opisano szczegółowo w podrozdziale 2.1.1.

PageRank (Brin i Page, 1998), (Baeza–Yates i Ribeiro–Neto, 1999). Dodatkowo w wyszukiwarce Google inaczej traktowany jest tekst zawarty w odsyłaczu hipertekstowym. Tekst ten wiązany jest ze stroną, na którą odsyłacz wskazuje, a nie ze stroną, na której ten odsyłacz znajduje się. Takie podejście dostarcza precyzyjniejszych informacji na temat strony wskazywanej niż sama strona, jak również umożliwia poindeksowanie stron niedostępnych dla wyszukiwarek tekstowych, np. obrazów, oprogramowania, baz danych. Doświadczenia przeprowadzone przez autorów Google pokazały, że wyniki wyszukiwania tego systemu są lepsze niż wyniki wyszukiwarek komercyjnych. W odpowiedziach nie pojawiały się odsyłacze nieaktualne, a dla pytań, dla których wyszukiwarki komercyjne nie zwróciły żadnej odpowiedzi, wyszukiwarka Google zaprezentowała kilka stron internetowych (Brin i Page, 1998).

Dla omówionej scentralizowanej architektury *crawler–indexer*, obecnie stosowanej w większości istniejących wyszukiwarek internetowych głównymi problemami są: rozmiar zbieranych danych, obciążenie łączy komunikacyjnych oraz serwerów WWW, utrzymywanie aktualności indeksów w dynamicznie zmieniającym się środowisku, czasochłonność przetwarzania (więcej czasu procesora dla tej samej ilości pytań podczas wyszukiwania) i większe wydatki na konserwację dla utrzymania pełniejszych, a przez to większych indeksów (Baeza–Yates i Ribeiro–Neto, 1999, str. 254). Największe z aktualnie działających wyszukiwarek (każda niezależnie) mają dostęp jedynie do ok. 34% poindeksowanych przez wszystkie wyszukiwarki razem zasobów WWW, co odpowiada ok. 16% wszystkich zasobów WWW (Lawrence i Giles, 1998), (Lawrence i Giles, 1998A).

Stosowaną metodą powiększenia obszaru wyszukiwania w sieci WWW jest skierowanie tego samego pytania do kilku wyszukiwarek równocześnie. Wyniki wyszukiwania z kilku wyszukiwarek internetowych są łączone z pominięciem powtarzających się stron. Odpowiedź prezentowana jest użytkownikowi w postaci jednej listy storn internetowych. Idea ta zastosowana jest w metawyszukiwarkach internetowych (Eztioni i Weld, 1994), (Selberg i Etzioni, 1995), (Lawrence i Giles, 1998A). Działania te zwiększyły ilość wyszukanych stron do 60% poindeksowanych zasobów sieci (Lawrence i Giles, 1998).

3. Stan badań w zakresie modelowania użytkownika

Celem tej części pracy jest przedstawienie problematyki z zakresu modelowania użytkownika. Pokazane zostanie, że jest to dziedzina bardzo rozległa, która posiada wiele różnorodnych zastosowań. Właśnie różnorodne zastosowania modelowania użytkownika były przyczyną sięgnięcia przez badaczy do osiągnięć sztucznej inteligencji, statystyki, psychologii, czy nauk kognitywnych i zaadaptowania ich na potrzeby modelowania użytkownika. Wyszukiwanie informacji, wspomaganie nauczania oraz wspomaganie podejmowania decyzji należą do najważniejszych dziedzin zastosowania modelowania użytkownika. Zasadniczo, na podstawie literatury można stwierdzić, że wszystkie trzy dziedziny korzystają z podobnych rozwiązań w procesie modelowania użytkownika, ze skutkiem pozytywnie ocenianym przez użytkownika.

W niniejszym opracowaniu, spośród trzech wymienionych dziedzin zastosowania modelowania użytkownika, w osobnym podrozdziale przedstawiono zagadnienia modelowania użytkownika w wyszukiwaniu informacji, wyróżniając tę problematykę z dwóch powodów. Po pierwsze, ponieważ przedstawione w niniejszej pracy rozwiązanie modelowania użytkownika dotyczy internetowego systemu wyszukiwania informacji. A po drugie, ponieważ w dziedzinie tej dysponuje się ograniczonymi danymi, tj. tylko terminami, które można wykorzystać w procesie modelowania.

3.1. Cel modelowania użytkownika

Adaptacja procesu wyszukiwania w sieci WWW

Podstawowym celem leżącym u podstaw procesu modelowania użytkownika jest osiągnięcie szeroko rozumianej adaptacji systemu do potrzeb użytkownika. Adaptacja powinna przynosić konkretne korzyści użytkownikowi podczas pracy z systemem. Potrzeba modelowania użytkownika ujawnia się w wielu sytuacjach. Wraz z rozwojem sieci WWW ważne stało się wspomaganie użytkownika w znalezieniu istotnych dla niego informacji. Utworzenie i korzystanie z modelu użytkownika w systemie wyszukiwania informacji w sieci WWW może służyć do selekcji interesujących użytkownika dokumentów (Benaki i inni, 1997), czy rekomendacji stron WWW (Akoulchina i Ganascia, 1997), (Pazzani i Billsus, 1999), (Billsus i inni, 2002). Długoterminowe potrzeby informacyjne, czyli zainteresowania użytkownika związane z powtarzającymi się w czasie wyszukiwaniami, zapamiętane w modelu tego użytkownika, umożliwiają filtrowanie nowych dokumentów umieszczanych w sieci WWW bez udziału użytkownika (Ambrosini i inni, 1997). Wprowadzenie do systemu

modelu użytkownika, w którym umieszczone są informacje o sposobie prowadzenia wyszukiwań w sieci WWW, umożliwia dostarczenie pomocy w znalezieniu istotnych informacji przez przypomnienie dotychczasowych ścieżek nawigacji w sieci WWW (Maglio i Barrett, 1997).

Adaptacja prezentowania informacji

Modelowanie użytkownika znalazło również zastosowanie w procesie prezentacji informacji według potrzeb i wymagań użytkownika. Dostosowanie to może dotyczyć zarówno sposobu prezentacji wykresów, uwzględniającej możliwości i preferencje użytkownika (Gutkauf i inni, 1997), jak i wybranie przez użytkownika sposobu prezentowania informacji (w postaci tekstu, wykresów lub schematów), (Kalyuga i inni, 1997), czy też prezentowania informacji hipermedialnej dostosowanej do zainteresowań i wiedzy użytkownika (De Carolis i Pizzutilo, 1997). W systemie wspomagania decyzji wykorzystano model użytkownika do dopasowania argumentów podejmowania decyzji zgodnie z przekonaniami użytkownika (Grasso, 1997) oraz do dostosowania informacji wspomagania decyzji zgodnie z cechami osobowościowymi i preferencjami użytkownika (Paranagama i inni, 1997)

Adaptacja interfejsu użytkownika

Jak wspomniano na początku, celem modelowania użytkownika jest osiągnięcie adaptacji systemu do potrzeb użytkownika. Element systemu, dzięki któremu adaptacja całego systemu jest postrzegana przez użytkownika to interfejs użytkownika. Dlatego też, w ostatnich latach powstało wiele prac z dziedziny adaptacji interfejsu do potrzeb i wymagań użytkownika. W systemie wspomagającym wyszukiwanie informacji w sieci WWW, opracowanym przez Maglio i Barrett'a, skróty przejścia pomiędzy stronami WWW, stosowane przez użytkowników podczas nawigacji pomiędzy stronami WWW, prezentowane były na podstawie poprzednich ścieżek przejść przebytych przez użytkownika (Maglio i Barrett, 1997). Uwzględniając: różny poziom możliwości, różne doświadczenia w pracy w sieci WWW, wiedzę i podstawy posiadane przez użytkownika, prezentowane elementy interfejsu (tj.: okna, ramki, formularze, przyciski) oraz porady interfejsowe dostosowano do poziomu użytkownika systemu InterBook (Brusilowski i Schwarz, 1997). Podobnym zagadnieniem zajęli się autorzy projektu informacji miejskiej AVANTI, w którym na podstawie modelu potrzeb użytkownika, informacja hipermedialna prezentowana na stronach WWW dostosowywana jest od zainteresowań, wiedzy, wieku czy nawet poziomu niesprawności ruchowej użytkownika (np. informacja o szerokości drzwi, podjazdach czy windach jest istotna dla osoby niepełnosprawnej, a informacja taka nie musi być prezentowana osobie bez ograniczeń ruchowych) (Fink i inni, 1997). Rozróżnienie, na podstawie modelu

użytkownika, pomiędzy doświadczonymi i niedoświadczonymi użytkownikami systemu komercyjnego (np. systemu zarządzania finansami) pozwala na ukierunkowanie pomocy i uproszczenie interfejsu dla nowych użytkowników (Strachan i inni, 1997). Rozwiązanie takie wpływa na poziom satysfakcji użytkownika podczas korzystania z systemu. W systemie wyszukiwani informacji, dzięki adaptacji interfejsu użytkownika, możliwe jest podpowiedzenie następnego kroku w procesie wyszukiwania (podczas dialogu wyszukiwawczego pomiędzy użytkownikiem a systemem) po nieprzewidzianym, niejednoznacznym akcie dialogu (ang. *dialog act*) użytkownika (Stein i inni, 1997).

Gromadzenie informacji o poziomie wiedzy użytkownika

Celem modelowania użytkownika jest również przekazanie informacji zwrotnej o wiedzy posiadanej przez użytkownika. Jest to szczególnie istotne w dziedzinie systemów uczących. Informacja zwrotna najczęściej zawiera informacje o postępach w nauce. W systemie SeeYourselfWrite, wspomagającym naukę języków obcych, model użytkownika wykorzystany został do dostarczenia uczniom informacji zwrotnej o popełnianych przez nich błędach podczas nauki pisania w języku obcym (Bull, 1997).

Modelowanie użytkownika znajduje również swoje uzasadnienie w dziedzinie wspomagania współpracy. Tworzone i przechowywane w systemie PHelpS (Peer Help System) modele użytkowników dla pracowników pewnego ośrodka umożliwiają wybór grupy współpracowników posiadających odpowiednią wiedzę (lub pomoc użytkownikowi w wyborze współpracowników), jeśli któryś z pracowników ośrodka zgłosi problem z wykonaniem powierzonego mu zadnia. Dodatkowo modele użytkowników ułatwiają i udostępniają komunikację pomiędzy określonymi grupami współpracowników (Collins i inni, 1997). W systemie wspomagającym powtarzanie posiadanej przez studentów wiedzy przed testem egzaminacyjnym, model użytkownika wykorzystany został do wskazania określonej formy współpracy podczas powtórek pomiędzy studentami (Bull i Smith, 1997). W zależności od poziomu wiedzy współpracujących studentów, polecana jest współpraca w formie nauki wspólnej, uczenia jednego studenta przez drugiego lub nauki indywidualnej.

Celem modelowania użytkownika jest również przewidywanie przyszłych zachowań. Przewidywanie to może dotyczyć zarówno poprawnych i błędnych odpowiedzi uczniów (Chiu i inni, 1997), jak i celów, akcji i położenia w przestrzeni agenta grającego w grę (Albrecht i inni, 1997).

3.2. Zawartość modelu użytkownika

Dziedzina zastosowania modelu użytkownika implikuje zazwyczaj rodzaj informacji o użytkowniku, które są gromadzone w modelu. Informacje znajdujące się w modelu

można podzielić na kilka grup: informacje o preferencjach i celach działania użytkownika, informacje o aspektach wiedzy i przekonań użytkownika oraz zaawansowaniu użytkownika w dziedzinie zastosowania systemu, charakterystyka osobista, czy historia interakcji użytkownika z systemem.

Informacje o preferencjach i celach działania użytkownika

Model użytkownika w systemach wyszukiwania informacji może zawierać zainteresowania użytkownika wyrażone podczas wyszukiwania w sieci WWW (Ambrosini i inni, 1997), preferencje dotyczące wiadomości sieciowych (Billsus i Pazzani, 1999), cele wyszukiwania w sieci WWW (Akoulchina i Granascia, 1997), czy też kontekst aktualnego łącza hipertekstowego, który odzwierciedla zainteresowania użytkownika (Staff, 1997). Paranagama, Burstein i Arnott sugerują, że w systemach wspomagania decyzji model użytkownika powinien zawierać stopień ważności, istotności atrybutów związanych z podejmowaniem decyzji (Paranagama i inni, 1997). W innych dziedzinach, gdzie konieczna jest adaptacja systemu do potrzeb użytkownika, jego model może zawierać preferencje dotyczące różnych aspektów graficznej prezentacji wykresów, uwzględniające np. poziom postrzegania kolorów przez konkretnego człowieka (Gutkauf i inni, 1997). W hipermedialnym systemie informacji miejskiej AVANTI, model użytkownika zawiera charakterystykę użytkownika związaną z informacjami prezentowanymi hipermedialnie. Charakterystyka ta dotyczy m.in. zainteresowań określonymi obiektami historycznymi, preferencji dotyczących sposobu prezentowania informacji hipermedialnej za pomocą wybranych mediów (tj. grafiki lub wideo), sprawności ruchowej (np. nie są prezentowane informacje o obiektach niedostępnych dla osób niepełnosprawnych) oraz główne cele uzasadniające korzystanie z systemu przez użytkowników (Fink i inni, 1997).

Informacje o aspektach wiedzy i przekonań użytkownika

Druga grupa informacji jakie mogą znaleźć się w modelu użytkownika to pewne aspekty wiedzy i przekonań użytkownika. W systemach edukacyjnych, wyposażonych w adaptacyjny interfejs, w modelu użytkownika reprezentowana jest wiedza użytkownika o elementach złożonego interfejsu systemu edukacyjnego, działającego w sieci WWW (Brusilowski i Schwarz, 1997), czy też wiedza lub braki w wiedzy ucznia systemu wspomagającego naukę języka obcego (Bull, 1997). W systemie wspomagania decyzji model użytkownika może zawierać informacje o wiedzy studenta dotyczącej określonych reguł rozwiązywania problemu (Corbett i Bhatnagar, 1997), natomiast w systemie wspomagającym współpracę pomiędzy pracownikami pewnej jednostki – informacje o umiejętności wykonania określonych zadań (Collins i inni, 1997). Generowanie dokumentów hipermedialnych, dostosowanych do wymagań

użytkownika, znalazło swoje zastosowanie w systemach edukacyjnych, szczególnie w różnego rodzaju instrukcjach, podręcznikach użytkowych. Utworzony obiekt hipermedialny – opis pewnego elementu lub urządzenia wraz z instrukcją wykorzystania i użytkowania tego elementu lub urządzenia musi być zrozumiały dla użytkownika, będącego równocześnie uczniem. Dlatego też model użytkownika w systemach tego rodzaju zawiera informacje o znajomości zagadnień dotyczących hiperterstu, hipermediów oprócz informacji o wieku, wykształceniu doświadczeniu w dziedzinie, której dotyczy nauka (De Carolis i Pizzutilo, 1997).

Informacje o zaawansowaniu użytkownika w dziedzinie zastosowania systemu

W modelu użytkownika reprezentowane jest również zaawansowanie w znaczeniu biegłości użytkownika w dziedzinie zastosowania systemu. Corbett i Bhatnagar, przedstawiając inteligentny system uczący ACT Programming Tutor (APT), proponują model użytkownika, który zawiera informacje o poszerzaniu się wiedzy studenta wraz z czasem spędzonym na pracy z systemem. System APT należy do grupy systemów wspomagających naukę programowania w językach Lisp, Pascal lub Prolog. Pozyskiwanie wiedzy przez studenta o regułach programowania i wykorzystaniu tych reguł, określane jest przez system podczas procesu śledzenia przyrostu wiedzy u studenta (ang. *knowledge tracing*). Pozyskiwana przez studenta wiedza jest związana z rozwiązywaniem problemu programistycznego (Corbett i Bhatnagar, 1997). W zaproponowanym przez Bull i Smith systemie wspomagającym powtarzanie posiadanej przez studentów wiedzy przed testem egzaminacyjnym, model użytkownika zawiera informacja o umiejętności posługiwania się określonymi zagadnieniami i pojęciami indywidualnie, i we współpracy z innymi studentami (Bull i Smith, 1997). Informacje o poziomie kompetencji w posługiwaniu się komputerem i systemem hipermedialnym są elementami modelu użytkownika w systemie informacji miejskiej AVANTI (Fink i inni, 1997). Podobnie, biegłość użytkownika systemu TIMS (Tax and Investment Management Strategizer) w dziedzinie strategii planowania finansowego i wykonywania ekspertyz finansowych oraz biegłość użytkownika w korzystaniu z tego systemu są informacjami zawartymi w modelu użytkownika systemu TIMS (Strachan i inni, 1997). W modelu użytkownika systemu SATELITE, aktywnie wspomagającego wyszukiwanie informacji przez użytkownika w sieci Internet, reprezentowane jest doświadczenie użytkownika w dziedzinie w której prowadzone jest wyszukiwanie i teoretyczna orientacja użytkownika w tej dziedzinie (Akoulchina i Ganascia, 1997).

Cechy osobowe

Czwarta grupa informacji, które mogą znaleźć się w modelu użytkownika to charakterystyka osobista użytkownika. Jest ona również istotna w procesie tworzenia

modelu użytkownika. W systemie PHelpS istotnymi są informacje o miejscu pracy w ramach instytucji, czy stanowisku zajmowanym przez potencjalnych współpracowników (Collins i inni, 1997). Dane te, przechowywane w modelu użytkownika, umożliwiają dokonanie przez system wyboru współpracownika/ów, którzy mogą udzielić użytkownikowi pomocy podczas wykonywania problematycznego zadania. Wybór dokonywany jest na podstawie modeli współpracowników i modelu użytkownika, któremu potrzebna jest pomoc. W edukacyjnym systemie hipermedialnym GeNet model użytkownika zawiera informacje o poziomie wykształcenia użytkownika, doświadczeniu w dziedzinie, której dotyczy hipermedialny podręcznik wygenerowany przez system GeNet oraz doświadczenie i znajomość zagadnień związanych z hipermediami (De Carolis i Pizzutilo, 1997). Paranagama, Brustein i Arnott w zrealizowanym systemie wspomagania decyzji zaproponowali wprowadzenie do modelu użytkownika informacji o typie osobowości użytkownika. Wprowadzenie tych informacji jest uzasadniona istnieniem powiązań pomiędzy osobowością, a procesem podejmowania decyzji (Paranagama i inni, 1997). W zaproponowanym przez Paranagama, Brustein'a i Arrott'a rozwiązaniu, preferencje podejmowania decyzji opisywane są przez wielokryteriowe metody podejmowania decyzji. Metody te bazują na dwóch podstawowych elementach: atrybutach i wagach. Atrybuty są czynnikami uwzględnianymi podczas procesu podejmowania decyzji. Czynniki sytuacyjne podejmowania decyzji są wprowadzane do wielokryteriowego modelu przez miary atrybutów, czyli kryteria. Nie każdy atrybut posiada tak samo wysokie znaczenie dla różnych osób podejmujących te same decyzje. Dlatego też, poszczególne osoby mogą mieć odmienne preferencje dla tych samych atrybutów. Wagi zostały wykorzystane do reprezentowania stopnia ważności atrybutu. Cały zestaw preferencji dla poszczególnych atrybutów tworzy dla konkretnej osoby model preferencji podejmowania decyzji. W literaturze uzasadniono, że model preferencji podejmowania decyzji jest związany z typem osobowości osoby podejmującej decyzje. Wynika stąd, że preferencje podejmowania decyzji mogą być przewidywane na podstawie informacji o osobowości (Paranagama i Burstein, 1996).

Historia interakcji użytkownika z systemem

Piąta grupa informacji, które mogą znaleźć się w modelu użytkownika to informacje uzyskane na podstawie śledzenia interakcji użytkownika z systemem. Analiza interakcji może dostarczyć istotnych informacji na temat użytkownika. Skutkiem tej hipotezy jest pojawienie się propozycji reprezentowania w modelu użytkownika jego historii interakcji z systemem. W systemach wspomagających wyszukiwanie informacji w sieci Internet zwrócono uwagę na zapamiętywanie i analizowanie historii nawigacji po stronach WWW (Weber i Sprecht, 1997), (Gori i inni, 1997), (Maglio i Barrett, 1997). Maglio i Barrett zasugerowali, że użytkownicy sieci WWW realizują wyszukiwania

opierając się na charakterystycznych dla siebie procedurach i schematach. Każdy z użytkowników, podczas różnych wyszukiwań, korzysta zazwyczaj z tego samego narzędzia – tej samej wyszukiwarki lub tego samego katalogu stron. Nie wynika to jednak z preferencji uzasadnionych merytoryczną wiedzą na temat jakości narzędzia stosowanego do wyszukiwania, ale z przyjętych schematów wykonywania wyszukiwania. Jako przykład takiego zachowania Maglio i Barrett opisali eksperyment, w którym brało udział 15 osób wykonując wyszukiwania w sieci WWW przez kilka kolejnych dni. Z obserwacji wysnuli oni wnioski, że każdy użytkownik posiada charakterystyczne dla siebie zachowania wyszukiwawcze. Autorzy twierdzą, że użytkownicy realizują (ang. *conceptualise*) zadanie wyszukiwawcze korzystając z najczęściej wykorzystywanej procedury będącej ciągiem czynności. Jeśli aktualnie wykonywane czynności wyszukiwania zaczynają odbiegać od standardowych wzorców, użytkownik powraca do takich zachowań wyszukiwawczych, które są zgodne z wzorcami standardowymi dla tego użytkownika. Na przykład, jeden z uczestników eksperymentu korzystał z wyszukiwarki AltaVista, jako punktu startowego swoich wyszukiwań w pierwszym dniu eksperymentu. W czasie szukania odpowiedzi dla kolejnych zadań podawanych przez prowadzącego eksperyment, zmienił narzędzie na katalog stron Yahoo! W katalogu tym użytkownik nie znalazł interesujących go informacji więc powrócił do korzystania z wyszukiwarki AltaVista. Eksperyment przewidywał, że następnego dnia użytkownicy dostaną do wykonania te same zadania wyszukiwawcze. Śledząc zachowania wyszukiwawcze użytkowników Maglio i Barrett zauważyli, że dnia następnego użytkownicy wykonują takie same czynności początkowe, rozpoczynając wyszukiwanie korzystają z tego samego narzędzia co dnia poprzedniego. Podczas tych eksperymentów zaobserwowali również, że użytkownicy nie są w stanie odtworzyć sekwencji stron WWW i odsyłaczy, po których przeszli docierając do rozwiązania zadania dnia poprzedniego. Do strony zawierającej istotne informacje (w tym przypadku – rozwiązanie zadania) użytkownik dociera dnia następnego przywołując strony specyficzne – tzw. „kamienie milowe". Autorzy stwierdzili, że użytkownicy Internetu realizują wyszukiwanie korzystając z charakterystycznych dla poszczególnych użytkowników procedur i stron istotnych – „kamieni milowych". Na podstawie opisanego eksperymentu Maglio i Barrett zasugerowali, że w tworzonym modelu użytkownika systemu wyszukiwania informacji w sieci WWW powinny zostać zawarte informacje o procedurach stosowanych podczas wyszukiwania oraz stronach istotnych.

Propozycje reprezentowania w modelu użytkownika historii interakcji z systemem dotyczą również systemu wyszukiwania informacji MIRACLE (Stein i inni, 1997). Model użytkownika zawiera historię aktów mowy (ang. *dialog acts*), które są elementami dialogu prowadzonego podczas wyszukiwania pomiędzy użytkownikiem a systemem. Na podstawie powyższej historii, w momencie pojawienia się ze strony użytkownika aktu, który ma niejednoznaczną interpretację, system może

wywnioskować jaką akcję podjąć, w jaki sposób dalej prowadzić dialog wyszukiwawczy. Brusilowski i Schwarz zaproponowali system InterBook, przeznaczony do tworzenia elektronicznych podręczników opartych o koncepcję hipertekstu oraz posiadających cechy adaptacji interfejsu dla wymagań określonego użytkownika. W systemie tym, model użytkownika zawiera historię korzystania przez użytkownika z elementów interfejsu wraz historią czytania wskazówek dotyczących tych elementów (Brusilowski i Schwarz, 1997). W systemie PHelpS wspomagającym szkolenia pracowników model użytkownika, będący modelem każdego z pracowników, zawiera historię wykonania przez pracownika określonych kroków aktualnego zadania w ramach szkolenia (Collins i inni, 1997). Albrecht, Zukerman, Nicholson i Bud zaproponowali podejście do problematyki rozpoznawania planów (ang. *plan recognition*) bazujące na dynamicznej sieci Bayes'owskiej (ang. *Dynamic Bayesian Network*). Sieć wykorzystana została do reprezentowania cech pewnej dziedziny, które są niezbędne do identyfikacji planów i celów użytkownika. Wyniki eksperymentalne przedstawione zostały na przykładzie przygodowej gry komputerowej, w której bierze udział wielu graczy. Każdy z graczy rywalizuje o ograniczone zasoby, aby osiągnąć określony cel. Gracze mogą przemieszczać się pomiędzy pomieszczeniami wykonując akcje. Wykonując akcje, dążą do określonego celu. Dla każdego z graczy przechowywane są informacje o akcjach wykonanych przez gracza i zaobserwowanych przez system oraz o położeniu gracza (w jakim pomieszczeniu znajduje się po wybraniu określonej akcji). Informacje te utożsamiane są z modelem użytkownika wykorzystywanym w tym systemie. Na podstawie dotychczas wykonanych akcji i miejsca, w którym znajduje się użytkownik podczas gry, system uczy się jakie akcje i położenia lub sekwencje akcji i położeń mogą doprowadzić do celu tj. zwycięstwa (Albrecht i inni, 1997).

3.3. Dane wejściowe wykorzystywane w konstrukcji modelu użytkownika

Preferencje i cele użytkownika

System może dostosować swoje działanie, uwzględniając charakterystyki poszczególnych użytkowników, tzn. na podstawie informacji o użytkowniku zgromadzonych w modelu użytkownika. Istotne jest więc prześledzenie na podstawie jakich faktów system może konstruować model użytkownika. W rozwiązaniach prezentowanych w literaturze można zauważyć kilka podstawowych źródeł informacji o użytkowniku. Najpowszechniej stosowanym źródłem informacji są wprost wyrażone preferencje i cele użytkownika. W systemie wyszukiwania informacji w sieci WWW użytkownik wyraża swoje preferencje przez postawienie pytania, wskazanie zbioru

dokumentów dla każdej dziedziny zainteresowań (Benaki i inni, 1997) lub wskazanie wprost kontekstu w dokumencie hipertekstowym (Staff, 1997). W hipermedialnym systemie informacji miejskiej są to preferencje i cele dotyczące prezentacji hipermediów (Fink i inni, 1997), czy też krytyka zaproponowanych rozwiązań w systemie wspomagania decyzji (Paranagama i inni, 1997).

Odpowiedzi z testów, kwestionariuszy

Drugim ważnym źródłem danych o użytkowniku, wykorzystywanych w modelu, mogą być odpowiedzi na testy, kwestionariusze lub ćwiczenia testowe przygotowane przez autorów systemów. Testy zawierają zestawy pytań, na które użytkownik udziela odpowiedzi podczas pierwszej sesji pracy z systemem (Benaki i inni, 1997), (Akoulchina i Ganascia, 1997), (Bull i Smith, 1997), lub mogą być testami opartymi na koncepcji gry, sprawdzającymi zdolności użytkownika (Gutkauf i inni, 1997). Danych do modelu użytkownika mogą dostarczać również kolejne kroki rozwiązywania problemu przez studenta (Conati i inni, 1997), czy zachowania, działania studenta, gdy ma możliwość zastosowania określonych reguł programowania (Corbett i Bhatnagar, 1997).

Współpraca użytkownika z systemem

Innym źródłem danych są akcje wykonywane przez użytkownika podczas pracy z systemem. Może to być wybór ustawień urządzeń w różnych środowiskach (Doux i inni, 1997), czy też korzystanie przez użytkownika z elementów interfejsu i czytanie wskazówek na temat tych elementów (Brusilovsky i Schwarz, 1997). W hipertekstowych systemach wyszukiwania informacji źródłem danych wejściowych może być fakt odwiedzenia przez użytkownika strony hipermedialnej (Staff, 1997), (Akoulchine i Ganascia, 1997), (Maglio i Barrett, 1997), (Gori i inni, 1997).

Charakterystyka osobista oraz ocena własna użytkownika

Źródłem danych wykorzystywanych w modelu użytkownika są również informacje dotyczące charakterystyki osobistej użytkownika oraz oceny własnej uzyskane od samego użytkownika. Są to zarówno ogólne dane o zajmowanym stanowisku w pracy, wykształceniu (Collins i inni, 1997), (Strachan i inni, 1997), jak i szczegółowe informacje związane z charakterystycznymi dla użytkownika możliwościami korzystania z hipermediów – informowanie o braku możliwości, ułomnościach (Fink i inni, 1997). Korzystanie z oceny własnej użytkownika jest częste w systemach wspomagających uczenie się lub nauczanie. W systemach tych danymi, które mogą być wprowadzone do modelu, jest ocena własnych kompetencji w dziedzinie uczenia oraz zaawansowania w użytkowaniu systemu (Strachan i inni, 1997), (De Carolis i Pizzutilo,

1997), czy też zgłoszenie pozytywnego zakończenia wykonywanych zadań (Collins i inni, 1997).

3.4. Metody konstruowania modelu użytkownika

W poprzednich podrozdziałach opisane zostały cele modelowania użytkownika, rodzaj informacji, które mogą być gromadzone w modelu o użytkowniku, metody wykorzystania informacji zgromadzonych w modelu użytkownika w celu dostosowania działania systemu do wymagań użytkownika oraz źródła danych wprowadzanych do modelu użytkownika. Na podstawie wymienionych danych, system przyjmuje hipotezy dotyczące użytkownika. Hipotezy te są podstawą konstruowania modelu. Proces konstruowania modelu przebiega według określonych zasad i technik wnioskowania. Techniki konstruowania modelu użytkownika, w celu uzyskania adaptacji systemu można podzielić na kilka, opisanych poniżej, grup.

Metody bazujące na technikach statystycznych

Do pierwszej grupy można zaliczyć podejścia bazujące na metodach Bayesowskich. W systemie *ACT Programing Tutor* wspomagającym nauczanie programowania w językach Lisp, Pascal i Prolog, Corbett i Bhatnagar zastosowali procedury Bayesowskie do obliczania prawdopodobieństwa faktu, że reguły programowania są znane studentowi. Hipoteza, że uczeń zna pewną regułę programowania jest wprowadzana do modelu użytkownika, a następnie jest podstawą prezentowania kolejnych tematów do nauki (Corbett i Bhatnagar, 1997).

W dziedzinie maszynowego uczenia oraz wnioskowania z niepewnością (ang. *reasoning under uncertainty*), w wielu podejściach (m.in. drzewa decyzyjne, sieci neuronowe, sieci Bayes'owskie) wykorzystywane są metody statystyczne. Przewidywania będące wynikiem zastosowania tych podejść mogą być wykorzystane do adaptacji działania systemu do potrzeb użytkownika. Można wyróżnić dwa podejścia do problemu adaptacji systemu. Podejście bazujące na analizie zawartości (ang. *content–based approach*), w którym przyjęto zasadę, że użytkownik wykazuje pewne określone zachowanie w danych okolicznościach i zachowanie to jest powtarzalne dla danych okoliczności. Zachowanie użytkownika może być więc przewidywane na postawie zachowania, które miało miejsce w przeszłości. W drugim podejściu bazującym na analizie współpracy (ang. *collaborative approach*) przyjęto, że użytkownicy należący do pewnej grupy, wykazują podobne zachowania w określonych okolicznościach. Zachowanie użytkownika może być więc przewidywane na postawie zachowania innych użytkowników należących do tej samej grupy. Kilka różnych metod statystycznych znalazło zastosowanie zarówno w podejściu bazującym na analizie zawartości, jak i analizie współpracy. Między innymi metoda *tf–idf* (ang. *term*

frequency–inverted document frequency), statystyczna metoda analizy treści dokumentów oraz nadawania wag terminom należącym do dokumentów. Metoda wykorzystywana jest powszechnie w procesie wyszukiwania informacji do znalezienia dokumentów odpowiadających pytaniu użytkownika (Salton i McGill, 1983). Przez Moukasa i Maes metoda *tf–idf* wykorzystana została w procesie rekomendacji dokumentów użytkownikowi, gdzie rekomendacja odbywała się na podstawie innych, podobnych, dokumentów interesujących użytkownika (Moukas i Maes, 1998).

Metody wykorzystujące techniki maszynowego uczenia

Druga grupa metod wykorzystywanych do konstruowania modelu użytkownika bazuje na technikach maszynowego uczenia się (Webb i inni, 2001), (Zukerman i Albrecht, 2001), (Zhang i Seo, 2001). W systemie wspomagania decyzji Paranagama, Burstein, Arnott wykorzystali sieć neuronową do aktualizacji profilu użytkownika (Paranagama i inni, 1997). Dla systemu wyszukiwania informacji, Ambrosini, Cirillo i Micarelli zaproponowali hybrydowe podejście do konstruowania modelu użytkownika, łączące modelowanie użytkownika przy użyciu *stereotypów* z technologią sieci neuronowych. Stereotyp jest opisem prototypowego użytkownika pewnej klasy (Rich, 1983). Rozważano zadanie automatycznego tworzenia stereotypu użytkownika na podstawie biblioteki stereotypów ustalonej przez ekspertów. Stereotypy mają strukturę ram i zawierają opis użytkownika w postaci zbioru wartości określonych atrybutów. Dla nowego użytkownika, sieci neuronowe zostały wykorzystane do klasyfikacji stereotypów z istniejącej biblioteki stereotypów. Stereotypy z biblioteki stereotypów wykorzystane zostały jako przypadki uczące, pozyskane od ekspertów (Ambrosini i inni, 1997). Natomiast dla internetowego systemu wyszukiwania informacji Gori, Maggini i Martinelli wykorzystali powrotną sieć neuronową (ang. *recurrent neuron networks*) do podsumowania nawigacji użytkownika w sieci WWW. Podsumowanie to tworzy model użytkownika (Gori i inni, 1997).

Istotą modelowania użytkownika w systemach wyszukiwania informacji jest zaproponowanie reprezentacji najlepiej wyrażającej rzeczywiste potrzeby informacyjne użytkownika. Postulat ten dotyczy również internetowych systemów wyszukiwania informacji, które są przedmiotem niniejszej pracy. Na potrzeby tworzenia reprezentacji zainteresowań użytkownika, istotnych informacji może dostarczyć analiza historii interakcji użytkownika z systemem (Stein i inni, 1997). Jednak, należy rozważyć, jakie rozwiązania są możliwe do zastosowania dla systemów wyszukiwania informacji w sieci WWW. Zastosowanie metod sztucznej inteligencji wymaga konstruowania bazy wiedzy, na podstawie której podejmowane są decyzje (Zukerman i Alberecht, 2001). Systemy, w których stosowane jest modelowania użytkownika, w procesie wnioskowania na podstawie obserwacji działań użytkownika najczęściej wykorzystują ręcznie utworzone bazy wiedzy. I tak na przykład, w niektórych systemach

rozpoznawania planów (ang. *plan recognition systems*) wykorzystywano ręcznie tworzone biblioteki planów, aby określić intencje lub preferencje użytkownika na podstawie wypowiedzi użytkownika (Carberry, 2001). Ręcznie tworzone bazy wiedzy są zazwyczaj budowane na podstawie szczegółowej analizy przypadków, które uważane są za reprezentatywne dla danego problemu. Jednakże tak tworzone bazy wiedzy posiadają dwa podstawowe ograniczenia: proces ręcznego konstruowania bazy jest bardzo kosztowny oraz utworzona w taki sposób baza nie jest rozszerzalna i modyfikowalna. Szczególnie mocno uwidaczniają się te ograniczenia wraz z pojawieniem się nowoczesnych technologii, jak np.: automatyczne przesyłanie poczty elektronicznej do kolejnego odbiorcy, wspomaganie edycji dokumentów, rekomendacja stron WWW lub filmów. Systemy te generują duże ilości danych, jak np. elektroniczne logi użytkowników, które są potencjalnym źródłem danych do tworzenia bazy wiedzy. Niestety dane te są zazwyczaj zniekształcane przez zakłócenia, takie jak np. przerwanie współpracy z systemem, błędne rozpoczęcie pracy. Podczas analizy przykładów na potrzeby tworzenia bazy wiedzy zakłócenia muszą zostać pominięte. Tak więc ręczne budowanie bazy wiedzy, która obejmuje niewiele przykładów jest znacznie mniej pracochłonne niż budowanie bazy wiedzy reprezentującej przypadki z wymienionych, nowych zastosowaniach. Dodatkowo zniekształcone dane wymagają osobnej analizy, co zwiększa trudności w tworzeniu bazy.

Problemy związane z tworzeniem bazy wiedzy, sieci semantycznych, czy neuronowych dla różnych dziedzin, problemy z interpretacją i analizą możliwych aktów mowy podczas dialogu wyszukiwawczego wpływają na stopień niepewności w modelowaniu użytkownika (Stein i inni, 1997). Należy więc szukać rozwiązań prostszych, takich, dla których istnieje szansa zastosowania w dynamicznym środowisku jakim jest sieć WWW. Obiecującą alternatywą stają się metody statystyczne. W metodach statystycznych zaobserwowane wyniki przykładowe wykorzystywane są do określenia nieznanych parametrów zależnych (Zukerman i Alberecht, 2001). W procesie modelowania użytkownika parametry te reprezentują aspekty przyszłego zachowania użytkownika, m.in. cele, preferencje, przyszłe akcje.

Metody wykorzystujące algorytmy genetyczne

Inną techniką bazującą na maszynowym uczeniu się, wykorzystaną do konstruowania modelu użytkownika, są algorytmy genetyczne. Moukas i Maes w swoich pracach podali propozycję wykorzystania sztucznego, ewolucyjnego ekosystemu do filtrowania i odkrywania informacji w sieci WWW (Moukas, 1996), (Moukas, 1997), (Moukas i Maes 1998). Zaproponowany system złożony jest z populacji współdziałających i współzawodniczących agentów. W systemie działają dwa rodzaje agentów podlegających ewolucji – agenci filtrujący informacje oraz agenci odkrywający informacje. Agent filtrujący informacje odpowiedzialny jest za

personalizację systemu, za śledzenie zainteresowań użytkownika i zmian w tych zainteresowaniach oraz za adaptację systemu zgodnie z zainteresowaniami użytkownika. Agent odkrywający informacje odpowiedzialny jest za obsługę źródeł informacji, tj. dokumentów sieciowych, adaptację do tych źródeł informacji oraz za znajdowanie i przesyłanie użytkownikowi aktualnych informacji zgodnych z zainteresowaniami. Agent filtrujący informacje pełni funkcję bardzo wyspecjalizowanego filtru, który ma zastosowanie tylko do bardzo wąskiej dziedziny zainteresowań użytkownika. Gdy zmieniają się zainteresowania, agenci obsługujący dawne zainteresowania zostają usuwani, a tworzeni są nowi. Nowi agenci ukierunkowani są na nowe zainteresowania dzięki zastosowaniu mechanizmów ewolucji i selekcji naturalnej. Ewolucja agentów kontrolowana jest przez współczynnik poziomu przystosowania pojedynczego agenta (ang. *individual fitness*) oraz przez współczynnik poziomu przystosowania całego systemu (ang. *overall fitness*) (Moukas, 1996). Na podstawie aktualnego poziomu przystosowania, agenci poddawani są rankingowi. Potomkowie mogą zostać utworzeni tylko dla określonej liczby agentów znajdujących się na początku listy rankingowej uwzględniającej całą populację. Liczba agentów, którzy mogą mieć potomków zależy wprost proporcjonalnie od liczby agentów, którzy zostaną usunięci ze względu na niski poziom przystosowania. Liczba ta jest zmienna i zależy od poziomu przystosowania całego systemu. Jeśli poziom przystosowania całego systemu zmniejsza się, przyspieszana jest ewolucja, aby przyspieszyć adaptację systemu do nowych zainteresowań użytkownika. Natomiast jeśli poziom przystosowania całego systemu zwiększa się, ewolucja jest zatrzymywana na stałym poziomie. Nowi agenci tworzeni są na podstawie operacji mutacji lub krzyżowania, albo z wykorzystaniem obu tych operacji. Wymienione operacje stosowane są tylko do genotypu agenta. Genotypem agenta filtrującego informacje jest ważony wektor słów kluczowych wyznaczonych z dokumentów interesujących dla użytkownika. Ewolucji nie podlega druga część agenta – fenotyp. Fenotyp agenta filtrującego informacje zawiera informacje o poziomie przystosowania agenta, o utworzeniu agenta przez użytkownika lub automatycznie (dla pierwszego przypadku dokumenty prezentowane przez agenta będą z większą pewnością odpowiadały zainteresowaniom użytkownika), oraz kody wykonywalności, które umożliwiają agentowi komunikowanie się z innymi agentami. Fenotyp jest traktowany jak swego rodzaju wzorzec, który zostaje wypełniony genotypem i następnie zastosowany.

Problematyka metod maszynowego uczenia się jest związana z technikami klasyfikacji, które również znalazły zastosowanie w procesie konstruowania modelu użytkownika (Cover i Hart, 1967). Doux, Laurent i Nadal zastosowali wariant algorytmu K–średnich (ang. *K–Means*) do klasyfikacji danych symbolicznych reprezentujących zachowania użytkowników. Zachowanie definiowane jest jako zbiór par – {środowisko, akcja}. Zastosowana metoda klasyfikacji, dzieli zbiór zachowań na K klas i wyznacza K typowych zachowań. Profil użytkownika zdefiniowany zostaje

przez zachowania prototypowe, czyli takie które są charakterystyczne dla użytkowników. (Doux i inni, 1997).

Metody oparte na koncepcji stereotypów

Trzecia grupa, którą można wyróżnić spośród metod wykorzystywanych do konstruowania modelu użytkownika bazuje na technikach opartych na koncepcji stereotypów. W literaturze dotyczącej problematyki modelowania użytkownika, stereotyp jest opisem użytkownika należącego do pewnej klasy użytkowników. Użytkownicy należący do jednej klasy posiadają wspólne zainteresowania, opisane zbiorem atrybutów (Rich, 1983). Stereotypy mają strukturę hierarchicznej bazy wierzy. Według koncepcji stereotypów, użytkownik zawsze opisany jest przez co najmniej jeden stereotyp, ponieważ atrybuty pozwalające przypisać użytkownikowi stereotyp są faktami – danymi o użytkowniku, które nie ulegają zmianie. Np. użytkownik pracujący w departamencie zarządzania przedsiębiorstwem zawsze będzie zainteresowany zagadnieniami związanymi z sukcesem w zarządzaniu, zarządzaniem zasobami ludzkimi (Benaki i inni, 1997a). Stereotypy są niezmienne w czasie. Najczęściej w systemie zgromadzony jest z góry określony zbiór stereotypów opisujących klasy użytkowników, którzy korzystają z systemu. Podczas pierwszej sesji pracy z systemem, model pojedynczego użytkownika tworzony jest na podstawie stereotypu. Stereotyp odpowiadający zainteresowaniom użytkownika zostaje przyporządkowany użytkownikowi na podstawie przynależności użytkownika do pewnej grupy użytkowników, np. departamentu firmy, grupy społeczności akademickiej, grupy zaawansowania w pracy z systemem. Modele użytkowników, którzy należą do jednej grupy zawierają ten sam stereotyp. W hypermedialnym systemie informacji miejskich AVANTI, model użytkownika powstaje przez przypisanie nowemu użytkownikowi cech opisujących grupę użytkowników hipermediów dostępnych w systemie, np. turystów, osób niewidomych (Fink i inni, 1997) lub przypisanie użytkownikowi zainteresowań związanych z siecią WWW na postawie stereotypu użytkownika, który został ustalony przez eksperta (Ambrosini i inni, 1997). W prototypowym systemie generowania dokumentów hypermedialnych GeNet, model użytkownika utworzony jest z wykorzystaniem prostego stereotypu, zawierającego takie dane o użytkowniku jak: wiek, poziom wykształcenia, doświadczenia w dziedzinie zastosowania aplikacji i w dziedzinie hipermediów (De Carolis i Pizzutilo, 1997). Benaki, Karkaletsis i Spyroupoulus zaproponowali schemat modelowania użytkownika, który został wykorzystany w projekcie ECRAN wydobywania informacji (ang. *information extraction*) z sieci WWW. Model użytkownika powstaje na podstawie danych osobowych użytkownika (tj. imienia, nazwiska, miejsca pracy, nazwy departamentu), stereotypu opisującego zainteresowania użytkownika pracującego w określonym departamencie przedsiębiorstwa oraz ocenionej (interesująca, nieinteresująca, bez

znaczenia) przez użytkownika wiedzy dziedzinowej, zawierającej kategorie dziedzinowe. W dziedzinie wiadomości o przedsiębiorstwie, kategoriami tymi są np. zarząd przedsiębiorstwa, rodzaje spółek, wyniki przedsiębiorstwa (Benaki i inni, 1997).

W zastosowaniu stereotypów dla bardzo szerokiego grona użytkowników WWW problematyczne wydaje się ustalanie skończonego zbioru stereotypów przez grupę ekspertów z racji na liczne różnorodne zainteresowania użytkowników, różne dziedziny wyszukiwania oraz potrzeba ogromniej liczby ekspertów ustalających stereotypy i przypisujących je dla użytkowników.

3.5. Metody korzystania z modelu użytkownika

W poprzednich podrozdziałach opisane zostały cele modelowania użytkownika oraz rodzaj informacji jakie mogą być gromadzone w modelu o użytkowniku. Na podstawie informacji zgromadzonych w modelu, system może dostosować swoje działanie do użytkownika, jeśli uwzględnione zostaną charakterystyki poszczególnych użytkowników. W tym celu, wykorzystywane są różne podejścia i techniki wnioskowania bazujące m.in. na osiągnięciach sztucznej inteligencji. Techniki wykorzystania modelu użytkownika w celu uzyskania adaptacji systemu można podzielić na kilka grup: techniki bazujące na logice, na twierdzeniach rachunku prawdopodobieństwa (metodach Bayesowskich), techniki wykorzystujące maszynowe uczenie, techniki bazujące na metodach reprezentacji wiedzy oraz techniki wykorzystujące reguły i heurystyki.

Metody bazujące na wnioskowaniu

Do pierwszej grupy można zaliczyć klasyczne techniki bazujące na logice. W systemie wyszukiwani informacji MIRACLE, model użytkownika zawiera historię interakcji z systemem, reprezentowaną jako dialog pomiędzy użytkownikiem a systemem (Stein i inni, 1997). Kolejne elementy dialogu interpretowane jako akty dialogu (ang. *dialog acts*). Ciąg aktów dialogu tworzy strategię wyszukiwania informacji. W systemie MIRACLE wykorzystano skrypty. Skryptem jest reprezentacja akcji, które są konieczne lub potrzebne do wykonania określonej strategii wyszukiwania lub innych powiązanych ze strategią zadań. Skrypty zawierają wszystkie możliwe akcje systemu i wszystkie zalecane akcje użytkownika, które mogą pojawić się na różnych poziomach prowadzenia dialogu pomiędzy systemem a użytkownikiem podczas procesu wyszukiwania. Skrypty mogą wywoływać podskrypty, które są odpowiedzialne za wykonanie podzadań. Skrypty reprezentowane są za pomocą formalizmu rekurencyjnej sieci przejść (ang. *recursive transition network – RTN*). Warunki początkowe określają, kiedy akt dialogu jest dostępny, a warunki końcowe gwarantują, że system wykona niezbędne akcje w procesie wyszukiwania. Podczas prowadzenia

dialogu wyszukiwawczego, ze strony użytkownika może pojawić się akt dialogu nieprzewidziany w skrypcie, według którego prowadzony jest aktualny dialog. Dla systemu jest to akt niejednoznaczny, po którym prowadzony podczas wyszukiwania dialog nie może być już kontynuowany według standardowego skryptu. W takiej sytuacji system uruchamia komponent odpowiedzialny za wnioskowanie abdukcyjne (ang. *abductive dialog component – ADC*). Analizując historię dialogu komponent ADC korzysta z wnioskowania abdukcyjnego w celu wygenerowania możliwych interpretacji tego nieprzewidzianego aktu dialogu, a następnie proponuje użytkownikowi kilka następnych akcji w dialogu wyszukiwawczym.

Metody bazujące na technikach statystycznych

Druga grupa technik stosowanych w procesie wykorzystania modelu użytkownika do adaptacji systemu bazuje na metodach Bayes'owskich. Corbett i Bhatnagar w systemie ACT Programming Tutor, wspomagającym naukę programowania w językach Lisp, Pascal i Prolog, zastosowali rozwiązania umożliwiające śledzenie poszerzania się wiedzy studenta podczas nauki. Przewidywanie przez system poziomu biegłości i opanowania reguł programowania przez użytkownika–studenta na podstawie poprzednich czynności studenta i poziomu jego wiedzy deklaratywnej bazuje na metodach wykorzystujących rachunek prawdopodobieństwa. W systemie zastosowano komponent śledzenia przyrostu wiedzy studenta (ang. *knowledge tracing*). Proces śledzenia przyrostu wiedzy studenta dotyczy wiedzy proceduralnej. Nauczane w systemie zasady programowania reprezentowane są w postaci reguł produkcji. Reguła produkcji, odzwierciedlająca jedną zasadę programowania, której nauczył się student, zostaje zapisana w modelu użytkownika. W ten sposób model reprezentuje wiedzę programistyczną posiadaną przez użytkownika. W systemie przyjęto dwustanowy model uczenia się przez studenta. Reguła produkcji może być w stanie „nauczona" oznaczającym, że użytkownik nauczył się pewnej zasady programowania lub w stanie „nienauczona". Podczas nauki studenta, system szacuje prawdopodobieństwo p(L) zdarzenia, że reguła przyjęła stan „nauczona". Jeśli zaistnieje możliwość zastosowania danej reguły produkcji, czyli użytkownik w procesie nauki będzie miał za zadanie wykorzystać zasadę programowania reprezentowaną przez daną regułę produkcji, oszacowanie prawdopodobieństwa p(L) wykonane przez system może zostać weryfikowane. Weryfikacja odbywa się na podstawie faktu, że użytkownik poprawnie lub niepoprawnie zastosował zasadę programowania, której reprezentacją jest reguła produkcji. Prawdopodobieństwo, że reguła produkcji jest w stanie „nauczona" zależy od sumy prawdopodobieństwa a posteriori zdarzenia, że reguła jest już w stanie „nauczona" oraz prawdopodobieństwa przejścia do stanu „nauczona", jeśli jeszcze nie jest w tym stanie. Podczas wykonywania zadań programistycznych przez użytkownika, prawdopodobieństwo przejścia reguły produkcji ze stanu „nienauczona" do stanu

„nauczona" nie jest zależne od faktu poprawnego lub niepoprawnego zastosowania tej reguły przez użytkownika–studenta. Schemat Bayesowski został wykorzystany do oszacowania prawdopodobieństwa a posteriori, które zależy od faktu, czy student wykonał poprawnie akcję polegającą na zastosowaniu określonej zasady programowania (Corbett i Bhatnagar, 1997).

W systemie wspomagania decyzji, zaproponowanym przez Conati, Gertner, VanLehn i Druzdzel, dane zgromadzone w modelu użytkownika wykorzystane zostały w sieci Bayesowskiej do przewidywania zachowania ucznia podczas rozwiązywania problemu (Conati inni, 1997). Albrecht, Zukerman, Nicholson i Bud zaproponowali wykorzystanie dynamicznej sieci Bayesowskiej (ang. *Dynamic Bayesian Network*) do przewidywania zmieniających się w czasie akcji i położenia opisujących użytkownika. Środowiskiem testowym wykorzystanym przez autorów jest gra komputerowa, w której wielu graczy rywalizuje o ograniczone zasoby, aby osiągnąć określony cel. Gracze mogą przemieszczać się pomiędzy pomieszczeniami wykonując akcje dążą do określonego celu. Informacje o akcjach i położeniu lub sekwencji akcji i położeń, które doprowadziły do celu, pozyskiwane są na podstawie poprzednich, zakończonych pomyślnie celów. Informacje te pozyskiwane są podczas fazy nauki i zamodelowane przy wykorzystaniu dynamicznej sieci Bayesowskiej. Węzłami sieci są zmienne opisujące pewną dziedzinę, czyli akcje, położenia i cele. Podczas fazy testów sieć Bayesowska służy do przewidywania przyszłych celów, akcji i położeń gracza. Jeśli gracz wykona pewną akcję, system uaktualnia sieć Bayesowską, a dokładnie prawdopodobieństwo tego, że gracz podejmuje próbę osiągnięcia każdego z celów, wykonania każdej z akcji oraz przejścia do każdego z możliwych położeń (Albrecht i inni, 1997).

Billsus i Pazzani zaproponowali model inteligentnego agenta, uczącego się zainteresowań użytkownika dotyczących wiadomości radiowych (Billsus i Pazzani, 1999). Zainteresowania użytkownika reprezentowane są przez dwa modele: model zainteresowań długoterminowych (ang. *long–term model*) oraz model zainteresowań krótkoterminowych (ang. *short–term model*). W modelu zainteresowań krótkoterminowych reprezentowana jest aktualna ścieżka zainteresowań (np. zainteresowanie wiadomościami dotyczącymi wydarzeń z jednego dnia). W modelu długoterminowym reprezentowane są ogólne preferencje dotyczące wiadomości radiowych, zebrane na podstawie obserwacji czynionych przez system przez dłuższy okres czasu. Jeśli nowa wiadomość nie jest związana z aktualną ścieżką zainteresowań to system nie jest w stanie zaklasyfikować tej wiadomości jako interesującej lub nieinteresującej na podstawie posiadanego modelu zainteresowań krótkoterminowych. Zamodelowanie ogólnych zainteresowań użytkownika umożliwia podjęcie decyzji, czy nowa wiadomość, która pojawiła się, a nie jest związana z aktualnie wskazaną ścieżka zainteresowań, może być interesująca dla użytkownika. Zaproponowano, aby powyższy problem rozwiązać wykorzystując algorytm ucznia się probabilistycznego – naiwny

klasyfikator Bayesowski (ang. *naive Bayesian classifier*). Zainteresowania ogólne reprezentowane są w modelu zainteresowań długoterminowych przez boolowskie wektory cech. Cechami są terminy ręcznie wybrane przez badaczy. Wybrali oni ok. 200 słów najczęściej pojawiających się w wiadomościach dotyczących polityki, biznesu, rozwoju technologii, przestępstw, kataklizmów i sportu. Billsus i Pazzani przyjęli założenie, że cechy są wzajemnie niezależne, jeśli należą do opisu klasy (klasa interesująca, nieinteresująca). Prawdopodobieństwo, że wiadomość należy do klasy *j* opisanej przez wartości cech, $p(klasy_j \mid f_1, f_2, \ldots, f_n)$ jest równe $p(klasy_j)\prod_i^n p(f_i \mid klasy_j)$. Prawdopodobieństwa $p(klasy_j)$ oraz $p(f_i|klasy_j)$ zostały określone na podstawie danych uczących, a dokładnie liczby wystąpień słów i klas w danych uczących. Ci sami autorzy wcześniej zastosowali techniki bazujące na metodach Bayesowskich do identyfikacji interesujących stron WWW (Billsus i Pazzani, 1997). Wadą zaproponowanego przez Billsusa i Pazzaniego rozwiązania jest konieczność ręcznego wyznaczenie terminów, które są cechami reprezentującymi długoterminowe zainteresowania użytkownika. W realiach środowiska WWW działanie takie jest niemożliwe do zrealizowania dla wszystkich użytkowników sieci WWW.

Metody wykorzystujące techniki maszynowego uczenia

Trzecia grupa technik wykorzystywania modelu użytkownika w celu uzyskania adaptacji systemu to techniki bazujące na maszynowym uczeniu się. Doux, Laurent i Nadal zaproponowali wykorzystanie klasyfikacji K–średnich (ang. *K–Means classification*) do określenia działania bliskiego do tego, które wybrał użytkownik. Zaproponowane rozwiązanie wykorzystane zostało do uproszczenia interakcji pomiędzy człowiekiem i maszyną przez zautomatyzowanie nadawania ustawień pewnemu urządzeniu, gdy ustawienia te wcześniej nadawane były ręcznie. Autorzy pracy zakładają, że interakcja pomiędzy maszyną i użytkownikiem ma miejsce, gdy użytkownik wykonuje akcję w celu nadania pewnych ustawień urządzeniu, odpowiadających preferencjom użytkownika. Wykonywana akcja zależy od warunków zewnętrznych, nazywanych przez autorów środowiskiem. Na przykład podczas oglądania telewizji akcja ustawienia kontrastu i jasności odbiornika telewizyjnego zależy m.in. od oświetlenia pomieszczenia. W pracy podjęto się scharakteryzowania działania użytkownika w celu automatycznego generowania akcji, która będzie najbardziej odpowiadać tej, którą wybrałby użytkownik. Użytkownicy najczęściej zaakceptują ustawienia urządzenia, które będą odpowiednio bliskie tym, które wybraliby sami. Dlatego też, działania użytkowników zostały poklasyfikowane tak, aby każdemu użytkownikowi odpowiadały akcje zdefiniowane dla prototypowego działania. Zaproponowany schemat klasyfikacji zachowań użytkowników oparty jest na algorytmie klasyfikacji K– średnich, zaadoptowanym do danych symbolicznych

reprezentujących działania użytkownika. Aby zdefiniować działania prototypowe, wszystkie zachowania użytkowników zostały pogrupowane w K klas i wyznaczono K zachowań typowych. Algorytm uczenia się bez nadzoru, działający dla danych symbolicznych, wykorzystany został do wygenerowania reprezentanta każdej klasy (Doux i inni, 1997).

Algorytm grupowania tekstów metodą najbliższego sąsiedztwa (ang. *Nearest–Neighbour Algorithm*) wykorzystany został do wyznaczenia modelu krótkoterminowych zainteresowań użytkownika dotyczących codziennych wiadomości (Billsus i Pazzani, 1999). Billsus i Pazzani zaprezentowali architekturę inteligentnego agenta informacyjnego, który będzie jednym z komponentów inteligentnego radia. Agent, uczący się zainteresowań użytkownika i wyszukujący wiadomości, ma możliwość syntetyzowania mowy, dzięki czemu użytkownik może słuchać wiadomości z jednego z sześciu dostępnych programów informacyjnych. Podczas słuchania użytkownik w dowolnym momencie przekazuje informację zwrotną dotyczącą słuchanej wiadomości. Informacja zwrotna nie jest tylko dwustanową opcją wiadomość interesująca/nieinteresująca. Użytkownik może dodatkowo zaznaczyć swoje preferencje wskazując, że zna już daną wiadomość lub chce dowiedzieć się więcej o wiadomości. Podczas pracy, po zakończeniu początkowej fazy nauki, agent może na żądanie użytkownika utworzyć indywidualny program informacyjny (dziennik). Celem takiego procesu adaptacji jest ułożenie sekwencji codziennych wiadomości według zainteresowania wyrażonego przez użytkownika. Agent uczy się zainteresowań użytkownika na podstawie dwóch zaproponowanych modeli użytkownika. Jeden model reprezentuje zainteresowania krótkoterminowe, drugi – długoterminowe. Rozróżnienie pomiędzy krótkoterminowym, a długoterminowym modelem ma istotne znaczenie w dziedzinach, w których istotny jest czynnik czasu (Chiu i Webb, 1998). Jeśli użytkownik jest zainteresowany pewną tematyką, najczęściej chce uzyskać kolejne wiadomości związane z tym tematem. Można powiedzieć, że użytkownik podąża pewną ścieżką tematyczną. Model krótkoterminowych zainteresowań użytkownika reprezentuje najbardziej aktualne zainteresowania dotyczące wiadomości z poprzednich kilku dni. Model jest budowany na podstawie 100 najbardziej aktualnych wiadomości wskazanych przez użytkownika, a więc zawiera informacje o aktualnych wydarzeniach związanych z zainteresowaniami użytkownika. Dzięki tym informacjom możliwe jest zidentyfikowanie innych wiadomości dotyczących tej samej ścieżki tematycznej oraz wiadomości, które już są znane użytkownikowi. Do zrealizowania tych zadań Billsus i Pazzani zastosowali algorytm klasyfikowania metodą najbliższego sąsiedztwa, już wcześniej wykorzystywany do klasyfikacji tekstów. Algorytm ten przechowuje wszystkie przykłady uczące, którymi są wiadomości ocenione przez użytkownika w skali od 0 do 1. Wiadomości reprezentowane są przez wektory, których współrzędne są wagą każdego słowa w wiadomości. Waga wyznaczana jest na podstawie schematu *tf–idf*. Aby zaklasyfikować nową wiadomość, algorytm dokonuje porównania danej

wiadomości ze wszystkimi sklasyfikowanymi wcześniej wiadomościami wykorzystując zdefiniowaną miarę podobieństwa. Dla tekstów w języku naturalnym przyjęto, powszechnie stosowaną w wyszukiwaniu informacji, miarę kosinusową (Salton, 1998). W efekcie określana jest najbardziej podobna wiadomość („najbliższy sąsiad") lub *k* najbardziej podobnych wiadomości („najbliższych sąsiadów") oraz wiadomości już znane użytkownikowi. Zaproponowany model krótkoterminowych zainteresowań użytkownika może reprezentować wiele dziedzin zainteresowań i łatwo adaptuje się do nowych zainteresowań. Podstawową zaletą wykorzystania do klasyfikowania wiadomości metody najbliższego sąsiedztwa jest możliwość określenia spośród nowych wiadomości tych, które należą do jednej ścieżki tematycznej na podstawie tylko jednej wiadomości wcześniej ocenionej przez użytkownika. Mankamentem przyjętego rozwiązania jest jednak obciążenie użytkownika obowiązkiem oceniania każdego przykładu uczącego oraz nadania mu wagi w skali 0 do 1, co wydaje się szczególnie kłopotliwe dla dużego zbioru przykładów uczących.

Metody bazujące na technikach reprezentacji wiedzy

Czwarta grupa technik stosowanych w procesie wykorzystania modelu użytkownika do adaptacji systemu bazuje na metodach reprezentacji wiedzy. Ambrosini, Cirillo i Micarelli zaproponowali architekturę adaptującego się do potrzeb użytkownika systemu filtrowania informacji w sieci WWW. System wykorzystuje sieć semantyczną w procesie filtrowania informacji, w celu uzyskania wartości relewancji dokumentu do zainteresowań użytkownika. Opisany algorytm MAF (ang. *Matching Algorithm for Filtering*) przypisuje każdemu wskazanemu przez użytkownika dokumentowi wagę odzwierciedlającą podobieństwo dokumentu do modelu użytkownika i pytania. Oprócz, tradycyjnie stosowanej w dziedzinie wyszukiwania informacji, metody określania podobieństwa na podstawie miary kosinusowej wektorów dokumentu i modelu użytkownika oraz pytania, dodatkowo w procesie obliczania podobieństwa uwzględniane są relacje semantyczne pomiędzy terminami znajdującymi się w dokumencie. Wykorzystując wiedzę zgromadzoną w sieci semantycznej, algorytm ten może identyfikować tematykę, która opisana jest przez kilka słów kluczowych (np. *network oriented languages)* i określić podobieństwo pomiędzy dokumentem a modelem i pytaniem. Sieć semantyczna tworzona jest w oparciu o informację zwrotną przekazaną przez użytkownika dla każdego z dokumentów, które zostały przekazane jako odpowiedź w wyniku procesu filtrowania informacji. Informacja zwrotna, reprezentacja dokumentu oraz pytanie są wykorzystane do modyfikacji modelu użytkownika. Modyfikacja modelu polega na dodaniu nowych terminów oraz uaktualnieniu wag terminów istniejących lub na usunięciu terminów, których waga jest poniżej pewnego progu. W systemie skorzystano również z, utworzonej specjalnie na potrzeby badań, bazy *Terms DataBase (TBS*). Baza posłużyła do określenia

semantycznego znaczenia terminu. Jeśli znaleziony przez system w dokumencie termin znajduje się w bazie *TBS*, zostaje on dodany do modelu użytkownika. Jeśli termin nie występuje się w bazie, wykorzystane zostają informacje zawarte w sieci semantycznej. Centralny węzeł sieci jest terminem, który reprezentuje potencjalne zainteresowania użytkownika. Powiązane z nim węzły reprezentują terminy współwystępujące z tym terminem w jednym dokumencie. Termin, który nie występował w bazie, a znajduje się w dokumencie wskazanym przez użytkownika, zostaje dołączony do modelu użytkownika jako termin współwystępujący. Termin ten zostaje dołączony do terminów modelu znajdujących się w dokumencie na podstawie ważonego powiązania semantycznego uzyskanego z sieci semantycznej. Taki model użytkownika umożliwia systemowi rozróżnienie na podstawie kontekstu pomiędzy różnymi znaczeniami terminu. Wiedza na temat dziedziny zainteresowań użytkownika jest pozyskiwana dynamicznie jako wynik wnioskowania na podstawie wymagań użytkownika dotyczących informacji (Ambrosini i inni, 1997). Problemem w zaproponowanym rozwiązaniu jest nałożony na użytkownika obowiązek ocenienia każdego dokumentu przekazanego mu w odpowiedzi przez system. Tym bardziej, że ocena ta jest podstawą do budowania sieci semantycznej, której zastosowanie jest kluczowe w rozwiązaniu zaproponowanym przez Ambrosiniego, Cirilla i Micarelliego

Metody wykorzystujące reguły i heurystyki

Najobszerniejsza grupa technik stosowanych w procesie wykorzystania modelu użytkownika do adaptacji systemu bazuje na regułach, heurystykach i procedurach jakościowych. Maglio i Barrett zastosowali procedury przetwarzania zapisanej historii odwiedzania stron WWW przez użytkownika. Procedury określone zostały na podstawie wniosków z przeprowadzonych eksperymentów. W eksperymentach zaobserwowali i wykazali, że użytkownicy szukający informacji w sieci WWW nie planują wyszukiwania, w pełnym tego słowa znaczeniu, ale opierają się na heurystykach oraz na lokalnym kontekście poszukiwanej informacji (sekwencji stron istotnych – „kamieniach milowych"). Autorzy sugerują, że użytkownicy szukający informacji przedkładają przeglądanie (ang. *browsing*) nad ukierunkowane i ustruktualizowane wyszukiwanie (ang. *searching*). Bazując na poczynionych obserwacjach, Maglio i Barrett zaproponowali mechanizm tworzenia i wykorzystania modelu użytkownika wspomagającego wyszukiwanie informacji w sieci WWW. Model użytkownika zawiera procedury charakterystyczne oraz strony istotne. Wykorzystywany jest przez dwóch osobistych agentów internetowych. Pierwszy z nich identyfikuje, na podstawie modelu, powtarzane przez użytkownika schematy przeglądania oraz podpowiada zbliżone schematy przeglądania nowym użytkownikom. Zidentyfikowany schemat postępowania zostaje następnie wykorzystany do utworzenia skrótu, dzięki któremu podczas następnego przeglądania użytkownik może przemieścić

się bezpośrednio z punktu startowego przeglądania do strony istotnej opisanej przez adres URL. Drugi agent, który ma doprowadzić do znalezienia informacji istotnej, identyfikuje sekwencje istotnych stron w procesie przeglądania. Jest to równoznaczne z identyfikowaniem indywidualnej historii interakcji użytkownika, czyli nieprzerwanej ścieżki przechodzenia pomiędzy stronami, którą podążał użytkownik (Maglio i Barrett, 1997), (Barrett i inni, 1997).

W systemie uczącym model użytkownika wykorzystywany jest w oparciu o zdefiniowane reguły. Reguły te umożliwiają proponowanie przez system kolejnej strony WWW, którą powinien odwiedzić użytkownik (Weber i Sprecht, 1997). Również Strachan, Anderson, Sneesby i Evans wykorzystują model użytkownika bazując na zdefiniowanych regułach. Reguły określają warunki udostępnienia nowym użytkownikom mniej skomplikowanego interfejsu oraz większej pomocy podczas wykonywania zadań. Opisane rozwiązanie zastosowane zostało w komercyjnym systemie planowania inwestycji finansowych i podatków *TIMS* (ang. *Tax and Investment Management Strategizer*), który wykorzystywany jest zarówno przez doświadczonych doradców finansowych, jak i doradców początkujących. W modelu użytkownika zawarta jest m.in. informacja o doświadczeniu użytkownika jako doradcy finansowego, doświadczeniu w pracy z systemem *TIMS* oraz o pracy z systemem Windows. Na podstawie informacji o poziomie doświadczenia posiadanego przez doradcę finansowego, uruchamiana jest demonstracja wspomagającą użytkownika podczas wykonania oceny finansowej podmiotu gospodarczego lub też dostarczana jest dodatkowa pomoc podczas procesu wprowadzania danych finansowych, analizy sytuacji finansowej oraz tworzenia planu finansowego w postaci podpowiedzi następnego kroku do wykonania (Strachan i inni, 1997). Wykorzystanie modelu użytkownika zaproponowane przez De Carolis i Pizzutilo również opiera się na stosowaniu odpowiednich reguł. Reguły te łączącą informacje o użytkowniku zgromadzone w modelu użytkownika z parametrami generowania hipertekstu, jak np. poziom szczegółowości informacji o przedmiocie nauki, sposób prezentacji strony (zastosowane media, rozmiar i rozmieszczenie elementów strony), sposób przechodzenia pomiędzy stronami (przechodzenie w głąb lub wszerz sterowane przez system, przejście wszystkich lub części stron hipertekstowych sterowane przez system, eksploracja stron hipertekstowych sterowana przez użytkownika). W procesie wykorzystania modelu użytkownika reguły zastosowane zostały również do generowania dokumentów hipermedialnych w systemie edukacyjnym (De Carolis i Pizzutilo, 1997). Innymi przykładami stosowania reguł w procesie wykorzystania modelu użytkownika do adaptacji systemu są: system wspomagający efektywne tworzenie wykresów (Gutkauf i inni, 1997), system PHelpS (Collins i inni, 1997), czy system wspomagający współpracę pomiędzy uczniami (Bull i Smith, 1997). W systemie wspomagającym tworzenie wykresów reguły uwzględniają przechowywane w modelu użytkownika preferencje i możliwości użytkowników (np. rozróżnianie kolorów przez

użytkownika), które są adekwatne do problemu projektowania wykresów. Na podstawie informacji zawartych w modelu system konstruuje wykres (Gutkauf i inni, 1997). W systemie PHelpS, wspomagającym znalezienie współpracownika posiadającego wiedzę na temat zadania wykonywanego przez użytkownika, reguły zastosowane zostały do znalezienia profilu potencjalnego współpracownika, który może udzielić pomocy użytkownikowi (Collins i inni, 1997). Natomiast w systemie PairSM wprowadzone zostały heurystyki, na podstawie których sugeruje jaka zalecaną formę współpracy pomiędzy uczniami podczas nauki (Bull i Smith, 1997). Reguły wyszukiwania w dokumentach hipertekstowych, uwzględniające kontekst, zastosowane zostały w procesie wykorzystania modelu użytkownika do adaptacji systemu hipertekstowego. Dla sieci WWW, kontekstem danej strony WWW są strony zawierające odsyłacze do danej strony (Staff, 1997).

Kolejnym przykładem metod wykorzystania modelu użytkownika do adaptacji systemu są podejścia opisane w pracach Moukas'a i Maes. Zastosowane przez nich reguły, bazujące na technikach stosowanych w wyszukiwaniu informacji oraz heurystykach, zastosowane zostały w procesie wykorzystania modelu użytkownika. Moukas'a i Maes zaproponowali ekosystem agentów filtrujących i odkrywających informacje. Agent filtrujący informacje wybiera dokument, który jest najbliższy do wektora reprezentującego zainteresowania użytkownika, zawartego w genotypie. Do określenia dokumentu najbardziej bliskiego do genotypu agenta, reprezentowanego przez ważony wektor słów kluczowych, wykorzystana została miara kosinusowa. Na podstawie tej miary określany jest poziom pewności, że dokument będzie interesujący dla użytkownika. Jeśli wektor reprezentujący dokument jest identyczny jak wektor zawarty w genotypie agenta, przyjmowane jest, że poziom pewności agenta jest równy 1.0. W następnym etapie system decyduje, które z dokumentów przekazanych przez agentów filtrujących informacje zostaną zaprezentowane użytkownikowi. Decyzję tę system podejmuje na podstawie rankingu przekazanych dokumentów odpowiedzi. Na początku listy znajdą się dokumenty, których reprezentacja jest najbliższa ważonemu wektorowi słów kluczowych należących do genotypu agenta o najwyższy poziom przystosowania (ang. *fitness*) (Moukas, 1996), (Moukas, 1997), (Moukas i Maes 1998).

3.6. Modelowanie użytkownika w systemach wyszukiwania informacji

Tradycyjnie przyjmuje się, że model użytkownika to reprezentacja preferencji podanych wprost przez użytkownika. Z definicją tą nie zgadzają się Bianchi–Berthouze, Berthouze i Kato, którzy opierając się na badaniach z nauki kognitywnej twierdzą, że model użytkownika nie może być konstruowany w oparciu o bezpośrednie, jawnie podane przez użytkownika informacje (Bianchi–Berthouze i inni, 1997), (Kim i inni, 2000). Autorzy ci postulują, aby konstruowanie modelu użytkownika opierało się w

głównej mierze na informacjach czerpanych z interakcji prowadzonej przez człowieka z systemem. W obszarze systemów wyszukiwania informacji istotne informacje uzyskiwane mogą być na podstawie czynności, jakie wykonał użytkownik ze znalezionym dokumentem. Ilość czasu poświęcona przez użytkownika na czytanie dokumentu może być jedną podstawowych informacji wpływających na ocenę, czy dokument jest interesujący dla użytkownika (Kelly i Belki, 2002). Wydrukowanie tego dokumentu świadczy, że jest on bardziej istotny, niż jeśli został by tylko zapisany na dysku. A taka czynność świadczy, że dokument jest bardziej istotny, niż jeśli byłby tylko przeczytany bez zapisania na dysku (Seo, Zhang 2000), (Nichols, 1997).

Bogate zestawienie interesujących pozycji literaturowych dotyczących tworzenia profilu użytkownika w oparciu o informację niejawną zawiera praca (Kelly i Teevan, 2003).

W klasycznych systemach wyszukiwania informacji przyjmowano początkowo, że potrzeba informacyjna użytkownika reprezentowana jest tylko przez pytanie zadane w określonym momencie czasu. Profil użytkownika, jako nowy element w systemach wyszukiwania informacji, pojawił się w systemach selektywnej dystrybucji informacji, w skrócie – systemach SDI. W systemach SDI profil zawierał informacje o zainteresowaniach i preferencjach użytkownika, które charakteryzowała niezmienność w pewnym dłuższym przedziale czasu oraz zostały ujawnione podczas wyszukiwania. Definiowanie profilu oraz korzystanie z niego w systemach SDI było znacznie prostsze niż jest teraz w internetowych systemach wyszukiwania informacji. W systemie SDI zmiany w bazie danych były ściśle określone, a pytanie użytkownika pozostawało bez zmiany pełniąc funkcję filtru (Daniłowicz, 1992). Natomiast w WWW różne są kolejno zadawane pytania i permanentnie zmienia się baza danych, czyli zasoby WWW, więc w konsekwencji musi dynamicznie zmieniać się również profil użytkownika (Aas, 1997), (Indyka-Piasecka i Daniłowicz, 2004).

3.6.1. Model użytkownika jako reprezentacja potrzeby informacyjnej

Omówiony w poprzednich podrozdziałach zakres problematyki modelowania użytkownika pokazuje, że jest to dziedzina mająca bardzo rozległe zastosowania. Różnorodność zastosowań wpłynęła na potrzebę korzystania z osiągnięć różnych dziedzin, m.in. sztucznej inteligencji, statystyki, psychologii, czy nauk kognitywnych. Problematyka dotycząca modelowania użytkownika w niniejszej pracy jest zawężona do modelowania użytkownika na potrzeby procesu wyszukiwania informacji.

Rozważmy, kiedy mamy do czynienia z procesem wyszukiwania informacji. Potrzeba wyszukania informacji powstaje, gdy użytkownik mający pewien problem lub zadanie do rozwiązania, czy cel do osiągnięcia, nie posiada wystarczającej wiedzy do

rozwiązania powstałej sytuacji. Aby rozwiązać taką problematyczną sytuację, użytkownik musi rozszerzyć swoją wiedzę przez sięgnięcie do zewnętrznych zasobów informacji. W klasycznym podejściu do wyszukiwania informacji, dostęp do zasobów informacji był możliwy tylko poprzez człowieka–pośrednika, obsługującego zasoby. Pośrednik pomagał i ułatwiał interakcję pomiędzy użytkownikiem a zasobami informacji. W latach obecnych, w dobie komputeryzacji i rozwoju Internetu, dostęp do zasobów informacji jest możliwy bez korzystania z pomocy pośrednika.

Z powyższych faktów wypływają cele, jakie powinien spełniać system, umożliwiający dostęp do informacji. System wyszukiwania informacji powinien wspomagać użytkownika na kilku poziomach. Na poziomie identyfikacji problemu wyszukiwawczego i osiągnięcia celu w postaci wyjaśnienia tego problemu. Na poziomie rozwiązania (ang. *resolution*) problematycznej sytuacji, w wyniku której użytkownik podjął proces wyszukiwania informacji. Na poziomie podnoszenia efektywności dostępu (interakcji) do zasobów informacyjnych oraz na poziomie określania informacji relewantnej.

W literaturze opisywane są dwa typy systemów umożliwiających dostęp do informacji: systemy wyszukiwania informacji oraz systemy filtrowania informacji (Belkin i Croft, 1992), (Macskassy, 2003). W systemach wyszukiwania informacji, poszukiwanie informacji przez użytkownika ma, najczęściej, charakter jednorazowych epizodów, wyszukiwań ad hoc, w których potrzeba informacyjna wyrażona jest krótkim pytaniem. W taki sposób rozumiane są tradycyjne systemy wyszukiwania informacji (Salton i McGill, 1983). W systemach filtrowania informacji, poszukiwanie informacji ma charakter wielokrotnego, powtarzającego się procesu. Potrzeba informacyjna wyrażona jest pytaniem złożonym z wielu terminów. Reprezentacja potrzeby informacyjnej w systemach filtrowania informacji nazywana jest profilem użytkownika (Daniłowicz, 1992).

Dla obu typów systemów wyszukiwania informacji można wymienić kilka wspólnych cech charakterystycznych. W systemach tych podstawowymi elementami są: użytkownik oraz zasoby informacji. W niektórych systemach istnieją również pośrednicy wspomagający użytkownika w procesie wyszukiwania. Proces wyszukiwania informacji jest procesem cyklicznym lub określonym przez kolejne kroki wyszukiwania. Jest to proces dynamiczny, w którym podczas interakcji zmienia się reprezentacja potrzeby informacyjnej użytkownika, a czasami również zasoby informacji.

Wymienić należy podstawowe problemy związane z korzystaniem przez użytkownika z systemu wyszukiwania informacji i będące w kręgu zainteresowań wielu prac badawczych. Zazwyczaj w systemach wyszukiwania informacji przyjęte jest założenie, że reprezentacja potrzeby informacyjnej jest formułowana przez użytkownika lub pośrednika wspomagającego użytkownika w procesie wyszukiwania informacji. W systemach tych nie jest uwzględniane zagadnienie trudności, jakie

stwarza użytkownikowi wyrażenie i sprecyzowanie potrzeby informacyjnej, a następnie przedstawienia jej w postaci reprezentacji akceptowalnej przez system wyszukiwania informacji. Najczęściej największe problemy w sformułowaniu poprawnej reprezentacji potrzeby informacyjnej mają użytkownicy o małym doświadczeniu, nie będący ekspertami w dziedzinie, w której dokonują wyszukiwania informacji. Do tej grupy zaliczyć można wielu spośród wszystkich użytkowników systemów wyszukiwani informacji. Bo czyż użytkownik nie korzysta z systemu najczęściej wtedy, gdy brakuje mu informacji na określony temat i właśnie nie jest jeszcze ekspertem w danej dziedzinie? Z powyższych rozważań można wysnuć wniosek, że wiedza, doświadczenie mają znaczący wpływ na ocenę informacji dostarczonej użytkownikowi przez system (np. na ocenę relewancji). Jednocześnie, zmienia się w czasie rozumienie przez użytkownika problemu informacyjnego oraz informacji, które są pomocne w rozwiązaniu tego problemu.

Przedstawione cechy charakteryzujące systemy wyszukiwania informacji oraz omówione problemy implikują określone podejście do wyszukiwania informacji. Wyszukiwanie informacji można traktować jako dialog, w którym użytkownik bierze udział przy tworzeniu reprezentacji swojej potrzeby informacyjnej. W systemie, informacje istotne dla użytkownika zmieniają się w czasie, dlatego też statyczna reprezentacja potrzeby informacyjnej nie jest wystarczająca. Dodatkowo, należy uwzględnić potrzeby informacyjne użytkownika kształtujące się zarówno na przestrzeni dłuższego czasu, jak i potrzeby krótkoterminowe. Implikuje to zastosowanie dwóch reprezentacji opisujących obie potrzeby lub jednej elastycznej reprezentacji uwzględniającej wymienione dwa rodzaje potrzeb.

W dziedzinie wyszukiwania informacji, proces tworzenia reprezentacji potrzeby informacyjnej użytkownika można przyjąć za równoważny z konstruowaniem modelu użytkownika. Dlatego też wyszukiwanie informacji jest z natury procesem, w którym dokonuje się modelowania użytkownika. Jednak wśród zagadnień dotyczących wyszukiwania informacji, nie poświęcano zbyt wiele uwagi problemowi jakości modelu użytkownika. Przyjmuje się zazwyczaj, że użytkownik formułując potrzebę informacyjną przedstawia swoje rzeczywiste zapotrzebowanie na informacje. Zazwyczaj jednak okazuje się, że reprezentacja potrzeby informacyjnej odbiega od wyobrażenia użytkownika o swojej potrzebie. Efektem tej rozbieżności są niezadowalające informacje, które użytkownik otrzymuje z systemu wyszukiwania informacji. Przyczynę takiej sytuacji upatruje się w niedoskonałości systemu wyszukiwania informacji, pomijając fakt, że istotny wpływ może mieć również formułowanie (reprezentacja) potrzeby informacyjnej. We współczesnych badaniach często podkreślane jest, że w wyszukiwaniu informacji reprezentacja potrzeby informacyjnej jest tylko częścią modelu użytkownika. Jednak prace nad reprezentacją innych aspektów modelu użytkownika jak np. informacji o ograniczeniach w liczbie elementów interfejsu, którymi jednocześnie może manipulować użytkownik

(Brusilowski i Schwartz, 1997), (Strachan i inni, 1997), zdolność postrzegania kolorów (Gutkauf i inni, 1997), czy poziom przyswojenia wiedzy w pewnej dziedzinie (De Carolis i Pizzutilo, 1997), (Corbett i Bhatnagar, 1997) są relatywnie mało rozwinięte ze względu na techniczne, praktyczne i eksperymentalne ograniczenia. W rozwiązaniach takich sugerowane jest, że dodanie innych, oprócz potrzeby informacyjnej, elementów do opisu użytkownika podniesie skuteczność wyszukiwania informacji z zastosowanie modelu. W procesie modelowania użytkownika w dziedzinie wyszukiwania informacji, problem reprezentacji samej potrzeby informacyjnej użytkownika powinien być traktowany priorytetowo. Propozycje rozszerzania modelu użytkownika wydają się być pewnego rodzaju wyjściem tymczasowym. W rozwiązaniu takim przyjęte jest założenie, że jeśli do uzyskania satysfakcji użytkownika nie wystarczy sama reprezentacja potrzeby informacyjnej, to może dodanie do modelu pewnych informacji dodatkowych o użytkowniku wpłynie na wzrost satysfakcji użytkownika. Założenie to wydaje się nie do końca uzasadnione, lecz skuteczne w konkretnych zastosowaniach. Jeśli jednak nie potrafimy dobrze opisać kluczowego elementu modelu użytkownika, czyli potrzeby informacyjnej, dodatkowe informacje wykorzystywane do opisu potrzeby informacyjnej są tylko uzupełnieniami. Uzasadnione jest więc prowadzenie dalszych prac badawczych nad rozwiązaniami, które umożliwią skonstruowanie reprezentacji potrzeby informacyjnej bardzo bliskiej rzeczywistemu zapotrzebowaniu użytkownika na informacje.

Podstawową reprezentacją potrzeby informacyjnej użytkownika w systemach wyszukiwania informacji jest pytanie zadane przez użytkownika. Opisane powyżej problemy użytkownika z wyrażeniem swojej potrzeby w postaci pytania skłoniły badaczy do poszukiwania rozwiązań, które wspomagałyby użytkownika w wyrażeniu tej potrzeby. Było to inspiracją dla metod modyfikacji pytania użytkownika. Modyfikacja pytania ma na celu doprowadzenie pytania do takiej postaci, która będzie lepiej odzwierciedlała rzeczywiste rzeczywiste zainteresowania użytkownika niż pytanie przed modyfikacją. W odpowiedzi na pytanie zmodyfikowane użytkownik uzyska więcej interesujących go informacji.

Można wyodrębnić trzy kategorie podejścia do modyfikacji pytania: podejście bazujące na sprzężeniu zwrotnym, podejście wykorzystujące informacje uzyskane po analizie dokumentów wyszukanych (zbiór tych dokumentów nazywany jest zbiorem lokalnym) oraz podejście wykorzystujące informacje uzyskane po analizie dokumentów kolekcji (zbiór tych dokumentów nazywany jest zbiorem globalnym).

3.6.2. Relewancyjne sprzężenie zwrotne

Relewancyjne sprzężenie zwrotne (ang. *relevance feedback*), lub krócej sprzężenia zwrotne, jest automatycznym procesem stosowanym w wyszukiwaniu informacji, umożliwiającym dynamiczne modelowanie potrzeby informacyjnej użytkownika.

Sprzężenie zwrotne, zaproponowane po raz pierwszy przez Rocchio dla wektorowego modelu wyszukiwania informacji, prowadzi do sformułowania lepszego pytania w kolejnych operacjach wyszukiwania (Rocchio, 1971). Podczas cyklu sprzężenia zwrotnego, system prezentuje użytkownikowi listę dokumentów wyszukanych, spośród których wybierane są przez użytkownika dokumenty relewantne. Z tych dokumentów relewantnych selekcjonowane są następnie terminy lub wyrażenia znaczące. W zmodyfikowanym pytaniu, wyselekcjonowanym terminom nadawane jest wyższe znaczenie poprzez zwiększenie wagi tych terminów. W efekcie, podczas wyszukiwania, tak utworzone nowe pytanie będzie bliższe dokumentom relewantnym, a dalsze dokumentom nierelewantnym.

Formułowanie pierwszego pytania nie jest zazwyczaj oczywiste dla użytkowników systemu wyszukiwania informacji. Przyczyna leży w braku szczegółowej wiedzy dotyczącej kolekcji, w której dokonywana jest operacja wyszukiwania oraz w nieznajomości środowiska systemu. Dlatego też, pierwsze sformułowanie pytania można potraktować jak swego rodzaju wyszukiwanie próbne, wykonane w celu uzyskania kilku dokumentów z przeszukiwanej kolekcji. Dokumenty te, przeanalizowane następnie przez użytkownika pod kątem zgodności z pytaniem (pod kątem relewancji), służą do utworzenia nowego, poprawionego pytania. Zmodyfikowane pytanie powinno dostarczyć większej liczby dokumentów relewantnych, niż pytanie zadane jako początkowe. Proces ten może się powtarzać w kolejnych operacjach wyszukiwania, aż do momentu, gdy potrzeba informacyjna użytkownika zostanie zaspokojona. Modyfikacja pytania obejmuje rozszerzenie pytania o terminy nowe oraz przypisanie nowych wag, wskazujących na ich istotność, terminom w pytaniu zmodyfikowanym (Salton i Buckley, 1990), (Baeza-Yates i Ribeiro-Neto, 1999).

W wektorowym modelu wyszukiwania informacji dokument d i pytanie q reprezentowane są przez ważone wektory przestrzeni n–wymiarowej, odpowiednio: $d = (d_1, d_2, ..., d_n)$, $q = (q_1, q_2, ..., q_n)$. Najczęściej stosowana miarą wyznaczania podobieństwa $Sim(d, q)$ pomiędzy pytaniem, a dokumentem jest miara kosinusowa:

$$Sim(d,q) = \frac{\sum_{i=1}^{n} d_i q_i}{\sqrt{\sum_{i=1}^{n} (d_i)^2 \sum_{i=1}^{n} (q_i)^2}}.$$

Zastosowanie sprzężenia zwrotnego w modelu wektorowym zakłada, że ważone wektory dokumentów zidentyfikowanych przez użytkownika jako relewantne (dla danego pytania) mają wysoką wartość miary podobieństwa $Sim(d, q)$ (Salton, 1971). Natomiast dla dokumentów nierelawantnych miara ta jest niska. U podstaw procesu sprzężenia zwrotnego leży założenie, że wartość podobieństwa pomiędzy pytaniem zmodyfikowanym a dokumentem relewantnym jest wyższa niż pomiędzy pytaniem przed modyfikacją a tym samym dokumentem.

W systemie wyszukiwania informacji, w którym miara podobieństwa wykorzystywana jest do określenia podobieństwa pomiędzy pytaniem i dokumentem, najlepsze pytanie wyszukujące dokumenty relewantne spośród wszystkich dokumentów kolekcji tworzone jest na podstawie następującego wzoru (Rocchio, 1971):

$$q_{opt} = \frac{1}{n} \sum_{\forall d \in D_{rel}} d - \frac{1}{N-n} \sum_{\forall d \notin D_{rel}} d \ ,$$

gdzie d – wektor dokumentu, n – liczba dokumentów relewantnych w kolekcji, N – liczba wszystkich dokumentów w kolekcji, D_{rel} – zbiór dokumentów relewantnych w kolekcji. Powyższa formuła nie może być jednak zastosowana dla pierwszego pytania, ponieważ zbiór dokumentów relewantnych nie jest znany przed wykonaniem operacji wyszukiwania. W rzeczywistości, zbiór dokumentów relewantnych jest dopiero poszukiwany. Powyższa formuła umożliwia tworzenie kolejnych pytań zmodyfikowanych na podstawie uzyskanej od użytkownika informacji o relewancji dokumentów wyszukanych. W takim przypadku, podczas analiz uwzględniane są wszystkich dokumenty relewantne i dokumenty nierelewantne znane użytkownikowi z odpowiedzi, a nie wszystkie dokumenty relewantne i nierelewantne w kolekcji. Dodatkowo, doświadczenia publikowane w literaturze pokazały, że terminy zawarte w pierwotnym pytaniu są istotne i dlatego powinny zostać zachowane w kolejnych modyfikacjach tego pytania (Salton i Buckley, 1990). W literaturze znane są trzy klasyczne formuły modyfikacji pytania (Ide, 1971):

Standard–Rocchio: $$q_{new} = \alpha q_{old} + \frac{\beta}{n_1} \sum_{\forall d \in D_r} d - \frac{\gamma}{n_2} \sum_{\forall d \in D_n} d \ ,$$

Ide–Regular: $$q_{new} = \alpha q_{old} + \beta \sum_{\forall d \in D_r} d - \gamma \sum_{\forall d \in D_n} d \ ,$$

Ide Dec–Hi: $$q_{new} = \alpha q_{old} + \beta \sum_{\forall d \in D_r} d - \gamma d_{top\_not\_rel} \ ,$$

gdzie D_r – zbiór dokumentów relewantnych według oceny użytkownika w zbiorze dokumentów wyszukanych, D_n – zbiór dokumentów nierelewantnych według oceny użytkownika wśród dokumentów wyszukanych, n_1 i n_2 liczba dokumentów odpowiednio w zbiorze D_r i D_n, α, β, γ – stałe przyjmujące wartości z przedziału $\langle 0,1 \rangle$[1], $d_{top\_not\_rel}$ – wektor dokumentu nierelewantnego, który znajdował się na najwyżej w rankingu.

Techniki sprzężenia zwrotnego, wykorzystywane do modyfikacji pytania, mogą skutkować zmianą wartości wag terminów z pytania pierwotnego, a poprzez to promowanie niektórych terminów z pytania i obniżanie znaczenia innych, lub mogą powodować wprowadzenie nowych terminów do pytania, które nie występowały w pytaniu pierwotnym (Müller i inni, 2000), (Croft i inni, 2001). Do technik sprzężenia zwrotnego mogą być zastosowane różne metody rozszerzania pytania (ang. *query*

[1] w oryginale Rocchio przyjął $\alpha = 1$, a Ide przyjął $\alpha = \beta = \gamma = 1$

expansion). Pierwszą możliwością jest rozszerzenie pytania pierwotnego o wszystkie terminy należące do wyszukanych dokumentów relewantnych. Innym rozwiązaniem jest rozszerzenie pytania tylko o niektóre terminy z wyszukanych dokumentów relewantnych (Spink i inni, 2000). Potwierdzono eksperymentalnie, że dobre wyniki wyszukiwania otrzymano rozszerzając pytanie o terminy najbardziej popularne, czyli takie, które występowały z największą częstością w poprzednio wyszukanych dokumentach relewantnych. Dobrymi terminami są również terminy o najwyższej wadze obliczonej według zaprezentowanych formuł modyfikacji wag terminów pytania z wykorzystaniem sprzężenia zwrotnego (Salton i Buckley, 1990).

Głównymi zaletami technik wykorzystujących sprzężenie zwrotne w modelu wektorowym jest ich prostota oraz pozytywne efekty stosowania podczas procesów wyszukiwania. Prostota uwidacznia się w tym, że waga terminu w pytaniu zmodyfikowanym wyliczana jest na podstawie zbioru dokumentów wyszukanych. Natomiast pozytywne efekty stosowania podczas wyszukiwania potwierdzone zostały eksperymentalnie (Chang, Cirillo, Razon, 1971), (Harman, 1992), (Buckley i inni, 1994), (Efthimiadis, 2000).

3.6.3. Modyfikacja pytania bazująca na analizie lokalnej

Wykorzystanie sprzężenia zwrotnego do modyfikacji pytania, a zatem reprezentowania potrzeby informacyjnej użytkownika, polega na uzyskaniu od użytkownika oceny początkowych dokumentów z rankingu dokumentów odpowiedzi. Ocena użytkownika dzieli dokumenty odpowiedzi na dwie grupy: dokumenty relewantne oraz dokumenty nierelewantne. Ocena dostarcza informacji, które mogą zostać wykorzystane do selekcji terminów do modyfikacji pytania. W kolejnym wyszukiwaniu, na pytanie zmodyfikowane użytkownik otrzymuje więcej dokumentów relewantnych. U podstawy opisanego podejścia leży założenie, że znane użytkownikowi dokumenty relewantne tworzą grupę (ang. *cluster*) oraz zawierają terminy, które mogą zostać wykorzystane do opisania grupy dokumentów relewantnych. Opis tej grupy dokumentów relewantnych powstaje w kolejnych cyklach, przy współpracy, czyli interakcji użytkownika.

Odmienny nurt badań nad modyfikacją pytania użytkownika stanowią podejścia, w których pytanie modyfikowane jest automatycznie – bez korzystania z informacji pochodzących od użytkownika. Celem tych podejść jest automatyczna identyfikacja terminów związanych (ang. *related terms*) z terminami pytania pierwotnego. Terminami związanymi są synonimy, formy morfologiczne dla których w wyniku stemmingu[1] został zidentyfikowany wspólny rdzeń lub terminy znajdujące się w dokumencie w otoczeniu terminów z pytania (tj. terminy, które znajdują się w dokumencie w

[1] Koncepcja operacji stemmingu opisana została w Rozdziale 2.1.1.

odległości co najwyżej *k* słów od terminu z pytania). Istnieją dwa podejścia w procesie automatycznej modyfikacji pytania: analiza lokalna i analiza globalna. W procesie modyfikacji pytania bazującej na analizie lokalnej, w celu pozyskania terminów do modyfikacji pytania, analizowane są dokumenty wyszukane przez system na pytanie *q*. Proces ten ma miejsce podczas obsługiwania pytania przez system. Posiada on cechy wspólne z procesem relewancyjnego sprzężenia zwrotnego, jednak podstawową różnicą jest fakt, że analiza lokalna wykonywana jest bez udziału użytkownika – automatycznie i uwzględniane są wszystkie dokumenty odpowiedzi. Przedstawione zostaną dwie strategie analizy lokalnej: grupowanie dokumentów lokalnych (ang. *local clustering*) oraz analiza lokalnego kontekstu (ang. *local context analysis*).

Grupowanie dokumentów lokalnych

Modyfikacja pytania, w której wykorzystywane są informacje uzyskane z procesu pogrupowania dokumentów, opiera się na utworzeniu ze zbioru dokumentów kolekcji macierzy podobieństwa terminów. Macierz podobieństwa opisuje powiązania pomiędzy terminami. Powiązanie pomiędzy dwoma terminami wyznaczane jest na podstawie liczby dokumentów, w których dwa terminy występują wspólnie. Jeśli termin z pytania znajduje się w macierzy, to terminy współwystępujące z nim w dokumentach kolekcji mogą zostać wykorzystane do modyfikacji pytania. Podstawową wadą tego rozwiązania jest brak możliwości wykorzystania macierzy podobieństwa do polepszenia efektywności wyszukiwania dla dowolnej kolekcji. Macierz utworzona dla pewnej kolekcji dokumentów nie musi dawać dobrych wyników modyfikacji pytania dla innej kolekcji. Macierz podobieństwa terminów jest bardzo mocno związana z kolekcją dokumentów i dlatego powiązania terminów występujące dla pewnej kolekcji nie koniecznie muszą być prawdziwe dla dokumentów innej kolekcji. Struktura reprezentująca powiązania globalne, jaką jest macierz podobieństwa, nie zawsze można zastosować z pozytywnym wynikiem do lokalnego kontekstu definiowanego przez aktualne pytanie. Należy więc zastosować strategie bazujące na analizie *dokumentów lokalnych*. Zbiorem dokumentów lokalnych nazywany jest zbiór dokumentów wyszukanych w odpowiedzi na aktualne pytanie użytkownika.

Jedną z pierwszych prac dotyczących modyfikacji pytania na podstawie grupowania dokumentów lokalnych opublikowali Attar i Fraenkel w 1977 r. Strategie modyfikacji pytania na podstawie lokalnego sprzężenia zwrotnego polegają na dodaniu do pytania terminów, które są związane z terminami pytania. W strategii grupowania dokumentów lokalnych, terminy związane to te, które znajdują się w klastrach dokumentów lokalnych. Klastry tworzone są w wyniku procesu grupowania zbioru dokumentów lokalnych, czyli zbioru dokumentów wyszukanych w odpowiedzi na aktualne pytanie. Attar i Fraenke zaproponowali trzy rodzaje klastrów tworzonych dla zbioru

dokumentów lokalnych: klaster powiązań, klaster metryczny i klaster skalarny (Attar i Fraenkel, 1977).

Klaster powiązań (ang. *association cluster*) tworzony jest na podstawie analizy współwystępowania terminów w dokumentach. Autorzy twierdzą, że terminy, które często występują razem w dokumentach powiązane są relacją synonimii. Klastry powiązań generowane są na podstawie lokalnej macierzy powiązań terminów (ang. *local stem–stem association matrix*). Element macierzy reprezentuje stopień powiązania $c_{u,v}$ terminu t_u i t_v. Wartość wyznaczana jest według następującego wzoru:

$$c_{u,v} = \sum_{d \in D_{odp}} f_{t_u} \times f_{t_v}$$

gdzie f_{t_u} to częstość występowania terminu t_u w dokumencie d należącym do dokumentów odpowiedzi D_{odp}.

Tworzenie klastra powiązań opiera się na częstości występowania par terminów w dokumencie, jednak nie uwzględniane jest położenie tych terminów w dokumencie. Dwa terminy znajdujące się w jednym zdaniu są bardziej powiązane ze sobą niż terminy znajdujące się w dużej odległości od siebie w różnych zdaniach, choć w tym samym dokumencie. Własność ta została uwzględniona podczas obliczania stopnia powiązania pomiędzy terminami w tym samym dokumencie w drugim zaproponowanym przez Attara i Fraenkela klastrze – klastrze metrycznym (ang. *metric cluster*). Wartość elementu $c_{u,v}$ metrycznej macierzy powiązań terminów (ang. *local stem–stem metric correlation matrix*) wyznaczana jest według poniższego wzoru:

$$c_{u,v} = \sum_{d \in D_{odp}} \frac{1}{odl(t_u, t_v)},$$

gdzie $odl(t_u, t_v)$ określa odległość pomiędzy terminami wyrażoną przez liczbę słów występujących pomiędzy terminem t_u a terminem t_v w tym samym dokumencie.

Trzecią zaproponowaną przez Attara i Freankela formą wydobywania zależności pomiędzy terminami w zbiorze dokumentów lokalnych jest porównanie otoczeń (ang. *neighbourhoods*) rozważanych dwóch terminów. Twierdzą oni, że terminy, które posiadają podobne otoczenia są dla siebie synonimami. Zależność ta jest nazywana pośrednią lub zależną od otoczenia. Podobieństwo otoczeń wyznaczane jest na podstawie stopnia powiązania $c_{u,v}$ terminu t_u z wszystkimi pozostałymi terminami reprezentowanymi w lokalnej macierzy powiązań oraz terminu t_v z wszystkimi pozostałymi terminami reprezentowanymi w lokalnej macierzy powiązań. Inaczej mówiąc, jest to porównanie wektorów $\vec{w}_u$, $\vec{w}_v$ terminów, reprezentujących jednocześnie wiersze lokalnej macierzy powiązań. Wektory mogą zostać porównane na podstawie jednej z miar podobieństwa, np. miary kosinusowej:

$$c_{u,v} = \frac{\vec{w}_u \bullet \vec{w}_v}{|\vec{w}_u| \times |\vec{w}_v|},$$

gdzie wektory $\vec{w}_u = (c_{u1}, c_{u2}, \ldots, c_{un})$, $\vec{w}_v = (c_{v1}, c_{v2}, \ldots, c_{vn})$ reprezentują wartości powiązań dla terminów t_u i t_v. Skalarna macierz powiązań, której element $c_{u,v}$ został zdefiniowany powyżej, wykorzystywana jest do utworzenia klastra skalarnego.

Opisane powyżej lokalne macierze powiązań terminów wykorzystywane są w procesie tworzenia klastrów terminów związanych. W tym celu definiowana jest funkcja $S_u(n)$, której argumentem jest u–ty wiersz lokalnej macierzy powiązań. Wynikiem funkcji $S_u(n)$ jest zbiór n największych wartości korelacji $c_{u,v}$, gdzie v zmienia się po wszystkich terminach lokalnej macierzy (tj. kolumnach macierzy) oraz $u \neq v$. Funkcja $S_u(n)$ definiuje lokalny klaster wokół terminu t_u. W zależności od przyjętej metody wydobywania zależności pomiędzy terminami, i utworzonej na tej podstawie lokalnej macierzy powiązań, macierzy metrycznej lub macierzy skalarnej otrzymujemy lokalny klaster powiązań, lokalny klaster metryczny lub lokalny klaster skalarny.

W koncepcji klastrów budowanych dla dokumentów lokalnych przyjęto założenie, że terminy, które należą do tego samego klastra są ze sobą powiązane. Opierając się na tym założeniu przyjęto, że do pytania pierwotnego mogą zostać dołączone terminy, które należą do tego samego klastra co termin pytania (lub terminy pytania). Terminy te nazywane są *sąsiadami* (terminów z pytania) i definiowane następująco: Termin t_u należący do klastra (o rozmiarze n), który jest powiązany z terminem t_v ($t_u \in S_u(n)$) nazywamy *sąsiadem* terminu t_u.

Termin t_v nazywany jest również *searchonymem* terminu t_u. Terminy będące sąsiadami są wzajemnie w relacji synonimii, jednak nie koniecznie są synonimami w sensie gramatycznym. Najczęściej terminy będące sąsiadami reprezentują różne słowa, które są powiązane poprzez wspólny kontekst aktualnego pytania. Lokalny aspekt tego powiązania jest odzwierciedlony przez fakt, że zarówno dokumenty jak i terminy uwzględniane w lokalnej macierzy powiązań są lokalne, czyli należą do zbioru dokumentów odpowiedzi. W szerszym rozumieniu, termin będący sąsiadem terminu z pytania jest istotnym wynikiem procesu grupowania dokumentów lokalnych. Termin taki może zostać wykorzystany do podjęcia wyszukiwania w obiecującym, jednak nie przewidzianym wcześniej kierunku, raczej niż uzupełnić pytanie jako termin synonimiczny – synonim.

Attar i Freankel zaproponowali dla modelu wektorowego własną metodę rozszerzenia pytania użytkownika q, wykorzystując terminy sąsiadujące z terminami pytania q (Attar, Freankel, 1977). Dla każdego terminu $t_v \in q$ wybieranych jest z klastra $S_v(n)$ m terminów sąsiadujących, a następnie terminy te dołączane są do pytania. Dodane terminy sąsiadujące przyczyniają się zazwyczaj do uzyskania nowych dokumentów relewantnych w odpowiedzi podczas wyszukiwania. Klaster $S_v(n)$ w zależności od zastosowanej wcześniej metody tworzenia klastrów może być klastrem powiązań, klastrem metrycznym lub klastrem skalarnym.

Opisane w literaturze przeprowadzane eksperymenty potwierdzają przydatność metod grupowanie dokumentów lokalnych w celu rozszerzania pytania użytkownika.

Przy czym wykorzystanie klastrów metrycznych daje lepsze wyniki niż wykorzystanie klastrów powiązań. Potwierdza to istnienie korelacji pomiędzy faktem istnienia powiązania dwóch terminów, a odległością w jakiej występują te dwa terminy w dokumencie.

Opisana metoda modyfikacji pytania bazuje na technikach grupowania dokumentów lokalnych (ang. *local clustering*) oraz wykorzystuje dokumenty znajdujące się na początku listy rankingowej dokumentów wyszukanych w odpowiedzi na pytanie użytkownika do utworzenia klastrów terminów sąsiadujących. Klastry terminów tworzone są na podstawie częstości współwystępowania terminów w analizowanych dokumentach. Z klastra utworzonego dla każdego terminu należącego do pytania pierwotnego, wybierane są najlepsze terminy do rozszerzenia tego pytania. Inne podejście, polegające na poszukiwaniu korelacji występujących pomiędzy terminami na podstawie analizy całej kolekcji dokumentów, nosi nazwę analizy globalnej. Techniki globalne zazwyczaj wykorzystują tezaurus, który służy do identyfikowania zależności pomiędzy terminami w całej kolekcji. Terminy są traktowane jako pojęcia, a tezaurus jest strukturą reprezentującą relacje, inaczej powiązania, pomiędzy pojęciami. Podczas tworzenia tezaurusa zazwyczaj uwzględniany jest kontekst mniejszy niż kontekst całego dokumentów oraz struktura analizowanej frazy. Poniżej omówiona zostanie kolejna metoda modyfikacji pytania, która wykorzystuje rozwiązania stosowane w analizie globalnej (tj. mniejszy kontekst oraz strukturę frazy) do lokalnego zbioru dokumentów odpowiedzi: analiza lokalnego kontekstu (ang. *local context analysis*).

Analiza lokalnego kontekstu

Podejście bazujące na analizie lokalnego kontekstu czerpie zarówno z koncepcji analizy lokalnej, jak i globalnej (Xu i Croft, 1996), (Belkin i inni, 2000). Xu i Croft przyjęli, że treść dokumentu reprezentowana jest nie przez pojedyncze słowa kluczowe, ale przez pojęcia. Pojęcie zdefiniowane jest przez grupę rzeczowników, tj. pojedynczy rzeczownik, dwa rzeczowniki lub więcej. Pojęcia do rozszerzenia pytania wybierane są z dokumentów znajdujących się na początku listy rankingowej odpowiedzi na podstawie częstości współwystępowania pojęć z terminami pytania, jak to miało miejsce dla analizy lokalnej. Analiza częstości współwystępowania odbywa się dla paragrafów, a nie dla całych dokumentów, jak to miało miejsce dla analizy globalnej. Przyjęto, że paragrafem jest fragment tekstu o stałej długości. Procedura analizy lokalnego kontekstu odbywa się w następujących krokach. Najpierw wyszukiwane jest na podstawie pytania początkowego n paragrafów o najwyższym rankingu. Aby to osiągnąć, dokumenty wyszukane na pytanie początkowe dzielone są na paragraf jednakowej długości (przyjęto rozmiar równy 300 słów), a następnie tworzony jest ranking tych paragrafów, tak jakby były one dokumentami. Następnie dla każdego pojęcia c, znajdującego się w paragrafach na początku rankingu, obliczana jest wartość

podobieństwa *sim(q,c)* całego pytania (nie pojedynczych terminów pytania) i pojęcia *c*. Do obliczenia podobieństwa wykorzystywany jest schemat ważenia terminów *tf–idf*[1]:

$$sim(q,c) = \prod_{k_i \in q} \left(\delta + \frac{\log\left(f(c,k_i) \times idf_c\right)}{\log n} \right)^{idf_i}$$

gdzie *n* – liczba paragrafów znajdujących się na początku rankingu, poddana analizie,

δ – stała równa 0.1, dzięki której podobieństwo *sim(q, c)* nie przyjmuje wartości zerowych[2],

$f(c, k_i)$ – funkcja, na podstawie której obliczany jest stopień korelacji pomiędzy pojęciem *c* i terminem pytania k_i :

$$f(c,k_i) = \sum_{j=1}^{n} pf_{i,j} \times pf_{c,j}$$

gdzie $pf_{i,j}$ – częstość występowania terminu k_i w *j*–tym paragrafie,

$pf_{c,j}$ – częstość występowania pojęcia *c* w *j*–tym paragrafie.

Funkcja $f(c, k_i)$ jest miarą korelacji pomiędzy terminami, zdefiniowaną dla klastrów powiązań, zaadaptowaną na potrzeby obliczenia korelacji na podstawie paragrafów, a nie całych dokumentów.

idf_i – odwrotna częstość dokumentowa terminu k_i jest obliczana wg wzoru:

$$idf_i = \max\left(1, \frac{\log_{10} N / np_i}{5}\right)$$

idf_c – odwrotna częstość dokumentowa pojęcia *c* jest obliczana wg wzoru:

$$idf_c = \max\left(1, \frac{\log_{10} N / np_c}{5}\right)$$

gdzie N – liczba paragrafów w całej kolekcji,

np_i – liczba paragrafów zawierających termin k_i,

np_c – liczba paragrafów zawierających pojęcie *c*.

Składnik idf_i wprowadzony we wzorze w potędze powoduję, że waga terminów pytania, występujących z niezbyt dużą częstością w paragrafach, jest podwyższana.

W ostatnim kroku, na podstawie wartości podobieństwa *sim(q,c)*, ustalany jest ranking pojęć i *m* pojęć znajdujących się na początku rankingu dodawanych jest do pytania początkowego *q*. Dla każdego pojęcia, które ma zostać dołączone do pytania obliczana jest waga według następującego wzoru:

$$w = 1 - 0{,}9 * i/m,$$

gdzie: *i* – pozycja pojęcia w utworzonym rankingu,

m – liczba pojęć dołączanych do pytania.

Terminom pytania początkowego przypisana zostaje waga równa 2.

[1] Schemat *tf–idf* ważenia terminów opisano w Rozdziale 2.1.1.

[2] Własność ta może być istotna jeśli podejście zostanie zastosowane dla modeli probabilistycznych.

Opisane rozwiązanie testowane było dla kolekcji testowej udostępnianej uczestnikom konferencji TREC. Nie potwierdzono jednak, czy analizę kontekstu lokalnego można zastosować do rozszerzania pytania z podobnie pozytywnym wynikiem dla innych kolekcji.

Przedstawione podejścia oraz weryfikujące je badania eksperymentalne potwierdzają pozytywne efekty zastosowania metody modyfikacji pytania bazującej na analizie lokalnej. Metoda ta wykorzystuje fakt, że powiązania występujące pomiędzy terminami należącymi do dokumentów odpowiedzi są prawdziwe również dla terminów pytania, ponieważ dotyczą tego samego kontekstu. W tym miejscu upatrywać można również słabości metod modyfikacji pytania bazujących na analizie globalnej. Powiązania istniejące pomiędzy terminami całej kolekcji mogą nie być prawdziwe dla terminów aktualnego pytania. Jest to również przyczyną problemów z wykorzystaniem dla nowej kolekcji struktury zależności pomiędzy terminami zbudowanej dla innej kolekcji.

Modyfikacja pytania bazująca na analizie lokalnej wymaga częstego dostępu do pełnego tekstu dokumentów wyszukanych jako odpowiedź na pytanie użytkownika. Dlatego też, zastosowanie strategii lokalnej, w jej oryginalnej postaci, do wyszukiwania w sieci WWW stwarza poważne problemy. Analiza dokumentów z sieci WWW w celu uzyskania struktury lokalnych powiązań pomiędzy terminami wykonywana po stronie użytkownika spowoduje znaczne obciążenie, a w efekcie zmniejszy poziom zadowolenia użytkownika z wyszukiwania. Również po stronie wyszukiwarki internetowej, analiza dokumentów odpowiedzi zajmuje dodatkowy czas procesora, co jest mało opłacalne, gdyż dzisiejsze wyszukiwarki internetowe czerpią główne zyski z obsługi maksymalnej liczby pytań w jednostce czasu.

3.6.4. Modyfikacja pytania bazująca na analizie globalnej

Omówione powyżej metody lokalnej analizy polegają na pozyskiwaniu informacji niezbędnych do modyfikacji pytania ze zbioru dokumentów wyszukanych. Rozwiązanie taki przyczynia się do poprawy efektywności wyszukiwania dla różnorodnych kolekcji dokumentów. Innym podejściem do modyfikacji pytania jest wykorzystanie informacji uzyskanej na podstawie analizy wszystkich dokumentów kolekcji. Metody bazujące na tej koncepcji określane są wspólnym mianem metod bazujących na analizie globalnej. Początkowe badania prowadzone w dziedzinie zastosowania metod analizy globalnej do wyszukiwania w dowolnych kolekcjach dokumentów nie przynosiły oczekiwanej poprawy efektywności wyszukiwania. Sytuacja ta zmieniła się wraz z pojawieniem się nowych propozycji i rozwiązań. Są to podejścia wykorzystujące struktury typu tezaurus, budowane na podstawie analizy całej kolekcji dokumentów. Struktura typu tezaurus opisuje relacje pomiędzy terminami kolekcji. Zaprezentowanie użytkownikowi utworzonej struktury umożliwia wybranie terminów do modyfikacji pytania.

Modyfikacja pytania na podstawie tezaurusa podobieństwa

Qiu i Frei zaproponowali metodę rozszerzania pytania na podstawie tezaurusa podobieństwa (ang. *similarity thesaurus*), automatycznie tworzonego dla całej kolekcji dokumentów (Qiu i Frei, 1993), (Qiu, 1996). Tezaurus podobieństwa tworzony jest na podstawie powiązań pomiędzy terminami, a nie na podstawie częstości współwystępowania, jak w przypadku opisanej powyżej macierzy. Zaprezentowano nowe rozwiązanie wybierania terminów do rozszerzenia pytania oraz nadawania wag tym terminom. Terminy do rozszerzenia pytania wybierane są na podstawie ich podobieństwa do całego pytania, a nie tylko do pojedynczego terminu pytania, jak to miało miejsce we wcześniejszych metodach rozszerzania pytania w procesie analizy globalnej.

Zaproponowany przez Qiu i Frei tezaurus podobieństwa, jak już wspomniano, tworzony jest na podstawie stwierdzonych istniejących relacji pomiędzy terminami. Relacje te nie są jednak określane bezpośrednio na podstawie częstości współwystępowania terminów w dokumentach. W prezentowanym podejściu terminy traktowane są jako pojęcia w przestrzeni pojęć. Przyjęto. że w przestrzeni tej każdy termin jest poindeksowany dokumentami, w których występuje. Tak więc terminy pełnią rolę jaką oryginalnie pełniły dokumenty, podczas gdy dokumenty traktowane są jako elementy indeksujące. W tak określonej przestrzeni przyjęte zostały następujące oznaczenia:

t – liczba terminów w kolekcji;

N – liczba dokumentów w kolekcji;

f_i – częstość występowania terminu t_i w dokumencie d;

k – liczba różnych terminów występujących w dokumencie d;

itf – odwrotna częstość występowania terminu dla dokumentu d;

Odwrotna częstość występowania terminu dla dokumentu d jest pojęciem analogicznym do definicji odwrotnej częstości dokumentowej zdefiniowanej dla przestrzeni dokumentów.

$$itf = \log \frac{t}{k}$$

W przestrzeni pojęć, każdemu terminowi t_i przyporządkowany jest wektor z przestrzeni dokumentów: $\vec{t_i} = (w_{i1}, w_{i2}, w_{i3}, ..., w_{iN})$, gdzie waga w_i związana jest z parą termin indeksowy t_i – dokument d (oznacza wagę terminu t_i w dokumencie d), obliczaną na podstawie następującego wzoru:

$$w_i = \frac{\left(0.5 + 0.5 \frac{f_i}{\max\left(f_i\right)}\right) itf}{\sqrt{\sum_{l=1}^{N} \left(0.5 + 0.5 \frac{f_{il}}{\max_l\left(f_{il}\right)}\right)^2 \left(itf_l\right)^2}}$$

gdzie max(f_i) jest funkcją obliczającą maksymalną częstość występowania f_i terminu t_i we wszystkich dokumentach kolekcji. Zamieszczony powyżej wzór jest analogiczny do schematu *tf–idf* ważenia terminów w przestrzeni dokumentów. Różnicą jest użycie odwrotnej częstości występowania terminu dla dokumentu zamiast, standardowo wykorzystywanej w schemacie *tf–idf*, odwrotnej częstości dokumentowej. Odwrotna częstość dokumentowa *idf* obliczana jest jako proporcja wszystkich dokumentów w kolekcji do liczby dokumentów, w których występuje termin dla którego obliczana jest waga. Zależność c_{uv} pomiędzy dwoma terminami t_u i t_v wyznaczana jest na podstawie następującego wzoru:

$$c_{uv} = \vec{t}_u \bullet \vec{t}_v = \sum_{\forall d_j} w_{uj} \times w_{vj}$$

Równanie to jest podobne do miary korelacji wykorzystywanej do obliczenia elementów skalarnej macierzy powiązań (zdefiniowanej w rozdziale 3.6.3). Podstawową różnicą jest potraktowanie we wzorze na obliczenie wagi w_{ij} terminu dokumentów, jako elementów indeksujących. Nie jak w modelu klasycznym, gdzie dokumenty opisywane są przez terminy, występujące w dokumentach z określoną częstością. Tezaurus podobieństwa dla kolekcji dokumentów budowany jest więc na podstawie obliczonej wartości korelacji c_{uv} dla każdej pary terminów indeksujących t_u, t_v należących do kolekcji. Proces ten jest obliczeniowo czasochłonny, jednak raz utworzony tezaurus podobieństwa może być cyklicznie aktualizowany.

Proces modyfikacji pytania na podstawie tezaurusa podobieństwa zbudowanego dla kolekcji rozpoczyna się od przedstawienia pytania użytkownika w postaci reprezentacji przyjętej dla przestrzeni pojęć. W przestrzeni tej każdy termin indeksowy reprezentowany jest jako *n*–wymiarowy wektor wag. Pytanie q reprezentowane jest jako wektor $\vec{q}$ w przestrzeni terminów i pojęć w następującej postaci:

$$\vec{q} = \sum_{t_i \in q} w_{iq} \vec{t}_i ,$$

gdzie w_{iq} – waga przypisana parze termin t_i – pytanie q, obliczana według formuły obliczania wagi termin indeksowy t_i – dokument d_j.

W drugim etapie, na podstawie tezaurusa podobieństwa obliczane jest podobieństwo $sim(q, t_v)$ pomiędzy każdym terminem t_v należącym do tezaurusa i będącym w korelacji z terminami pytania, a całym pytaniem q według następującej formuły:

$$sim(q, t_v) = \vec{q} \bullet \vec{t}_v = \sum_{t_u \in q} w_{uq} \times c_{uv} ,$$

gdzie c_{uv} jest wartością korelacji pomiędzy terminami t_u, t_v.

Termin, który ma zostać wybrany do rozszerzenia pytania może mieć dużą wartość podobieństwa do całego pytania, podczas gdy odległość tego terminu od poszczególnych terminów pytania może być dość duża. Dlatego też terminy wybrane według opisanej metody z tezaurusa podobieństwa do rozszerzenia pytania mogą być inne niż terminy, które zostałyby wybrane po zastosowaniu metody analizy globalnej

(która wykorzystuje podobieństwo analizowanego terminu do poszczególnych terminów pytania w celu wybrania terminów do rozszerzenia pytania).

W ostatnim etapie, pytanie *q* rozszerzane jest o *r* terminów znajdujących się na początku rankingu terminów, utworzonego na podstawie wartości podobieństwa $sim(q, t_v)$. Nowe pytanie oznaczane jest *q'*. Waga terminu w pytaniu *q'* obliczana jest według poniższego wzoru:

$$w_{vq'} = \frac{sim(q, t_v)}{\sum_{t_u \in q} w_{uq}}.$$

Rozszerzone pytani *q'* jest następnie wykorzystane do wyszukania nowych dokumentów. Opisane podejście wykazuje polepszenie efektywności wyszukiwania w granicach 20% w stosunku do przedstawionego wcześniej podejścia bazującego na analizie globalnej (Qiu i Frei, 1993).

Modyfikacja pytania na podstawie tezaurusa statystycznego

Crouch i Yang zaproponowali metodę rozszerzania pytania użytkownika na podstawie tezaurusa statystycznego (ang. *global statistical thesaurus*) (Crouch, 1990), (Crouch i Yang, 1992). Jest to technika należąca do grupy metod analizy globalnej, ponieważ tezaurus budowany jest na podstawie całej kolekcji dokumentów. Tezaurus statystyczny zawiera klasy, które grupują terminy będące ze sobą w korelacji (ang. *correlated terms*) w kontekście całej kolekcji dokumentów. Terminy będące w korelacji mogą zostać wykorzystane do rozszerzenia oryginalnego pytania użytkownika. Jednak, aby rozwiązanie to było efektywne, terminy wybrane do rozszerzenia pytania, muszą mieć wysoką wartość dyskryminacyjną, czyli musza być to terminy o małej częstości występowania w całej kolekcji dokumentów (Salton i inni, 1975). Pojawia się jednak problem, że dla terminów o małej częstości występowania trudno jest zastosować efektywnie procedury grupowania, ponieważ terminy te występują w niewielu dokumentach kolekcji. Dlatego też zaproponowano rozwiązanie, w którym najpierw dokumenty kolekcji grupowane są w klasy, a następnie wybrane są z pogrupowanych dokumentów terminy o niskiej częstości występowania. Wybrane w ten sposób terminy definiują klasy tezaurusa.

Na potrzeby zbudowania tezaurusa statystycznego zastosowany został algorytm grupowania metodą najdalszego sąsiedztwa (ang. *complete link clustering*) (Voorhees, 1986). W wyniku zastosowania tego algorytmu, kolekcja dokumentów podzielona zostaje na grupy zawierające mało dokumentów, ale mocno ze sobą wzajemnie związanych (tj. podobnych dokumentów). Podobieństwo dwóch grup dokumentów (ang. *cluster*) jest zdefiniowane jako minimum z podobieństwa wszystkich par dokumentów, gdzie każdy dokument z pary nie należy do tej samej grupy. W celu policzenia podobieństwa pomiędzy dokumentami zastosowano miarę kosinusową (Salton, 1983). Przyjęte w algorytmie kryterium minimalnego podobieństwa w procesie

wyznaczania grup dokumentów, skutkuje tym, że otrzymane w wyniku grupy dokumentów są mało liczne, a dokumenty należące do grupy mają wysoką wartość podobieństwa pomiędzy sobą wzajemnie.

W procesie tworzenia tezaurusa statystycznego, grupowanie metodą najdalszego sąsiedztwa zastosowane jest dla całej kolekcji dokumentów. W wyniku powstaje hierarchia grup dokumentów, w której wszystkie dokumenty należące do jednej grupy związane są wzajemnie pewnym minimalnym podobieństwem. Dla zamieszczonego rysunku podobieństwo dwóch klas opisywane jest funkcją: $sim(C_u, C_v) = 0.15$, $sim(C_{u+v}, C_z) = 0.11$, gdzie C_{u+v} oznacza grupę po połączeniu dwóch grup C_u i C_v. Im grupa znajduje się wyżej w hierarchii, tym podobieństwo pomiędzy dokumentami tej grupy jest mniejsze, ponieważ grupa wyżej w hierarchii zawiera więcej dokumentów i reprezentuje słabsze powiązania. Dlatego też grupy zawierające dokumenty najbardziej podobne znajdują się na dole hierarchii.

Na podstawie zbudowanej hierarchii grup dokumentów, wybierane są terminy, które utworzą poszczególne klasy tezaurusa podobieństwa. W pierwszym kroku ustalane są wartości parametrów: progu dla klasy *PK* (ang. *threshold class*), liczba dokumentów w klasie *LDK* oraz minimalna odwrotna częstość dokumentowa *MIDF*. W drugim kroku tworzenia tezaurusa na bazie hierarchii dokumentów, na podstawie parametru *PK* określane jest, które grupy dokumentów wybrane zostaną do utworzenia klas tezaurusa. Wartość parametru *PK* nie powinna przewyższać wartości podobieństwa $sim(C_u, C_v)$ dwóch grup dokumentów, aby z dokumentów należących do tych grup mogły zostać wybrane terminy do utworzenia klas tezaurusa. I tak, dla zamieszczonej przykładowej hierarchii grup dokumentów jeśli $PK = 0.14$, wtedy w dalszej analizie uwzględniona zostanie klasa C_{u+v}, natomiast dla $PK = 0.10$ – klasy C_{u+v} i C_{u+v+z}. W kolejnym kroku, na podstawie wartości parametru *LDK*, odrzucane są te grupy, których liczna dokumentów przekracza przyjętą wartość *LDK*. Po zastosowaniu parametrów *PK* oraz *LDK*, dla każdej z uzyskanych grup dokumentów przeprowadzana jest na podstawie parametru *MIDF* selekcja terminów. Parametr *MIDF* określa minimalną wartość odwrotnej częstości dokumentowej, którą musi posiadać termin, aby zostać wybranym do klasy tezaurusa. Zapobiega to pojawieniu się w klasie tezaurusa terminów ogólnych (o wysokiej odwrotnej częstości występowania *idf*), które nie są dobrymi synonimami.

Zbudowany według powyższego schematu tezaurus może zostać wykorzystany w procesie rozszerzania pytania wyszukiwawczego. W tym celu wyznaczana jest średnia waga terminu w_i^C dla każdej klasy C tezaurusa:

$$w_i^C = \frac{\sum_{i=1}^{|C|} w_{i,C}}{|C|}$$

gdzie $|C|$ to liczba terminów w klasie C, $w_{i,C}$ – waga terminu w klasie C.

Opisana metoda została zweryfikowana eksperymentalnie (Crouch i Yang, 1992). Eksperymenty przeprowadzano dla 4 kolekcji testowych (ADI, Medlars, CACM, ISI) i pokazano możliwość polepszenia efektywność wyszukiwania, jeśli zastosowana

zostanie metoda analizy globalnej, wykorzystująca tezaurus podobieństwa, zbudowany z wykorzystaniem algorytmu grupowania metodą najszerszego sąsiedztwa.

Podstawową wadą zaproponowanego rozwiązania jest konieczność ustalenia parametrów *PK*, *LDK*, *MIDF*, od których zależy jakość budowanego tezaurusa. Wartość progu *PK* jest uzależniona od konkretnej kolekcji dokumentów i może być trudno ustalić ją poprawnie. Dlatego też, aby ustalić parametr *PK*, niezbędna jest wcześniejsza analiza hierarchii grup dokumentów. Zbyt wysoka wartość *PK* spowoduje, że klasy tezaurusa będą zawierały niewiele terminów. Natomiast zbyt niska wartość *PK* skutkuje powstaniem niewielkiej liczby klas tezaurusa (zawierających dużo terminów). Podobne problemy wiążą się z koniecznością eksperymentalnego wyznaczania parametru *LDK* oraz *MIDF*, indywidualnie dla każdej kolekcji, dla której ma zostać zastosowana metoda rozszerzania pytania na podstawie tezaurusa statystycznego.

Podsumowując przedstawione trzy metody modyfikacji pytania, można stwierdzić, że sprzężenie zwrotne ma kilka zalet w stosunku do pozostałych metod modyfikacji pytania tj. modyfikacji pytania na podstawie analizy lokalnej i analizy globalnej. Sprzężenie zwrotne ukrywa przed użytkownikiem szczegóły procesu modyfikacji pytania, gdyż użytkownik tylko ocenia dokumenty pod względem relewancji. Wprowadzenie sprzężenia zwrotnego dzieli cały proces wyszukiwania na sekwencje krótszych wyszukiwań, pomyślane tak, aby stopniowo zbliżać się do poszukiwanej dziedziny. W końcu sprzężenie zwrotne umożliwia sterowanie procesem modyfikacji pytania tak, aby podnosić znaczenie pewnych terminów, a obniżać znaczenie innych.

3.6.5. Reprezentowanie powiązań pomiędzy terminami

Macierz współwystępowania terminów

Macierz współwystępowania terminów jest strukturą, która reprezentuje powiązania terminów w pewnym zbiorze dokumentów. Zbiorem tym może być cała kolekcja dokumentów systemu wyszukiwania informacji lub tylko pewien podzbiór tych dokumentów, np. dokumenty odpowiedzi. Licząc częstość współwystępowania każdej pary terminów ze zbioru dokumentów otrzymujemy wagi współwystępowania terminów. Wagi te są elementami macierzy współwystępowania. Macierz współwystępowania terminów opisano i stosowano w literaturze dotyczącej wyszukiwania informacji. Zastosowano ją m.in. do rozszerzania pytania użytkownika (Qiu, 1996) oraz do reprezentowania zainteresowań użytkownika (Asnicar i Tasso, 1997).

Stosowanie macierzy współwystępowania terminów, zbudowanej dla całej kolekcji dokumentów, do rozszerzania czy modyfikacji pytania użytkownika może być jednak

przyczyną pogorszenia wyników wyszukiwania. Eksperymenty przeprowadzone przez Peata i Willetta pokazują, że rozszerzanie pytania terminami z macierzy współwystępowania terminów powoduje dołączanie do pytania początkowego terminów o wysokiej częstości występowania w dokumentach, a niekoniecznie terminów, które polepszają wyniki wyszukiwania (Peat i Willett, 1991). Peat i Willett uzasadniają fakt dołączania do pytania początkowego terminów o wysokiej częstości występowania tym, że proces wybierania terminów z macierzy współwystępowania do rozszerzenia pytania promuje terminy często współwystępujące z terminami pytania początkowego. A te ostatnie są terminami o wysokiej częstości występowania w dokumencie.

Stosowanie macierzy współwystępowania terminów, zbudowanej dla całej kolekcji dokumentów, do modyfikacji pytania użytkownika może być również przyczyną złego kierunku modyfikacji pytania oraz nieprawidłowego reprezentowania zainteresowań użytkownika. Przez zły kierunek modyfikacji pytania rozumiemy inną niż przyjęta przez użytkownika interpretację pytania (terminów pytania). Przyczyną są wieloznaczności słów. Te same słowa, w zależności od dziedziny, mogą równie często współwystępować z różnymi słowami, tworząc połączenia o diametralnie różnym znaczeniu. Konieczne jest więc zawężenie możliwych interpretacji pytania. Stwierdzamy, że taką właściwość ma macierz współwystępowania terminów SM^{loc} zbudowana tylko dla dokumentów relewantnych odpowiedzi.

Efektem proponowanego w pracy procesu modyfikacji pytania użytkownika ma być pytanie lepiej odzwierciedlające rzeczywiste zainteresowania użytkownika niż pytanie początkowo sformułowane przez użytkownika. W pracy przyjęto, że istotnymi są informacje o współwystępowaniu terminów w dokumentach relewantnych odpowiedzi, ponieważ są to dokumenty wskazane przez użytkownika jako zgodne z jego zainteresowaniami. Przeanalizowano możliwości wykorzystania do modyfikacji pytania w internetowym systemie wyszukiwania informacji proponowane w literaturze metody wyznaczania powiązań pomiędzy terminami. Opis przeprowadzonej analizy zamieszczono poniżej.

Przyjmujemy, że macierz SM^{loc} reprezentuje zależność współwystępowania terminów należących do dokumentów relewantnych odpowiedzi. Macierz tworzona jest tylko dla terminów znaczących tz_i [1] wyznaczonych z dokumentów relewantnych odpowiedzi, a nie dla wszystkich terminów z tych dokumentów:

$$SM^{loc} = \begin{pmatrix} 1 & \theta(tz_1, tz_2) & \dots & \theta(tz_1, tz_n) \\ \theta(tz_2, tz_1) & 1 & & \theta(tz_2, tz_n) \\ \dots & & & \\ \theta(tz_n, tz_1) & \theta(tz_n, tz_2) & \dots & 1 \end{pmatrix}$$

[1] Definicję terminu znaczącego tz_i podano w Rozdziale 4.6.1.

Macierz SM^{loc} jest macierzą symetryczną: $\theta(tz_j, tz_i) = \theta(tz_i, tz_j)$, $\theta(tz_i, tz_i) = 1$. Elementem macierzy jest wartość miary podobieństwa terminów liczona na podstawie współwystępowania terminów znaczących tz_i i tz_j w dokumentach relewantnych. Jako funkcję podobieństwa terminów zaproponowano tutaj miarę Dice'a:

$$\theta(tz_j, tz_i) = \frac{2 * \sum_{k=1}^{|D_{rel}|} \left|\{tz_i\} \cap T_{D_k}\right| * \left|\{tz_j\} \cap T_{D_k}\right|}{\sum_{k=1}^{|D_{rel}|} \left|\{tz_i\} \cap T_{D_k}\right| + \sum_{k=1}^{|D_{rel}|} \left|\{tz_j\} \cap T_{D_k}\right|} \qquad (3.6.5.1)$$

W literaturze proponowane jest również stosowanie miary kosinusowej, czy miary Tanimoto (Peat i Willett, 1991).

Modyfikacja macierzy SM^{loc} ma miejsce zawsze, jeśli wyznaczony zostanie termin znaczący tz_i z dokumentów relewantnych odpowiedzi. Modyfikacja polega na uaktualnieniu wartości podobieństwa dla terminów znaczących, które pojawiły się w dokumentach relewantnych. Modyfikacji podlegają podobieństwa tych terminów, które przekazane zostały po kolejnym wyszukiwaniu dokumentów jako terminy znaczące. Wartość podobieństwa tych terminów uaktualniana jest na podstawie poniższego wzoru:

$$\theta(tz_j, tz_i) = \theta(tz_j, tz_i) + \frac{2 * \sum_{k=1}^{|D_{rel}|} \left|\{tz_i\} \cap T_{D_k}\right| * \left|\{tz_j\} \cap T_{D_k}\right|}{\sum_{k=1}^{|D_{rel}|} \left|\{tz_i\} \cap T_{D_k}\right| + \sum_{k=1}^{|D_{rel}|} \left|\{tz_j\} \cap T_{D_k}\right|} \qquad (3.6.5.2)$$

W procesie wyznaczania terminów znaczących tz_i może się zdarzyć sytuacja, że w zbiorze terminów znaczących nie będzie terminów z pytania użytkownika[1]. Istnieje kilka możliwych interpretacji braku terminu z pytania użytkownika w zbiorze terminów znaczących, otrzymanych w procesie analizy dokumentów relewantnych odpowiedzi.

W takiej sytuacji, termin z pytania można wyjątkowo dołączyć do macierzy SM^{loc}. Jednak wprowadzenie terminu nie będącego terminem znaczącym do macierzy SM^{loc} budzi wątpliwość, czy uzasadnione jest wykorzystanie miary współwystępowania tego terminu z innymi terminami do selekcji terminów z profilu. Miara współwystępowania wyznacza terminy z profilu, które następnie zostaną wykorzystane do zastąpienia terminów kolejnego pytania użytkownika. Jeśli termin z pytania nie został wyselekcjonowany jako termin znaczący oznacza to, że nie jest on istotny w słownictwie charakterystycznym dla dziedziny zainteresowań użytkownika.

Zaproponowana macierz współwystępowania SM^{loc} może być profilem użytkownika, który oprócz terminów charakterystycznych dla dziedziny zainteresowań użytkownika – terminów znaczących, zawiera informacje dotyczące podobieństwa terminów znaczących tz_i. Widoczny jest jednak brak możliwości określenia w pewnych

[1] terminy pytania nie spełniły warunków nałożonych na wartości wag, aby termin mógł być zaakceptowany jako termin znaczący (podrozdział 4.7.2)

sytuacjach powiązania pomiędzy terminami zadanymi przez użytkownika w pytaniu, a zbiorem terminów znaczących tz_i z dokumentów relewantnych dla tego pytania.

Macierz wystąpień terminów

Analiza możliwości zastosowania macierzy współwystępowania terminów SM^{loc} do modyfikacji pytania użytkownika doprowadziła autora pracy do wniosku, że istotne jest powiązanie pomiędzy całym pytaniem użytkownika, a terminami znaczącymi z dokumentów relewantnych odpowiedzi. Wiedza o tym powiązaniu umożliwi poprawną modyfikację pytania niezależnie od faktu uznania, lub nie, pojedynczego terminu z pytania za termin znaczący tz_i.

Macierz wystąpień terminów AM^{loc} reprezentuje zależności pomiędzy terminami t_j z pytania użytkownika, a terminami znaczącymi tz_i, wyznaczonymi z dokumentów relewantnych odpowiedzi:

$$AM^{loc} = \begin{pmatrix} \theta'(t_1, tz_1) & \theta'(t_1, tz_2) & \dots & \theta'(t_1, tz_n) \\ \theta'(t_2, tz_1) & \theta'(t_2, tz_2) & & \theta'(t_2, tz_n) \\ \dots & & & \\ \theta'(t_m, tz_1) & \theta'(t_m, tz_2) & \dots & \theta'(t_m, tz_n) \end{pmatrix}$$

Związek terminów określony jest dla wszystkich par terminów, z których jeden termin występuje w pytaniu użytkownika, a drugi w zbiorze terminów znaczących otrzymanych w wyniku analizy dokumentów wyszukanych na to pytanie. Macierz wystąpień terminów AM^{loc} umożliwia zidentyfikowanie związku pomiędzy terminami w znaczeniu przyjętym przez użytkownika, a terminami w znaczeniu powszechnie stosowanym w pewnej dziedzinie. Wyznaczenie związku pomiędzy terminami t_j i tz_i umożliwia skorygowanie użycia przez użytkownika terminu lub terminów niestosowanych w pewnej dziedzinie przez zastąpienie tych niestosowanych terminów w pytaniu terminami stosowanymi w pewnej dziedzinie. Dla użytkownika nie będącego ekspertem w danej dziedzinie wiedzy, terminy niestosowane mają taki sam sens jak terminy stosowane mają nadany w dokumentach z dziedziny zainteresowań użytkownika. Terminy te różni je tylko to, że tradycyjnie pewne terminy są stosowane w określonych dziedzinach, a inne nie są stosowane. Terminy stosowane i niestosowane są syntaktycznie różne, ponieważ pochodzą z różnych zakresów słownictwa. Dla terminów tych system wyszukiwania informacji zwraca różne zbiory dokumentów jako odpowiedź na syntaktycznie różne, lecz semantycznie takie same pytania.

Rozważmy następujący przykład obrazujący opisany problem. Jeśli użytkownik zadał pytanie: „szukanie informacji” (ang. *information search*) myśląc o zagadnieniach związanych z wyszukiwaniem informacji, pytanie użytkownika powinno zostać skorygowanie i zmodyfikowane z wykorzystaniem macierzy wystąpień terminów AM^{loc}. Pytanie zmodyfikowane będzie miało następującą postać: „wyszukiwanie

informacji" (ang. *information retrieval*). Przesłankami do modyfikacji pytania jest fakt, że w dziedzinie związanej z wyszukiwaniem informacji termin „szukanie" nie jest powszechnie stosowany oraz fakt, że termin „informacji" najczęściej występuje w tej dziedzinie z terminem „wyszukiwanie".

Oznaczamy przez:

T – zbiór wszystkich terminów,

T_Q – zbiór terminów t_j użytych przez użytkownika we wszystkich pytaniach, $T_Q \subseteq T$,

$$T_Q = T_{q_1} \cup T_{q_2} \cup ... \cup T_{q_l},$$

gdzie T_{q_1} – zbiór terminów t_j użytych przez użytkownika w pytaniu q_1,

T_D – zbiór terminów znaczących tz_i wyznaczonych z dokumentów relewantnych odpowiedzi w kolejnych wyszukiwaniach, $T_D \subseteq T$,

$$T_D = T_{D_1} \cup T_{D_2} \cup ... \cup T_{D_l},$$

gdzie T_{D_1} – zbiór terminów znaczących tz_i wyznaczonych z dokumentów relewantnych odpowiedzi w pierwszym wyszukaniu,

Zbiory terminów T_Q i T_D mogą posiadać część wspólną: $T_Q \cap T_D \neq \varnothing$. Im lepiej użytkownik potrafi za pomocą prawidłowej terminologii dziedzinowej określić swoje potrzeby, tym więcej jest terminów należących do części wspólnej obu zbiorów. Jeśli $T_Q \cap T_D = \varnothing$ to oznacza, że terminy użyte przez użytkownika w pytaniu nie są terminami powszechnie stosowanymi w dziedzinie zainteresowań użytkownika.

Elementem macierzy wystąpień terminów AM^{loc} jest wartość miary współwystępowania terminów t_j i tz_i. Przez współwystępowanie terminów rozumiemy tutaj pojawienie się terminu t_j w pytaniu oraz terminu tz_i w zbiorze terminów znaczących. Miara takiego współwystępowania wzrasta, jeśli w kolejnym procesie wyszukiwania w pytaniu pojawi się termin t_j, a w zbiorze terminów znaczących – termin tz_i. Miara liczona jest na podstawie wystąpienia terminów w pytaniu q_k i zbiorze terminów znaczących wyznaczonych z dokumentów relewantnych wyszukanych w odpowiedzi na pytanie q_k. Miarą współwystępowania terminów t_j i tz_i jest funkcja θ':

$$\theta': T_Q \times T_D \rightarrow \Re_+$$

taka, że

$$\theta'(t_j, tz_i) = \frac{2 * \sum_{k=1}^{l} \left| \{t_i\} \cap T_{q_k} \right| * \left| \{tz_j\} \cap T_{D_k} \right|}{\sum_{k=1}^{l} \left| \{t_i\} \cap T_{q_k} \right| + \sum_{k=1}^{l} \left| \{tz_j\} \cap T_{D_k} \right|}.$$

Funkcję θ' można zapisać rekurencyjnie. Przyjmuje ona następującą postać:

$$\theta'^{(1)}(t_j, tz_i) = \frac{2 * \left(\left| \{t_i\} \cap T_{q_1} \right| * \left| \{tz_j\} \cap T_{D_1} \right| \right)}{\left| \{t_i\} \cap T_{q_1} \right| + \left| \{tz_j\} \cap T_{D_1} \right|}$$

$$\theta^{'(k)}(t_j, tz_i) = \frac{2*\left(\sum_{k=1}^{l-1}\left|\{t_i\}\cap T_{q_k}\right|*\left|\{tz_j\}\cap T_{D_k}\right| + \left|\{t_i\}\cap T_{q_l}\right|*\left|\{tz_j\}\cap T_{D_l}\right|\right)}{\sum_{k=1}^{l-1}\left|\{t_i\}\cap T_{q_k}\right| + \left|\{t_i\}\cap T_{q_l}\right| + \sum_{k=1}^{l-1}\left|\{tz_j\}\cap T_{D_k}\right| + \left|\{tz_j\}\cap T_{D_l}\right|}$$

Poprawne skonstruowanie macierzy wystąpień terminów AM^{loc} możliwe jest jednak tylko przy założeniu, że użytkownik będzie formułował pytania składające się z jednego terminu. Możliwe jest wtedy wyznaczenie miary wystąpienia terminu t_j z pytania z terminami znaczącymi tz_i. Zastosowanie macierzy wystąpień terminów AM^{loc} jest problematyczne dla pytań złożonych, czyli pytań z operatorami koniunkcji lub alternatywy. Dla tego rodzaju pytań nie jest możliwe wyróżnienie w zbiorze terminów znaczących tych, które są związane z każdym pojedynczym terminem pytania złożonego.

3.6.6. Podsumowanie metod modelowania użytkownika w systemach wyszukiwania informacji

Na podstawie cytowanej literatury można stwierdzić, że poważną wadą powszechnie stosowanego do reprezentowania potrzeb informacyjnych użytkownika profilu w postaci pojedynczego wektora przestrzeni n–wymiarowej jest brak możliwości wykorzystania tak zdefiniowanego profilu dla pytań dotyczących różnych zainteresowań użytkownika. W literaturze profil w postaci pojedynczego wektora rozumiany był jako reprezentacja wszystkich zainteresowań użytkownika. Jeśli zachodziła potrzeba wykorzystania profilu do modyfikacji kolejnego pytania użytkownika pojawiał się poważny problem, które współrzędne wektora, reprezentujące terminy, odnoszą się do poszczególnych zainteresowań użytkownika. Autor pracy zauważył, że opisany problem rozwiązałaby taka reprezentacja profilu, w której uwzględniona byłaby struktura powiązań pomiędzy terminami.

Dalsza analiza literatury z zakresu reprezentowania powiązań pomiędzy terminami pokazała, że powiązania terminów wyznaczane są zazwyczaj albo na podstawie całej kolekcji dokumentów, albo na podstawie wszystkich dokumentów odpowiedzi. Niestety, wyznaczane są w ten sposób również powiązania niekoniecznie istniejące w dziedzinie zainteresowań użytkownika, co jest poważną wadą obu wspomnianych rozwiązań. Przyczyny mogą być następujące. Po pierwsze, w zbiorze dokumentów odpowiedzi mogą zaleźć się zarówno dokumenty relewantne do zainteresowań użytkownika, jak i nierelewantne. Tak więc wyznaczone może zostać powiązanie pomiędzy terminem z dokumentu relewantnego oraz terminem z dokumentu nierelewantnego. Po drugie, cała kolekcja dokumentów dotyczy zazwyczaj wielu dziedzin tematycznych – wyznaczone może zostać powiązanie pomiędzy terminem z jednej dziedziny tematycznej oraz terminem z innej dziedziny tematycznej. Szczególnie uwidacznia się ten problem podczas wyszukiwania w sieci WWW, w której

użytkownicy nie mają prostej metody ograniczenia zakresu wyszukiwania do wybranej dziedziny tematycznej.

Jak już wspomniano powyżej, w literaturze przedmiotu, tradycyjnie profil użytkownika reprezentowany jest w postaci jednego wektora w przestrzeni *n*–wymiarowej. W takim profilu ujawniają się jednak istotne wady zarówno na poziomie tworzenia, jak i wykorzystania profilu.

Pierwszym problemem jest fakt, że modyfikacje profilu, w postaci jednego wektora w przestrzeni *n*–wymiarowej, w wyniku pojawiania się różnorodnych pytań z różnych dziedzin mogą przebiegać w nieprzewidywalnych kierunkach. Nieprzewidziany kierunek modyfikacji polega na nieproporcjonalnym wzmacnianiu w wektorze profilu wag terminów, które pojawiają się w odpowiedzi na wiele różnych pytań użytkownika, niezależnie od dziedziny wyszukiwania. Wielokrotne pojawianie się takiego terminu powoduje nieproporcjonalne zwiększanie wagi jednego terminu w profilu, co jest poważną wadą, a w efekcie wybieranie tego terminu do modyfikacji kolejnego pytania, jako terminu istotnego z racji na „sztucznie" podwyższoną wagę tego terminu w profilu. W rzeczywistości termin ten nie jest terminem szczególnie istotnym w reprezentowaniu określonej dziedziny zainteresowań użytkownika, a wysoką wagę uzyskał tylko dlatego, że jest terminem powszechnym, tj. używanym we wszystkich dziedzinach zainteresowania.

Kolejnym nierozwiązanym problemem, zidentyfikowanym na podstawie analizy literatury, jest wspomaganie niedoświadczonego użytkownika w formułowaniu pytania do systemu wyszukiwania informacji. Znany jest fakt, że pierwotne pytanie formułowane jest przez niedoświadczonego użytkownika na podstawie jego nikłej wiedzy z dziedziny wyszukiwania. Na takie pytanie zazwyczaj w odpowiedzi przekazanych jest użytkownikowi wiele dokumentów niezgodnych z jego rzeczywistą potrzebą. Zastosowanie dla takiego zbioru dokumentów odpowiedzi metody modyfikacji pytania na podstawie analizy lokalnej, czyli analizy dokumentów odpowiedzi, nie spowoduje poprawy wyników następnego wyszukiwania.

W klasycznych systemach wyszukiwania informacji przyjęte jest założenie, że termin zadany przez użytkownika w pytaniu funkcjonuje w takim samym sensie w pytaniu (jest w takim samym sensie), jak został użyty przez autora dokumentu w tekście tego dokumentu. Nie jest to założenie zawsze prawdziwe, szczególnie dla wyszukiwań w sieci WWW, gdzie każde pytanie analizowane jest dla bardzo szerokiego zakresu tematycznego dokumentów kolekcji. Wadą wyszukiwania dokumentów bez uwzględnienia sensu terminów z pytania jest odpowiedź zawierająca dokumenty relewantne oraz dokumenty poindeksowane terminami z pytania, ale użytymi w odmiennym sensie niż termin zadany przez użytkownika w pytaniu. Metodą, która umożliwi sprecyzowanie sensu terminu w pytaniu, a tym samym ograniczenie liczby dokumentów nierelewantnych w odpowiedzi, jest podanie kontekstu terminu. Kontekst wprowadzony może być przez dołączenie do pytania innych terminów, które często

współwystępują z tym terminem w określonej dziedzinie tematycznej. Dołączanie kolejnych terminów do pytania prowadzi do sprecyzowania sensu terminu początkowego, a tym samym sprecyzowania całego pytania. W wyniku kolejnego precyzowania pytania, odpowiedź będzie węższa (będzie zawierała mniej dokumentów, wśród których będzie więcej dokumentów odpowiadających zainteresowaniom użytkownika).

Analiza literatury oraz zidentyfikowanie na podstawie analizy wszystkich opisanych dotychczas problemów pozwala postawić tezę, że połączenie metod relewancyjnego sprzężenia zwrotnego oraz analizy lokalnej w procesie modyfikacji pytania użytkownika jest obiecującym rozwiązaniem podniesienia satysfakcji użytkownika podczas wyszukiwania w sieci WWW.

4. Model użytkownika

4.1. Koncepcja modelu użytkownika

Zanim przejdziemy do precyzyjnych definicji proponowanych rozwiązań (podrozdziały 4.2–4.9), zmierzających do poprawy efektywności i skuteczności wyszukiwania, przeanalizujmy podstawowe idee i pojęcia w sposób mniej sformalizowany. Powinno to ułatwić późniejszą lekturę szczegółowych definicji.

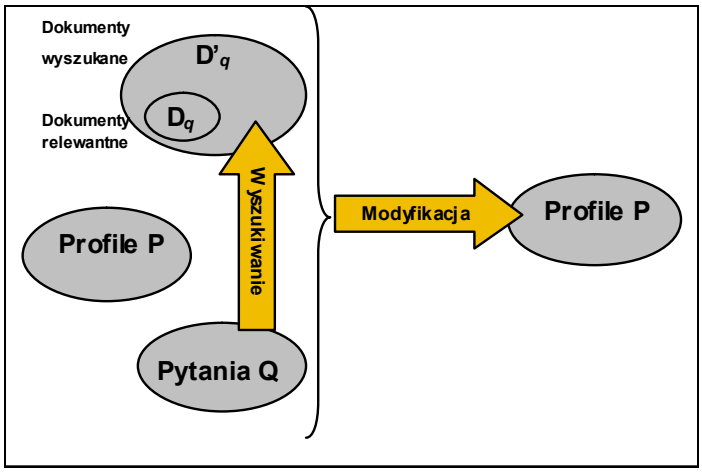
Można zaobserwować, że użytkownicy korzystający z internetowych systemów wyszukiwawczych zazwyczaj formułują pytania dotyczące różnorodnej tematyki – nie koncentrują się tylko na jednej wybranej tematyce, jak to miało miejsce w klasycznych systemach wyszukiwania informacji (Internet zawiera niezmiernie zróżnicowane zasoby informacyjne). Wiadomo również, że w klasycznych systemach wyszukiwania informacji do reprezentowania zainteresowania użytkownika stosowany był, z sukcesem, profil będący pojedynczą strukturą – najczęściej pojedynczym wektorem w pewnej przestrzeni. Struktura taka była wystarczająca, ponieważ wyszukiwania zazwyczaj prowadzone były przez użytkownika w jednej dziedzinie tematycznej. Ze względu na różnorodność zainteresowań, prezentowanych przez użytkownika systemu internetowego, a tym samym różnorodność pytań stawianych do systemu, zaistniała potrzeba rozszerzenia profilu tak, aby skutecznie reprezentował wszystkie tematy zainteresowań ujawniane i wykorzystywane w trakcie pracy z siecią.

Wykorzystanie w systemach internetowych profilu reprezentowanego przez pojedynczy wektor powoduje, że profil zawsze podlega modyfikacji w wyszukiwaniu, niezależnie od dziedziny tematycznej, której dotyczy bieżące pytanie użytkownika. W efekcie powstaje profil, w którym reprezentowane są wszystkie zainteresowania użytkownika, jednocześnie jednak nie ma możliwości wykorzystania zgromadzonych informacji do wspomagania wyszukiwania w jednej wybranej dziedzinie zainteresowań. Modyfikacja różnorodnych pytań za pomocą profilu w postaci pojedynczego wektora, wspólnego dla wszystkich zainteresowań użytkownika, będzie powodować niepożądany wpływ zainteresowań z jednej dziedziny na wynik wyszukiwania w innej dziedzinie zainteresowań.

Opisane powyżej problemy stały się inspiracją do zaproponowania odmiennej koncepcji profilu, w której różne zainteresowania użytkownika reprezentowane są w różnych *subprofilach* – częściach składowych *struktury złożonego profilu*. Każde pytanie użytkownika powiązane jest tylko z jednym subprofilem, który zawiera reprezentację o konkretnej dziedziny zainteresowań użytkownika (Indyka-Piasecka, 2002) (Indyka-Piasecka i Piasecki, 2003) (Indyka-Piasecka i Daniłowicz, 2004). Zapamiętane pytanie wskazuje na zainteresowanie użytkownika pewną

dziedziną tematyczną. Użytkownik formułując pytanie posługuje się *swoim własnym słownictwem*. Nie zawsze musi być ono prawidłowe, w sensie powszechnie stosowanych terminów oraz może opisywać dobrze daną dziedzinę tylko i wyłącznie z subiektywnego punktu widzenia danego użytkownika. Natomiast zadaniem subprofilu powiązanego z konkretnym pytaniem jest obiektywny opis *tej samej dziedziny tematycznej*, ale z zastosowaniem słownictwa powszechnie stosowanego w tej dziedzinie w sieci. Oczywiście, wyłania się tu poważny problem sposobu identyfikacji *obiektywnie* 'odpowiedniego' słownictwa.

Ponieważ profil łączy jednoznacznie *subiektywnie* sformułowane pytanie użytkownika z *obiektywnym* opisem w subprofilu, można powiedzieć, że zaproponowany w niniejszej pracy *profil użytkownika* jest strukturą opisującą *translację* pomiędzy terminologią wykorzystywaną przez użytkownika w pytaniu, a słownictwem powszechnie stosowanym w danej dziedzinie zainteresowań użytkownika. Użytkownik używa oczywiście tych samych *wyrazów* (w sensie napisów), które występują również w sieci, np. wyrazów języka polskiego, jednak *sensy* przypisywane przez użytkownika używanym przez siebie wyrazom mogą się różnić od sensów przypisywanych powszechnie tym samym wyrazom (w sensie napisów) w sieci Internet. W profilu translacja pomiędzy terminami użytkownika a terminami z sieci jest wyrażona poprzez przyporządkowanie *wzorcowi pytania* użytkownika *subprofilu*, wyznaczonego w procesie analizy dokumentów relewantnych odpowiedzi. Proces ten stanowi mechanizm identyfikacji obiektywnie 'odpowiedniego' słownictwa. Wzorcem pytania jest pytanie, które zostało zadane przez użytkownika co najmniej raz do internetowego systemu wyszukiwania informacji. Tak więc subprofil reprezentuje dziedzinę zainteresowań użytkownika, natomiast wzorzec pytania użytkownika, powiązany z tym subprofilem, jest swego rodzaju etykietą, wskaźnikiem, według którego użytkownik subiektywnie identyfikuje pewną dziedzinę.



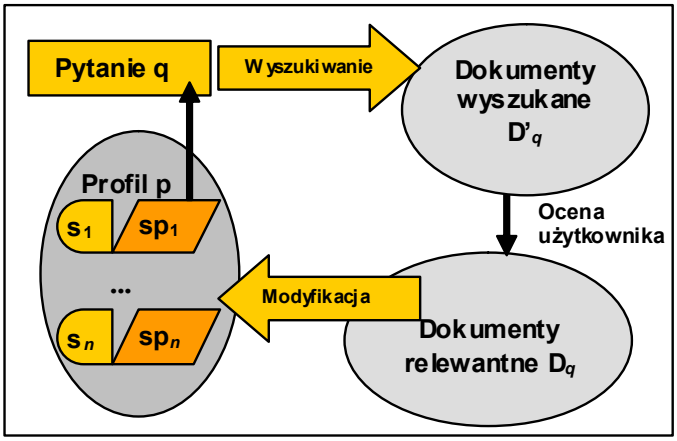
Rysunek 4.1: Model internetowego systemu wyszukiwania informacji z zastosowaniem profilu użytkownika.

W internetowym systemie wyszukiwania informacji, użytkownik stawia pytanie, w odpowiedzi na które otrzymuje od systemu zbiór dokumentów. Intencją użytkownika jest uzyskanie dokumentów z interesującej go tematyki. Oczekuje więc, że w odpowiedzi systemu znajdzie właśnie takie dokumenty. Często jednak oczekiwanie to nie jest zaspokojone, tzn. dokumenty relewantne są trudne do odnalezienia w odpowiedzi składającej się z wielu tysięcy dokumentów lub znalezione dokumenty nie w pełni satysfakcjonują użytkownika. Dzieje się tak, ponieważ pytanie zadane przez użytkownika jest bardzo często zbyt ogólne lub użyte są w pytaniu słowa, które nie są powszechnie stosowane w tematyce, której dotyczy wyszukiwanie. Istotną pomocą w sformułowaniu pytania może być dodatkowa informacja zgromadzona w profilu – opisująca zainteresowania użytkownika, a przechowywana w internetowym systemie wyszukiwawczym. Założono, że dobrym źródłem informacji o zainteresowaniach użytkownika mogą być dokumenty, które użytkownik wskaże wśród dokumentów odpowiedzi jako zgodne z jego zainteresowaniami. Dokumenty wskazane przez użytkownika jako dobrze opisujące jego bieżące zainteresowania, czyli *dokumenty relewantne*, są bardzo istotnym źródłem informacji, ponieważ zostały ocenione przez samego użytkownika. Jednocześnie, dokumenty te zawierają terminy, które opisują bieżące zainteresowania użytkownika (tzn. te, których dotyczy pytanie) i terminy te są stosowane w danej dziedzinie tematycznej. Terminy kluczowe wyselekcjonowane z dokumentów relewantnych, wykorzystane do zmodyfikowania pytania, w istotny sposób mogą polepszyć wyniki wyszukiwania, co zostało pokazane w eksperymentach przeprowadzonych w ramach pracy i zaprezentowanych w Rozdziale 5. Tym samym, spełnione zostało podstawowe oczekiwanie użytkownika, że otrzyma on z internetowego systemu wyszukiwawczego większą liczbę dokumentów relewantnych w mniej obszernej odpowiedzi.

Zaproponowany w pracy profil użytkownika, składający się z subprofili i wzorców, służy do *modyfikacji pytania* sformułowanego przez użytkownika w internetowym systemie wyszukiwawczym. Pytanie modyfikowane jest automatycznie na podstawie informacji zgromadzonych w profilu, a następnie zadawane do systemu wyszukiwawczego. Modyfikacja zadanego pytania realizuje *procedurę translacji* pomiędzy terminami użytkownika a terminami ‘odpowiednimi’ dla wyszukiwania w sieci Internet. Część pytań nie podlega jednak modyfikacji – upraszczając chwilowo, możemy je określić jako ‘niepodobne’ do żadnego z pytań zadanych uprzednio. Zasada podobieństwa zostanie określona precyzyjnie w dalszej części rozdziału.

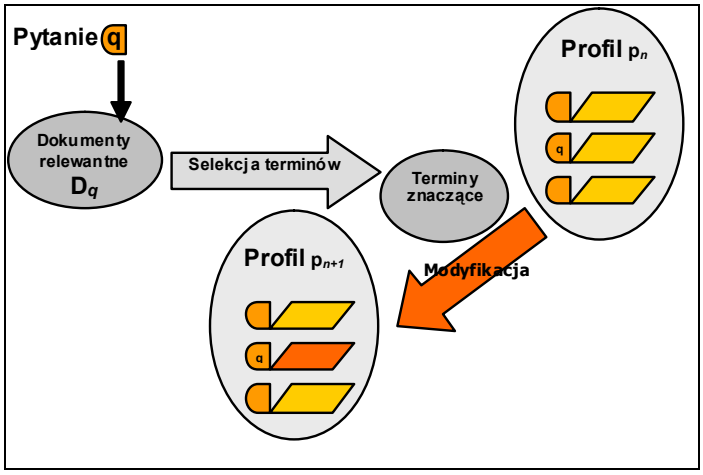
Użytkownik ocenia otrzymaną odpowiedź i, jeśli nie jest ona w pełni satysfakcjonująca, aby polepszyć wyniki kolejnego wyszukiwania, wskazuje nowe dokumenty relewantne w odpowiedzi. Ze wskazanych dokumentów relewantnych automatycznie wyselekcjonowane są terminy kluczowe, istotne w terminologii stosowanej w danej dziedzinie zainteresowań użytkownika. Terminy te zapamiętywane

są w odpowiednim subprofilu i wykorzystywane do zmodyfikowania jednego z kolejnych pytań zadanych przez użytkownika (Rysunek 4.2).



Rysunek 4.2: Schemat procesu wyszukiwania z wykorzystaniem profilu.

Z dokumentów relewantnych wskazanych przez użytkownika w odpowiedzi wybierane są terminy kluczowe, nazywane w pracy *terminami znaczącymi*. W metodzie selekcji terminów wykorzystano koncepcję: *terminów dyskryminacyjnych*, wprowadzoną przez Saltona i McGilla (Salton i McGill, 1983) oraz *wskaźnika ważności* dla terminu, zaproponowanego przez Goldberga (Goldberg, 1996). Szczegóły metody selekcji terminów zostały opisane w podrozdziałach 4.7.2, 4.7.3. Wyselekcjonowane terminy znaczące wprowadzane są do subprofilu, a następnie wykorzystane do zmodyfikowania pytania użytkownika. Schemat modyfikacji profilu użytkownika obrazuje Rysunek 4.3.



Rysunek 4.3: Schemat modyfikacji profilu użytkownika.

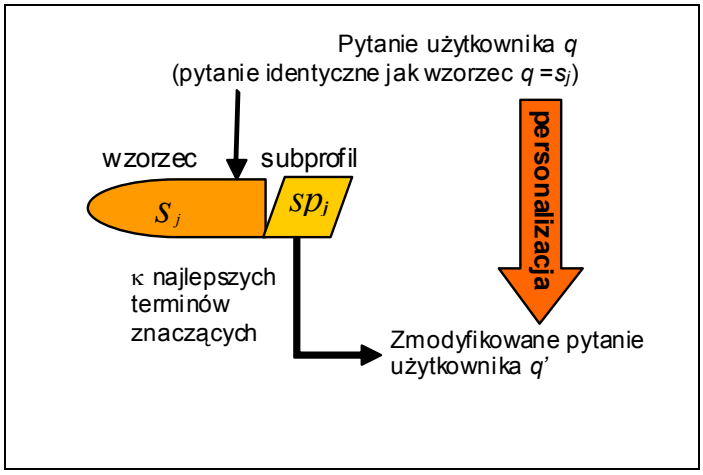
Jak wspomniano wcześniej, profil jest strukturą składającą się z: subprofili oraz wzorców pytań – identyfikujących poszczególne subprofile. W wyniku jednego wyszukiwania modyfikowany jest zawsze tylko jeden subprofil, ten który jest

identyfikowany przez wzorzec pytania identyczny z pytaniem zadanym przez użytkownika. Natomiast do modyfikacji pytania zadanego przez użytkownika może być wykorzystany więcej niż jeden subprofil. W najprostszej sytuacji, jeśli użytkownik podczas kolejnego wyszukiwania zada pytanie identyczne z zadanym już uprzednio, to przed przekazaniem do internetowego systemu wyszukiwawczego, pytanie zostanie zmodyfikowane przez terminy zgromadzone w subprofilu. Zastosowany zostanie subprofil oznaczony wzorcem pytania identycznym z powtórzonym pytaniem – identyczność wzorca pytania z zadanym pytaniem jednoznacznie wskazuje subprofil, z którego terminy mają być użyte do zmodyfikowania pytania. W bardziej skomplikowanym przypadku, brak jest wzorca pytania (przypisanego do dowolnego subprofilu) identycznego z zadanym pytaniem, ale część terminów z zadanego pytania występuje w już zapamiętanych wzorcach. Tak więc, istniejące subprofile zostają użyte do sformułowania *hipotezy* dotyczącej prawdopodobnej, kontekstowej translacji terminów użytkownika na podstawie terminów z subprofili oznaczonych wzorcami *podobnymi* do zadanego pytania. Do kwestii tej wrócimy w dalszej części rozdziału.

Modyfikacja pytania na podstawie informacji zgromadzonych w profilu użytkownika realizuje *proces personalizacji wyszukiwania*. Jak to było wspomniane, w opracowanym modelu pytanie może zostać zmodyfikowane w dwóch przypadkach:

- gdy pytanie użytkownika jest *identyczne* jak wzorzec pytania istniejący w profilu,
- lub gdy jest *podobne* do jednego lub kilku wzorców w profilu.

Schemat zaproponowanych dwóch modyfikacji pytania obrazują, odpowiednio, rysunki 4.4 i 4.5.



Rysunek 4.4: Modyfikacja pytania użytkownika identycznego jak wzorzec pytania w profilu.

W każdym z wymienionych dwóch przypadków, z odpowiedniego subprofilu wybierane są najlepsze terminy znaczące, które następnie zastępują terminy z pytania sformułowanego przez użytkownika. Pytanie zmodyfikowane, w zamierzeniu lepiej opisujące zainteresowania użytkownika, zadawane jest do internetowego systemu

wyszukiwania informacji i, w wyniku wyszukiwania, użytkownik otrzymuje bardziej satysfakcjonującą odpowiedź.

W profilu nigdy nie występują dwa takie same wzorce pytań. Każdy subprofil oznaczony jest przez dokładnie jeden wzorzec. Ideą wykorzystania profilu o strukturze subprofili i wzorców pytań przypisanych tym subprofilom jest powiązanie słownictwa użytkownika, wyrażonego przez sformułowane pytanie, zapamiętane w profilu jako wzorzec pytania, ze słownictwem stosowanym w dziedzinie zainteresowań użytkownika, reprezentowanym przez subprofil. Bardzo często można zaobserwować, że użytkownik rozpoczynając wyszukiwania w pewnej dziedzinie swoich zainteresowań na początku zadaje pytanie, które jest swego rodzaju pytaniem próbnym, testowym. Po przeanalizowaniu dokumentów otrzymanych w odpowiedzi na to pierwsze pytanie, weryfikuje słownictwo użyte w postawionym wcześniej pytaniu i zmienia pytanie tak, aby otrzymać więcej dokumentów z interesującej go dziedziny oraz aby zmniejszyć liczbę dokumentów odpowiedzi w ogóle.

Zaproponowane w pracy rozwiązanie w pewnym stopniu modeluje tę opisaną powyżej taktykę wyszukiwania stosowaną przez użytkowników internetowego systemu wyszukiwania informacji. W wyszukiwaniu wspomaganym przez wykorzystanie profilu, rolą użytkownika jest zadanie pytania początkowego oraz ocena i wskazanie dokumentów relewantnych w kolejnych odpowiedziach systemu. Natomiast analiza dokumentów relewantnych, modyfikacja pytania oraz gromadzenie terminów znaczących w odpowiednim subprofilu reprezentującym wybraną dziedzinę (wykorzystywanych następnie do modyfikacji pytania) wykonywana jest automatycznie.

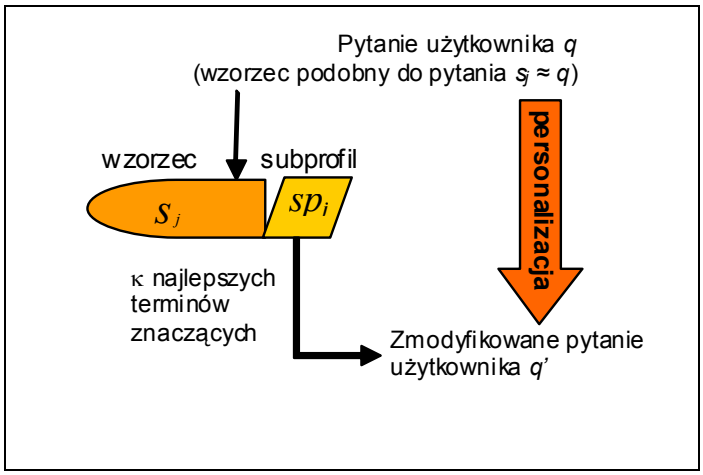
Możliwość stopniowego douczania się przez system, jakiej informacji użytkownik szuka zadając konkretne pytanie istnieje wtedy, gdy pytanie nie zmienia się w czasie. Taka sytuacja może mieć miejsce, ponieważ zasoby sieci Internet cały czas się zmieniają i użytkownik często wraca do pytań zadawanych wcześniej. Jednak równie często użytkownik formułuje nowe pytania, dotyczące kolejnych zagadnień. Pojawia się więc pytanie: czy profil może tu pomóc w poprawie wyszukiwania dla nieznanych pytań? Możliwość taka istnieje. Należy zauważyć, że poszczególne subprofile, łącząc wzorce pytań z wektorem wag terminów określają znaczenie konkretnego podzbioru terminów użytkownika. Wzorzec pytania przypomina frazę języka naturalnego, której sens, w kontekście systemu wyszukiwania informacji, jest opisany wektorem subprofilu. Znaczenie każdego terminu użytkownika ze wzorca pytania jest określone poprzez subprofil w kontekście pozostałych terminów użytkownika (czyli całej ‘frazy’) tam występujących.

Wraz ze wzrostem liczby subprofili w profilu, zwiększa się prawdopodobieństwo, że nowe pytanie użytkownika będzie zawierało podzbiory terminów już występujące w profilu jako wzorce pytań powiązane z pewnymi subprofilami. W takim przypadku, znaczenie pewnych podzbiorów terminów z nowego pytania jest już określone

w poszczególnych subprofilach. Porównując dla poszczególnych terminów z nowego pytania zawarte w tych subprofilach wektory wag, możemy spróbować sformułować przypuszczenie, jaki jest sens danego terminu z nowego pytania użytkownika (w kontekście znanych sensów 'fraz'). Oczywiście należy zachować tu daleko idącą ostrożność – podobieństwo wyrażone przez wysokość wag pewnych terminów z systemu wyszukiwawczego jest jedynie wskazówką, że dany termin użytkownika może być tak rozumiany, czyli może być zastąpiony podzbiorem terminów z sieci Internet, określonym przez podobieństwo występowania wysokich wag w subprofilach identyfikowanych wzorcami podobnymi do pytania.

Precyzyjny opis procesu formułowania *hipotezy*, co do prawdopodobnego znaczenia terminów użytkownika, poprzedźmy definicją pojęcia *podobieństwa wzorca pytania powiązanego z subprofilem do pytania użytkownika*. Wzorzec pytania jest podobny do pytania użytkownika, jeśli wzorzec jest wyrażeniem składowym jednego z wariantów wyrażenia boolowskiego stanowiącego pytanie.

Jeśli wzorzec jest podobny do pytania użytkownika, to formułujemy hipotezę mówiącą, że terminy z subprofilu powiązanego ze wzorcem podobnym mogą być użyte do opisania potrzeby informacyjnej użytkownika reprezentowanej przez pytanie. W pierwszym kroku znajdowane są w profilu wszystkie wzorce podobne do zadanego pytania. Jeśli jest więcej niż jeden taki wzorzec, to tworzony jest wektor będący sumą wektorów wag, czyli subprofili powiązanych ze wzorcami podobnymi. W następnym kroku, dla każdego terminu w pytania użytkownika wyznaczane są terminy z wektora zsumowanego. Dla każdego terminu z pytania rozważamy tylko sumę wag z tych subprofili podobnych do pytania, których wzorce zawierają rozpatrywany termin (za każdym razem szukamy reprezentacji sensu konkretnego terminu użytkownika). Terminy o najwyższych wagach w sumarycznych wektorach będą zastępowały terminy pytania użytkownika. Sprawdzane jest również, czy wyznaczony termin występuje we wszystkich subprofilach sumowanych w przypadku danego terminu użytkownika.



Rysunek 4.5: Modyfikacja pytania użytkownika, gdy w profilu istnieje wzorzec podobny do pytania.

Jeśli warunek ten jest spełniony, to termin jest dołączany do pytania zmodyfikowanego. W ostatnim kroku, do pytania zmodyfikowanego dołączane są pozostałe terminy z pytania użytkownika, które nie występowały we wzorcach podobnych. W ostatnim kroku, do pytania zmodyfikowanego dołączane są pozostałe terminy z pytania użytkownika, które nie występowały we wzorcach podobnych.

Schemat modyfikacji pytania użytkownika, gdy w profilu istnieje wzorzec podobny do pytania użytkownika obrazuje rysunek 4.5.

Definicje wprowadzonych tu pojęć i metod zostaną przedstawione w kolejnych podrozdziałach.

Wzorce i subprofile

Zaproponowany w pracy model użytkownika wprowadza strukturę subprofili i wzorców pytań. Każdy z subprofili reprezentuje inne zainteresowania użytkownika. Dodatkowo, metoda tworzenia i modyfikacji subprofilu zapewnia, że istnieje powiązanie pomiędzy terminami subprofilu, wyznaczonymi z dokumentów relewantnych odpowiedzi oraz terminami identyfikującego go wzorca pytania, złożonego ze słów stosowanych przez użytkownika. Reprezentacja zainteresowań użytkownika jest wyznaczana na podstawie wskazanych przez użytkownika kilku dokumentów spośród dokumentów odpowiedzi. Wskazane dokumenty ograniczają możliwe zakresy dziedzinowe stosowania terminu. Ma to istotny wpływ na terminy, które zostaną wyznaczone jako reprezentujące zainteresowania użytkownika.

Na podstawie analizy ograniczeń i możliwości zastosowania do modyfikacji pytania opisanej w literaturze macierzy współwystępowania terminów SM^{loc} (opisanej w podrozdziale 3.6.5) oraz opracowanej na jej podstawie przez autora pracy macierzy wystąpień terminów AM^{loc} (opisanej również w podrozdziale 3.6.5), w niniejszej pracy zaproponowano strukturę, która reprezentuje związek pomiędzy całym pytaniem użytkownika a zbiorem terminów znaczących z dokumentów relewantnych danego pytania. Określenie powiązania pomiędzy całym pytaniem użytkownika a zbiorem terminów znaczących umożliwi poprawną modyfikację kolejnego pytania niezależnie od faktu uznania terminu z pytania za termin znaczący (co było nierozwiązanym problemem w macierzy współwystępowania terminów[1] SM^{loc}) oraz niezależnie od formy pytania (pytanie złożone lub pytanie proste – ograniczenie macierzy wystąpień terminów[2] AM^{loc}).

Zaproponowana w pracy struktura profilu zawiera wzorce pytań użytkownika oraz subprofile, które składają się na profil użytkownika zdefiniowany i opisany w Rozdziale 4. Elementami profilu są pary: wzorzec pytania – subprofil. Subprofil jest

[1] Macierz współwystępowania terminów SM^{loc} omówiono w podrozdziale 3.6.5.

[2] Macierz wystąpień terminów AM^{loc} omówiono w podrozdziale 3.6.5.

reprezentowany przez *n*–wymiarowy wektor wag, gdzie waga $w_{j,i}^{(k)}$ terminu znaczącego tz_i w subprofilu wyznaczana jest na podstawie równania (4.6.1.1). Termin t_i zostaje wybrany terminem znaczącym tz_i ze zbioru dokumentów relewantnych dla pytania q_j. Kolejne pozycje wektora subprofilu odpowiadają terminom ze zbioru *T*. Jeśli termin t_i nie jest terminem znaczącym tz_i dla tego subprofilu, waga $w_{j,i}^{(k)}$ dla tego terminu w wektorze subprofilu jest równa zero. Każdy subprofil powiązany jest z jednym wzorcem pytania oraz identyfikowany jest przez ten wzorzec pytania.

Oznaczenie pytania q_j indeksem *j* nie implikuje, że pytanie q_j musi pojawić się podczas *j*–tego procesu wyszukiwania. Takie samo pytanie q_j może pojawić się podczas dowolnego z kolejnych wyszukiwań i może powtarzać się dowolną liczbę razy w kolejnych wyszukiwaniach. Natomiast te same indeksy wzorca pytania i pytania oznaczają, że wzorzec pytania jest identyczny jak pytanie zadane przez użytkownika.

Istotnym założeniem dla zaproponowanej struktury łączącej wzorce pytań i subprofile, reprezentujące zainteresowania użytkownika, jest możliwość zastosowania tej struktury w rzeczywistym systemie wyszukiwania informacji w sieci WWW. W przypadku opisanej macierzy współwystępowania terminów[1] SM^{loc} problemem było zastosowanie macierzy jako profilu użytkownika, realizującego wyszukiwania w sieci WWW w różnych dziedzinach tematycznych. Korzystanie z tak zdefiniowanego profilu wiąże się z jednoczesnym uaktualnianiem wszystkich miar współwystępowania terminów znaczących w macierzy bez możliwości zaznaczenia, dla jakiej dziedziny wykonywane było wyszukiwanie. W efekcie gubiona jest istotna informacja o współwystępowaniu terminów w pewnej dziedzinie wiedzy. Co więcej, informacja o współwystępowaniu terminów w jednej dziedzinie wiedzy najczęściej nie ma zastosowania dla innej dziedziny.

4.2. Model systemu

System wyszukiwania informacji zawiera zbiór dokumentów *D*, zbiór profili użytkowników *P*, zbiór pytań *Q* oraz zbiór terminów *T*.

Określona jest funkcja wyszukiwawcza:

$$\omega: Q \rightarrow 2^D.$$

Wprowadzimy następujące oznaczenia dla pojęć wykorzystywanych w pracy:

- $D_q' = \omega(q,D)$ – *odpowiedź* systemu na pytanie *q* (zbiór dokumentów wyszukanych),
- $D_q \subseteq D_q'$ – zbiór *dokumentów relewantnych* w odpowiedzi systemu na pytanie *q*,

[1] Macierz współwystępowania terminów SM^{loc} omówiono w podrozdziale 3.6.5.

- $T = \{t_1, t_2, ..., t_n\}$ – zbiór wszystkich terminów należących do *dokumentów poindeksowanych* w systemie wyszukiwania informacji. Zakładamy, że w zbiorze T nie ma terminu, który nie występowałby w jakimkolwiek dokumencie zbioru D. Zbiór T nazywamy *słownikiem*.
- T_d – zbiór terminów należących do dokumentu d.

Zakładamy, że w systemie wyszukiwania informacji nie ma dokumentów nie poindeksowanych żadnym terminem indeksowym.

Profilem nazywamy obiekt $p \in P$. Profil użytkownika wyznacza *funkcja modyfikacji profilu* π:

$$p_0 = \varnothing,$$
$$p_m = \pi(q_m, D_q, p_{m-1}),$$

gdzie:

- q_m – m-te pytanie użytkownika,
- p_0 – początkowy profil użytkownika,
- p_m – profil użytkownika po modyfikacji.

Funkcja modyfikacji profilu π odwzorowuje pytanie użytkownika, zbiór dokumentów relewantnych w odpowiedzi systemu na pytanie użytkownika oraz profil użytkownika – stan przed postawieniem przez użytkownika pytania do systemu – w profil użytkownika – stan po analizie odpowiedzi na pytanie użytkownika.

W tym miejscu zebrano wszystkie podstawowe pojęcia modelu. Umieszczono tu również pojęcie profilu użytkownika. Jednak ze względu na obszerną definicję proponowanego profilu, definicja ta zaprezentowana zostanie w podrozdziale 4.6. W podrozdziale 4.8 opisano metodę tworzenia i modyfikacji profilu użytkownika.

4.3. Struktura obiektów

Obiekt x definiujemy jako n-wymiarowy wektor:

$$x = (x_1, x_2, ..., x_n),$$

gdzie $x_i \in \Re_+$.

Dla modelu możliwe są dwie następujące interpretacje wektora x:

1. Jeśli $x \in D$, to x jest reprezentacją dokumentu. Obiekt $x \in D$ będziemy oznaczać w dalszej części pracy przez d. Współrzędna x_i jest *wagą terminu* t_i w dokumencie d (stopniem reprezentacji terminu t_i w dokumencie d).
2. Jeśli $x \in SP$, gdzie SP oznacza zbiór subprofili, to x jest reprezentacją zainteresowania użytkownika wyrażonego w subprofilu. Obiekt $x \in SP$ będziemy oznaczać w dalszej części pracy przez sp. Definicję subprofilu

podajemy w podrozdziale 4.6. Współrzędna x_i jest wagą terminu t_i w subprofilu *sp*.

4.4. Reprezentacja dokumentu

Dokument d reprezentowany jest przez n-wymiarowy wektor wag terminów indeksowych zawartych w tym dokumencie, tzn. $d = (d_1, d_2, ..., d_n)$. Wartość d_i określa stopień reprezentacji przez termin t_i treści dokumentu d, gdzie $t_i \in T$. Jeżeli termin nie występuje w treści dokumentu, to oczywistym jest, że nie można w takim przypadku mówić o stopniu reprezentowania przez taki termin treści dokumentu. W pracy przyjęto, że terminowi, który nie występuje w treści dokumentu przypisujemy w danym dokumencie wagę równą zero. Waga terminu t_i występującego w dokumencie d (czyli stopień reprezentacji terminu t_i w dokumencie d lub ważność terminu reprezentującego dokument) określana jest według miary *tf–idf*. Dla terminów należących do dokumentu waga terminu obliczana jest według następującego wzoru:

$$d_i = \begin{cases} \dfrac{0{,}5 + 0{,}5\dfrac{tf(d, t_i)}{\max tf(d)} idf(t_i)}{\sqrt{\sum_{m=1}^{n}\left(0{,}5+0{,}5\dfrac{tf(d, t_m)}{\max tf(d)} idf(t_m)\right)^2}}, & t_i \in T_d \\ 0, & t_i \notin T_d \end{cases} \qquad (4.4.1)$$

t_m – termin należący do dokumentu d,

$tf(d, t_i)$ – *częstość* terminu t_i w dokumencie d,

$idf(t_i)$ – *odwrotna częstość dokumentowa* – liczba dokumentów kolekcji (oznaczanej symbolem ND) do liczby dokumentów kolekcji zawierających termin t_i, (oznaczanej symbolem nd_i), tj.:

$$idf(t_i) = \log\frac{ND}{nd_i}, \text{ gdzie } t_i \in T,$$

$\max tf(d)$ – *maksymalna częstość* terminu w dokumencie d (czynnik normalizacji).

4.5. Reprezentacja pytania

W modelu przyjęto, że pytanie kierowane przez użytkownika do systemu jest dowolnym wyrażeniem boolowskim. Dowolne pytanie w postaci boolowskiej przekształcane jest do dysjunkcyjnej postaci normalnej:

$$q = q_1 \vee q_2 \vee q_3 \vee \ldots \vee q_m,$$

gdzie:

$q_j=r_1 \wedge r_2 \wedge r_3 \wedge \ldots \wedge r_n$ – jest koniunkcją terminów niezanegowanych $r_i=t_i$, terminów zanegowanch $r_i=\neg t_i$, lub jedynek logicznych, $r_i=1$, reprezentujących terminy, które nie występują w danym pytaniu.

Rozpatrzmy przykładowe pytanie: $q_a = \neg(t_1 \vee \neg t_2) \vee t_3 \wedge t_4$. Pytanie to można przekształcić do następującej postaci: $q_a = (\neg t_1 \wedge t_2) \vee (t_3 \wedge t_4)$. Z kolei przykładowe pytanie: $q_b = t_5 \wedge t_6 \wedge \neg t_7 \vee t_8 \wedge t_9$ można przekształcić do postaci: $q_b = (t_5 \wedge t_6 \wedge \neg t_7) \vee (t_8 \wedge t_9)$.

Własność 4.5.1 Jeśli pytanie zawiera operator alternatywy, to poszczególne sekwencje koniunkcji jednego lub wielu terminów oraz zanegowanych terminów połączone operatorem alternatywy traktowane są jako osobne pytanie, czyli

$$\omega((q_1 \vee q_2),D) = \omega(q_1,D) \cup \omega(q_2,D).$$

Każde pytanie ‘składowe’ definiowane jest poprzez pojedynczą sekwencję koniunkcji: terminów i zanegowanych terminów. Do internetowego systemu wyszukiwania informacji zadawane są pytania w postaci koniunkcji terminów, a następnie łączone są odpowiedzi i użytkownikowi prezentowany jest do oceny relewancji jeden zbiór dokumentów odpowiedzi.

Weźmy przykładowe pytanie $q_A = t_1 \wedge t_2 \vee t_3 \wedge t_4$. Pytanie przed skierowaniem go do internetowego systemu wyszukiwania informacji zostanie przekształcone do następujących dwóch pytań składowych: $q_C = t_1 \wedge t_2$, $q_D = t_3 \wedge t_4$. Kolejne przykładowe pytanie

$q_B = t_5 \wedge t_6 \wedge \neg t_7 \vee t_8 \wedge t_9$ również zostanie przekształcone do dwóch pytań: $q_E = t_5 \wedge t_6 \wedge \neg t_7$ oraz $q_F = t_8 \wedge t_9$.

Zamieszczone powyżej przykłady mają przybliżyć ideę przekształcenia dowolnego pytania zadanego przez użytkownika. Wszystkie przykłady zawierają alternatywę dwóch wyrażeń koniunkcyjnych, dzięki czemu przykłady są proste i jasne. W definicji pytania nie ma jednak ograniczenia co do liczby wyrażeń koniunkcyjnych, z których może być zbudowane pytanie użytkownika.

4.6. Profil użytkownika

Proces wyznaczania profilu użytkownika należy do procedur systemu wyszukiwania informacji. Profil użytkownika powstaje na podstawie informacji uzyskanych ze zweryfikowanej przez użytkownika odpowiedzi na postawione pytanie.

Weryfikacja pytania polega na wskazaniu przez użytkownika dokumentów relewantnych wśród dokumentów odpowiedzi.

Poniżej sprecyzujemy pojęcia stosowane do definiowania profilu użytkownika.

Terminem znaczącym tz_i nazywamy termin, który należy do każdego ze wskazanych przez użytkownika dokumentów relewantnych w odpowiedzi, tzn. termin ten posiada niezerowe wagi w tych dokumentach oraz termin ten został wybrany w *procesie selekcji* terminów znaczących, jako termin dobrze opisujący dziedzinę zainteresowań użytkownika. *Metoda selekcji* terminów znaczących[1] tz_i zostanie szczegółowo opisana w podrozdziale 4.7.2.

Wzorcem pytania użytkownika s nazywamy wyrażenie boolowskie odpowiadające koniunkcyjnemu wyrażeniu w pytaniu zadanym przez użytkownika:

$$s_j = r_1 \wedge r_2 \wedge r_3 \wedge \ldots \wedge r_n ,$$

gdzie na człon r_i nałożone są takie same warunki jak w przypadku pytania, zdefiniowanego w podrozdziale 4.5, tzn. $r_i = t_i$ lub $\neg t_i$ lub 1.

Wzorzec pytania użytkownika s_j *identyfikuje* subprofil sp_j. Związek pomiędzy pytaniem użytkownika, wzorcem pytania i subprofilem jest szczegółowo wyjaśniony w opisie metody tworzenia i modyfikacji profilu użytkownika w podrozdziale 4.8.

Subprofil użytkownika sp $\in SP$ jest n–wymiarowym wektorem wag terminów, które należą do dokumentów relewantnych:

$$sp_j^{(k)} = (w_{j,1}^{(k)}, w_{j,2}^{(k)}, w_{j,3}^{(k)}, \ldots, w_{j,n}^{(k)}) ,$$

gdzie:

SP	–	zbiór subprofili,
k	–	numer ostatniej modyfikacji subprofilu użytkownika,
$w_{j,i}^{(k)}$	–	waga terminu znaczącego $tz_i \in T$, po k–tej modyfikacji w subprofilu sp_j.

Waga $w_{j,i}^{(k)}$ obliczona jest na podstawie częstości wystąpienia terminu znaczącego tz_i w dokumentach relewantnych w k–tym wyszukiwaniu. Waga uwzględnia również częstość występowania terminu we wszystkich dokumentach kolekcji, jak i liczbę dotychczas wykonanych modyfikacji subprofilu (tj. dotychczasowe k-1 wyszukiwań z wykorzystaniem tego subprofilu). Waga terminu znaczącego tz_i w subprofilu modyfikowana jest tylko wtedy, gdy w procesie wyszukiwania wykorzystany był dany subprofil, a termin tz_i wystąpił we wszystkich dokumentach relewantnych wskazanych przez użytkownika.

W tym miejscu chcemy zaznaczyć, że zarówno subprofil sp_j, jak i dokument d są wektorami przestrzeni n-wymiarowej, jak to zostało zdefiniowane w podrozdziale 4.3, gdzie $n = |T|$. Inne jest jednak znaczenie współrzędnych każdego z tych wektorów oraz

[1] *Terminy znaczące* odróżniamy od słów kluczowych należących do dokumentów relewantnych poprzez dodanie indeksu z do oznaczenia terminu.

sposób obliczania wartości współrzędnych, czyli wag terminów. Dla dokumentu d wartość współrzędnej wektora to waga terminu t_i w dokumencie d. Waga ta obliczana jest jednokrotnie dla danego dokumentu na podstawie funkcji (4.4.1). Dla subprofilu sp_j wartość współrzędnej wektora to stopień reprezentacji przez termin znaczący tz_i zainteresowań użytkownika. Waga terminu tz_i w subprofilu jest obliczana na podstawie podanej poniżej funkcji (4.6.1.1) i systematycznie modyfikowana (na podstawie tej samej funkcji) po każdym wyszukiwaniu, w którym został wykorzystany dany subprofil.

Profil użytkownika $p \in P$ jest zbiorem par:

$$p = \{\langle s_1, sp_1\rangle, \langle s_2, sp_2\rangle, \ldots, \langle s_l, sp_l\rangle\}$$

gdzie:

s_j – wzorzec pytania użytkownika;

sp_j – subprofil użytkownika; subprofil identyfikowany jest przez wzorzec pytania użytkownika.

Zdefiniujmy pojęcia niezbędne w dalszych rozważaniach.

Definicja 4.6.1 Pytanie q_i jest *identyczne* jak wzorzec s_j, jeśli są to identyczne wyrażenia boolowskie.

Definicja 4.6.2 Wzorzec s_j jest *podobny* do pytania q_i, jeśli wzorzec s_j jest podwyrażeniem pytania q_i, tzn. wyrażeniem składowym jednego z wariantów wyrażenia boolowskiego stanowiącego pytanie q_i.

Przykład 4.6.1 Jeśli pytanie jest w postaci: $q_1 = t_1 \wedge \neg t_2 \wedge t_3$, a wzorcami są: $s_2 = t_2 \wedge t_4$, $s_3 = t_1 \wedge \neg t_2 \wedge t_5$, to wzorce s_2 i s_3 nie są podobne do pytania q_1.

Przykład 4.6.2 Jeśli pytanie jest w postaci: $q_3 = t_1 \wedge t_2 \wedge t_3$, a wzorcami są: $s_4 = t_1 \wedge t_2$,
$s_5 = t_1 \wedge t_3$, $s_6 = t_2$. Wzorce s_4, s_5, s_6 są podobne do pytania q_3.

Definicja 4.6.3 *Frazą* nazywamy koniunkcję terminów znaczących tz_i, które należą[1] do jednego lub wielu subprofili oraz wyznaczone zostały do modyfikacji pytania na podstawie profilu użytkownika.

4.6.1. Waga terminu znaczącego w profilu

Profil użytkownika zawiera tylko terminy znaczące tz_i, wyselekcjonowane spośród wszystkich terminów t_i należących do dokumentów relewantnych. Dokumenty relewantne zostały wskazanych przez użytkownika w odpowiedziach na kolejne pytania.

Waga terminu tz_i w profilu p wyznaczana jest na podstawie poniższej funkcji, inspirowanej pracą (Daniłowicz, 1998):

$$w_{j,i}^{(k)} = \frac{1}{k}\left((k-1)\, w_{j,i}^{(k-1)} + wz_i^{(k)}\right), \qquad (4.6.1.1)$$

gdzie:

- k – numer modyfikacji subprofilu (po k-tym wyszukiwaniu)
- i – indeks terminu w słowniku T,
- j – indeks subprofilu,
- $w_{j,i}^{(k)}$ – waga terminu znaczącego tz_i w profilu po k–tej modyfikacji subprofilu identyfikowanego przez wzorzec s_j (po k–tym wyszukiwaniu dokumentów z wykorzystaniem danego subprofilu);
- $wz_i^{(k)}$ – waga terminu znaczącego tz_i wyznaczona dla zbioru dokumentów relewantnych w k–tym wyszukiwaniu i wykorzystana w procesie selekcji terminów znaczących z tych dokumentów relewantnych[2].

4.7. Reprezentowanie dziedziny zainteresowań użytkownika – analiza dokumentów odpowiedzi

Profil użytkownika powstaje na podstawie informacji dostarczonych przez użytkownika poprzez wskazanie dokumentów interesujących go pomiędzy dokumentami odpowiedzi. Zakładamy, że działanie użytkownika ogranicza się do binarnej oceny każdego dokumentu jako: interesujący lub nieinteresujący. Dokumenty odpowiedzi zawierają dokumenty wyszukane przez internetowy system wyszukiwawczy. Nie wszystkie jednak dokumenty wyszukane przez system odpowiadają rzeczywistym zainteresowaniom użytkownika. Wskazując, według

[1] Tzn. posiadających niezerowe wagi w odpowiednim subprofilu.

[2] Metodę wyznaczania wagi wzi oraz jej wykorzystania opisano w podrozdziałach: 4.7.1 i 4.7.2.

własnej oceny, wśród dokumentów odpowiedzi dokumenty interesujące, użytkownik wskazuje prawdziwą dziedzinę swoich zainteresowań. Przypomnijmy, że w niniejszej pracy zbiór dokumentów wskazanych przez użytkownika nazywamy zbiorem dokumentów relewantnych i oznaczamy D_q.

Na potrzeby wyszukiwaniu informacji w sieci WWW, dziedzina zainteresowań użytkownika może być reprezentowana i identyfikowana przez słowa kluczowe, znajdujące się we wskazanych dokumentach relewantnych. Istotnym elementem utworzenia proponowanej reprezentacji zainteresowań użytkownika, czyli profilu użytkownika, jest proces analizy dokumentów odpowiedzi, w celu określenia tych terminów, które są kluczowe w dokumentach relewantnych i tym samym są *hasłami kluczowymi* w dziedzinie zainteresowań użytkownika. Celem procesu analizy dokumentów odpowiedzi jest identyfikacja słownictwa stosowanego w pewnej dziedzinie wiedzy będącej dziedziną zainteresowań użytkownika oraz wyznaczenie reprezentacji zainteresowań użytkownika na podstawie dokumentów relewantnych. W zaproponowanym modelu pojedyncze zainteresowanie użytkownika reprezentowane jest przez n-wymiarowy wektor wag terminów, czyli subprofil sp_j. Natomiast wszystkie ujawnione przez użytkownika zainteresowania reprezentowane są przez strukturę wiążącą wzorce pytań użytkownika, oznaczane s_j, oraz subprofile, oznaczane sp_j. Strukturę tę nazywamy profilem użytkownika i oznaczamy p. Wzorzec pytania s_j jest identyczny z pewnym zadanym przez użytkownika pytaniem q. Wzorzec pytania jest unikalny w profilu, natomiast pytania identyczne z pytaniem q mogą wielokrotnie pojawiać się w trakcie wyszukiwania. Każdy wzorzec pytania s_j identyfikuje jednoznacznie określony subprofil sp_j, czyli jedną dziedzinę zainteresowań użytkownika. Powiązanie pomiędzy wzorcami i subprofilami jest przechowywane w strukturze profilu. Szczegóły dotyczące związku pomiędzy pytaniem użytkownika, wzorcem pytania i subprofilem są opisane w podrozdziale 4.8, poświęconym metodom tworzenia i modyfikacji zaproponowanego w pracy profilu użytkownika.

4.7.1. Nadanie wag terminom należącym do dokumentów relewantnych

Tradycyjnie w wyszukiwaniu informacji, istotność terminu w dokumencie wyznaczana jest na podstawie wagi tego terminu. Im wyższa jest waga terminu tym jest on bardziej istotny. Każdemu terminowi w dokumencie należącym do systemu wyszukiwania informacji przypisana jest waga d_i według schematu *tf–idf* (wzór (4.4.1)). Waga ta pozwala na wyznaczenie *terminów indeksowych*, dobrze opisujących treść danego dokumentu. Terminy te dobrze opisując treść danego dokumentu umożliwiają, w procesie wyszukiwania, selekcję tego dokumentu spośród innych dokumentów kolekcji, jeśli terminy te przekazane zostaną w pytaniu do systemu.

W pracy przyjęto, że terminy istotne, należące do dziedziny zainteresowań użytkownika, będą wybierane z pomiędzy wszystkich terminów dziedziny przede wszystkim w oparciu o wagi tych terminów w poszczególnych dokumentach relewantnych[1]. Aby ocenić na ile termin t_i dobrze reprezentuje daną dziedzinę zainteresowań użytkownika, musimy ustalić wagę terminu t_i uwzględniającą stopień reprezentowania przez termin t_i treści *wszystkich* dokumentów relewantnych D_q wskazanych przez użytkownika pośród dokumentów wyszukanych. Wagę tę oznaczamy symbolem wz_i'. Waga wz_i' jest jednym z kryteriów wyboru spośród wszystkich terminów t_i należących do dokumentów relewantnych, terminów dobrze reprezentujących dziedzinę zainteresowań użytkownika, czyli *terminów znaczących* tz_i. Tylko terminy znaczące mogą pojawić się w subprofilu użytkownika, odpowiednim dla danej dziedziny zainteresowań, co oznacza modyfikację wag tych terminów w jednym, wybranym subprofilu.

W pracy postawiono hipotezę, że terminy znaczące, uzyskane w zaproponowanym procesie konstruowania i wykorzystania profilu użytkownika, wprowadzone do pytania zmodyfikowanego, zadanego następnie do internetowego systemu wyszukiwawczego prowadzą do dostarczenia większej liczby dokumentów relewantnych w kolejnych wyszukiwaniach prowadzonych przez użytkownika. Hipoteza ta została zweryfikowana eksperymentalnie.

Termin t_i może występować w więcej niż jednym dokumencie relewantnym, dlatego też na wagę wz_i' terminu t_i mają wpływ wagi d_i tego terminu w każdym ze wskazanych dokumentów relewantnych odpowiedzi. Zaproponowano trzy metody wyznaczenia wagi terminu t_i, które uwzględniają powyższe wymagania: waga wz_i' wyrażona jako *minimum, średnia* oraz *maksimum* wag terminu t_i w dokumentach relewantnych.

Waga wz_i' określa istotność terminu znaczącego tz_i w zbiorze wskazanych przez użytkownika dokumentów relewantnych. Dlatego też przyjęto, że waga wz_i' będzie stanowić jedno z kryteriów wyboru terminów znaczących. Szczegóły dotyczące wykorzystania wagi wz_i' jako wspomniane kryterium opisano w podrozdziale 4.7.2.

Rozważmy, jaki będzie wpływ zastosowania, jako kryterium nadawania wagi terminom w dokumentach relewantnych, każdej z wymienionych powyżej metod wyznaczenia wagi wz_i' na zbiór terminów dobrze reprezentujących dziedzinę zainteresowań użytkownika.

1. Waga wz_i' wyrażona jest jako *minimum* wag terminu t_i w dokumentach relewantnych.

[1] Oprócz wag, jest też brany pod uwagę *wskaźnik ważności terminu*, co zostanie dokładnie opisano w dalszej części pracy.

$$wz_i' = \min_{d \in D_q} d_i \qquad (4.7.1.1)$$

d_i – waga terminu t_i w dokumencie $d \in D_q$, wyznaczona na podstawie wzoru (4.4.1).

Jeżeli za wagę wz_i' terminu w zbiorze dokumentów relewantnych przyjmiemy minimum wag d_i tego terminu w dokumentach relewantnych odpowiedzi i ustalony zostanie pewien próg τ, to terminy, których waga jest wyższa od zadanego progu zostaną uznane za opisujące dokument z pewną minimalna dobrocią. Warunek przyjęcia za wagę terminu t_i minimum z wag d_i, pozwala wyeliminować terminy, które 'słabo' opisują chociażby jeden dokument relewantny. Po zastosowaniu dla wag tych terminów progu τ, wybrane zostaną tylko terminy dobrze reprezentujące *wszystkie* dokumenty relewantne. Terminowi reprezentującemu dokumenty przypisana zostaje, w postaci wagi minimalnej, najmniejsza ważność reprezentowania treści dokumentu lub inaczej najmniejszy (minimalny) stopień reprezentacji treści w dowolnym z dokumentów relewantnych. Jeśli tak ustalona waga wz_i' terminu t_i będzie większa od danego progu τ, to termin ten będzie rozważany jako ewentualny termin dobrze reprezentujący dziedzinę zainteresowań użytkownika. Powyższa metoda daje ostre kryterium wyboru terminów będących reprezentatywnymi dla dokumentów relewantnych. Waga wz_i' jest wykorzystywana następnie do selekcji terminów w procesie wyznaczania terminów znaczących.

2. Waga wz_i' wyrażona jest jako *średnia* sumy wag terminu t_i w dokumentach relewantnych:

$$wz_i' = \frac{1}{N_{Rel}} \sum_{d \in D_q} d_i \qquad (4.7.1.2)$$

d_i – waga terminu t_i w dokumencie $d \in D_q$, wyznaczona na podstawie wzoru (4.4.1),

N_{Rel} – liczba dokumentów relewantnych wskazanych przez użytkownika wśród dokumentów wyszukanych.

Załóżmy, że zastosowana zostanie metoda wyznaczenia wagi terminu t_i, w której waga wz_i' liczona jest jako średnia sumy wag terminu t_i w dokumentach relewantnych. Tak obliczona waga terminu będzie wyższa niż obliczona według metody wagi minimalnej, a co za tym idzie, przy tym samym progu τ, więcej terminów zostanie uznanych za dobrze reprezentujące treść dokumentów relewantnych. W efekcie, podczas selekcji terminów dobrze reprezentujących dziedzinę zainteresowań użytkownika spośród powyższych terminów, wyznaczony może zostać szerszy zbiór terminów, w którym pojawią się terminy mniej istotne dla dziedziny zainteresowania użytkownika.

3. Waga wz_i' wyrażona jest jako *maksimum* wag terminu t_i w dokumentach relewantnych.

$$wz_i' = \max_{d \in D_q} d_i \qquad (4.7.1.3)$$

d_i – waga terminu t_i w dokumencie $d \in D_q$, wyznaczona na podstawie wzoru (4.4.1).

Jeśli waga terminu policzona zostanie według trzeciej metody, tj. jako maksimum wag terminu t_i w dokumentach relewantnych, podczas selekcji terminów dobrze reprezentujących dziedzinę zainteresowań użytkownika otrzymamy najszerszy zbiór terminów. Poszerzanie tego zbioru terminów może spowodować pojawianie się w zbiorze terminów mało istotnych w reprezentowaniu treści większości dokumentów, natomiast bardzo specyficznych dla jednego z nich. Często może to być spowodowane obecnością fragmentów dokumentu odmiennych treściowo od wybranego całego zbioru dokumentów relewantnych. Dodatkowo negatywnym efektem powiększania zbioru terminów może być eliminacja innych dokumentów relewantnych z odpowiedzi lub nawet odpowiedź pusta, po dłuższej ewolucji subprofilu, z powodu 'nagłego' (tj. z konkretnego, kolejnego wyszukiwania) przedostania się do subprofilu dużej ilości terminów rzadkich.

Przytoczone powyżej argumenty sugerują, że najlepszą metodą nadania wagi terminowi t_i, należącemu do dokumentów relewantnych jest waga wyznaczana na podstawie wzoru opartego na minimum, tj. (4.7.1.1).

Wzory (4.7.1.1), (4.7.1.2) i (4.7.1.3) inspirowane są badaniami nad grupowaniem kolekcji dokumentów (Voorhees, 1992). Kryterium wyboru wagi minimalnej, przez pewną analogię, przypomina grupowanie *metodą najdalszego sąsiedztwa* (ang. *complete link clustering*), w którym poprzez przyjęcie za podobieństwo dokumentów minimum z podobieństwa par dokumentów, utworzone zostają małe grupy dokumentów, w których dokumenty są ze sobą mocno powiązane. Dla całej kolekcji dokumentów, metoda ta powoduje, że utworzona hierarchia grup dokumentów jest szeroka, rozbudowana bardziej w szerz niż w głąb. Taka struktura hierarchii oznacza, że hierarchia reprezentuje dużo ograniczonych klas znaczeniowych (klas znaczenia terminów).

W metodzie najdalszego sąsiedztwa, w pierwszym kroku określane jest podobieństwo pomiędzy wszystkimi parami klastrów, a następnie łączone są ze sobą dwa klastry o największej wartości podobieństwa. Procedura powtarza się aż do momentu, gdy wszystkie klastry zostaną włączone do hierarchii klastrów. Jako podobieństwo pomiędzy dwoma klastrami przyjmowane jest minimum z podobieństw pomiędzy parami dokumentów, z których każdy dokument należy do innego klastra (Baeza-Yates, Ribeiro-Neto, 1999, str. 135). W wyniku przyjęcia kryterium

minimalnego podobieństwa powstaje hierarchia małych, mocno związanych grup dokumentów.

Grupowanie dokumentów *metodą najbliższego sąsiedztwa* (ang. *single link clustering*) realizowane jest w takich samych krokach jak metoda grupowania metodą najdalszego sąsiedztwa. Inne jest jedynie kryterium obliczania podobieństwa pomiędzy dwoma klastrami. Jako podobieństwo pomiędzy dwoma klastrami przyjmowane jest maksimum z podobieństw pomiędzy parami dokumentów z dwóch klastrów. W wyniku zastosowania tej metody powstaje hierarcha dużych, słabo związanych grup dokumentów.

4.7.2. Selekcja terminów znaczących z dokumentów relewantnych

Wybór *terminów znaczących* tz_i, dobrze opisujących dziedzinę zainteresowań użytkownika, z dokumentów relewantnych jest jednym z głównych zagadnień, które wymagają rozstrzygnięcia w ramach pracy.

Najczęściej opisywaną w literaturze i stosowaną w klasycznym wyszukiwaniu informacji techniką wyboru pewnej podgrupy terminów jest zastosowanie progu dla wagi terminu, przy czym wartość progu jest z góry ustalona i niezmienna. Terminy, których waga przekracza ustalony próg są dołączane do podgrupy. Taka metoda była stosowana do wyznaczania terminów dyskryminujących z dokumentów, należących do tematycznych kolekcji utworzonych ze ściśle wyselekcjonowanych dokumentów (ang. *authoritative documents*). Natomiast kolekcje dokumentów w sieci WWW mają całkiem odmienną specyfikę. Kolekcje te charakteryzuje ogromna różnorodność tematyczna, duża zmienność w czasie, zarówno pod względem ilości terminów, jak i dokumentów. W takich kolekcjach istotność terminu wyrażona przez wagę terminu zmienia się wraz z modyfikacją kolekcji, tj. po dodaniu nowych dokumentów do kolekcji. Zastosowanie dla kolekcji WWW klasycznej metody wyznaczenia terminów dyskryminujących, tj. na podstawie progów wyrażanych przez raz ustalone i stałe wartości liczbowe, nie da oczekiwanego zbioru terminów znaczących. Dlatego też do wyznaczenia zbioru terminów znaczących w pracy zaproponowano progi, które przyjmują postać wielostopniowego kryterium, a ich wartości nie są stałe, ale wyznaczane na podstawie funkcji uwzględniających dynamikę zmian wag terminów w kolekcji.

Zaproponowany w pracy sposób wyboru terminów znaczących jest procesem wielostopniowym, w którym nadawanie wagi wz_i' terminom należącym do dokumentów relewantnych[1] jest jednym z etapów wyznaczenia zbioru terminów

[1] Proces opisano w podrozdziale 4.7.1.

znaczących tz_i z dokumentów relewantnych. W niniejszej pracy nowością jest propozycja zastosowania dwóch kryteriów wyboru terminów znaczących:

1. Pierwszym z nich jest waga wz_i' terminu t_i w dokumentach relewantnych, wyznaczona na podstawie wag tego terminu w każdym z dokumentów relewantnych $d \in D_q$.
2. Drugim kryterium wyboru terminów znaczących jest miara *wskaźnika ważności terminu*, oznaczana cv_i (ang. *cue validity*) (Goldberg, 1996), (Weiss, 1997), (Kazienko, 2000). Termin t_i ma tym wyższą wartość cv_i im jest bardziej charakterystyczny dla grupy dokumentów relewantnych odpowiedzi i im rzadziej pojawia się w pozostałych dokumentach kolekcji. Wartość cv_i jest tym większa im większa jest wartość stosunku częstości występowania terminu t_i w grupie relewantnych dokumentów do częstości występowania tego terminu we wszystkich dokumentach.

Tradycyjnie w literaturze terminom indeksowym przypisywane są wagi według schematu *tf–idf* (Salton i inni, 1975), (Salton, Buckley, 1988), (Rao, 1988), (Rao, 1988a). Schemat *tf–idf* dostarcza informacji o *dyskryminatywności* terminów należących do pewnej grupy dokumentów. Kryterium *cv* niesie informacje na temat *reprezentatywności terminu dla danej grupy*. Termin jest reprezentatywny, jeśli należy do słownictwa charakterystycznego dla określonej grupy dokumentów, a nie należy do słownictwa charakterystycznego dla pozostałych dokumentów, z których wydzielona została grupa pierwsza.

Autor pracy sądzi, że obiecujące jest połączenie dwóch powyższych kryteriów w rodzaj dwustopniowego filtru. W eksperymentach przeprowadzonych w ramach pracy połączono dwa opisane powyżej, a cytowane również w literaturze, kryteria: *tf–idf* oraz *cv*. Skonstruowano w ten sposób kryterium będące *sumą ważoną*. W efekcie połączenia omówionych metod ważenia terminów należących do dokumentów relewantnych, spośród wszystkich terminów należących do dokumentów relewantnych, wybierane są tylko terminy należące do słownictwa stosowanego w dziedzinie zainteresowań użytkownika. Waga terminu kandydata do zbioru terminów znaczących wyznaczana jest na podstawie poniższego wzoru:

$$wz_i = \alpha\, wz_i' + \beta\, cv_i \qquad (4.7.2.1),$$

gdzie α i β są współczynnikami umożliwiającymi określenia wpływu każdego z członów składowych na końcową dla danego wyszukiwania wagę terminu – kandydata do zbioru terminów znaczących.

Optymalne wartości współczynników α i β zostały wyznaczone drogą eksperymentalną, pozwalającą skonstruować efektywny filtr terminów. Proces wyznaczania wartości współczynników oraz ich wartości optymalne zostały opisane w Rozdziale 5.

Po zastosowaniu opisanych powyżej kryteriów: *tf–idf* i *cv* terminy należące do dokumentów relewantnych mają przypisane wagi, dzięki którym można ustalić *ranking* tych terminów dla danego zbioru dokumentów relewantnych. Typując z tak ustalonego rankingu wyróżniającą się *grupę czołową* otrzymaliśmy terminy będące kandydatami do zbioru terminów znaczących. Terminy te stanowią podzbiór terminów, które zostały wyznaczone na podstawie kryteriów *tf–idf* i *cv*.

Wyróżniająca się w rankingu grupa czołowa jest wyznaczona na podstawie dynamicznego progu τ – nazwanego w pracy *współczynnikiem istotności* ι. Współczynnik ten umożliwia wydzielenie terminów, które są kandydatami do zbioru terminów znaczących. W przeprowadzonych w pracy eksperymentach[1], *współczynnikiem istotności* ι jest współczynnik *ŚR*. We współczynniku *ŚR* porównywana jest waga pojedynczego terminu ze średnią wagą wszystkich terminów z analizowanych dokumentów relewantnych. Do wyróżniającej się grupy czołowej należeć będą te terminy, biorąc od terminów najwyżej w rankingu, których waga jest wyższa od średniej wagi wszystkich terminów z dokumentów relewantnych.

W literaturze opisywane są metody wyznaczania terminów dyskryminacyjnych dla klasycznych kolekcji dokumentów. Dowiedziono tam eksperymentalnie, że termin można uznać za dobry dyskryminator na podstawie liczby dokumentów kolekcji, w których występuje analizowany termin (Salton, 1988), (Voorhess, 1992). Liczba ta oznaczana jest zazwyczaj przez *df* (ang. *document frequency*). Eksperymenty przeprowadzane były dla tematycznych kolekcji, utworzonych ze ściśle wyselekcjonowanych dokumentów. Eksperymenty pokazały, że jeśli termin występuje w przedziale 1–10% dokumentów kolekcji to można go uznać za dobry dyskryminator. Jeśli termin występuje w mniej niż 1% dokumentów kolekcji, czyli bardzo rzadko, to uważany jest za słaby dyskryminator. Jeśli natomiast termin występuje w więcej niż 10% dokumentów kolekcji, uważany jest za zły dyskryminator, ponieważ występuje często w różnych dokumentach i nie można na jego podstawie wyróżnić zbioru dobrych dokumentów relewantnych.

W niniejszej pracy jako drugie kryterium wyznaczenia wyróżniającego się zbioru terminów czołowych zastosowano miarę *df*. W ten sposób wyeliminowane zostały terminy, które są mało istotne, spośród terminów, które są istotne w dokumentach relewantnych i są jednocześnie dobrymi dyskryminatorami w całej kolekcji dokumentów. Dolny i górny próg miary *df* pomiędzy df_{min}=1%, a df_{max}=10% liczby dokumentów kolekcji ustalony został przez Saltona i Buckleya dla kolekcji dokumentów, które były kolekcjami zawierającymi dokumenty dotyczące ściśle określonej tematyki. W środowisku sieci WWW, kolekcja dokumentów nie posiada takiej cechy. Dokumenty są związane z różnorodną tematyką, dodatkowo są w różnych

[1] Ogólną koncepcję eksperymentalnej weryfikacji profilu opisano podrozdziale 5.2, natomiast szczegóły przeprowadzonych eksperymentów opisano w podrozdziale 5.4.3.

językach. Dlatego przeniesienie bezpośrednio wartości miary *df* zaproponowanej w literaturze na grunt eksperymentów wykonywanych w ramach niniejszej pracy nie przyniosło oczekiwanych efektów. W części eksperymentalnej niniejszej pracy zweryfikowano dolny i górny próg miary *df*, czyli wartości df_{min} i df_{max}. Optymalne wartości tych progów, przyjęte w eksperymentach przedstawiono w Rozdziale 5.

Poniżej opisany zostanie proces selekcji terminów znaczących. W procesie selekcji wykorzystano najpierw kryterium wyboru terminów na podstawie wagi terminów, a następnie kryterium liczby dokumentów kolekcji *df* oraz współczynnika istotności ι. Proces jest realizowany w następujących krokach:

1. Użytkownik weryfikuje odpowiedź internetowego systemu wyszukiwania informacji przez zaznaczenie w odpowiedzi dokumentów relewantnych.
2. Obliczana jest waga d_i dla każdego terminu należącego do dokumentów relewantnych. Wagi liczone są według schematu *tf–idf* (wzór (4.4.1)), gdzie liczbę dokumentów, w których występuje dany termin określamy na podstawie analizy wszystkich dokumentów kolekcji (tj. bazy danych wyszukiwarki).
3. Każdemu terminowi t_i, który należy do *wszystkich* dokumentów relewantnych, przypisywana jest waga wz_i' równa minimum z wag d_i terminu t_i w dokumentach relewantnych. Termin t_i jest dalej analizowany jako potencjalny termin znaczący.
4. Do wyznaczonego zbioru potencjalnych terminów znaczących zastosowane zostaje kryterium *df*. W dalszej analizie uwzględniane są tylko te terminy, dla których wartość *df* mieści się w przedziale pomiędzy df_{min} a df_{max} .
5. Dla wszystkich wybranych w kroku 4 terminów t_i wyznaczony zostaje wskaźnik ważności cv_i (ang. *cue validity*).
6. Dla terminów t_i wyznaczonych w kroku 4 obliczana jest waga wz_i wyrażona wzorem (4.7.2.1).
7. Dla terminów t_i z kroku 6 zastosowany zostaje próg τ, noszący w pracy nazwę współczynnika istotności ι. Jeżeli waga wz_i terminu jest większa od współczynnika istotności ι – ustalonego dla powstałego rankingu terminów, termin ten jest dobrym ***terminem znaczącym*** $\boldsymbol{tz_i}$. Tym samym, waga wz_i jest podstawą wyboru terminów znaczących.

Przyjęto założenie, że w zbiorze terminów znaczących mogą wystąpić tylko te terminy, które znajdują się we wszystkich dokumentach relewantnych. Konstruując kryteria wyboru terminów znaczących postawiono sobie za cel znalezienie tylko tych terminów, które na pewno opisują dziedzinę zainteresowania użytkownika i umożliwią wyszukanie wszystkich dokumentów relewantnych. Włączenie do pytania zmodyfikowanego terminów znaczących, które reprezentują tylko pewien podzbiór dokumentów relewantnych wskazanych przez użytkownika (czyli włączenie terminów występujących tylko w niektórych spośród dokumentów relewantnych) może spowodować, że w kolejnym wyszukiwaniu nie zostaną wyszukane interesujące dla

użytkownika dokumenty, które opisane są terminami znaczącymi, należącymi do pozostałych dokumentów relewantnych w stosunku do wspomnianego wyżej podzbioru dokumentów relewantnych. Terminy znaczące, które nie należą do wszystkich dokumentów relewantnych nie powinny znaleźć się w pytaniu. Jeśliby takie terminy znalazły się w pytaniu zmodyfikowanym, to pojawia się również problem ze zinterpretowaniem, w jakim stopniu terminy te reprezentują zainteresowania użytkownika jeśli znajdują się tylko w części dokumentów relewantnych.

4.7.3. Terminy znaczące w profilu

Do profilu użytkownika powinny zostać dołączone tylko terminy dobrze opisujące dziedzinę zainteresowania użytkownika. Będą to terminy należące do dokumentów relewantnych odpowiedzi, które są dobrymi dyskryminatorami wyróżniającymi dokument relewantny spośród innych dokumentów kolekcji, a jednocześnie są terminami reprezentatywnymi dla całej grupy dokumentów relewantnych i występującymi rzadko w pozostałych dokumentach odpowiedzi. Warunki te spełniają terminy znaczące tz_i wyznaczone w procesie selekcji na podstawie kryteriów uwzględniających wagę wz_i' oraz wskaźnik ważności cv_i.

W podrozdziałach 4.7.1 i 4.7.2 opisano sposób wyboru terminów znaczących tz_i spośród wszystkich terminów t_i należących do dokumentów relewantnych. Selekcja wykonywana jest na podstawie wagi wz_i' terminów t_i (wzór (4.7.1.1)) oraz wartości wskaźnika ważności cv_i. W wyniku selekcji otrzymujemy zbiór terminów reprezentujących zainteresowanie użytkownika. Wyznaczone terminy znaczące są w profilu użytkownika dołączane do subprofilu sp_j reprezentującego określone zainteresowanie użytkownika, czyli identyfikowanego obsługiwanym właśnie pytaniem użytkownika. Waga wz_i terminu znaczącego tz_i określa istotność terminu znaczącego w zbiorze wskazanych przez użytkownika dokumentów relewantnych w jednym wyszukiwaniu. W profilu natomiast, a precyzyjnie w subprofilu sp_j, waga terminu znaczącego tz_i powinna uwzględniać również istotność tego terminu w reprezentowaniu danego zainteresowania użytkownika ujawnianego podczas kolejnych wyszukiwań. Jeśli dany termin będzie często pojawiał się w zbiorach terminów znaczących dla kolejnych wyszukiwań dotyczących danej dziedziny oznacza to, że jest on istotny dla reprezentowania zainteresowania użytkownika związanego z tą dziedziną. Wynika stąd, że waga terminu znaczącego tz_i w subprofilu, po kolejnym wyszukiwaniu dokumentów i analizie dokumentów odpowiedzi powinna uwzględniać zarówno stopień reprezentacji przez termin tz_i treści dokumentów relewantnych D_q, znalezionych w k-tym wyszukiwaniu (waga $wz_i^{(k)}$), jak i częstość pojawiania się tego terminu w zbiorach terminów znaczących, wyznaczanych po kolejnych wyszukiwaniach. Własność tą

posiada waga $w_{j,i}^{(k)}$ terminu, wyznaczana na podstawie wzoru (4.6.1.1) zainspirowanego pracą (Daniłowicz, 1998).

Waga $w_{j,i}^{(k)}$ obliczana jest na podstawie wagi $wz_i^{(k)}$ terminu znaczącego tz_i i jest normalizowana liczbą wszystkich wykonanych do tej pory selekcji terminów znaczących, tj. $(k-1)$.

W wyniku procesu selekcji terminów znaczących oraz nadania wag tym terminom, otrzymujemy subprofil identyfikowany wzorcem pytania identycznym z obsługiwanym właśnie pytaniem użytkownika (reprezentujący określone zainteresowanie użytkownika), gdzie subprofil jest wektorem terminów tz_i z przypisanymi im wagami $w_{j,i}^{(k)}$.

4.8. Modyfikacja profilu użytkownika

Zaproponowany w niniejszej pracy profil użytkownika jest strukturą opisującą translację pomiędzy terminologią wykorzystywaną przez użytkownika w pytaniu, a słownictwem powszechnie stosowanym w dziedzinie zainteresowań użytkownika. W profilu translacja ta jest wyrażona poprzez przyporządkowanie wzorcowi pytania użytkownika s_j subprofilu sp_j wyznaczonego w procesie analizy dokumentów relewantnych odpowiedzi.

W pracy przyjęte zostały następujące oznaczenia (część przypomnianych):

- q – pytanie użytkownika skierowane do systemu,
- D_q' – zbiór dokumentów zawartych w odpowiedzi na pytanie q użytkownika, gdzie $D_q' \subseteq D$,
- D_q – zbiór dokumentów wskazanych przez użytkownika jako dokumenty relewantne w zbiorze dokumentów odpowiedzi po wyszukiwaniu dokumentów na pytanie q, gdzie $D_q \subseteq D_q'$

Przypomnijmy, że w niniejszej pracy profilem użytkownika wyznaczonym na podstawie odpowiedzi systemu nazywamy reprezentację p_m: pytania q, zbioru dokumentów relewantnych D_q oraz istniejącego profilu p_{m-1}. Po każdym pytaniu użytkownika, wyszukiwaniu dokumentów i weryfikacji odpowiedzi przez użytkownika, profil podlega modyfikacji według poniższej procedury:

$$p_0 = \varnothing,$$
$$p_m = \pi(q_m, D_q, p_{m-1}),$$

gdzie:

- p_0 – profil początkowy; profil ten jest pusty,
- p_m – profil zmodyfikowany po zadaniu m pytań przez użytkownika i analizie zbioru dokumentów relewantnych po m–tym wyszukiwaniu.

Modyfikacja profilu użytkownika następuje po każdym pytaniu q zdanym przez użytkownika do internetowego systemu wyszukiwawczego oraz na podstawie analizy wyników wyszukiwania dla tego każdego kolejnego pytania użytkownika. Istotą pojedynczej modyfikacji profilu jest modyfikacja odpowiedniego subprofilu. Natomiast modyfikacja subprofilu, wg. funkcji 4.6.1.1, ma miejsce tylko w przypadku pojawienia się pytania zgodnego ze wzorcem (zależy od k). Czyli, jeśli zadane przez użytkownika pytanie q jest identyczne z istniejącym w profilu wzorcem pytania s_j, identyfikującym[1] subprofil sp_j, to modyfikowany jest tylko subprofil sp_j. Jeśli nie istnieje wzorzec pytania identyczny z pytaniem, natomiast pytanie użytkownika q jest podobne do jednego lub kilku wzorców pytań z profilu to poza modyfikacją pytania użytkownika do profilu dodawany jest nowy subprofil oraz identyfikujący go nowy wzorzec pytania, identyczny z pytaniem q. Jeśli nie zachodzi żaden z powyższych przypadków, tzn. w profilu nie ma ani wzorca identycznego ani wzorców podobnych, to automatycznie do profilu jest dodawany nowy wzorzec pytania i nowy subprofil utworzony na podstawie odpowiedzi systemu na niezmienione pytanie q. W każdym z trzech powyższych przypadków poprzez modyfikację subprofilu ma miejsce również modyfikacja profilu użytkownika. Szczegółowy opis procedury tworzenia i modyfikacji subprofilu zawiera podrozdział 4.8.1.

Tradycyjna reprezentacja zainteresowań użytkownika w postaci wektora przestrzeni n–wymiarowej (profilu) stwarza również problemy na poziomie wykorzystania profilu do modyfikacji pytania. Pytanie użytkownika w danej chwili dotyczy tylko jednej dziedziny zainteresowań. Tak więc z profilu o strukturze pojedynczego wektora przestrzeni n-wymiarowej, reprezentującego wszystkie zainteresowania należy wybrać tylko terminy, które są związane z aktualnie zadanym pytaniem. Aby uzyskać takie terminy, konieczna jest wiedza na temat powiązań terminów należących do pytania z terminami w profilu oraz terminów z profilu między sobą. Informacje te można uzyskać z utworzonej dla kolekcji dokumentów macierzy podobieństwa (Qiu, 1996) lub sieci semantycznej (Davies i inni, 1997). Główną wadą takiego rozwiązania jest potrzeba przechowywania i zarządzania dwoma strukturami – profilem użytkownika oraz strukturą przechowującą informacje o powiązaniach terminów. Dodatkowo dla internetowych systemów wyszukiwania informacji problemem jest uzyskanie i zarządzanie odpowiednio dużą macierzą podobieństwa, czy siecią semantyczną dla kolekcji jaką jest zbiór dokumentów w sieci WWW.

Przedstawione powyżej problemy nie pojawią się dla zaproponowanego w pracy profilu użytkownika p. W modelu przyjęto następującą koncepcję modyfikacji wag terminów w profilu: w każdym kolejnym wyszukiwaniu modyfikowane są wagi tylko tych terminów, które należą do jednego subprofilu sp_j identyfikowanego przez wzorzec

[1] Identyfikacja subprofilu sp_j przez wzorzec pytania s_j to powiązanie jednego subprofilu z jednym wzorcem pytania.

pytania użytkownika s_j, a nie wagi wszystkich terminów we wszystkich subprofilach. Natomiast w momencie wykorzystania profilu w celu zmodyfikowania pytania istnieje bezpośrednia translacja pomiędzy aktualnym pytaniem użytkownika q a terminami znaczącymi dziedziny, z którą pytanie jest związane. Translacja ta jest reprezentowana przez przypisanie do każdego subprofilu sp w profilu p jednego, unikalnego wzorca pytania s_j identyfikującego ten subprofil oraz identycznego z zadanym pytaniem q. Nowy subprofil oraz identyfikujący go wzorzec pytania dodawane są do profilu użytkownika tylko wtedy, gdy nowe pytanie zadane przez użytkownika jest inne niż jakikolwiek istniejący w profilu wzorzec pytania. Liczba subprofili równa jest liczbie *różnych* pytań zadanych przez użytkownika na przestrzeni czasu korzystania z systemu. Jednak pytania użytkownika kierowane do wyszukiwarki internetowej często powtarzają się ze względu na stałość pewnych zainteresowań użytkownika oraz pojawianie się nowych dokumentów w sieci WWW. Przypadku powtarzającego się pytania nowy subprofil nie jest dodawany.

Istotą zastosowania zaproponowanego profilu użytkownika jest aktywna interakcja podczas wyszukiwań w przeciągu pewnego okresu czasu pomiędzy użytkownikiem a internetowym systemem wyszukiwania informacji, poszerzonym o profil. Oznacza to, że profil zastosowany w systemie wyszukiwania informacji będzie wykorzystywany przez użytkownika przez pewien okres czasu podczas kolejnych wyszukiwań. Istnienie profilu użytkownika w przeciągu pewnego czasu może wiązać się z problemem rozrastania się profilu o kolejne subprofile. Precyzyjniej, rozrastanie to będzie polegało na zwiększaniu się liczby przechowywanych wzorców pytań oraz subprofili, identyfikowanych przez te wzorce. Proponowaną w pracy metodą ograniczenia rozrastania się profilu użytkownika jest weryfikacja ze względu na częstość korzystania z określonego subprofilu. Jeśli subprofil jest często wykorzystywany do modyfikacji pytań użytkownika to oznacza, że reprezentuje aktualne zainteresowania użytkownika. Subprofil taki wraz z identyfikującym go wzorcem pytania będą przechowywane w profilu użytkownika. W przeciwnym przypadku, tzn. jeśli subprofil dawno nie był wykorzystywany do modyfikacji pytania użytkownika, zostanie on usunięty z profilu, jako że reprezentuje on stare, nieaktualne już zainteresowania użytkownika. Granice czasowe aktualności subprofilu mogą być wyznaczone eksperymentalnie.

Ograniczeniem liczby przechowywanych subprofili mogą być tylko ograniczenia techniczne, co może się wiązać z koniecznością usuwania z profilu subprofili. Autor pracy sądzi jednak, że powstawanie nawet bardzo dużej liczby subprofili podczas długiej współpracy użytkownika z systemem wyszukiwania informacji, poszerzonym o profil, nie jest istotnym problemem dla mocy obliczeniowej, czy zasobów dyskowych dzisiejszych komputerów. Profil jest zazwyczaj wielkości kilkudziesięciu kilobajtów, więc biorąc pod uwagę fakt, ze każdy z użytkowników przechowuje profil lokalnie na swoim komputerze, usuwanie dawno nieużywanych subprofili z profilu użytkownika będzie bardzo sporadyczne.

4.8.1. Modyfikacja subprofilu użytkownika

Modyfikacja subprofilu s*p* ma miejsce zawsze, jeśli wyznaczony zostanie termin znaczący tz_i ze zbioru dokumentów relewantnych odpowiedzi na pytanie użytkownika q. Modyfikacji podlegają wagi tych terminów, które zostały wyselekcjonowane po kolejnym wyszukiwaniu dokumentów jako terminy znaczące tz_i. Modyfikacja polega na uaktualnieniu wagi $w_{j,i}^{(k)}$ terminu tz_i w subprofilu identyfikowanym przez wzorzec pytania s_j. Przy czym wzorzec s_j, identyfikujący modyfikowany subprofil, musi być identyczny z zadanym przez użytkownika pytaniem q_j. Wzorzec pytania s_j nie podlega zmianom. Wartość wagi terminu znaczącego uaktualniana jest na podstawie wzoru (4.6.1.1). We wzorze tym $w_{j,i}^{(k)}$ reprezentuje wagę terminu tz_i modyfikowanego w subprofilu identyfikowanym przez wzorzec pytania s_j w k–tej iteracji, tzn. po zadaniu k–ty raz pytania q. W jednym procesie wyszukiwania uaktualniane są wagi terminów znaczących tz_i wyznaczonych w k–tej selekcji, czyli takich, które wyznaczone zostały ze zbioru relewantnych dokumentów odpowiedzi na k–ty raz zadane pytanie q. Inaczej mówiąc, w jednym procesie wyszukiwania modyfikowane są wagi terminów znaczących tz_i tylko w jednym subprofilu sp_j, który jest identyfikowany przez wzorzec pytania s_j identyczny z pytaniem q. Modyfikacja po każdej selekcji terminów znaczących wag *wszystkich terminów* we *wszystkich subprofilach* zaproponowanego profilu spowodowałaby zniekształcenie reprezentacji sensu terminu znaczącego tz_i dla wielu pytań (różnych od obsługiwanego w danym momencie).

Nadmierne powiększanie się subprofilu ograniczone jest przez ustalenie maksymalnej liczby terminów znaczących tz_i, które mogą być wprowadzone do subprofilu. Liczba została wyznaczona eksperymentalnie (np. warunkowana jest maksymalną, sensowną w praktyce długością zmodyfikowanego pytania – pytanie powyżej pewnej długości przestaje zwracać jakiekolwiek rezultaty w wielu wyszukiwarkach).

4.9. Wykorzystanie profilu użytkownika

W profilu użytkownika znajdują się terminy tylko z dokumentów relewantnych wskazanych przez użytkownika w odpowiedzi. Terminy te są dobrymi dyskryminatorami wyróżniającymi wskazany dokument relewantny spośród innych dokumentów kolekcji, a jednocześnie terminy te są reprezentatywne dla całej grupy wskazanych dokumentów relewantnych. Terminy znaczące tz_i reprezentowane są w profilu w postaci struktury n-wymiarowych wektorów wag. Pojedynczy wektor wag terminów znaczących tz_i w profilu nazywamy *subprofilem*. Każdy subprofil identyfikowany jest przez wzorzec pytania s_j. Wzorzec s_j odpowiada pytaniu zadanemu przez użytkownika. W profilu użytkownika wzorce pytań są unikalne. Jeśli użytkownik powtórnie zada takie samo pytanie, a w profilu istnieje już wzorzec tego pytania, to nie

następuje dodanie do profilu kolejnego, takiego samego wzorca pytania, ale modyfikowany jest subprofil, identyfikowany tym wzorcem, na podstawie poprzedniego stanu tego subprofilu i wag terminów znaczących wyselekcjonowanych z dokumentów relewantnych odpowiedzi.

Wykorzystanie profilu p możliwe jest po zadaniu przez użytkownika pytania do systemu wyszukiwawczego. W procesie wykorzystania profilu podstawowym staje się problem wyboru z subprofilu tych terminów znaczących tz_i, które będą terminami dobrymi do zmodyfikowania kolejnego pytania użytkownika, czyli do zastąpienia tego pytania użytkownika pytaniem zmodyfikowanym. Z subprofilu należy wybrać terminy znaczące o najwyższych wagach, ponieważ są to terminy najlepiej reprezentujące dziedzinę zainteresowania użytkownika. Ograniczenie liczby terminów w pytaniu zmodyfikowanym jest szczególnie istotne, gdy subprofil jest wykorzystywany przez użytkownika od dłuższego czasu, co oznaczać może, że jest znacznie rozbudowany – zawiera wiele terminów znaczących. W takiej sytuacji wybierana jest z subprofilu ograniczona liczba terminów.

Jeśli użytkownik zadaje pytanie q_j po raz pierwszy, do profilu dołączany jest wzorzec s_j odpowiadający temu pytaniu oraz subprofil wyznaczony po analizie dokumentów relewantnych odpowiedzi. Jeśli kolejne zadane przez użytkownika pytanie q_k jest takie samo jak wcześniej zadane pytanie q_j, profil użytkownika zostaje wykorzystany do modyfikacji aktualnie zadanego pytania q_k. Tak więc nowością w zaproponowanym w pracy modelu jest wyszukiwanie, które odbywa się zarówno na podstawie pytania zadanego przez użytkownika, jak i zbioru dokumentów relewantnych wskazanych przez użytkownika. Taki proces wyszukiwania opisuje funkcja wyszukiwawcza zdefiniowana w podrozdziale 4.2.

Szczegółowy opis modyfikacji pytania użytkownika z wykorzystaniem profilu zamieszczono poniżej w podrozdziałach: 4.9.1 oraz 4.9.2.

Zmodyfikowane pytanie użytkownika jest kierowane do systemu wyszukiwawczego. W wyniku procesu wyszukiwania, w profilu uaktualniany jest odpowiedni subprofil. Po każdym postawieniu przez użytkownika kolejnego pytania, które jest takie samo jak pytanie q_j, subprofil identyfikowany przez wzorzec pytania s_j lepiej opisuje dziedzinę zainteresowań użytkownika wyrażoną tym pytaniem. W niniejszej pracy postawiono tezę, że kolejne wyszukiwania realizowane z wykorzystaniem systematycznie i automatycznie modyfikowanego subprofilu, identyfikowanego przez wzorzec s_j, prowadzą do zawężenia pytania, zmniejszenia liczby dokumentów odpowiedzi oraz zwiększenia liczby dokumentów relewantnych w odpowiedzi przekazywanych użytkownikowi. Teza została potwierdzona w ramach przeprowadzonych w pracy eksperymentów, opisanych w Rozdziale 5.

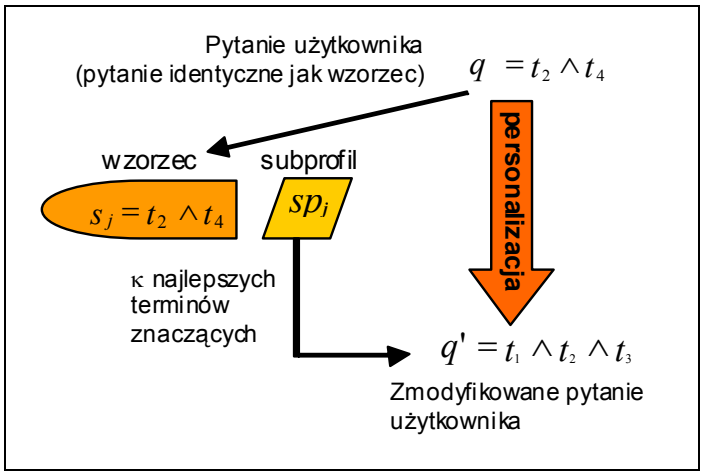
Zauważmy, że modyfikacja pytania na podstawie profilu użytkownika, nie prowadzi do pogorszenia wyników wyszukiwania. Uzasadnieniem jest fakt, że w danym subprofilu modyfikowane są wagi tylko tych terminów znaczących tz_i, które znajdują

się we wszystkich dokumentach relewantnych. Przecięcie zbiorów terminów z dokumentów relewantnych wyznacza zbiór terminów znaczących, których wagi są modyfikowane w subprofilu. W procesie wyszukiwania odpowiedź dla pytania, zmodyfikowanego takim zbiorem terminów znaczących, nie będzie gorsza niż odpowiedź dla pytania początkowego – zadanego przez użytkownika. Oznacza to, że w odpowiedzi na pytanie zmodyfikowane nie pojawią się nowe dokumenty nierelewantne, inne niż te, które znalazły się w odpowiedzi na pytanie początkowe, pod warunkiem, że w trakcie wyszukiwania użytkownik nie zmienił swojego rozumienia dziedziny zainteresowań i dokumenty, które poprzednio wskazał jako relewantne, teraz takimi również są. Odpowiedź na pytanie zmodyfikowane może być jedynie znacznie okrojona w stosunku do odpowiedzi na pytanie początkowe. W najbardziej pesymistycznym przypadku, odpowiedź dla pytania zmodyfikowanego będzie identyczna jak dla pytania początkowego. Własność ta została wykazana eksperymentalnie. Opis przeprowadzonych eksperymentów zawiera Rozdział 5.

4.9.1. Modyfikacja pytań identycznych

W procesie wykorzystania profilu użytkownika, najpierw uruchomiona zostaje procedura znalezienia wśród wzorców pytań s_j w profilu p wzorca identycznego do zadanego pytania q. Jeśli w profilu istnieje taki wzorzec s_j, pytanie q zostaje zmodyfikowane przez zastąpienie terminów pytania q terminami znaczącymi tz_i z subprofilu sp_j identyfikowanego przez wzorzec s_j. Powstaje w ten sposób pytanie zmodyfikowane q' (*Rysunek 4.6*). W pytaniu zmodyfikowanym znajdują się tylko te terminy znaczące tz_i, których waga $w_{j,i}^{(k)}$ w subprofilu sp_j jest większa od progu τ_{profil}. Próg τ_{profil} wyznaczany jest dynamicznie[1]. Wartość progu może być inna dla każdego subprofilu oraz zmienia się po każdej modyfikacji profilu. Pytanie zmodyfikowane zostaje przekazane do internetowego systemu wyszukiwawczego.

[1] Wartość progu τ_{profil} wyznaczana jest na podstawie współczynnika ŚR, którego koncepcję opisano w podrozdziale 4.7.2.



Rysunek 4.6: Modyfikacja pytania użytkownika identycznego jak wzorzec pytania istniejący w profilu.

Użytkownik otrzymuje nową odpowiedź w postaci listy dokumentów i weryfikuje tę odpowiedź. Następnie ma miejsce automatyczna selekcja terminów znaczących z dokumentów relewantnych wskazanych przez użytkownika podczas weryfikacji. Proces obsługi pytania użytkownika kończy modyfikacja wykorzystanego subprofilu. Zmodyfikowany subprofil zostanie wykorzystany ponownie, m.in. wtedy, gdy użytkownik sformułuje pytanie identyczne z istniejącym w profilu wzorcem pytania s_j.

Pytania w postaci koniunkcji terminów identyczne jak wzorzec pytania

Procedura 4.9.1.1

Procedura modyfikacji pytania, dla którego w profilu istnieje wzorzec identyczny, realizowana jest w następujących krokach:

1. Znalezienie wzorca s_j z profilu p, identycznego jak pytanie q.
2. Zmodyfikowanie pytania q przez zastąpienie terminów pytania q koniunkcją terminów znaczących tz_i, których wagi w subprofilu sp_j, identyfikowanym przez wzorzec s_j, są powyżej progu τ_{profil}.
3. Pytanie zmodyfikowane q' jest zadawane do internetowego systemu wyszukiwania informacji.

Stwierdzenie, że użytkownik otrzymuje odpowiedź, która jest mniej liczna niż odpowiedź na pytanie początkowe i zawiera więcej dokumentów związanych z zainteresowaniami użytkownika można uzasadnić w następujący sposób. Pytanie zmodyfikowane zawiera tylko terminy znaczące uzyskane ze wskazanych przez użytkownika dokumentów relewantnych. Wykorzystywana metoda selekcji terminów znaczących zapewnia, że są to terminy precyzyjniej opisujące dziedzinę zainteresowań niż terminy pytania przed modyfikacją. Dlatego też odpowiedź na pytanie zmodyfikowane, czyli pytanie zawierające tylko terminy znaczące, zawiera więcej

dokumentów związanych z zainteresowaniami użytkownika. Powyższe stwierdzenie zostało potwierdzone eksperymentami przeprowadzonymi w ramach pracy.

Przykład 4.9.1

Załóżmy, że zbiór T zawiera następujące terminy: t_1, t_2, t_3, t_4, t_5. Profil jest następujący:

$$p = \{\langle s_1, (0.3, 0.2, 0.0, 0.05, 0.2)\rangle,$$
$$\langle s_2, (0.0, 0.3, 0.4, 0.1, 0.0)\rangle,$$
$$\langle s_3, (0.5, 0.1, 0.0, 0.1, 0.3)\rangle,$$
$$\langle s_4, (0.0, 0.3, 0.4, 0.4, 0.05)\rangle,$$
$$\langle s_5, (0.0, 0.0, 0.2, 0.4, 0.5)\rangle\}$$

gdzie współrzędne wektorów subprofili odpowiadają kolejno terminom: t_1, t_2, t_3, t_4, t_5.

Subprofile identyfikowane są przez wzorce s_1, s_2, s_3, s_4, s_5, gdzie:

$s_1 = t_1 \wedge t_2$,

$s_2 = t_2 \wedge t_3$,

$s_3 = t_1$

$s_4 = t_2 \wedge t_3 \wedge t_4$

$s_5 = t_4 \wedge t_5$

Użytkownik zadał pytanie $q = t_1 \wedge t_2$. Próg τ_{profil} wynosi: $\tau_{profil} = \acute{S}R = 0.1875$.

Modyfikacja pytania $q = t_1 \wedge t_2$ przebiega według opisanej powyżej procedury 4.9.1.1 w następujących krokach:

1. W profilu wzorcem identycznym jak pytanie q jest wzorzec s_1.
2. Analizujemy subprofil sp_1. Terminami znaczącymi, których wagi są wyższe od progu $\tau_{profil} = 0.1875$ są terminy: t_1, t_2 oraz t_5. Terminy t_1, t_2 oraz t_5 zastąpią terminy pytania użytkownika q.
3. Pytanie zmodyfikowane jest w postaci: $q' = t_1 \wedge t_2 \wedge t_5$. Pytanie zmodyfikowane q' jest zadawane do internetowego systemu wyszukiwania informacji. Użytkownik otrzymuje odpowiedź, która jest co najwyżej tak liczna jak odpowiedź na pytanie q, zazwyczaj jednak znacznie mniej liczna, i zawiera więcej dokumentów związanych z jego zainteresowaniami. Oba te stwierdzenia ilustrowane są wynikami eksperymentów zaprezentowanych w Rozdziale 5.

Pytania zawierające negację terminów identyczne jak wzorzec pytania

Procedura 4.9.1.2

Procedura modyfikacji pytania, dla którego istnieje w profilu wzorzec identyczny, a które zawiera terminy zanegowane, realizowana jest w następujących krokach:

1. Znalezienie wzorca s_j z profilu p, identycznego jak pytanie q (porównywane są wyrażenia boolowskie[1]).
2. Zastąpienie terminów pytania q koniunkcją terminów zanegowanych z tego pytania q (jest to de facto pozostawienie terminów zanegowanych bez zmian, przyjęte rozwiązanie zostanie wyjaśnione poniżej) oraz terminów znaczących tz_i, których wagi w subprofilu sp_j, identyfikowanym przez wzorzec s_j, są powyżej progu τ_{profil}.
3. Pytanie zmodyfikowane q' jest zadawane do internetowego systemu wyszukiwania informacji. Użytkownik otrzymuje odpowiedź, która jest mniej liczna niż odpowiedź na pytanie q i zawiera więcej dokumentów związanych z jego zainteresowaniami.

Pozostawienie terminów zanegowanych z pytania użytkownika może budzić wątpliwości. Warto jednak zauważyć, że w ogromnej większości internetowych systemów wyszukiwania informacji przyjęto rozwiązanie (zgodne ponadto z boolowskim modelem wyszukiwania), w którym zanegowanie terminu w pytaniu jest jednoznaczne z '*wykluczeniem'* z odpowiedzi dokumentów, które zawierają ten termin. Ten rodzaj negacji możemy określić jako *negację techniczną*. Pozostawienie terminów zanegowanych w pytaniu zmodyfikowanym powoduje, że dokumenty zawierające te terminy zostaną wyeliminowane z wyników wyszukiwania, terminy zanegowane nigdy nie przedostaną się do subprofilu. Ponadto identyfikacja wzorca pytania z pytaniem zawsze spowoduje pozostawienie niezmienionych terminów zanegowanych. Osiągamy w ten sposób zachowanie systemu, które jest zgodne z oczekiwaniami typowego użytkownika.

Potraktowanie negacji terminu jako negacji technicznej (w sensie określonym powyżej) nie jest jedynym możliwym postępowaniem. Bowiem użycie zaprzeczenia w języku naturalnym np. *szukam informacji o sztucznej inteligencji, ale nie o sieciach neuronowych* może oznaczać, że interesują nas dokumenty dotyczące aspektów sztucznej inteligencji, które jednak nie *koncentrują* się na sieciach neuronowych. Nie znaczy to jednak, że nie chcemy kategorycznie, aby terminy „sieć neuronowa" nie wystąpił w wyszukanym dokumencie. Przecież, może się tam znaleźć na zasadzie odniesienia, porównania, a cały dokument nadal będzie *dotyczył* innej dziedziny sztucznej inteligencji niż sieci neuronowe. Zaprzeczenie w języku naturalnym precyzuje tematykę interesujących nas dokumentów. Ten drugi rodzaj negacji nazwiemy *negacją semantyczną*. W zaproponowanym w pracy profilu możliwe jest modelowanie negacji semantycznej. W tym celu w pytaniu użytkownika negacja i termin pod negacją traktowane są jako jeden symbol – trafiający do wzorca pytania s_j. Sens pytania identycznego ze wzorcem pytania zawierającym terminy zanegowane zostałby następnie opisany poprzez wagi terminów znaczących w subprofilu. Subprofil ten

[1] Porównywanie z dokładnością do kolejności członów koniunkcji.

zostałby użyty do modyfikacji ponownie zadanego pytania zawierającego negację niektórych terminów. W efekcie modyfikacji powinny zostać znalezione dokumenty zgodne z zainteresowaniami użytkownika – nie znaczy to jednak, że dokumenty te nie zawierałyby koniecznie zanegowanych terminów. Spełniałyby za to o wiele istotniejsze kryterium z punktu widzenia użytkownika – dotyczyłyby precyzyjnie pożądanej tematyki.

Negacja semantyczna nie została zaimplementowana w systemie w sposób omówiony powyżej. Z przyczyn omawianych wcześniej, w systemie została zaimplementowana negacja techniczna. Jednak nie ma istotnych przeszkód technicznych, aby w kolejnych wersjach systemów nie wprowadzić negacji semantycznej jako alternatywy dostępnej dla użytkownika.

Stwierdzenie to można uzasadnić tym, że pytanie zmodyfikowane zawiera terminy znaczące uzyskane ze wskazanych przez użytkownika dokumentów relewantnych. Wykorzystywana metoda selekcji terminów znaczących zapewnia, że są to terminy precyzyjniej opisujące dziedzinę zainteresowań niż terminy pytania przed modyfikacją. Jednocześnie terminy, których wykluczenia poprzez zastosowanie ‘negacji’ domaga się użytkownik, nadal pozostają wykluczone. Zakładamy tutaj, że użytkownik jest przyzwyczajony do praktyki internetowych systemów wyszukiwania informacji, w których negacja oznacza wykluczenie z odpowiedzi dokumentów zawierających ten termin.

Podsumowując dyskusję dotyczącą negacji uwzględnianej w systemie oraz komentując konstrukcję algorytmu, w internetowym systemie wyszukiwania informacji intencją użytkownika, który używa w swoim pytaniu termin zanegowany jest wykluczenie tematyki opisywanej przez ten termin z odpowiedzi. Tak więc operator negacji jest używany przez użytkowników systemów internetowych w sensie boolowskim. Wprowadzenie do pytania zmodyfikowanego, oprócz terminów znaczących z subprofilu, terminów zanegowanych z pytania początkowego ma zapewnić to wykluczenie. Terminy znaczące wybrane z odpowiedniego subprofilu dobrze opisują tę tematykę, która jest reprezentowana przez terminy nie zanegowane z pytania początkowego, jednak nie zapewniają, że w odpowiedzi na pytanie zmodyfikowane nie pojawią się dokumenty zawierające terminy, które zostały zanegowane przez użytkownika w pytaniu początkowym. Pojawienie się takich dokumentów, pomimo zanegowania terminu w pytaniu, może być nie do zaakceptowania przez użytkownika. Dzięki operacji dołączenia terminów zanegowanych z pytania początkowego do pytania zmodyfikowanego sytuacja taka nie będzie miała miejsca. Dlatego też odpowiedź na pytanie zmodyfikowane, czyli pytanie zawierające terminy znaczące oraz terminy zanegowane z pytania użytkownika, zawiera zazwyczaj więcej dokumentów związanych z zainteresowaniami użytkownika.

Modyfikację pytania zawierającego negację, rozumianą w sensie negacji technicznej, obrazuje zamieszczony poniżej przykład.

Przykład 4.9.2

Załóżmy, że zbiór T zawiera następujące terminy: t_1, t_2, t_3, t_4, t_5. Profil jest następujący:

$$p = \{\langle s_1, (0.3, 0.2, 0.0, 0.1, 0.2)\rangle,$$
$$\langle s_2, (0.0, 0.3, 0.4, 0.1, 0.0)\rangle,$$
$$\langle s_3, (0.5, 0.1, 0.0, 0.1, 0.3)\rangle,$$
$$\langle s_4, (0.0, 0.2, 0.3, 0.4, 0.2)\rangle,$$
$$\langle s_5, (0.0, 0.0, 0.2, 0.4, 0.5)\rangle\}$$

gdzie współrzędne wektorów subprofili odpowiadają kolejno terminom: t_1, t_2, t_3, t_4, t_5.

Subprofile identyfikowane są przez wzorce s_1, s_2, s_3, s_4, s_5, gdzie:

$s_1 = t_1 \wedge t_2$,

$s_2 = t_2 \wedge t_3$,

$s_3 = t_1 \wedge \neg t_3$,

$s_4 = t_2 \wedge t_3 \wedge t_4$

$s_5 = t_4 \wedge t_5$

Użytkownik zadał pytanie zawierające termin zanegowany: $q = t_1 \wedge \neg t_3$. Próg τ_{profil} wynosi: $\tau_{profil} = \acute{S}R = 0.25$.

Modyfikacja pytania $q = t_1 \wedge \neg t_3$ przebiega według opisanej powyżej procedury 4.9.1.2 w następujących krokach:

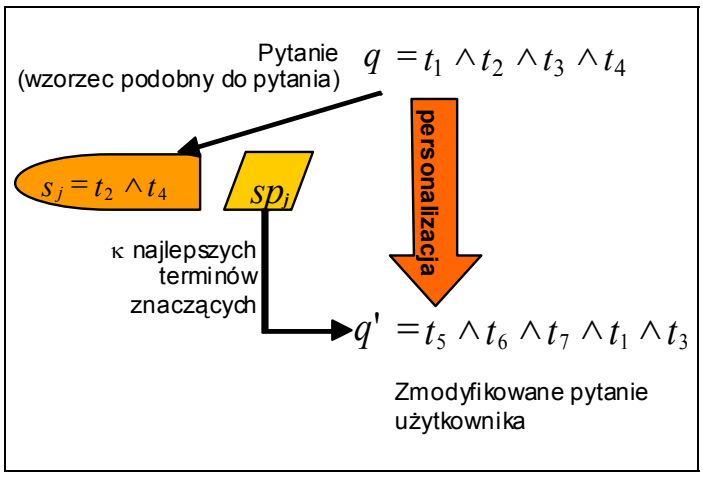
1. W profilu wzorcem identycznym jak pytanie q jest wzorzec s_3.
2. Analizujemy subprofil sp_3. Terminami znaczącymi, których wagi są wyższe od progu $\tau_{profil} = 0.25$ są terminy: t_1 oraz t_5. W pytaniu zmodyfikowanym q' włączony będzie również termin $\neg t_3$ z pytania q. Terminy t_1, t_5 oraz $\neg t_3$ zastąpią terminy pytania użytkownika q.
3. Pytanie zmodyfikowane ma postać: $q' = t_1 \wedge \neg t_3 \wedge t_5$. Pytanie zmodyfikowane q' jest zadawane do internetowego systemu wyszukiwania informacji. Użytkownik otrzymuje odpowiedź, która jest mniej liczna niż odpowiedź na pytanie q i zawiera więcej dokumentów związanych z jego zainteresowaniami.

4.9.2. Modyfikacja pytań podobnych

Pytanie q jest modyfikowane z wykorzystaniem profilu p, jeśli istnieje wzorzec s_j , który jest identyczny jak pytanie q lub, przynajmniej jeden wzorzec, jest podobny do

pytania q^1. Jeśli w profilu istnieją wzorce, które są podobne do pytania q i nie ma wzorca identycznego z zadanym pytaniem, analizowane są terminy znaczące tz_i wszystkich subprofili identyfikowanych przez wzorce, które są podobne do aktualnego pytania q. Terminy znaczące tz_i, należące do subprofili identyfikowanych przez wzorce podobne do pytania q, mogą zostać wykorzystane w procesie modyfikacji pytania q. Pytanie podobne modyfikowane jest na podstawie opisanej poniżej *hipotezy o podobieństwie sensów terminów.*

W pracy postawiono hipotezę podobieństwie podobieństwie sensów terminów. Hipoteza o podobieństwie sensów terminów mówi, że jeśli wzorzec jest podobny do pytania, tzn. przypomnijmy że terminy wzorca są podzbiorem terminów pytania, to terminy ze wzorca pytania zostały wcześniej użyte przez użytkownika w sensie zbliżonym do ich sensu w pytaniu bieżącym. Sens ten jest w przybliżeniu opisany poprzez subprofil – wskazuje on, które terminy z systemu wyszukiwawczego[2] i w jakim stopniu reprezentują sens określonych terminów użytkownika z pytania, a dokładniej określonego wyrażenia składowego[3] pytania. Na podstawie podobieństwa sensów (fragmentu pytania i wzorca pytania), stawiamy hipotezę, że terminy z określonego subprofilu będą dobrymi terminami do zmodyfikowania pytania.



Rysunek 4.7: Modyfikacja pytania użytkownika, gdy w profilu istnieje wzorzec podobny do pytania.

[1] Definicje wzorca identycznego i wzorca podobnego podano w podrozdziale 4.6, w definicjach 4.6.1 oraz 4.6.2.

[2] Warto tu przypomnieć, że użytkownik posługuje się tymi samymi napisami do formułowania swoich pytań jak napisy reprezentujące graficznie terminy z systemu wyszukiwawczego, ale użytkownik może przypisywać odmienne sensy używanym przez siebie napisom niż sensy przypisane do tych samych napisów w kolekcji dokumentów – tu sieci WWW.

[3] Wyrażenia składowego identycznego z wzorcem podobnym do pytania identyfikującym dany subprofil.

Terminy subprofilu oraz wzorca pytania, a także pytania użytkownika, należą do słownika *T*, rozumianego jako zbiór napisów. Jednak użytkownik może przypisywać używanym przez siebie terminom inne sensy (np. na skutek nieścisłej wiedzy lub węższego ich rozumienia) niż są przypisywane tym samym terminom w danej dziedzinie zainteresowań w dokumentach w systemie. Hipoteza pozwala przybliżyć sens, w perspektywie użytkownika, części terminów użytych w pytaniu i zamienić je, w zmodyfikowanym pytaniu, na terminy oddające sens zbliżony (do zamierzonego przez użytkownika) w kontekście systemu wyszukiwania informacji[1].

Pytanie zmodyfikowane (częściowo, w aspekcie tych terminów, których dotyczyła hipoteza), utworzone z wykorzystaniem hipotezy, jest zadawane do internetowego systemu wyszukiwania informacji. Użytkownik otrzymuje dokumenty odpowiedzi, które może ocenić pod względem zgodności z własną potrzebą informacyjną. Ocena ta, będąca weryfikacją odpowiedzi przez użytkownika, jest potwierdzeniem poprawności postawionej hipotezy.

Hipoteza zostaje zastosowana w przypadku nowych, nieznanych dotąd w profilu pytań, dla których jednak próbujemy się ‘domyśleć’, w jakim sensie użytkownik użył danych terminów. Im większy jest profil, tym więcej wiemy o słownictwie używanym przez użytkownika, tzn. jakie sensy przypisuje określonym grupom wyrazów. W profilu każde powiązanie: wzorzec pytania – subprofil, stanowi jakby pozycję w leksykonie znaczeń słownictwa, a dokładniej – w leksykonie znaczeń ‘fraz’ używanych przez użytkownika.

Postępowanie z pytaniami podobnymi do wzorca pytania

Procedura 4.9.2.1

Modyfikacja pytania podobnego realizowana jest w kolejnych krokach:

1. Wyszukanie w profilu wszystkich wzorców s_j podobnych do pytania q.
2. Przepisanie do zmodyfikowanego pytania wszystkich zanegowanych terminów z początkowego pytania zadanego przez użytkownika (postępowanie analogiczne jak w przypadku obsługi pytań identycznych z pewnym wzorcem;

[1] W podobny sposób działa subprofil w przypadku identyczności pytania ze wzorcem identyfikującym ten subprofil: terminy użytkownika zostają zastąpione terminami z subprofilu, wyrażającymi w kontekście systemu ten sam sens co terminy użytkownika wyrażają w kontekście jego rozumienia danej dziedziny zainteresowań. Użytkownik może również używać swoich terminów nieprecyzyjnie: terminów reprezentujących pojęcia ogólne, o szerokim użyciu w języku, może używać w znacznie węższym sensie określającym bardzo konkretne informacje. Takie postępowanie użytkownika może wynikać z przeświadczenia o konieczności bardo ogólnego formułowania pytań do wyszukiwarek internetowych.

podobnie jak tam, zachowujemy wprowadzone przez użytkownika wykluczenie terminów przez negację z dokumentów odpowiedzi).

3. Dołączenie do pytania zmodyfikowanego wszystkich terminów z początkowego pytania użytkownika, które *nie występują w żadnych* wzorcach podobnych do pytania początkowego.
4. Dla każdego terminu *tp* z pytania początkowego, który występuje w chociażby jednym wzorcu podobnym do pytania początkowego, wykonujemy następujące kroki:

4.1. Sumowane są wagi terminów znaczących tz_i, $tz_i \in sp_j$, dla subprofili o wzorcach s_j podobnych do pytania takich, że s_j zawiera termin *tp* (rozważamy tylko te subprofile, które dostarczają opisu sensu danego terminu użytkownika). Wynikiem jest n–wymiarowy wektor $R = (r_1, r_2, ..., r_n)$, będący sumą[1] wektorów wag terminów znaczących z tych subprofili.

4.2. Wybranie tych terminów znaczących z wektora R, które zostaną użyte do zastąpienia rozważanego terminu *tp* z początkowego pytania użytkownika. Jako terminy do modyfikacji pytania są rozpatrywane terminy znaczące tz_i, których wagi w wektorze R są powyżej progu τ_{profil}. Jednocześnie dla zastępowanego w danym momencie terminu *tp* bierzemy pod uwagę tylko kilka terminów znaczących o najwyższej wartości wagi w wektorze R. Uznajemy, że te terminy znaczące, które uzyskują najwyższą wartość wagi najlepiej oddają sens rozpatrywanego terminu *tp*, stanowią rodzaj ‘części wspólnej’ sumowanych subprofili (a to, co łączy sumowane subprofile to fakt występowania w identyfikujących je wzorcach terminu *tp*).

5. Pytanie użytkownika przed modyfikacją jest w postaci koniunkcji: terminów i zanegowanych terminów, dlatego pytanie zmodyfikowane również jest w postaci koniunkcyjnej. Pytanie zmodyfikowane jest zadawane do internetowego systemu wyszukiwawczego.

Użytkownik powinien otrzymać odpowiedź leżącą bliżej jego rzeczywistych zainteresowań, określonych na podstawie analizy sensu użytych przez niego terminów. Warto tu podkreślić, że hipoteza sensu terminów użytkownika (ich sensu dla użytkownika) jest formułowana na podstawie wiedzy o rozumieniu sensu poszczególnych terminów przez użytkownika – wiedzy zapisanej w profilu.

[1] Sumę tę można by było uczynić sumą ważoną w zależności od długości podobnego wzorca s_j, biorą pod uwagę obserwację, iż im dłuższy jest wzorzec, tym samym im większą ‘część’ pytania on pokrywa, tym dokładniejszy jest opis specyficznego użycia rozważanego terminu użytkownika *tp* w pytaniu (mamy zbliżony kontekst użycia terminu *tp* określony pozostałymi terminami ze wzorca, występującymi jednocześnie w pytaniu). Mechanizm takiej sumy ważonej nie jest rozpatrywany w badaniach eksperymentalnych, ale może stanowić ciekawe rozwinięcie technik proponowanych w niniejszej pracy.

Przykład 4.9.2

Załóżmy, że zbiór T zawiera następujące terminy: t_1, t_2, t_3, t_4, t_5. Profil jest następujący:

$$p = \{\langle s_1, (0.3, 0.2, 0.0, 0.1, 0.2)\rangle,$$
$$\langle s_2, (0.0, 0.3, 0.4, 0.1, 0.0)\rangle,$$
$$\langle s_3, (0.5, 0.1, 0.0, 0.1, 0.3)\rangle,$$
$$\langle s_4, (0.0, 0.2, 0.3, 0.4, 0.2)\rangle,$$
$$\langle s_5, (0.0, 0.0, 0.2, 0.4, 0.5)\rangle\}$$

gdzie współrzędne wektorów subprofili odpowiadają kolejno terminom t_1, t_2, t_3, t_4, t_5.

Subprofile identyfikowane są przez wzorce s_1, s_2, s_3, s_4, s_5, gdzie:

$s_1 = t_1 \wedge t_2$,

$s_2 = t_2 \wedge t_3$,

$s_3 = t_1$,

$s_4 = t_2 \wedge t_3 \wedge t_4$,

$s_5 = t_4 \wedge t_5$,

Użytkownik zadał pytanie $q = t_1 \wedge t_2 \wedge t_3$. Próg $\tau_{profil} = \acute{S}R$. Liczba terminów zastępujących każdy termin pytania: $u = 2$.

Modyfikacja pytania podobnego przebiega w następujących krokach opisanych powyżej:

1. Wzorcami podobnymi do pytania q są: s_1, s_2 i s_3.
2. Puste pytanie zmodyfikowane jest postaci: $q' = 1$.
3. W początkowym pytaniu użytkownika brak terminów zanegowanych.
4. Wszystkie terminy z pytania początkowego występują we wzorcach podobnych, więc będzie dla nich przeprowadzony proces zastępowania.
5. W wektorze R sumujemy wagi terminów znaczących zawartych w subprofilach, których wzorce są podobne do pytania q,:
   a) Wykonujemy kroki 4.1 i 4.2 dla terminu t_1 z pytania użytkownika q. Wzorce podobne do pytania q, które zawierają termin t_1 to wzorce s_1 i s_3. Tworzymy wektor R_{t1} sumując wagi terminów znaczących z subprofili identyfikowanych przez wzorce s_1 i s_3 – $R_{t1} = (0.8, 0.3, 0.0, 0.2, 0.5)$. Terminami, których wagi są powyżej przyjętego progu $\tau_{profil} = \acute{S}R = 0.45$ są terminy t_1, t_5, t_2 według malejących wartości wag. Dla terminu t_1 z pytania q zostaną wybrane z subprofilu dwa terminy znaczące o najwyższych wagach, czyli t_1 i t_5, co jest maksymalną, ustaloną liczbą u terminów zastępujących każdy termin pytania q.
   b) Rozpoczyna się analiza dla kolejnego terminu pytania użytkownika q, czyli terminu t_2. Wykonujemy kroki 4.1 i 4.2 dla terminu t_2 z pytania użytkownika q. Wzorce podobne do pytania q, które zawierają termin t_2 to

wzorce s_1 i s_2. Tworzymy wektor R_{t2} sumując wagi terminów znaczących z subprofili identyfikowanych przez wzorce s_1 i s_2 – $R_{t2} = (0.3, 0.5, 0.4, 0.1, 0.2)$. Terminami, których wagi są powyżej przyjętego progu $\tau_{profil} = \acute{S}R = 0.30$ są terminy t_2, t_3, t_1 według malejących wartości wag. Dla terminu t_2 z pytania q zostaną wybrane z subprofilu dwa terminy znaczące o najwyższych wagach, czyli t_2 i t_3, co jest maksymalną, ustaloną liczbą u terminów zastępujących każdy termin pytania q. Proces wyznaczania terminów do zastąpienia terminu t_2 z pytania zadanego przez użytkownika zostaje zakończony, ponieważ maksymalna liczna u terminów znaczących, zastępujących terminy pytania użytkownika została ustalona na wartość 2.

c) Rozpoczyna się analiza dla kolejnego terminu pytania użytkownika, czyli terminu t_3. Wykonujemy kroki 4.1 i 4.2 dla terminu t_3 z pytania użytkownika q. Wzorzec podobny do pytania q, który zawiera termin t_2 to wzorzec s_2. Wektor R_{t3} jest identyczny jak subprofil identyfikowany przez wzorzec s_2 – $R_{t3} = (0.0, 0.3, 0.4, 0.1, 0.0)$. Terminami, których wagi są powyżej przyjętego progu $\tau_{profil} = \acute{S}R = 0.26$ są terminy t_3, t_2 według malejących wartości wag. Dla terminu t_3 z pytania q zostaną wybrane z subprofilu dwa terminy znaczące o najwyższych wagach, czyli t_3 i t_2, co jest maksymalną, ustaloną liczbą u terminów zastępujących każdy termin pytania q.

6. Terminami do zastąpienia aktualnego pytania są: t_1 i t_5 (odpowiednio dla terminu t_1), t_2 i t_3 (dla terminu t_2), t_3 i t_2 (dla terminu t_3). Pytanie zmodyfikowane zawiera terminy: t_1, t_2, t_3, t_5 i jest w postaci: $q = t_1 \wedge t_2 \wedge t_3 \wedge t_5$. Pytanie zmodyfikowane jest zadawane do internetowego systemu wyszukiwania informacji.

W kroku 4.2 zamieszczonego algorytmu wybranie są terminy znaczące z wektora R. Terminy te są następnie użyte do zastąpienia terminu t_i z pytania użytkownika. Jako terminy do modyfikacji pytania są rozpatrywane terminy znaczące tz_i, których wagi w wektorze R są powyżej progu τ_{profil}. Jednocześnie dla zastępowanego w danym momencie terminu t_i użytkownika bierzemy pod uwagę tylko kilka terminów znaczących o najwyższej wartości wagi w wektorze R. Uznajemy, że te terminy znaczące, które uzyskują najwyższą wartość wagi najlepiej oddają sens rozpatrywanego terminu t_i, stanowią rodzaj ‘części wspólnej’ sumowanych subprofili (a to, co łączy sumowane subprofile to fakt występowania w identyfikujących je wzorcach terminu t_i z pytania użytkownika). Ponieważ nie wiemy, które terminy z pytania początkowego są istotniejsze dla użytkownika, a także na skutek różnorodności wzorców podobnych (do pytania początkowego) pod względem długości i zawartości (różne wyrażenia, zbudowane z różnych terminów) trudno jest

porównywać wektory sumy wag uzyskane dla rozpatrywanych poszczególnych terminów użytkownika. Rozsądne wydaje się dążenie do zapewnienia w zmodyfikowanym pytaniu równomiernej reprezentacji każdego zastępowanego terminu użytkownika (przypomnijmy, że zastępowany jest, dlatego że wydaje nam się na podstawie analizy subprofili, że sens przypisany do tego terminu przez użytkownika oddają w wyszukiwarce lepiej terminy znaczące tz_i wyznaczone na podstawie analizy subprofili).

Od strony technicznej, zapewnienie równomiernej reprezentacji każdemu zastępowanemu terminowi zostało osiągnięte przez dążenie do wybrania dla każdego zastępowanego terminu użytkownika identycznej liczby terminów znaczących zastępujących go. Liczba ta może być mniejsza, gdy w wektorze sumy nie będzie dostatecznej liczby terminów znaczących o wagach powyżej wyznaczanego progu τ_{profil}. W szczególnym przypadku termin użytkownika możemy pozostawić bez zmian, jeżeli nie ma terminów znaczących powyżej progu (bardzo mało prawdopodobne). Konkretna liczba zastępujących terminów została dobrana eksperymentalnie i wynosiła 2. Zadaniem modyfikacji pytania w oparciu o hipotezę jest jedynie ukierunkować wstępnie proces wyszukiwania dla nowego, nieznanego dotąd pytania, także zmodyfikowane pytanie nie może być zbyt szczegółowe, aby nie zawęzić nadmiernie odpowiedzi.

Pytanie zmodyfikowane jest następnie zadawane do internetowego systemu wyszukiwawczego. Jeśli pytanie użytkownika q jest podobne do jednego lub kilku wzorców pytań z profilu to poza modyfikacją pytania użytkownika do profilu dodawany jest nowy subprofil oraz identyfikujący go nowy wzorzec pytania, identyczny z pytaniem q. Szczegółowy opis procedury tworzenia i modyfikacji subprofilu zawiera podrozdział 4.8.1, a definicję profilu początkowego podrozdział 4.2.

4.9.3. Pozostałe przypadki relacji pytanie – profil

Jeśli nowe pytanie użytkownika nie jest podobne do żadnego z wzorców s_j identyfikujących subprofile w profilu użytkownika, terminy z pytania nie zostaną zastąpione w aktualnym pytaniu przez żadne terminy znaczące tz_i z profilu użytkownika. Możemy przypuszczać, że użytkownik zainteresował się nową tematyką i na podstawie profilu nic nie wiemy o sposobie formułowania przez użytkownika zapytań w tej nowej tematyce. Rozpoczyna się od początku proces tworzenia subprofilu opisującego tę nową dziedzinę zainteresowań użytkownika, reprezentowaną przez wektor terminów znaczących.

Jeśli wzorzec s_j, identyfikujący subprofil w profilu p, podobny jest tylko częściowo do pytania q, co oznacza, że terminy pytania występują we wzorcu s_j z innymi terminami niż podane w pytaniu przez użytkownika – wzorzec s_j nie stanowi wyrażenia składowego q, to nie ma podstaw do wykorzystania subprofilu odpowiadającego takiemu wzorcowi do modyfikacji pytania. Uzasadnieniem jest fakt, że jeśli terminy z pytania występują we wzorcu w otoczeniu innych terminów niż w pytaniu, to oznacza, że termin we wzorcu jest użyty w całkiem innym sensie niż sens terminu w pytaniu użytkownika.

Podsumujmy przedstawione w tym rozdziale rozważania na temat wykorzystania profilu użytkownika. Modyfikacja pytania polega na zastąpieniu terminów z pytania użytkownika terminami, które zostały znalezione w subprofilu *sp*. Terminy z subprofilu przyjmują postać *frazy*.

Jeżeli pytanie było złożone z terminów i dla tego pytania znaleziono subprofile podobne, pytanie jest modyfikowane według procedury opisanej powyżej. Jeżeli w subprofilu *sp* zostały znalezione terminy znaczące tz_i do zastąpienia pytania użytkownika.

Początkowe pytanie użytkownika może zawierać operator alternatywy. Wtedy pytanie traktowane jest jako dwa osobne pytania. Jeśli profil użytkownika p zawiera odpowiednie subprofile identyfikowane przez wzorce identyczne lub podobne do pytania użytkownika, możliwe jest, aby każde z powstałych pytań zastąpić frazą i utworzyć w ten sposób pytanie zmodyfikowane.

Pytanie zmodyfikowane zadawane jest do internetowego systemu wyszukiwawczego, a odpowiedź systemu jest weryfikowana przez użytkownika. Na podstawie oceny użytkownika modyfikowany jest odpowiedni subprofil. Uruchomiona zostaje procedura wybierania terminów znaczących tz_i, które mają zostać dołączone do subprofilu. Procedura wybierania terminów znaczących opisana została w podrozdziale 4.7.2.

Podsumowując, profil użytkownika p, a precyzyjniej subprofil użytkownika, jest uaktualniany po wyselekcjonowaniu z dokumentów relewantnych *nowego* terminu znaczącego tz_i lub po wyselekcjonowaniu *po raz kolejny* terminu znaczącego znajdującego się już w subprofilu.

5. Eksperymentalna weryfikacja modelu

5.1. Założenia weryfikacji modelu

Zaproponowany w niniejszej pracy profil użytkownika ma służyć personalizacji wyszukiwania informacji w sieci WWW. Przyjęto założenie, że personalizacja wyszukiwania ma miejsce podczas automatycznego formułowania zmodyfikowanego pytania użytkownika oraz prezentowania podpowiedzi dotyczących pytania postawionego przez użytkownika, gdy korzysta on z wyszukiwarki internetowej. Podpowiedź, realizowana jako modyfikacja pytania postawionego przez użytkownika, powstaje na podstawie analizy dotychczasowej pracy użytkownika z wyszukiwarką internetową.

Celem zastosowania zaproponowanego profilu jest dostarczanie użytkownikowi odpowiedzi zawierającej coraz więcej dokumentów relewantnych, w stosunku do wszystkich dokumentów odpowiedzi, w kolejnych cyklach wyszukiwania: pytanie–odpowiedź. Przeprowadzone eksperymenty mają potwierdzić powyższą tezę.

5.2. Koncepcja symulacyjnej weryfikacji profilu

Proponowany profil został pomyślany jako narzędzie wspierające wyszukiwanie w sieci WWW. Naturalnym sposobem weryfikacji założonych własności profilu byłoby przeprowadzenie eksperymentów z udziałem użytkowników. W niewielkim zakresie, nieformalne eksperymenty z udziałem użytkowników zostały przeprowadzone. Niestety eksperymenty takie pochłaniają dużo czasu i do sprawdzenia własności modelu wymagałby zaangażowania znacznej liczby ochotników (dysponujących znaczną ilością wolnego czasu). Rozwiązaniem, które może dostarczyć dużej ilości danych weryfikujących przydatność modelu w stosunkowo krótkim czasie wydała się być metoda symulacyjna (przedstawiona poniżej). Zaproponowana metoda symulacji eksperymentów bazuje na uprzednich decyzjach konkretnych użytkowników w oznaczeniu relewancji dokumentów. Poproszeni o to użytkownicy, określili dla wybranych przez siebie dziedzin zainteresowań zbiory dokumentów relewantnych pochodzących z obszernej kolekcji wykorzystanej w eksperymentach. W trakcie eksperymentu symulowane było jedynie zadawanie kolejnych pytań przez

użytkownika[1], natomiast decyzje o relewantności konkretnych dokumentów względem dziedziny zainteresowań pochodziły bezpośrednio od potencjalnych użytkowników.

Symulacyjna weryfikacja zaproponowanego profilu użytkownika oraz wykorzystania tego profilu miała na celu pokazanie, że dla dowolnego pytania[2], na podstawie wskazanych w odpowiedzi dokumentów relewantnych, kolejne modyfikacje pytania w oparciu o tworzony profil użytkownika, prowadzą do uzyskania przez użytkownika odpowiedzi, w której liczba dokumentów relewantnych opisujących dziedzinę zainteresowań użytkownika jest większa niż w odpowiedzi na pytanie początkowe.

Eksperymenty symulują zachowanie użytkownika podczas wyszukiwania. W rzeczywistym zastosowaniu użytkownik zadaje do wyszukiwarki pytanie. Następnie weryfikuje otrzymaną odpowiedź wskazując dokumenty relewantne. Na podstawie informacji zawartych we wskazanych dokumentach, automatycznie tworzony i modyfikowany jest profil użytkownika. Profil wykorzystywany jest następnie do modyfikacji kolejnych pytań stawianych przez użytkownika do wyszukiwarki.

W eksperymentach przyjęto, że zadawane będą pytania, których terminy są *losowane*. Pytanie to oznaczono jako q_{random} (Rysunek 5-1). W ten sposób symulowane są fakty, że użytkownik może sformułować pytanie dotyczące dowolnej dziedziny tematycznej oraz, że formułując je może być bardzo daleki od zrobienia tego poprawnie[3]. Na potrzeby przeprowadzanych eksperymentów utworzone zostały zbiory testowe, które zawierają dokumenty relewantne. Testowe zbiory dokumentów relewantnych uzyskane zostały od rzeczywistych użytkowników, którzy wykonywali wyszukiwanie dla pewnej dziedziny, dla której wcześniej sformułowali bardzo szczegółowe pytanie, a następnie wskazali dokumenty relewantne w odpowiedzi na to szczegółowe pytanie, tj. weryfikowali odpowiedź systemu. Wyznaczone zbiory dokumentów relewantnych symulowały następne weryfikacje dokumentów relewantnych w odpowiedzi internetowego systemu wyszukiwania informacji[4],.

[1] W seriach nawet po kilkadziesiąt kolejnych pytań, czego trudno byłoby oczekiwać od rzeczywistych użytkowników - ludzi.

[2] Pewne ograniczenia nałożone na dowolność pytania zostaną wyjaśnione w następnych akapitach.

[3] Tzn. używając terminów odpowiednich dla wyszukiwarki i indeksu kolekcji, prowadzących do odnalezienia wszystkich relewantnych dokumentów oraz umieszczenia ich na wysokich pozycjach w rankingu dokumentów odpowiedzi.

[4] Stosownie do dziedziny zainteresowań, której zadawane, losowane, pytanie ma dotyczyć. W konkretnym eksperymencie zakładamy, że losowane pytania dotyczą określonej dziedziny zainteresowań.

Tak więc wykorzystanie testowego zbioru dokumentów relewantnych w odniesieniu do odpowiedzi na pytanie losowe symulowało *weryfikację odpowiedzi* przez użytkownika. Dokumenty ze zbioru testowego identyfikowane były w odpowiedzi na pytanie losowe jako *zbiór dokumentów relewantnych w tej odpowiedzi*, a następnie wyznaczone z tych dokumentów terminy znaczące modyfikowały subprofil użytkownika.

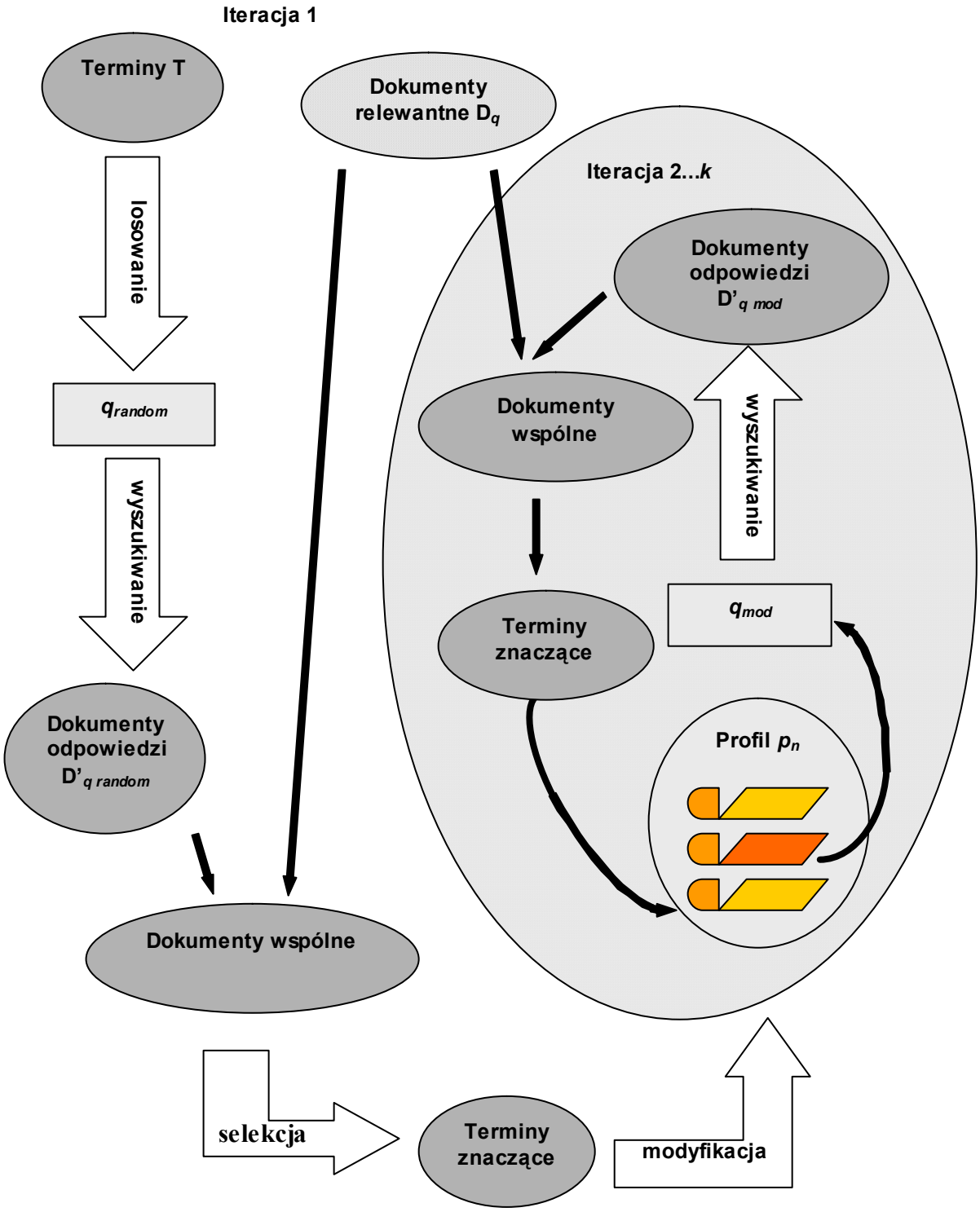
W zastosowanej metodzie weryfikacji, terminy pytania początkowego były losowane ze zbioru terminów T – terminów należących do dokumentów poindeksowanej kolekcji testowej. Eksperymenty przeprowadzono dla pytań, w których terminy wylosowane nie były zanegowane. Następnie losowe pytanie zadawane było do wyszukiwarki. Otrzymywaliśmy listę dokumentów odpowiedzi. Automatycznie sprawdzano, czy w odpowiedzi na pytanie losowe znajdują się dokumenty z aktualnie wykorzystywanego (w danym eksperymencie) zbioru dokumentów testowych.

Ustalone zostały następujące typy zbiorów dokumentów testowych:

- *gęste zbiory* dokumentów relewantnych, zawierające dokumenty bliskie sobie znaczeniowo, wszystkie dobrze opisujące pewną dziedzinę zainteresowań użytkownika wspólną dla całego zbioru – decyzję o bliskości znaczeniowej dwóch dokumentów i ich jednorodności tematycznej podejmował zawsze człowiek, tzn. (najczęściej) użytkownik formułujący zbiór testowy lub (w przypadku innych typów zbiorów) autor pracy,
- *rzadkie zbiory* dokumentów relewantnych dla różnych dziedzin tematycznych, poszczególne dokumenty są słabo znaczeniowo związane ze sobą nawzajem, oraz
- *mieszane zbiory* dokumentów, gdzie mamy gęsty podzbiory dokumentów relewantnych dla pewnej dziedziny zainteresowań (głównej dla całego zbioru) połączony z podzbiorem rzadkim dokumentów relewantnych dla dziedzin innych niż podzbiór gęsty.

Kryterium oceny podobieństwa znaczeniowego w oparciu o decyzję użytkownika jest, z formalnego punktu widzenia, skrajnie subiektywne i niemierzalne. Warto jednak zauważyć, że jest to właśnie takie kryterium oceny, jakiemu podlega ocena relewancji dokumentów podczas wyszukiwania w wyszukiwarce internetowej. Użytkownik ocenia przydatność odszukanego dokumentu na podstawie analizy jego *treści* (*znaczenia tekstu*). Zamodelowanie tego procesu w eksperymencie jest obecnie po prostu niemożliwe.

Jednak gęste zbiory testowe dobrze opisują rzeczywiste decyzje, które by podjął użytkownik – autor danego zbioru – podczas pracy z wyszukiwarką. Utworzenie zbiorów testowych w oparciu o dowolną miarę formalna podobieństwa dokumentów (np. cosinusową) spowodowałoby, że przestałyby one wyrażać rzeczywistą ocenę relewancji dokonywaną przez człowieka. Warto również podkreślić, że poszczególne zbiory testowe zostały utworzone przez różnych użytkowników dla różnych dziedzin. Trudno więc o jakąś z góry przyjętą tendencję w ich konstrukcji.



Rysunek 5-1: Schemat przebiegu eksperymentów dla symulacyjnej weryfikacji profilu.

Pomiędzy dokumentami odpowiedzi identyfikowana była część wspólna ze zbiorem dokumentów testowych, danym dla każdego eksperymentu[1]. Z dokumentów należących do części wspólnej wyznaczane były terminy znaczące tz_i. Terminy znaczące modyfikowały następnie profil użytkownika, a dokładnie odpowiedni subprofil. Na podstawie subprofilu zadane początkowe pytanie losowe było automatycznie zmodyfikowane. Pytanie zmodyfikowane było przekazywane w kolejnej iteracji do wyszukiwarki i zwracany był nowy zbiór dokumentów odpowiedzi. Zbiór ten był ponownie automatycznie sprawdzany pod kątem pokrywania się z danym zbiorem dokumentów testowych. Po analizie dokumentów testowych występujących w odpowiedzi na pytanie następowała kolejna iteracja modyfikacji profilu oraz modyfikacji pytania na podstawie zmodyfikowanego profilu.

[1] Symuluje to decyzję użytkownika, co do relewancji poszczególnych dokumentów odpowiedzi

5.3. Program eksperymentów

Przyjęty został następujący plan eksperymentów:

1. Implementacja profilu użytkownika, procedur modyfikacji i wykorzystania profilu. System podpowiedzi oparty na profilu wspomaga użytkownika w formułowaniu pytania skierowanego do wyszukiwarki internetowej przez proponowanie modyfikacji pytania.
2. Wyznaczenie wartości parametrów dla profilu:
   - współczynników α i β – we wzorze sumy ważonej określającej wagi potencjalnych terminów znaczących z dokumentów relewantnych odpowiedzi[1],
   - współczynnika istotności ι – określającego grupę czołową we zbiorze potencjalnych terminów znaczących[1],
   - progu τ_{profil} – dla selekcji terminów znaczących z subprofilu do modyfikacji.
3. Weryfikacja dolnego i górnego progu *df* (ang. *document frequency*) oraz kolejności zastosowania w procesie wyboru terminów znaczących: współczynnika istotności ι oraz progu *df*.
4. Przeprowadzenie eksperymentów weryfikujących zaproponowany profil użytkownika poprzez:
   - symulację procesu wyszukiwania dokumentów,
   - symulację wskazania dokumentów relewantnych[2] oraz
   - symulację procesu tworzenia, modyfikacji i wykorzystania profilu użytkownika. Podczas kolejnych cykli wyszukiwania (iteracji) mają miejsce kolejne modyfikacje pytania zadanego przez użytkownika na podstawie profilu.
5. Zebranie wyników wyszukiwań i ocena przeprowadzonych eksperymentów symulacyjnych na podstawie zaproponowanej miary efektywności wyszukiwania.

Weryfikację efektów zastosowania zaproponowanego profilu użytkownika przeprowadzono w internetowym systemie wyszukiwania informacji. Opracowany został system podpowiedzi *Profiler*, który umożliwia:

1. Pobranie od użytkownika pytania i przekazanie pytania do wyszukiwarki internetowej, która zwraca zbiór dokumentów odpowiedzi.

[1] Współczynniki α i β oraz współczynnik istotności ι zdefiniowano w podrozdziale 4.7.2.

[2] Zgodnie z tym, co było powiedziane w podrozdziale 5.2, symulowanie wskazania zachodzi w oparciu o zbiory testowe przygotowane uprzednio przez użytkowników.

2. Przeglądnięcie przez użytkownika wszystkich dokumentów odpowiedzi na zadane pytanie, a następnie wskazanie dokumentów relewantnych wśród dokumentów odpowiedzi.
3. Budowanie profilu użytkownika na podstawie wskazanych przez użytkownika dokumentów relewantnych.
4. Modyfikację pytania użytkownika na podstawie istniejącego profilu.
5. Przekazanie pytania zmodyfikowanego do wyszukiwarki internetowej, a następnie przedstawienie użytkownikowi odpowiedzi na pytanie zmodyfikowane.

System *Profiler* akceptuje pytanie w postaci listy słów, która przekazywana jest do wyszukiwarki internetowej *Netoskop*. Pytanie obsługiwane jest przez wyszukiwarkę. Wyszukiwarka zwraca jako odpowiedź na zadane pytanie te dokumenty, które zawierają wszystkie terminy zadane w pytaniu. Pytanie w postaci listy terminów jest traktowane jako pytanie będące koniunkcją terminów. Użytkownik systemu ocenia dokumenty odpowiedzi. Zaznacza te dokumenty, które według własnej oceny uważa za relewantne do posiadanej potrzeby informacyjnej. Wskazane dokumenty poddawane są analizie, w wyniku której wyznaczane są terminy znaczące tz_i. Tylko terminy znaczące wprowadzane są do profilu reprezentującego zainteresowania użytkownika. Profil użytkownika wykorzystywany jest w następujących sytuacjach:

- gdy kolejne pytanie zadane przez użytkownika jest identyczne z pytaniem zadanym we wcześniejszych wyszukiwaniach lub
- gdy pytanie jest podobne do wcześniej zadanego pytania.

W przypadku identyczności pytania, zostaje ono zmodyfikowane na podstawie informacji zawartych w profilu. Pytanie zmodyfikowane jest uszczegółowieniem rzeczywistej potrzeby informacyjnej użytkownika. Natomiast w przypadku podobieństwa, pytanie zmodyfikowane jest formułowane jako hipoteza, postawiona na podstawie istniejącego profilu oraz początkowego pytania użytkownika, dotycząca możliwej rzeczywistej potrzeby informacyjnej użytkownika wyrażonej w postaci pytania początkowego. Każde zmodyfikowane pytanie zostaje przekazane do wyszukiwarki internetowej, a użytkownik otrzymuje odpowiedź w postaci listy dokumentów.

Miary efektywności wyszukiwania

Miary efektywności wyszukiwania wykorzystywane najczęściej w klasycznych systemach wyszukiwania informacji opisano w Rozdziale 2.1.3. W niniejszej pracy jako miarę efektywności wyszukiwania przyjęto dokładność obciętą dla pierwszych 10, 20 i 30 dokumentów odpowiedzi. Dokładność obcięta obliczana jest według następującego wzoru:

$$Dokl_m = \frac{|Rel \cap Wysz_m|}{|Wysz_m|},$$

gdzie m= 10, 20 , 30, a $|Wysz_m|$ oznacza odpowiednio m pierwszych wyszukanych dokumentów odpowiedzi.

Dokładności obcięte $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ obliczano dla każdego zadanego do wyszukiwarki pytania, tj. zarówno dla pytania początkowego, jak i dla wszystkich pytań zmodyfikowanych.

Porównano dokładności obcięte $Dokl^{it}{}_{10}$, $Dokl^{it}{}_{20}$, $Dokl^{it}{}_{30}$ dla każdego kolejnego pytania zmodyfikowanego w każdej kolejnej iteracji cyklu wyszukiwania[1] w stosunku do dokładności obciętych $Dokl^{1}{}_{10}$, $Dokl^{1}{}_{20}$, $Dokl^{1}{}_{30}$ pytania początkowego:

$$POP^{it}_{10} = \frac{Dokl^{it}_{10} - Dokl^{1}_{10}}{Dokl^{1}_{10}} * 100\%,$$

gdzie:

$Dokl^{1}_{10}$ – dokładność obcięta dla pierwszych 10 dokumentów odpowiedzi w 1-szej iteracji modyfikacji pytania,

it – kolejna iteracja modyfikacji pytania i wyszukiwania dokumentów,

$Dokl^{it}_{10}$ – dokładność obcięta dla pierwszych 10 dokumentów odpowiedzi w kolejnej iteracji it modyfikacji pytania.

Analogicznie mamy:

$$POP^{it}_{20} = \frac{Dokl^{it}_{20} - Dokl^{1}_{20}}{Dokl^{1}_{20}} * 100\%, \text{ oraz } POP^{it}_{30} = \frac{Dokl^{it}_{30} - Dokl^{1}_{30}}{Dokl^{1}_{30}} * 100\%.$$

Miary poprawy efektywności wyszukiwania $POP^{it}{}_{i}$ określają na ile wyniki wyszukiwania dla każdego zmodyfikowanego pytania w kolejnej iteracji modyfikacji pytania są lepsze niż wyniki wyszukiwania dla pytania początkowego. Poprawę efektywności wyszukiwania obliczano w trzech kategoriach zdefiniowanych powyżej, ponieważ za miarę dokładności przyjęto dokładności obcięte. Na podstawie miary *POP* określono procent pytań zmodyfikowanych, które polepszają efektywność wyszukiwania w przeprowadzonych eksperymentach.

Na potrzeby, przeprowadzonej w ramach eksperymentów, symulacyjnej weryfikacji profilu porównano również liczbę dokumentów relewantnych wyszukanych $|Rel \cap Wysz|$ w odpowiedzi na kolejne zmodyfikowane pytanie do liczby wszystkich dokumentów relewantnych w ustalonych zbiorach testowych $|Rel_{test}|$:

[1] Cykl wyszukiwania rozumiany jest tutaj jako kolejno wykonane następujące czynności: zadanie pytania przez użytkownika, zmodyfikowanie pytania (jeżeli spełnione są odpowiednie warunki), uzyskanie odpowiedzi od systemu internetowego (stąd indeks *it*).

$$\%DR = \frac{|Rel \cap Wysz|}{|Rel_{test}|} * 100\%$$

Miarę *%DR* cechuje pewne podobieństwo do klasycznej miary kompletności *Kom.* Jednak podstawową różnicą jest fakt, że dla kompletności znana jest liczba wszystkich dokumentów relewantnych w kolekcji, natomiast miara *%DR* uwzględnia tylko pewien podzbiór dokumentów relewantnych, będący podzbiorem wszystkich dokumentów relewantnych kolekcji. Miara *%DR* określa skuteczność zaproponowanego profilu oraz metod wykorzystania profilu w dostarczaniu użytkownikowi nowych dokumentów relewantnych w wykonywanych kolejnych wyszukiwaniach. Aby pokazać tę własność wystarczy pokazać, że ma to miejsce dla każdego testowego, dowolnie wybranego zbioru dokumentów relewantnych. Nie jest celem natomiast pokazanie, że zastosowanie zaproponowanego w pracy profilu użytkownika prowadzi do wyszukania wszystkich dokumentów relewantnych w kolekcji sieci WWW. Ocena kompletności wyszukiwania w sieci WWW, z racji na dużą częstość modyfikacji kolekcji WWW oraz jej liczność, jest jedynie szacowana (Choroś, 2002).

5.4. Opis przeprowadzonych eksperymentów

Eksperymenty polegały na wykonaniu wyszukiwań w sieci WWW dla *pytań początkowych*, tj. pytań wylosowanych oraz pytań zadanych przez użytkownika, a następnie dla *pytań zmodyfikowanych* na podstawie tworzonego profilu użytkownika. Przeanalizowano wyniki wyszukiwania – zbiory dokumentów wyszukanych odpowiedzi. Przeprowadzone eksperymenty były podstawą opracowania wniosków dotyczących zaproponowanej koncepcji tworzenia i wykorzystania profilu użytkownika.

5.4.1. System podpowiedzi *Profiler*

Do przeprowadzenia eksperymentów niezbędne było przeniesienie działania systemu podpowiedzi *Profiler* na stronę serwera oraz skorzystanie z indeksów wyszukiwarki internetowej. Uzyskano możliwość dostępu do indeksów wyszukiwarki (zbiór opisywany w modelu jako słownik *T*) dzięki nawiązaniu współpracy z firmą Poland.com, która była właścicielem komercyjnej wyszukiwarki *Netoskop*.

Przygotowanie i testowanie środowiska niezbędnego do przeprowadzania eksperymentów wymagało czasu i wiązałoby się z koniecznością przerywania normalnego funkcjonowania wyszukiwarki komercyjnej. Dlatego też na potrzeby eksperymentów została uruchomiona kopia komercyjnej wyszukiwarki internetowej. Rozwiązanie takie było koniecznością, aby właściciele wyszukiwarki nie ponosili strat ze względu na przerywaną i spowolnioną pracę wyszukiwarki. Przeniesienie

środowiska wyszukiwarki wiązało się z kilkoma problemami. Sprzęt komputerowy, który autorowi pracy udostępniła uczelnia był skromniejszy niż sprzęt, na którym oryginalnie była uruchomiona wyszukiwarka. Skutkiem tego jest sytuacja, że wykorzystywana w eksperymentach kolekcja poindeksowanych dokumentów wyszukiwarki testowej jest podzbiorem zasobów wyszukiwarki komercyjnej. Podobnie indeks wykorzystywanej do testów wyszukiwarki jest podzbiorem indeksu wyszukiwarki komercyjnej. Na potrzeby eksperymentów przeprowadzonych w pracy wylosowano zbiór pytań testowych, dla których w ograniczonym, z konieczności, zbiorze poindeksowanych dokumentów będzie wyszukany niepusty zbiór dokumentów odpowiedzi. Terminy z dokumentów odpowiedzi dla pytań testowych występowały również w innych dokumentach kolekcji, co dawało szansę uzyskania odpowiedzi dla dowolnej modyfikacji pytania początkowego. Tym samym, można było badać proces modyfikacji pytania bez wychodzenia poza zakres poindeksowanej kolekcji, gdzie nie byłoby możliwe jego dalsze śledzenie.

5.4.2. Kolekcja testowa

Eksperymenty wykonano dla kolekcji testowej liczącej 41355 dokumentów (ponad 365 000 terminów) uzyskanych z sieci WWW na przestrzeni 3 miesięcy. Dla kolekcji testowej wykonano do 50 kolejnych modyfikacji pojedynczego pytania testowego, co jest wystarczające na potrzeby przeprowadzanej weryfikacji. Początkowo wydawało się, że na potrzeby wstępnej weryfikacji procedur tworzenia i wykorzystania profilu użytkownika odpowiednia będzie kolekcja zawierająca kilka tysięcy dokumentów. Opisywane w literaturze eksperymenty przeprowadzane były dla kolekcji liczących właśnie ok. 4 – 5 tysięcy dokumentów. Takie kolekcje początkowo poindeksowano w wyszukiwarce dla eksperymentów przeprowadzonych w niniejszej pracy, jednak eksperymenty wstępne wykluczyły możliwość zastosowania kolekcji o tak małej liczności. Istotną różnicą jest fakt, że kolekcje wykorzystywane do większości eksperymentów opisywanych w literaturze były dobierane treściowo i dotyczyły jednej dziedziny tematycznej. W kolekcji tego rodzaju, już przy kilku tysiącach dokumentów, rozkład częstości występowania terminów pozwala na stosowanie schematu *tf–idf* i wyznaczenie na jego podstawie terminów indeksowych opisujących treść dokumentów. Terminom z dokumentów kolekcji przypisane zostały wagi według schematu *tf–idf*, który jest uznawany i stosowany w systemach wyszukiwania informacji. Okazało się jednak, że dla pozyskanej z sieci WWW kolekcji o liczności kilku tysięcy dokumentów, występują znaczne odchylenia w rozkładzie częstości występowania terminów, a tym samym w wartościach wag, jakie są przypisywane terminom. Kolekcja tworzona na podstawie dokumentów sieci WWW zawiera dokumenty w pewien sposób losowo pozyskiwane z sieci WWW. Generowanie kolekcji dokumentów sieci WWW rozpoczyna się od podania punktu początkowego (adresu

URL dokumentu startowego). Następnie do kolekcji automatycznie dołączane są wszystkie dokumenty, na które wskazują odsyłacze umieszczone w dokumencie podanym jako punkt początkowy. Automatyczne dołączanie dokumentów do kolekcji na podstawie istniejących pomiędzy nimi odsyłaczy nie daje możliwości ograniczenia dziedziny tematycznej generowanej kolekcji. Dlatego też, jeśli kolekcja dokumentów z sieci WWW jest zbyt mała, częstość występowania terminów w kolekcji nie odzwierciedla rzeczywistej istotności terminów w danej dziedzinie tematycznej. W kolekcji liczącej kilka tysięcy dokumentów niektóre dziedziny mogą nie być w ogóle reprezentowane przez żadne dokumenty lub tylko przez niewielką liczbę dokumentów. Aby uniknąć opisanych problemów z sieci WWW zebrano i poindeksowano kolekcję liczącą 41355 dokumentów (ponad 365 000 terminów), na której przeprowadzano eksperymenty.

5.4.3. Symulacyjna weryfikacja profilu

Eksperymenty wstępne

W początkowej fazie eksperymentów wyznaczono wartości parametrów[1]: współczynnika istotności ι, progów df_{min}, df_{max} oraz progu τ_{profil}. Uwzględniając wagi terminów w dokumentach, parametr ι oraz progi df_{min}, df_{max} decydują o liczbie terminów znaczących, które zostaną wprowadzone do profilu użytkownika po analizie dokumentów testowych. Parametr τ_{profil} decyduje natomiast o liczbie terminów znaczących z profilu, jakie zostaną włączone do zmodyfikowanego pytania. Wartości parametrów wyznaczono eksperymentalnie, na podstawie wyszukiwań przeprowadzonych dla różnych kombinacji wartości tych parametrów. Eksperymenty pokazały, że największa efektywność wyszukiwania została osiągnięta dla następujących wartości parametrów: $\iota = \acute{S}R^2$, $\tau_{profil} = \acute{S}R$. Przypomnijmy, że we współczynniku $\acute{S}R$ porównywana jest waga pojedynczego terminu ze średnią wagą wszystkich terminów z analizowanych dokumentów relewantnych (w przypadku parametru ι) lub średnią wagą wszystkich terminów z analizowanego subprofilu (w przypadku parametru τ_{profil}). Do wyróżniającej się grupy czołowej należeć będą te terminy, biorąc w kolejności terminy najwyżej w rankingu, których waga jest wyższa od średniej wagi wszystkich terminów z dokumentów relewantnych lub z danego subprofilu.

[1] Dokładny opis znaczenia parametrów ι, df_{min}, df_{max} znajduje się w podrozdziale 4.7.2.

[2] Współczynnik $\acute{S}R$ omówiono w podrozdziale 4.7.2.

Współczynniki α i β decydują o wpływie wag składowych na końcową wagę terminu kandydata do zbioru terminów znaczących. Wagami składowymi[1] są: waga wyznaczona na podstawie schematu *tf–idf* oraz waga wyznaczona na podstawie wskaźnika ważności terminu *cv*. Na podstawie przeprowadzonych eksperymentach wstępnych ustalono, że wartości współczynników: $\alpha = 0.01$ i $\beta = 10.0$, zapewniają selekcję dobrych terminów znaczących. Dolny i górny próg *df* ustalono eksperymentalnie odpowiednio na $df_{min} = 2.5\%$, $df_{max} = 10\%$ liczby dokumentów kolekcji.

Po wyznaczeniu wartości parametrów przeprowadzono wyszukiwania dla ustalonych pytań testowych. Pytania testowe były pytaniami wylosowanymi. Dla każdego pytania testowego wykonywanych było 50 kolejnych iteracji modyfikacji pytania na podstawie tworzonego i cyklicznie modyfikowanego profilu. Podczas każdej iteracji, profil użytkownika był modyfikowany wyznaczonymi z dokumentów testowych terminami znaczącymi tz_i.

W wyniku wyszukiwań oraz analizy wyszukanych dokumentów relewantnych powstał profil zawierający subprofile. Wzorce pytań związane z subprofilami odpowiadały wylosowanym pytaniom. Każdy wzorzec pytania identyfikował subprofil zawierający wyznaczone terminy znaczące.

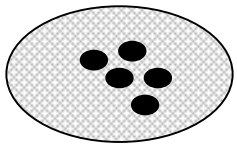
Dla każdej iteracji modyfikacji pytania losowego badano wartości: dokładności obciętej $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$, poprawy efektywności wyszukiwania *POP* oraz liczby wyszukanych dokumentów relewantnych *%DR*. Wyniki przeanalizowano i przedstawiono w postaci wykresów. Przykłady przeprowadzonych eksperymentów zamieszczono w Załączniku A.

Analiza wyników symulacji dla gęstych zbiorów dokumentów relewantnych[2]

Przeprowadzone wyszukiwania dla gęstych zbiorów dokumentów relewantnych (zdefiniowanych w 5.2, przypomnijmy, zbiorów dokumentów blisko związanych ze sobą znaczeniowo, dotyczących jednej dziedziny zainteresowania użytkownika), mają potwierdzić, że proces modyfikacji pytania na podstawie tworzonego profilu użytkownika w kolejnych wyszukiwaniach prowadzi do ograniczenia liczby dokumentów odpowiedzi oraz do zwiększenia udziału dokumentów relewantnych w zbiorach dokumentów odpowiedzi. Jest to istota zastosowania profilu użytkownika. Takiej własności profilu oczekuje również użytkownik korzystający z internetowego systemu wyszukiwania informacji z profilem.

[1] Przyjęty w pracy schemat obliczania wagi terminu kandydata do zbioru terminów znaczących omówiono w podrozdziale 4.7.2.

[2] Przykładowe z przeprowadzonych eksperymentów umieszczono w Załączniku A.



Rysunek 5-2: Schemat testowego zbioru dokumentów gęstych.

W eksperymentach przeprowadzonych dla testowych zbiorów dokumentów relewantnych wykonano wyszukiwania dla 50 wylosowanych pytań testowych. Przed losowaniem pytania przyjmowano z góry, do której dziedziny zainteresowań ma ono się odnosić. Dziedziny zainteresowań były jednoznacznie powiązane z odpowiednimi zbiorami dokumentów testowych. Dla każdego pytania testowego wykonano 50 kolejnych iteracji modyfikacji pytania. Modyfikacje wykonywane były na podstawie tworzonego profilu.

Dla wszystkich 50 pytań losowych porównano liczbę dokumentów dostarczanych w odpowiedzi na kolejne zmodyfikowane pytania. Wyniki przedstawia w postaci wykresu Rysunek 5-11. Dla wszystkich pytań losowych obliczano również miarę *%DR*, opisującą procent dokumentów relewantnych (wg danego zbioru dokumentów testowych), które były znajdowane w kolejnych zmodyfikowanych pytaniach. Wyniki przedstawia w postaci wykresu Rysunek 5-12. Z 50-ciu wykonanych iteracji, na wykresach przedstawiono pierwszych 20-cia iteracji, ponieważ powyżej 20-tej iteracji następowała stabilizacja pytania oraz stabilizacja zbioru relewantnych dokumentów wyszukanych.

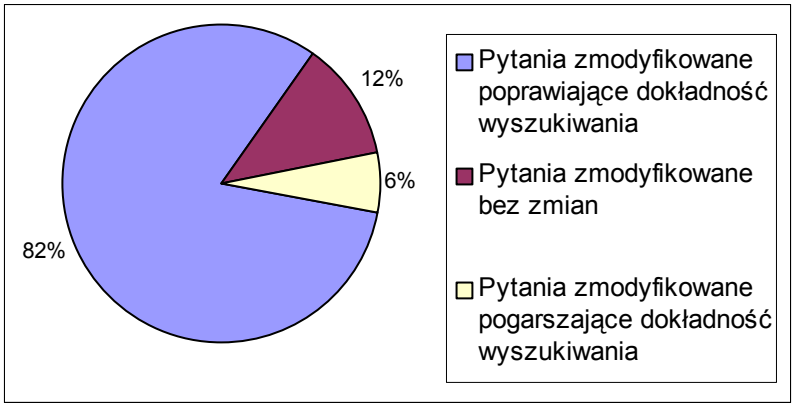
Dla każdego pytania wylosowanego oraz pytania zmodyfikowanego obliczono dokładności obcięte $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz wartość poprawy wyszukiwania *POP* uzyskaną dla każdego pytania zmodyfikowanego w stosunku do początkowego pytania wylosowanego. Przykładowe z przeprowadzonych eksperymentów, pokazujące zależności pomiędzy liczbą dokumentów odpowiedzi D'_q, dokładnościami obciętym $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz liczbą dokumentów relewantnych w odpowiedzi D_q zamieszczono w Załączniku A niniejszej pracy.

Wyniki przeprowadzonych wyszukiwań podzielono na trzy grupy:

1. wyszukiwania, w których dokładność rosła dla odpowiedzi na kolejne pytania zmodyfikowane,
2. wyszukiwania, w których dokładność nie zmieniała się dla odpowiedzi na kolejne pytania zmodyfikowane,
3. wyszukiwania, w których dokładność malała dla odpowiedzi na kolejne pytania zmodyfikowane.

Należy szerzej wyjaśnić przyczyny braku poprawy dokładności wyszukiwania w dwóch ostatnich grupach. Wyszukiwania, w których zaobserwowano pogorszenie wyników po kolejnych modyfikacjach pytania charakteryzowały się wzrostem ogólnej liczby dokumentów odpowiedzi. Były to wyszukiwania, w których pytanie

zmodyfikowane niekoniecznie było pytaniem bardziej ogólnym niż pytanie przed modyfikacją, a pomimo to, w odpowiedzi na pytanie zmodyfikowane internetowy system wyszukiwania informacji zwracał odpowiedź zawierającą więcej dokumentów. Oznacza to, że dziedzina, której dotyczyło pytanie zmodyfikowane jest reprezentowana w kolekcji przez więcej dokumentów, niż zostało to uwzględnione we zbiorze dokumentów testowych[1]. Przyczyną pogorszenia dokładności wyszukiwania jest właśnie większa odpowiedź, jak również przesunięcie dokumentów relewantnych na niższą niż 30-ta (ostatnia badana) pozycja w odpowiedzi. W zaproponowanej metodzie personalizacji wyszukiwania nie zajmowano się problematyką związaną z rankingiem dokumentów odpowiedzi. Autor pracy jest przekonany, że zastosowanie odpowiednich metod rankingu, uwzględniających w ustaleniu pozycji dokumentu jego podobieństwo do odpowiedniego subprofilu, polepszyłoby wyniki dokładności wyszukiwania i zwiększyło satysfakcję użytkownika (w eksperymentach automatycznie stosowany był ranking komercyjnej wyszukiwarki uwzględniający częstość występowania terminów i miejsce występowania terminów pytania w poszczególnych dokumentach – koncepcja oceny dokumentów w rankingu stosowanym w wyszukiwarce Netoskop była zupełnie inna niż koncepcja zaproponowanej oceny terminów profilu).



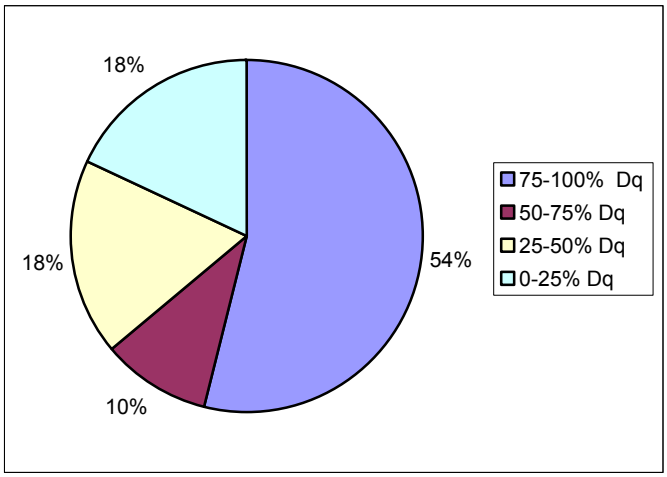
Rysunek 5-3: Zestawienie wszystkich pytań zmodyfikowanych obrazujące poprawę wyników wyszukiwania z wykorzystaniem profilu dla testowego zbioru dokumentów gęstych.

Przeprowadzone wyszukiwania potwierdziły zakładane polepszanie efektów wyszukiwania dla większości pytań zmodyfikowanych w stosunku do losowych pytań początkowych. Zestawienie wyników zawiera Rysunek 5-3. Dla 82% pytań początkowych, w kolejnych iteracjach wzrastały wartości dokładności obciętej $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ dla odpowiedzi na kolejne zmodyfikowane pytania oraz liczba

[1] Zbiorze utworzonym przez pewnego użytkownika na podstawie zadanego przez niego pytania i selekcji dokumentów z pomiędzy dokumentów odpowiedzi. Żaden użytkownik przygotowujący zbiór testowy nie przejrzał całej kolekcji. Przy liczbie dokumentów przekraczającej 40000 trudno było oczekiwać od użytkowników aż tak ogromnego zaangażowania w eksperyment.

dokumentów relewantnych *%DR* w odpowiedzi. Oznacza to wzrost liczby znalezionych dokumentów relewantnych w odpowiedzi na kolejne zmodyfikowane pytanie. Jednocześnie dla tych pytań zmniejszała się liczba wszystkich dokumentów znajdowanych w odpowiedzi na kolejne zmodyfikowane pytanie.

Zestawiono również przeprowadzone wyszukiwania pod kątem procentu znalezionych dokumentów relewantnych z gęstego zbioru dokumentów testowych. Wyniki zawiera Rysunek 5-4. Dla 54% pytań początkowych, w wyniku kolejnych modyfikacji pytania na podstawie profilu, w odpowiedzi na pytanie zmodyfikowane znajdowanych było 100% dokumentów relewantnych z przygotowanych zbiorów testowych (warto podkreślić, że działo się to dla *wylosowanego* pytania początkowego, czyli dla przypadku odpowiadającego skrajnej niewiedzy użytkownika formułującego pytanie początkowe).



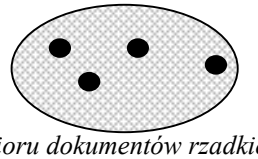
Rysunek 5-4: Liczba pytań testowych dla gęstego zbioru dokumentów testowych w podziale na procent znalezionych dokumentów relewantnych D_q w odpowiedziach na kolejne pytania zmodyfikowane (miara %DR).

Eksperymenty iteracyjnej modyfikacji pytania, przeprowadzone dla ustalonych gęstych zbiorów dokumentów relewantnych potwierdzają również zgodność zaproponowanego modelu z intuicją wykorzystania profilu użytkownika w systemie wyszukiwania informacji. Korzystając z systemu wyszukiwania informacji w sieci WWW, umożliwiającego personalizację wyszukiwania dzięki zastosowaniu profilu użytkownika, użytkownik oczekuje, że po pewnym czasie współpracy z systemem będzie otrzymywał w odpowiedzi na pytania z określonej dziedziny zainteresowań coraz więcej dokumentów na interesujący go temat, a odpowiedź ogólnie będzie coraz mniej liczna. Taką funkcjonalność zapewnia zaproponowany profil użytkownika, co potwierdziły przeprowadzone eksperymenty.

Analiza wyników symulacji dla rzadkiego zbiorów dokumentów testowych[1]

Celem drugiego etapu przeprowadzanych eksperymentów było potwierdzenie tezy, że jeśli użytkownik wskaże zbiór dokumentów relewantnych zawierających dokumenty, z których każdy (lub po kilka) reprezentuje różne dziedziny zainteresowania użytkownika to zaproponowany profil użytkownika doprowadzi do modyfikacji pytania, w odpowiedzi, na które znajdą się dokumenty reprezentujące jedną z tych dziedzin zainteresowania, a nie wszystkie wskazane dziedziny.

W tej części eksperymentów badano wyniki wyszukiwania dla pytań losowych oraz dla ustalonych zbiorów dokumentów testowych, zawierających dokumenty relewantne, opisujące różne dziedziny zainteresowania użytkownika, tj. dla rzadkiego zbioru dokumentów testowych. Na schemacie zbiór dokumentów rzadkich przedstawiamy jako dokumenty znacznie odległe od siebie.



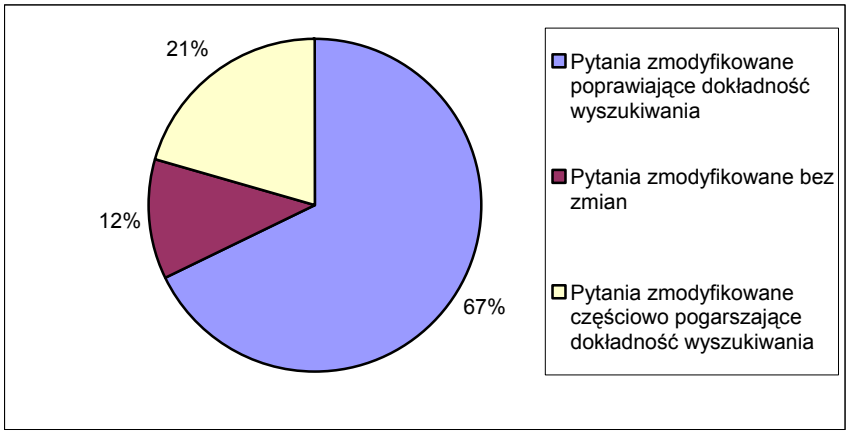
Rysunek 5-5: Schemat testowego zbioru dokumentów rzadkich.

Przeprowadzono wyszukiwania dla 38 pytań losowych. Dla każdego pytania testowego wykonywanych było 50 kolejnych iteracji modyfikacji pytania na podstawie tworzonego profilu. Parametry profilu pozostały niezmienione w stosunku do eksperymentów przeprowadzonych dla gęstego zbioru dokumentów testowych.

Dla wszystkich 38 pytań losowych porównano liczbę dokumentów dostarczanych w odpowiedzi na kolejne zmodyfikowane pytania. Wyniki przedstawia w postaci wykresu Rysunek 5-13. Dla wszystkich pytań losowych obliczono również miarę *%DR*, opisującą procent dokumentów relewantnych, które były znajdowane w kolejnych zmodyfikowanych pytaniach. Wyniki przedstawia w postaci wykresu Rysunek 5-14. Z 50-ciu wykonanych iteracji, na wykresach przedstawiono pierwszych 20-cia iteracji, ponieważ powyżej 20-tej iteracji następowała stabilizacja pytania oraz stabilizacja zbioru relewantnych dokumentów wyszukanych.

Dla każdego pytania wylosowanego oraz pytania zmodyfikowanego obliczono dokładności obcięte $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz wartość poprawy wyszukiwania *POP* dla każdego pytania zmodyfikowanego w stosunku do początkowego pytania wylosowanego. Przykładowe z przeprowadzonych eksperymentów, pokazujące zależności pomiędzy liczbą dokumentów odpowiedzi D'_q, dokładnościami obciętym $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz liczbą dokumentów relewantnych w odpowiedzi D_q, zamieszczono w Załączniku A niniejszej pracy.

[1] Przykładowe z przeprowadzonych eksperymentów umieszczono w Załączniku A.



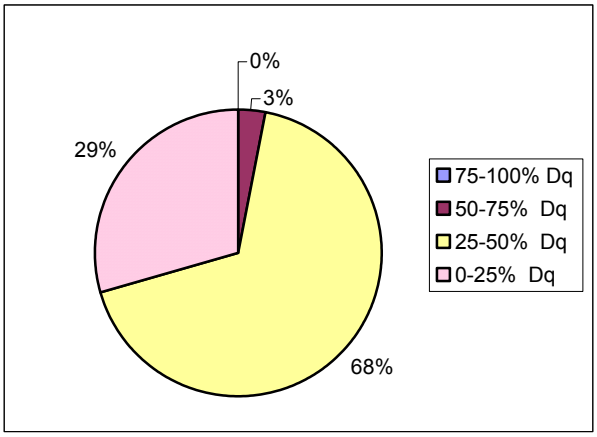
Rysunek 5-6: Zestawienie wszystkich pytań zmodyfikowanych obrazujące poprawę wyników wyszukiwania z wykorzystaniem profilu dla testowego zbioru dokumentów rzadkich.

Podobnie jak dla testowego gęstego zbioru dokumentów relewantnych, analizę wyników wyszukiwania przeprowadzono w trzech grupach. Zestawienie wyników zawiera Rysunek 5-6. W eksperymentach dla rzadkiego zbioru dokumentów testowych więcej jest o 15% cykli wyszukiwań, dla których zmniejsza się dokładność wyszukiwania. W tych eksperymentach modyfikacja profilu i, na jego podstawie, pytania przebiegała na podstawie tylko jednego dokumentu ze zbioru testowego, znalezionego w zbiorze dokumentów odpowiedzi. Wynika to z przyjętego, w tej części eksperymentów, małego podobieństwa dokumentów we zbiorze dokumentów testowych. Dlatego w wyniku modyfikacji pytanie często ulegało uogólnieniu. Odpowiedź na takie pytanie jest bardziej liczna, a ponieważ w metodzie nie jest stosowany inny niż standardowy ranking wyszukiwarki, dokumenty relewantne mogły znajdować się na dalszych niż 30-ta (ostatnia badana) pozycja.

Dla dokumentów ze zbioru rzadkiego, proces modyfikacji pytania na podstawie profilu użytkownika prowadzi do precyzowania pytania w jednej z dziedzin reprezentowanych w zbiorze rzadkim. Dziedzina, której będzie dotyczyć uszczegóławianie jest losową z pośród reprezentowanych w zbiorze rzadkim – wynika to z losowo wybieranego pytania początkowego. Przeprowadzone wyszukiwania pokazały, że w ramach dziedziny, w której następuje precyzowanie pytania również ma miejsce polepszenie efektów wyszukiwania dla każdego kolejnego pytania zmodyfikowanego w stosunku do losowego pytania początkowego (w odniesieniu do pewnego podzbioru zbioru rzadkiego). W kolejnych iteracjach wzrastają wartości dokładności obciętej $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ dla odpowiedzi na kolejne zmodyfikowane pytania, a tym samym wzrasta liczba dokumentów relewantnych *%DR* w odpowiedzi. Znajdowane są wszystkie dokumenty relewantne z jednej dziedziny, jednak nigdy nie znalezione zostały w odpowiedzi wszystkie dokumenty z zbioru testowego

dokumentów rzadkich[1], co oznacza, że modyfikacja pytania nie prowadzi do zmiany dziedziny wyszukiwania. Zmniejsza się łączna liczba wszystkich dokumentów odpowiedzi na kolejne modyfikacje pytania.

Rysunek 5-7 zawiera zestawienie wyników wyszukiwania, porównujące procent znalezionych dokumentów relewantnych z rzadkiego zbioru dokumentów relewantnych.



Rysunek 5-7: Liczba pytań testowych dla rzadkiego zbioru dokumentów testowych w podziale na procent znalezionych dokumentów testowych w odpowiedziach na kolejne pytania zmodyfikowane.

W eksperymentach, dla żadnego z pytań początkowych, w wyniku kolejnych modyfikacji pytania na podstawie profilu, w odpowiedzi na pytanie zmodyfikowane nigdy nie zostało znalezionych 75%–100% dokumentów z przygotowanych zbiorów testowych.

Eksperymenty przeprowadzone dla rzadkich zbiorów dokumentów testowych potwierdziły tezę mówiącą, że zastosowanie zaproponowanego profil użytkownika do modyfikacji pytania użytkownika prowadzi do poprawy wyników wyszukiwania tylko w jednej z dziedzin reprezentowanych w zbiorze testowym. Modyfikacja pytania nie prowadzi do uogólnienia pytania i znajdowania dokumentów z różnych dziedzin tematycznych.

Analiza wyników symulacji dla mieszanego zbioru dokumentów [2]

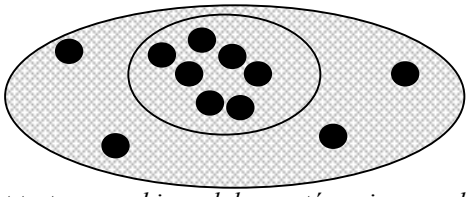
Celem trzeciego etapu eksperymentów było zweryfikowanie tezy mówiącej, że jeśli w kolekcji dokumentów zawierającej dokumenty relewantne, reprezentujące pewną dziedzinę zainteresowania użytkownika oraz dokumenty niezwiązane z tą dziedziną, użytkownik wskaże kilka dokumentów relewantnych, to wykorzystanie

[1] Ponieważ były tam rozmyślnie umieszczone dokumenty słabo powiązane z resztą zbioru.

[2] Przykładowe z przeprowadzonych eksperymentów umieszczono w Załączniku A.

zaproponowanego profilu użytkownika do personalizacji pytania doprowadzi do znalezienia innych dokumentów relewantnych z interesującej użytkownika dziedziny, natomiast dokumenty niezwiązane nie będą wyszukiwane w odpowiedzi na pytanie zmodyfikowane.

W tym etapie eksperymentów badano wyniki wyszukiwania dla kolejnych pytań losowych oraz dla ustalonych mieszanych zbiorów dokumentów testowych, które obok jądra – gęstego podzbioru – zawierały również dodatkowe rzadkie podzbiory dokumentów. Podzbiór gęsty dokumentów opisywał jedną dziedzinę zainteresowania użytkownika. Natomiast dokumenty z podzbiorów rzadkich były dokumentami słabo związanymi z dziedziną zainteresowania użytkownika opisywaną przez dokumenty gęstego podzbioru. Dokumenty z podzbiorów rzadkich były również słabo związane nawzajem ze sobą. Na schemacie zbiór dokumentów mieszanych przedstawiamy jako skupisko dokumentów bliskich sobie oraz pojedyncze dokumenty znacznie odległe od siebie nawzajem i od ‘jądra’.



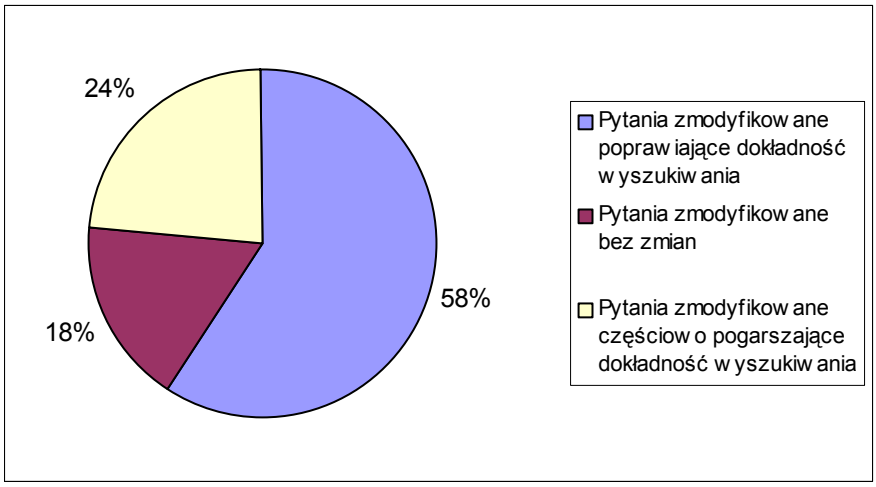
Rysunek 5-8: Schemat testowego zbioru dokumentów mieszanych.

W tym etapie przeprowadzono wyszukiwania dla 17 pytań losowych. Dla każdego pytania testowego wykonywanych było 50 kolejnych iteracji modyfikacji pytania na podstawie tworzonego profilu. Parametry profilu pozostały niezmienione w stosunku do eksperymentów przeprowadzonych dla gęstych i rzadkich zbiorów dokumentów testowych.

Dla wszystkich 17 pytań losowych porównano liczbę dokumentów dostarczanych w odpowiedzi na kolejne zmodyfikowane pytania. Wyniki przedstawia w postaci wykresu Rysunek 5-15. Dla wszystkich pytań losowych porównano również procent dokumentów relewantnych *%DR*, które były znajdowane w kolejnych zmodyfikowanych pytaniach. Wyniki przedstawia w postaci wykresu Rysunek 5-16. Z 50-ciu wykonanych iteracji, na wykresach przedstawiono pierwszych 20-cia iteracji, ponieważ powyżej 20-tej iteracji następowała stabilizacja pytania oraz stabilizacja zbioru relewantnych dokumentów wyszukanych.

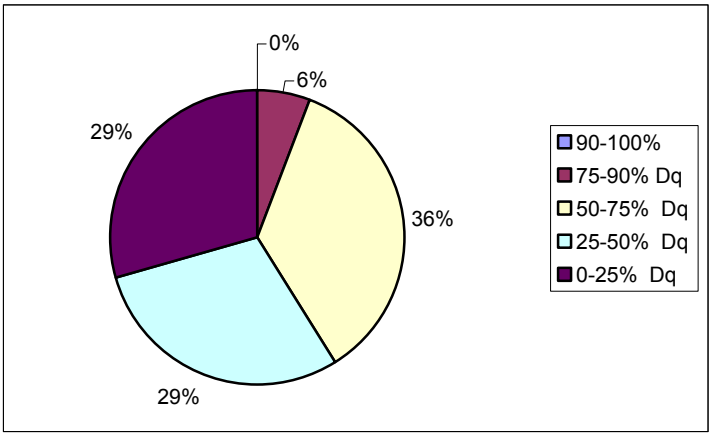
Dla każdego pytania wylosowanego oraz pytania zmodyfikowanego obliczono dokładności obcięte $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz wartość poprawy wyszukiwania *POP* dla każdego pytania zmodyfikowanego w stosunku do początkowego pytania wylosowanego. Przykładowe z przeprowadzonych eksperymentów, pokazujące

zależności pomiędzy liczbą dokumentów odpowiedzi D'_q, dokładnościami obciętym $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ oraz liczbą dokumentów relewantnych w odpowiedzi D_q, zamieszczono w Załączniku A.



Rysunek 5-9: Zestawienie wszystkich pytań zmodyfikowanych obrazujące poprawę wyników wyszukiwania z wykorzystaniem profilu dla testowego zbioru dokumentów mieszanych.

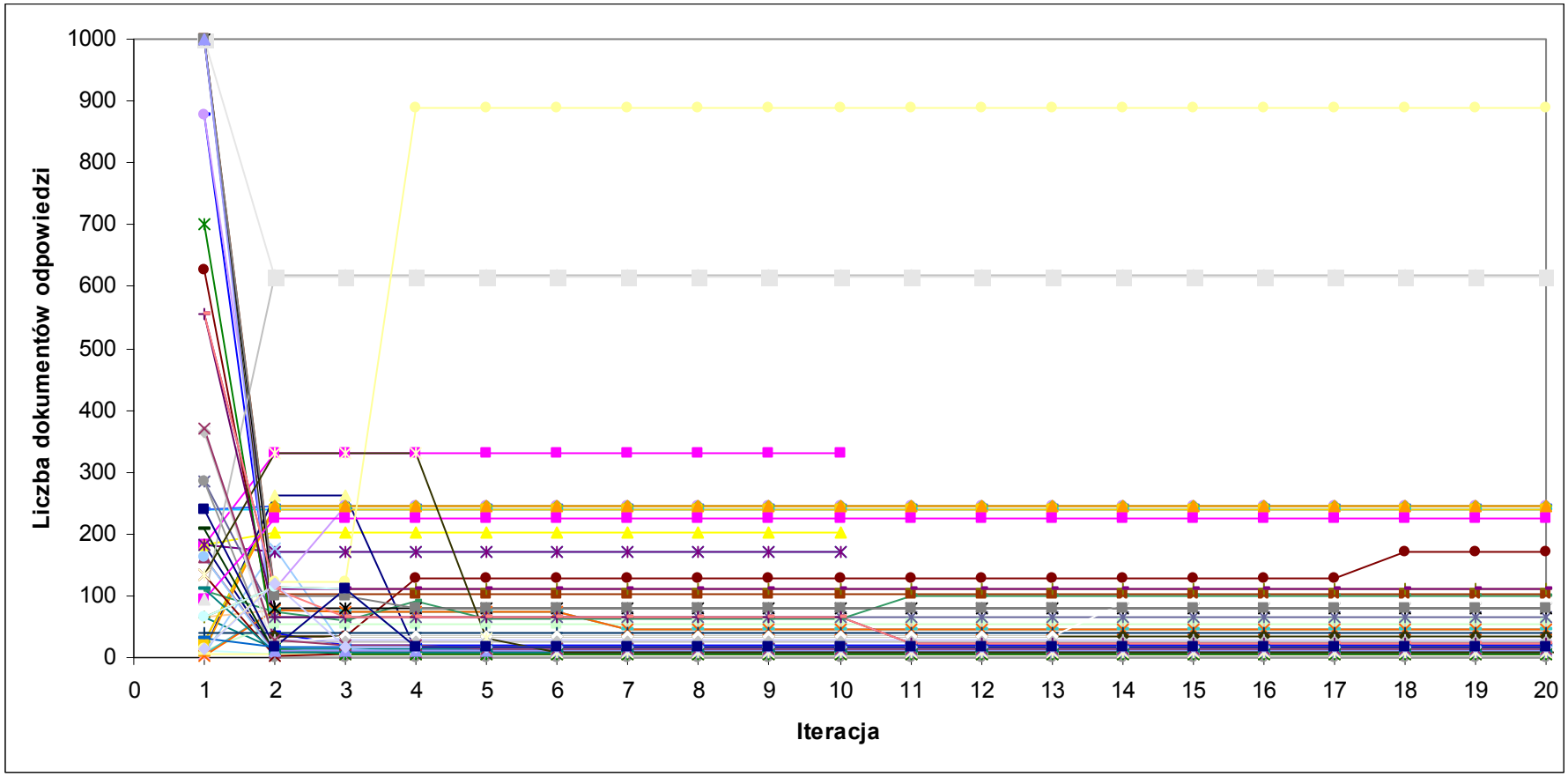
Podobnie jak dla poprzednich eksperymentów, analizę wyników wyszukiwania przeprowadzono w trzech grupach. Zestawienie wyników zawiera Rysunek 5-9. W eksperymentach dla mieszanego zbioru dokumentów testowych również jest więcej cykli wyszukiwań, dla których zmniejsza się dokładność wyszukiwania, niż dla wyszukiwań dla gęstego zbioru dokumentów testowych. Pogorszenie wyników wyszukiwania miało miejsce w sytuacji, gdy w odpowiedzi na pytanie losowe znaleziony został dokument z rzadkiego podzbioru, należącego do zbioru mieszanego. W tych eksperymentach modyfikacja profilu i, na jego podstawie, pytania przebiegała na podstawie tylko jednego dokumentu ze zbioru testowego, znalezionego w zbiorze dokumentów odpowiedzi. Wynika to z przyjętego, w tej części eksperymentów, małego podobieństwa części dokumentów ze zbioru mieszanego do 'jądra zbioru'. Dlatego w wyniku modyfikacji pytanie często ulegało uogólnieniu. Odpowiedź na takie pytanie jest bardziej liczna, a ponieważ w metodzie nie jest stosowany ranking dokumentów inny niż standardowo ustalony w komercyjnej wyszukiwarce *Netoskop*, dokumenty relewantne znajdowały się na dalszych niż 30-ta (ostatnia badana) pozycja.



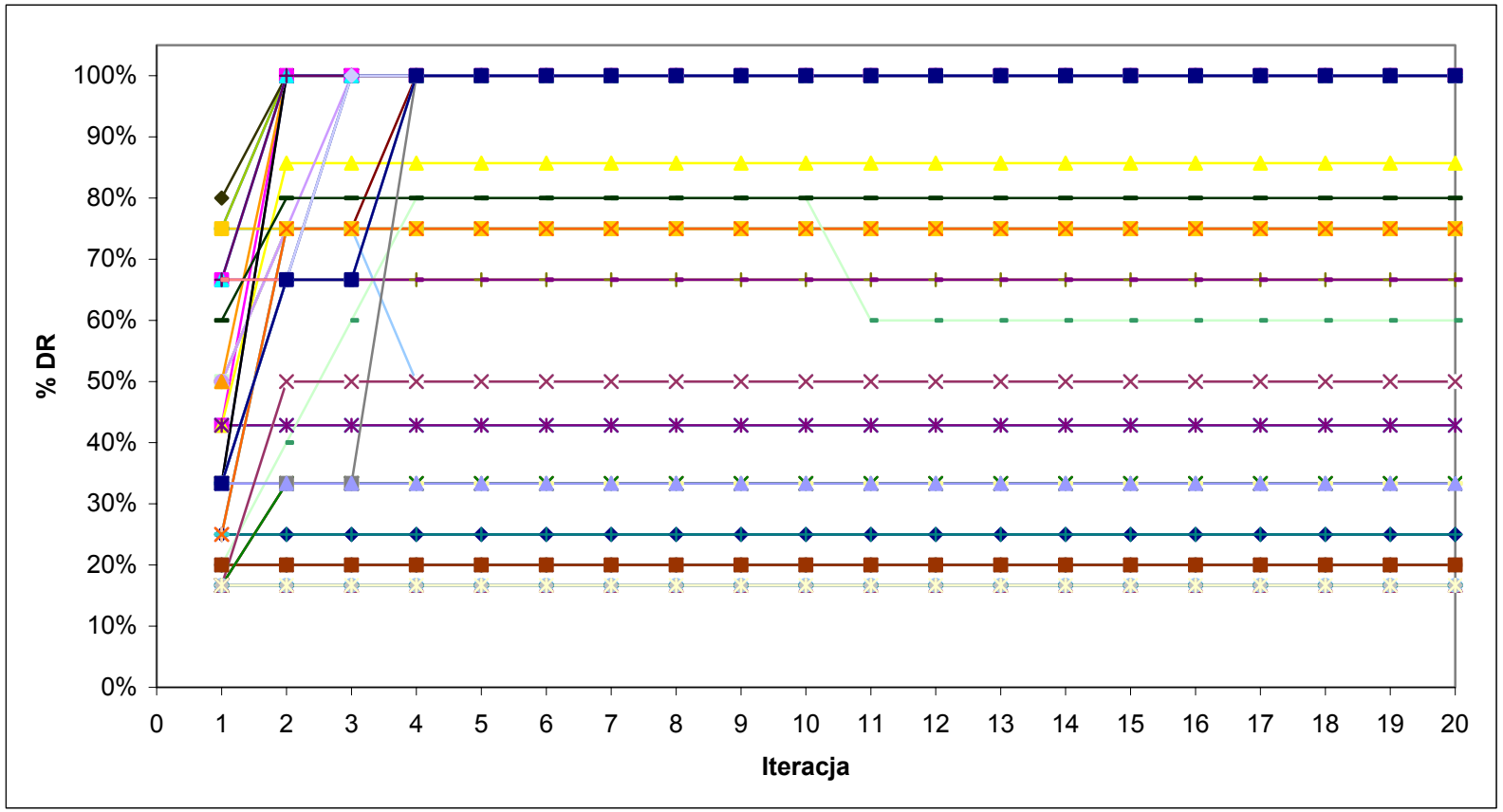
Rysunek 5-10: Liczba pytań testowych dla mieszanego zbioru dokumentów testowych w podziale na procent znalezionych dokumentów testowych w odpowiedziach na kolejne pytania zmodyfikowane.

Dla zbiorów mieszanych proces modyfikacji pytania na podstawie profilu użytkownika często prowadzi do znajdowania innych dokumentów relewantnych, związanych z dokumentami z podzbioru gęstego (‘jądra’). Przeprowadzone wyszukiwania pokazały, że w ramach dziedziny, w której następuje precyzowanie pytania również ma miejsce polepszanie efektów wyszukiwania (w odniesieniu do ‘jądra’ zbioru mieszanego) dla każdego kolejnego pytania zmodyfikowanego w stosunku do losowego pytania początkowego. W kolejnych iteracjach wzrastają wartości dokładności obciętej $Dokl_{10}$, $Dokl_{20}$, $Dokl_{30}$ dla odpowiedzi na kolejne zmodyfikowane pytania. Oznacza to wzrost liczby znajdowanych dokumentów relewantnych. Nigdy jednak nie zostały znalezione w odpowiedzi wszystkie dokumenty z mieszanego zbioru dokumentów testowych, co oznacza, że modyfikacja pytania nie prowadzi do zmiany dziedziny wyszukiwania[1]. W każdej kolejnej iteracji modyfikacji pytania zmniejsza się liczba wszystkich dokumentów znajdowanych jako odpowiedź na pytanie zmodyfikowane.

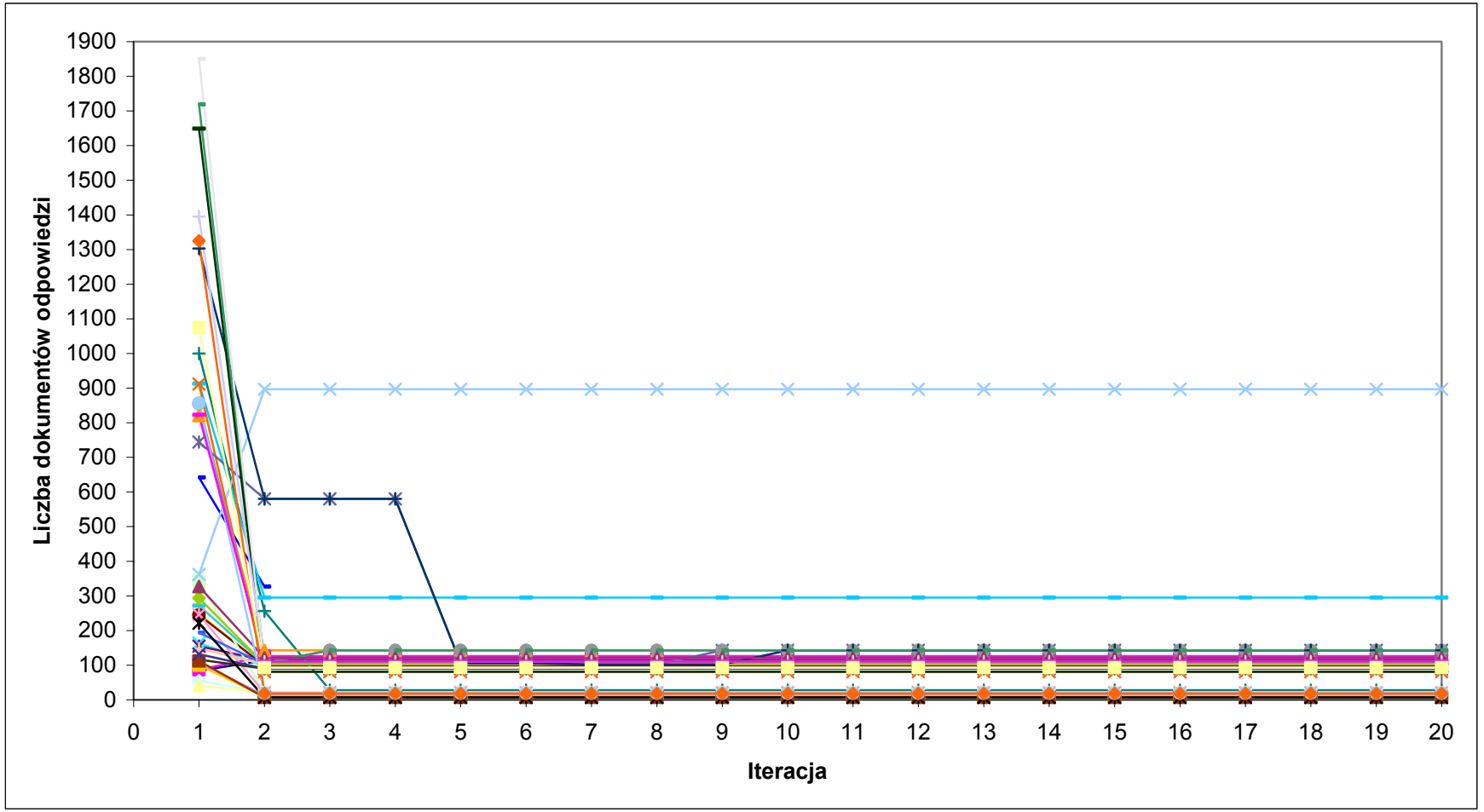
[1] Odnalezienie poprzez zmodyfikowane pytanie dokumentów z podzbioru rzadkiego zbioru mieszanego, które są słabo związane znaczeniowo z ‘jądrem’, sygnalizowałoby zmianę lub rozszerzenie dziedziny zainteresowań wyrażanej przez pytanie zmodyfikowane w stosunku do dziedziny założonej dla wylosowanego pytania początkowego. Dziedzina założona dla pytania początkowego odpowiada z definicji ‘jądru’ zbioru mieszanego.



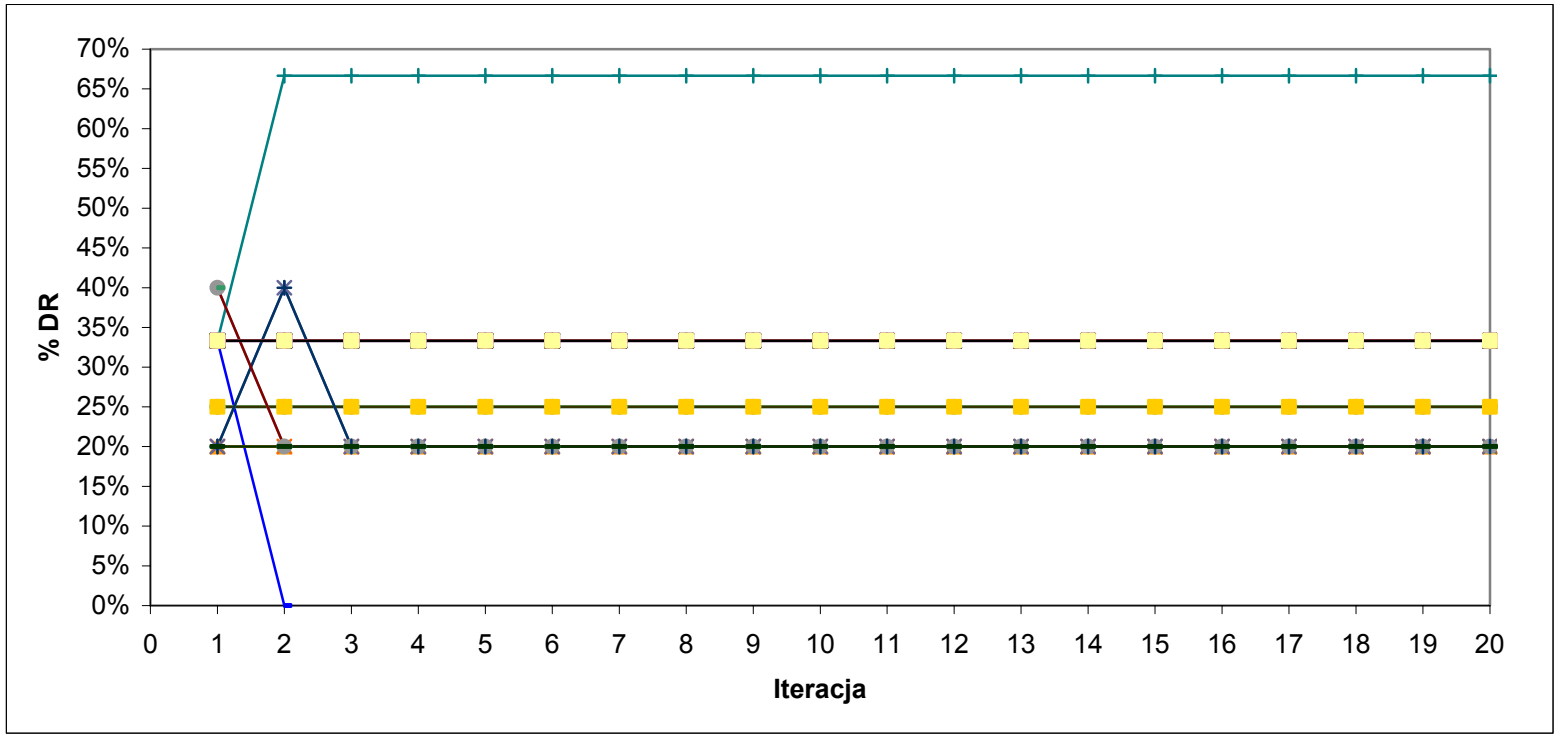
Rysunek 5-11: Liczba dokumentów zwracanych w odpowiedzi na pytanie losowe i pytania zmodyfikowane w kolejnych iteracjach dla gęstych zbiorów dokumentów relewantnych.



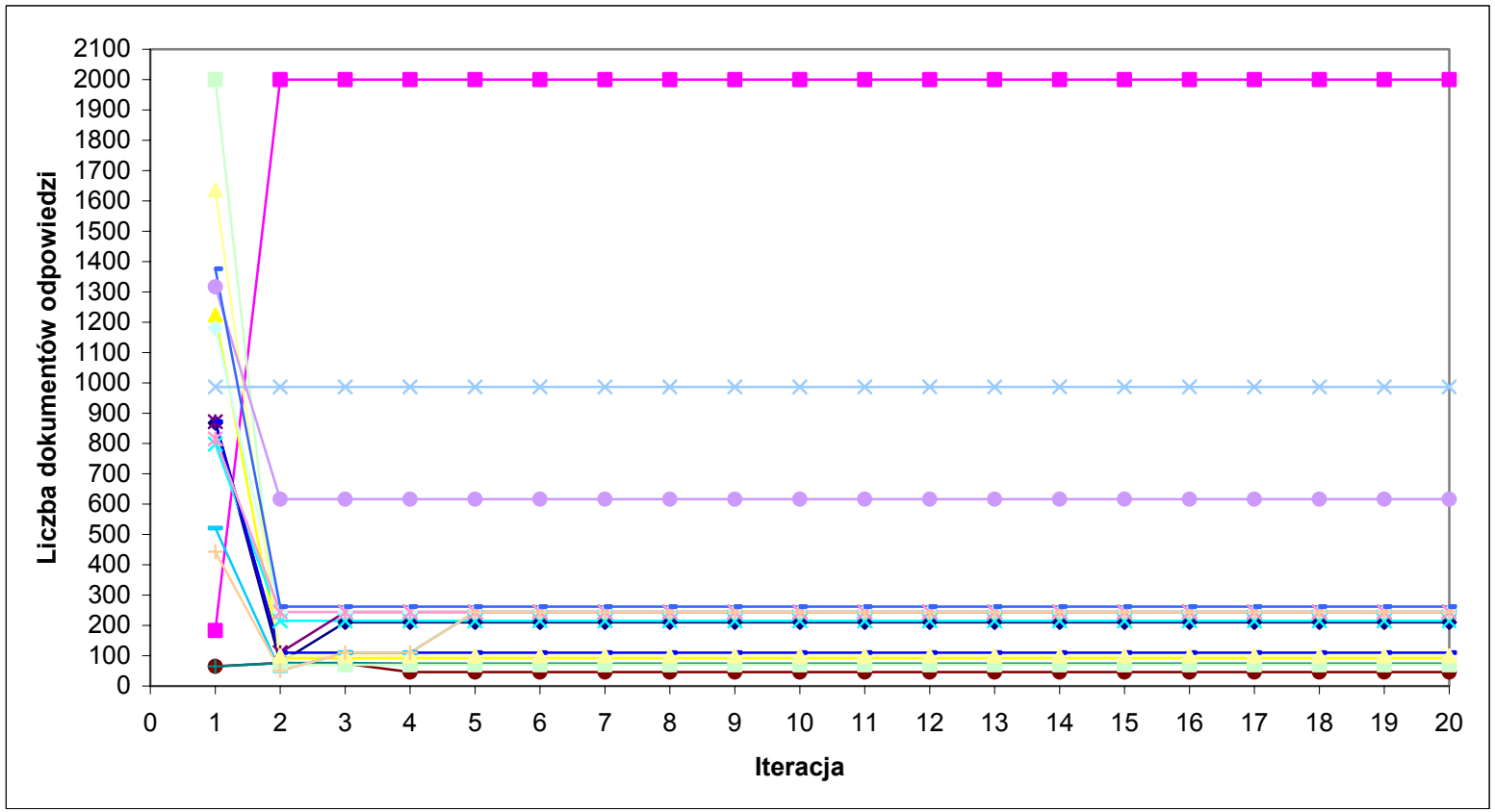
Rysunek 5-12: Zestawienie liczby dokumentów relewantnych wyszukanych w stosunku do liczby wszystkich dokumentów relewantnych dla gęstych zbiorów dokumentów relewantnych.



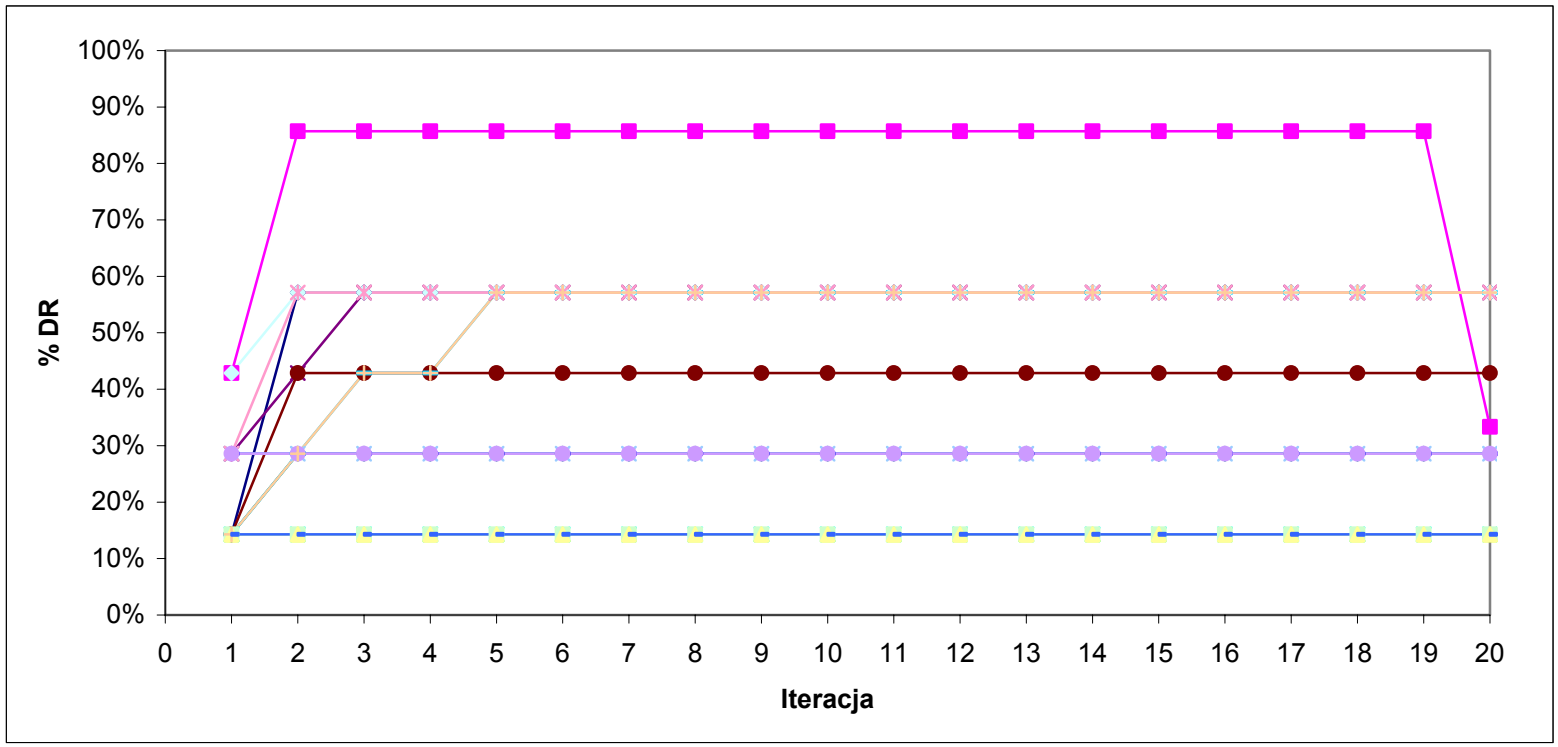
Rysunek 5-13: Liczba dokumentów zwracanych w odpowiedzi na pytanie losowe i pytania zmodyfikowane w kolejnych iteracjach dla rzadkich zbiorów dokumentów.



Rysunek 5-14: Zestawienie liczby dokumentów relewantnych wyszukanych w stosunku do liczby wszystkich dokumentów relewantnych dla rzadkich zbiorów dokumentów.



Rysunek 5-15: Liczba dokumentów zwracanych w odpowiedzi na pytanie losowe i pytania zmodyfikowane w kolejnych iteracjach dla mieszanych zbiorów dokumentów.



Rysunek 5-16: Zestawienie liczby dokumentów relewantnych wyszukanych w stosunku do liczby wszystkich dokumentów relewantnych dla mieszanych zbiorów dokumentów.

5.5. Wnioski z eksperymentów

Zaproponowany w niniejszej pracy profil użytkownika ma służyć personalizacji wyszukiwania informacji w sieci WWW. Personalizacja wyszukiwania ma miejsce podczas formułowania zmodyfikowanego pytania użytkownika oraz podczas prezentowania podpowiedzi dotyczących pytania postawionego przez użytkownika, gdy korzysta on z wyszukiwarki internetowej z profilem. Podpowiedź, będąca zmodyfikowaniem pytania postawionego przez użytkownika, powstaje na podstawie analizy dotychczasowej pracy użytkownika z wyszukiwarką internetową. Celem zastosowania niniejszego profilu jest podnoszenie satysfakcji użytkownika z wyszukiwania przez dostarczanie użytkownikowi odpowiedzi zawierającej coraz więcej dokumentów relewantnych w stosunku do wszystkich dokumentów odpowiedzi w kolejnych cyklach wyszukiwania, przy jednoczesnym zmniejszeniu ogólnej liczby dokumentów odpowiedzi.

Aby zweryfikować powyższe tezy, przygotowano środowisko testowe. Wykorzystano wyszukiwarkę internetową Netoskop oraz zbudowano system modelowania użytkownika *Profiler*. Przygotowano testową kolekcję poindeksowanych dokumentów oraz pytania testowe uzyskane od ekspertów. Dla pytań testowych uzyskanych od ekspertów utworzono zbiory dokumentów testowych: relewantnych (gęstych), słabo powiązanych (rzadkich) oraz relewantnych zawierających podzbiory dokumentów słabo związanych (mieszanych). Następnie losowano pytania przekazywane do wyszukiwarki, które zawierały terminy ze zbioru T. W ramach symulacyjnej weryfikacji zaproponowanego profilu użytkownika, wybieranie ze zbioru dokumentów odpowiedzi dokumentów relewantnych należących do zbioru testowego jest symulacją wskazywania przez użytkownika dokumentów relewantnych w odpowiedzi zwracanej przez internetowy system wyszukiwania informacji.

Celem eksperymentów przeprowadzonych w ramach pracy było wykazanie, że na podstawie automatycznej analizy dokumentów relewantnych z dziedziny zainteresowań użytkownika, zaproponowana metoda tworzenia i modyfikacji profilu oraz metoda modyfikacji pytania użytkownika doprowadzą do takiego zmodyfikowania początkowego pytania użytkownika, że w kolejnych wyszukiwaniach użytkownik otrzymuje coraz więcej dokumentów relewantnych przy zmniejszającej się ogólnej liczbie dokumentów odpowiedzi. O niepogarszaniu wyników wyszukiwania mówi własność 4.8.1 z podrozdziału 4.8, która została potwierdzona przez przeprowadzone eksperymenty.

Przeprowadzone eksperymenty pokazały również, że zaproponowany profil jest zgodny z ogólną intuicją profilu reprezentującego zainteresowania użytkownika oraz intuicją wykorzystania profilu w procesie wyszukiwania informacji:

1. Profil reprezentuje zainteresowania użytkownika ujawnione podczas interakcji użytkownika z internetowym systemem wyszukiwania informacji. Różne zainteresowania reprezentowane są przez różne subprofile użytkownika.
2. Profil tworzony i modyfikowany jest automatycznie na podstawie obserwacji interakcji użytkownika z systemem. Proces ten wymaga minimalnej ingerencji ze strony użytkownika.
3. Korzystając z systemu wyszukiwania informacji, umożliwiającego personalizację wyszukiwania dzięki zastosowaniu profilu użytkownika, użytkownik oczekuje, że po pewnym czasie współpracy z systemem, w odpowiedzi na kolejne pytania zmodyfikowane, będzie otrzymywał rosnącą liczbę dokumentów na interesujący go temat, a całkowita liczna dokumentów odpowiedzi zmniejszy się. Modyfikacja pytania nie zmienia dziedziny tematycznej, w której zostało zadane pytanie początkowe. Również nie są gubione raz znalezione dokumenty relewantne z danej dziedziny.

Wyznaczana liczba terminów z profilu do modyfikacji pytania użytkownika (polegającej na zastępowaniu terminów pytania) jest niezbyt duża, zazwyczaj w granicach 3–4 terminów. Jest to liczba terminów wystarczająca do zawężenia pytania początkowego, ale nie jest na tyle duża, aby terminy te zadane jako pytanie spowodowały, że system przekaże odpowiedź pustą. Zauważono jednak, że jeśli pytanie zmodyfikowane było dłuższe niż 6 terminów, to najczęściej wyszukiwarka internetowa zwracała odpowiedź pustą (dla wykorzystywanej ograniczonej kolekcji i dla wykorzystywanej konkretnej, komercyjnej wyszukiwarki).

Na początku powstawania profilu, gdy dołączane są pierwsze terminy do subprofilu, terminów w subprofilu jest niewiele (ok. 3–4). Im więcej dokumentów relewantnych jest wskazanych po pierwszym wyszukiwaniu, tym jest mniej wspólnych terminów znaczących w tych dokumentach. Wyznaczone podczas pierwszego wyszukiwania terminy znaczące mają wysokie wagi. Wysoka wartość wagi wskazuje, że do subprofilu włączone zostały terminy istotne, dobrze reprezentujące dziedzinę zainteresowania użytkownika. Podczas procesu wybierania terminów z subprofilu do modyfikacji pytania początkowego, w przypadku małej liczby terminów znaczących w subprofilu, pewne problemy powoduje zastosowanie współczynnika istotności ι, obliczanego jako średnia wag terminów w profilu. Jeśli terminów w subprofilu było niewiele (ok. 3–4 terminy), a ich wagi przyjmowały wartości podobnego rzędu, żadne terminy z subprofilu mogą nie zostać wybrane do modyfikacji pytania. Dlatego w procesie wybierania terminów z subprofilu, dodatkowo oprócz współczynnika istotności ι, zastosowano warunek określający liczbę wybieranych terminów. Uzasadnieniem przyjęcia takiego rozwiązania jest fakt, że wszystkie terminy, które zostały dodane do subprofilu po pierwszym pytaniu są terminami dobrymi.

Analiza wyników eksperymentów pokazała, że termin, który pojawia się w subprofilu po raz pierwszy, po wykonanych już kilku wyszukiwaniach

z wykorzystaniem danego subprofilu, ma w tym subprofilu niską wagę. Jest to pozytywne zachowanie profilu dla terminów nowych. Termin pojawiający się po raz pierwszy w subprofilu może być terminem przypadkowo użytym przez autora wskazanego dokumentu relewantnego. Jeżeli zaistniała taka sytuacja, jest mało prawdopodobne, że ten sam termin pojawi się w kolejnych wyszukiwaniach w innych dokumentach dotyczących tej samej dziedziny. Waga takiego terminu w subprofilu powinna maleć po każdym kolejnym wyszukiwaniu. I tak się właśnie dzieje w zaproponowanym profilu użytkownika. Jeśli natomiast termin będzie się pojawiał podczas kolejnych procesów wyszukiwawczych, realizowanych dla tego samego pytania, waga tego terminu wzrasta i termin ten jest wybierany podczas kolejnej iteracji tworzenia pytania zmodyfikowanego. Ciągły wzrost wagi nowego, w pewnym momencie, terminu w subprofilu oznacza pewną zmianę kierunku zainteresowań użytkownika na taki, w którym używany jest ten termin lub, że jest to nowy termin, który zaczął funkcjonować w słownictwie danej dziedziny.

Eksperymenty potwierdziły również, że zaproponowana na potrzeby tworzenia, wykorzystania i modyfikacji profilu użytkownika metoda selekcji terminów znaczących tz_i, które dobrze opisują dziedzinę zainteresowań użytkownika jest dobra i skuteczna. Wyselekcjonowane terminy znaczące są terminami charakterystycznymi dla dziedziny zainteresowań użytkownika, ale równocześnie rzadko występują w innych dziedzinach (czyli słabo opisują inne dziedziny), reprezentowanych w danej kolekcji dokumentów. Za własność tą odpowiadają waga terminów *tf–idf* oraz wskaźnik ważności terminów *cv*. Waga *tf–idf* wyznacza terminy będące dobrymi terminami indeksowymi w całej kolekcji, natomiast wskaźnik ważności *cv* wyznacza terminy charakterystyczne dla dokumentów relewantnych w odpowiedzi, które to terminy równocześnie rzadko pojawiają się w pozostałych dokumentach odpowiedzi. Zastosowana, nowa metoda selekcji terminów znaczących gwarantuje, że w odpowiedzi na kolejne zmodyfikowane pytanie użytkownik otrzymuje odpowiedź, w której zwiększa się udział dokumentów relewantnych opisujących dziedzinę zainteresowań użytkownika w stosunku do wszystkich dokumentów odpowiedzi.

Dodatkowo eksperymenty pokazały, że dla wyszukiwania informacji w sieci WWW istnieje wprost proporcjonalna zależność pomiędzy licznością kolekcji, a polepszeniem efektywności wyszukiwania. Dla większych kolekcji rzadko ma miejsce sytuacja, że zadane przez użytkownika pytanie nie zostanie zmodyfikowane z powodu pustej odpowiedzi internetowego systemu wyszukiwania informacji.

Analizując zagadnienie dynamiki kolekcji dokumentów w internetowym systemie wyszukiwania informacji w kontekście zaproponowanego w pracy modelu profilu użytkownika należy zauważyć kilka podstawowych trudności jakie pojawiają się w przypadku eksperymentalnej weryfikacji modelu w środowisku dynamicznym. Po pierwsze nie istnieje opracowany model zmian w dokumentach kolekcji WWW oraz zmian samej kolekcji, do którego można by odnieść zaproponowany w pracy model

profilu. Po drugie procesy zamian w kolekcji dokumentów WWW są wolne – trwają latami i są nieprzewidywalne – nie wiadomo, która część WWW ulegnie zmianie, a więc nie ma pewności, że dokumenty losowo wybrane na potrzeby eksperymentów akurat będą podlegały modyfikacji. Mając na uwadze powyższe trudności, eksperymentalną weryfikację profilu użytkownika zaplanowano i zrealizowano dla politematycznej kolekcji dokumentów, zgromadzonych z sieci WWW w pewnym ograniczonym interwale czasowym.

Zauważmy, że można wyodrębnić trzy możliwe zależności pomiędzy modyfikacją kolekcji a modyfikacją i wykorzystaniem profilu użytkownika:

1. kolekcja nie ulega modyfikacji pomiędzy kolejnymi wyszukiwaniami (czyli pomiędzy kolejnymi modyfikacjami profilu użytkownika),
2. kolekcja ulega modyfikacji pomiędzy kolejnymi wyszukiwaniami,
3. kolekcja ulega modyfikacji pomiędzy kolejnymi iteracjami (czyli modyfikacjami jednego początkowego pytania użytkownika na podstawie danych zgromadzonych w profilu).

W pierwszym przypadku wyszukiwanie odbywa się analogicznie, jak w klasycznych systemach wyszukiwania informacji ze stałą kolekcją dokumentów. Ta właśnie sytuacja była badana w przeprowadzonych eksperymentach.

W drugim przypadku, jeśli częstość dokumentowa (*idf*) dla terminu rośnie tzn., że dodawane są do kolekcji dokumenty zawierające ten termin. Należy rozważyć w jakiej sytuacji użytkownik może otrzymać pustą odpowiedź z internetowego systemu wyszukiwania informacji, rozszerzonego o profilowanie zainteresowań użytkownika?

Jeśli czynnik *idf* dla terminu rośnie tzn., że w nowej kolekcji pojawiają się nowe dokumenty zawierające dany termin. Jednocześnie jeśli waga tego terminu w subprofilu (zbudowanym na podstawie poprzednich wyszukiwań) jest niska, termin ten nie zostanie wytypowany do pytania zmodyfikowanego, wtedy nowy dokument może nie zostać wyszukany, a z drugiej strony wzrost czynnika *idf* dla terminu sugeruje wzrost istotności terminu w kolekcji. Taka sytuacja musiałaby jednocześnie oznaczać, że ta sama dziedzina opisana jest kompletnie różnymi słowami od słów stosowanych w dotychczasowej kolekcji, wtedy nowy dokument na interesujący użytkownika temat nie byłby znaleziony.

Jest to problematyczny przypadek krytyczny, w którym istotność terminu w reprezentowaniu dziedziny była niedoszacowana w profilu w poprzednich wyszukiwaniach, a znaczenie tego terminu w opisie dziedziny aktualnie rośnie. Teoretycznie jest to sytuacja możliwa, ale praktycznie mało prawdopodobne, aby język zmieniał się diametralnie pomiędzy kolejnymi wyszukiwaniami. Oraz mało prawdopodobne, aby zakresy słownictwa z danej dziedziny przed modyfikacją kolekcji i po jej modyfikacji były rozłączne, jeśli modyfikacja kolekcji następują pomiędzy kolejnymi wyszukiwaniami wykonywanymi w sensownych odstępach czasu.

Oczywiście wyszukiwania wykonywane przez użytkownika po wielu latach od poprzedniego nie może zagwarantować pokrywania się słownictwa. Jest to najbardziej niekorzystna sytuacja. Długi czas nie używania profilu jest najbardziej niekorzystnym przypadkiem dla jego działania. Jest możliwe, że po długim czasie słowa, które były używane w danej dziedzinie w przeszłości przestały być kompletnie stosowane obecnie. Jednak zmiana słownictwa w języku nie zachodzi gwałtownie. Słownictwo ulega modyfikacji stopniowo. Gdy pojawia się nowe słowo, to stare (opisujące tę samą tematykę) nie znika od razu ze słownictwa, ale może być powoli wypierane. Słowa funkcjonują razem, a jeśli nowe przyjmie się – może nastąpić zastąpienie.

Dynamika języka nie jest tak znaczna „w czasie rzeczywistym", aby stare słowo było całkowicie zastąpione przez nowe słowo, a nowe nie było w ogóle używane ze starym słowem, czyli aby nowe dokumenty (tzn. dodane lub o zmodyfikowanej treści) z pewnej tematyki w kolekcji gubione były całkowicie po zadaniu pytania utworzonego na podstawie profilu z niedalekiej przeszłości.

Największym zmianom w kolekcji mogą ulegać nazwy własne: nazwy firm, produktów, technologii ale nie tematyka, w której te nazwy funkcjonują, np. dział gospodarki lub nauki. I tak np. jeśli ktoś poszukuje informacji na temat „odtwarzaczy muzyki" to mogą to być zarówno odtwarzacze CD, jak i Mp3 pewnego lub innego producenta, ale cały czas będzie to informacja dotycząca odtwarzaczy muzyki.

W tym samym drugim przypadku, rozważmy sytuację gdy częstość dokumentowa (*idf*) dla terminu maleje co oznacza, że z kolekcji usunięte zostały dokumenty zawierające dany termin. Należy rozważyć kiedy użytkownik może otrzymać pustą odpowiedź z internetowego systemu wyszukiwania informacji, rozszerzonego o profilowanie zainteresowań użytkownika? Odpowiedź na pytanie zmodyfikowane na podstawie profilu mogłaby być pusta, gdyby w pytaniu znalazły się terminy o wysokiej wadze w subprofilu, ale dla których czynnik *idf* zmalał do zera.

Jednak terminy, które całkowicie znikają ze słownictwa stosowanego w dokumentach kolekcji muszą być terminami bardzo wąskimi, np. nazwami własnymi technologii lub produktu, które całkowicie nie są już stosowane, czy użytkowane w dziedzinie, a co więcej nie są już używane w nowotworzonych dokumentach. W takim przypadku pytanie zadane do systemu wyszukiwania informacji z nową kolekcją dokumentów, z której usunięte zostały dokumenty dotyczące np. tej szczególnej technologii lub produktu, dałoby odpowiedź pustą. Autor pracy mając na uwadze ten mankament, skonstruował metodę wyboru terminów do profilu tak, aby terminy bardzo szczegółowe i wąskie nie były pomijane, a wybierane tylko terminy ogólniej opisujące pewną tematykę. Powszechną praktyką jest archiwizowanie niektórych zasobów sieci WWW w innych lokalizacjach, dzięki czemu często dokumenty usunięte z jednej lokalizacji są możliwe do odnalezienia w innej.

W trzecim przypadku kolekcja ulega modyfikacji pomiędzy jedną a drugą iteracją (modyfikacją pytania na podstawie profilu użytkownika). Iteracje następują

w sekundowych odstępach czasu, więc kompensują modyfikacje kolekcji. Dodane lub zmodyfikowane dokumenty są już po następnej modyfikacji pytania wyszukiwane i, jeśli zostaną ocenione jako relewantne przez użytkownika, wykorzystane do budowania profilu użytkownika. Wzrost lub spadek czynnika *idf* (częstości dokumentowej) jest na bieżąco weryfikowany (wykorzystywany) przez użytkownika podczas kolejnych iteracji wyszukiwania.

Najbardziej niekorzystny przypadek ma miejsce w sytuacji całkowitego usunięcia z sieci WWW serwisu opisującego bardzo specyficzną dziedzinę, firmę, produkt, nazwy, które użytkownik zadaje jako pytanie do internetowego systemu wyszukiwania informacji, a co więcej opis taki znajduje się tylko w jednym dokumencie w całej sieci WWW. W takim przypadku żadna metoda modelowania zainteresowań użytkownika nie będzie pomocna i zawsze odpowiedź na takie pytanie będzie odpowiedzią pustą.

Z drugiej jednak strony, jeśli użytkownik zadaje tak szczegółowe pytania do wyszukiwarki, co oznacza, że ma bardzo dokładnie sprecyzowaną potrzebę informacyjną, to wspomaganie jego wyszukiwania proponowanym w pracy narzędziem jakim jest profil użytkownika może być uznane za nadmiarowe przez użytkownika. Doświadczony użytkownik nie potrzebuje tak znacznego wspomagania, które mogłoby być przy takiej wiedzy użytkownika potraktowane jako utrudnienie a nie wspomaganie.

Dynamika sieci WWW wiąże się głównie z dynamiką dodawania, modyfikacji i usuwania dokumentów z kolekcji. Natomiast nie jest związana ze szczególną dynamiką zamian w obrębie języka stosowanego do tworzenia tych dokumentów. Zmiany ilościowe w internetowej kolekcji dokumentów nie przekładają się na gwałtowne zmiany w zasobach języka, który jest używany do tworzenia tych dokumentów. W eksperymentach weryfikowano zaproponowaną w pracy koncepcję tworzenia i wykorzystani profilu użytkownika, w której to koncepcji chcemy skorzystać z relacji językowych. Jest to heurystyka, która w większości przypadków prowadzi do poprawy wyników wyszukiwania. W ramach tej heurystyki określamy relację pomiędzy słownictwem użytkownika a słownictwem w dokumentach z pewnej dziedzinie opisanej w systemie wyszukiwania informacji. Z dużą dozą pewności można sądzić, że relacja ta jest niezależna od zmiany liczby dokumentów w kolekcji.

Dynamika kolekcji internetowego systemu wyszukiwania informacji jest również związana z pojawianiem się dokumentów z nowej dziedziny, która nie była reprezentowana w kolekcji przed modyfikacją. Jeśli użytkownik zainteresuje się tą dziedziną (zadając odpowiednie pytanie i wskazując dokumenty relewantne w odpowiedzi), do wielotematycznego profilu dodana zostanie odpowiednia reprezentacja tej dziedziny w postaci subprofilu.

Przedstawiony powyżej wywód uzasadniający przeniesienie wyników eksperymentów wykonanych w środowisku statycznych dokumentów WWW na dynamiczną kolekcję w sieci WWW zaznacza sytuacje krytyczne, takie w których korzystanie z profilu może stać się bezużyteczne. Pomimo koniecznego ograniczenia

eksperymentów jednak uprawniony jest wniosek, że zastosowanie zaproponowanego profilu użytkownika w procesie wyszukiwania w systemie internetowym wpłynie na wzrost satysfakcji użytkownika z wyników tego wyszukiwania.

Zaproponowany w pracy model jest koncepcją rozwojową. Aby ostatecznie zweryfikować sytuacje krytyczne potrzebne wydaje się prowadzenie dalszych eksperymentów w tym kierunku. W przeprowadzonych eksperymentach potwierdzono korzyści z używania profilu przez użytkownika podczas wyszukiwania w rozległych oraz wielotematycznych zasobach sieci WWW zgromadzonych dla pewnego momentu w czasie.

Przytoczone powyżej wnioski można uznać za słuszne również w kontekście zmian zachodzących w języku. Język ewoluuje, ale nie na tyle dynamicznie, aby reprezentacja zainteresowań użytkownika zgromadzona w profilu zdewaluowała się pomiędzy kolejnymi modyfikacjami kolekcji. Zasoby języka (słownictwo) zmieniają się zazwyczaj płynnie niż gwałtownie, a stare słownictwo jest zastępowane przez nowe (jeśli to ostanie zostanie przyjęte) z fazą przejściową funkcjonowania obu jednocześnie w języku. Jeśli użytkownik regularnie wykorzystuje profil podczas wykonywanych wyszukiwań to tym samym, za sprawą zaproponowanej w pracy nowej metody selekcji i ważenia terminów, pojawiające się zamiany słownictwa wprowadzane są do profilu.

6. Podsumowanie

W pracy opracowano profil użytkownika, reprezentujący zainteresowania użytkownika korzystającego z internetowego systemu wyszukiwania informacji oraz procedury automatycznego tworzenia i modyfikacji profilu na podstawie pytań kierowanych przez użytkownika do systemu oraz dokumentów zwracanych w odpowiedzi systemu, ocenionych przez użytkownika. Zaproponowany profil jest niezależny od systemu wyszukiwania informacji, od przyjętego modelu tego systemu, czy realizacji systemu. Profil oraz mechanizmy jego tworzenia, modyfikacji i wykorzystania mogą być niezależnym, oddzielnym elementem, dołączonym zarówno do boolowskiego, jak i wektorowego systemu wyszukiwania informacji (działanie profilu nie zależy od sposobu wyboru dokumentów odpowiedzi przez system wyszukiwania). Jednak szczególnie istotne zastosowanie profilu użytkownika autor pracy widzi w obszarze personalizacji wyszukiwania informacji w sieci WWW. Personalizacja wyszukiwania z zastosowaniem profilu ma miejsce podczas formułowania zmodyfikowanego pytania użytkownika, gdy korzysta on z internetowego systemu wyszukiwania, czyli wyszukiwarki internetowej połączonej z modułem profilu. Pytanie zmodyfikowane, będące podpowiedzią modyfikacji pytania postawionego przez użytkownika, powstaje na podstawie analizy dotychczasowej interakcji użytkownika z wyszukiwarką internetową. Zastosowanie zaproponowanego profilu użytkownika podnosi satysfakcję użytkownika oraz zwiększa efektywność wyszukiwania. Użytkownikowi dostarczane są odpowiedzi z wyszukiwarki internetowej zawierające coraz więcej dokumentów relewantnych w stosunku do wszystkich dokumentów odpowiedzi w kolejnych cyklach wyszukiwania, przy jednoczesnym zmniejszeniu ogólnej liczby dokumentów odpowiedzi, tj. skróceniu odpowiedzi.

Zaproponowany profil użytkownika oprócz klasycznych systemów wyszukiwania informacji ma szczególne zastosowanie dla internetowych systemów wyszukiwania, gdzie użytkownicy są najczęściej nowicjuszami w dziedzinie wyszukiwania, a kolekcja dokumentów jest zmienna. Indeksowanie dokumentów z sieci WWW jest wykonywane praktycznie ciągle, aby zachować aktualność indeksów dla zmieniających się zasobów sieci WWW. Profil użytkownika budowany na podstawie informacji zawartych we wskazanych dokumentach relewantnych jest wykorzystywany do kolejnych wyszukiwań w zmienionej kolekcji. Pytanie zmodyfikowane, zadane do nowej kolekcji, spowoduje znalezienie innych, nowych dokumentów relewantnych, jeżeli takie pojawią się w nowej kolekcji.

Przeprowadzone eksperymenty, których celem była weryfikacja zaproponowanego profilu użytkownika, potwierdzają zgodność profilu z intuicją wykorzystania profilu użytkownika w systemie wyszukiwania informacji. Korzystając z systemu wyszukiwania informacji w sieci WWW, umożliwiającego personalizację wyszukiwania dzięki zastosowaniu profilu, użytkownik oczekuje, że po pewnym czasie współpracy z systemem będzie otrzymywał w odpowiedzi na pytania z określonej dziedziny zainteresowań coraz więcej dokumentów na interesujący go temat, a odpowiedź ogólnie będzie coraz mniej liczna. Taką funkcjonalność zapewnia zaproponowany profil użytkownika, co potwierdziły przeprowadzone eksperymenty.

Zaproponowany w pracy profil użytkownika ma szczególnie istotne zastosowanie w internetowych systemach wyszukiwania informacji. Profil służy do personalizacji wyszukiwania poprzez modyfikowanie pytania użytkownika na podstawie analizy interakcji użytkownika z systemem. Prowadzi to do adaptacji systemu na poziomie formułowania pytania w kierunku wyznaczonym przez dziedzinę zainteresowania użytkownika.

Dodatkowe polepszenie wyszukiwania mogłaby również przynieść personalizacja prezentowania dokumentów wyszukanych przez system w odpowiedzi na zmodyfikowane pytanie użytkownika. Autor pracy widzi możliwości zastosowania rankingu dokumentów odpowiedzi, gdzie kryterium uporządkowania dokumentów byłoby podobieństwo dokumentu odpowiedzi do subprofilu lub zmodyfikowanego pytania użytkownika. Na początku tak utworzonego rankingu znalazłyby się dokumenty najlepiej opisujące dziedzinę zainteresowania użytkownika.

Drugim obszarem, w którym autor pracy widzi możliwości rozwoju w zastosowaniu zaproponowanego profilu użytkownika jest ocena dokumentów odpowiedzi i wskazywanie przez użytkownika dokumentów relewantnych w odpowiedzi. W przyjętym w pracy rozwiązaniu użytkownik przegląda poszczególne pozycje odpowiedzi – dokumenty i ocenia każdy z nich (nawet nie koniecznie go otwierając). Jednak w procesie analizy dokumentów odpowiedzi w celu wyznaczenia terminów znaczących, które zostają dołączone do profilu użytkownika można uwzględnić nie tylko fakt, że użytkownik ocenił dokument jako relewantny. Istotnych informacji o tym, czy pewien dokument jest interesujący dla użytkownika dostarcza również zachowanie użytkownika, a dokładnie operacje, jakie wykona z dokumentem. Jeśli użytkownik wydrukuje dokument to oznacza, że jest on bardziej istotny niż dokument, który zostanie tylko zapamiętany na dysku. Jako najmniej istotny można uważać dokument, z którym użytkownika tylko zapozna się podczas przeglądania odpowiedzi, ale nie zapamięta na dysku.

Kolejna propozycja rozszerzenia zauważona przez autora dotyczy metody wykorzystania profilu użytkownika. Istnieje możliwość zastosowania tezaurusa podobieństwa (ang. *similarity thesaurus*), zbudowanego dla kolekcji dokumentów, zaproponowanego w pracy Qiu (Qiu, 1996). Tezaurus ten zawiera wartości podobieństwa terminów należących do dokumentów kolekcji. Termin do zmodyfikowania pytania wyznaczany jest z tezaurusa na podstawie łącznego podobieństwa tego terminu do wszystkich terminów z pytania[1]. Jeżeli wartość tego podobieństwa przekroczy próg τ_{podob}, to termin jest dobrym terminem do dołączenia do nowego pytania. Inną możliwością jest dołączenie do pytania *k* terminów z profilu o najwyższych wartościach podobieństwa.

[1] Podobieństwo terminu do całego pytania określane jest na podstawie podobieństw do każdego z terminów pytania i przyjętej metody określania łącznego podobieństwa wynikowego.

7. Załącznik A – Zestawienie przeprowadzonych eksperymentów

7.1. Zestawienie przeprowadzonych eksperymentów

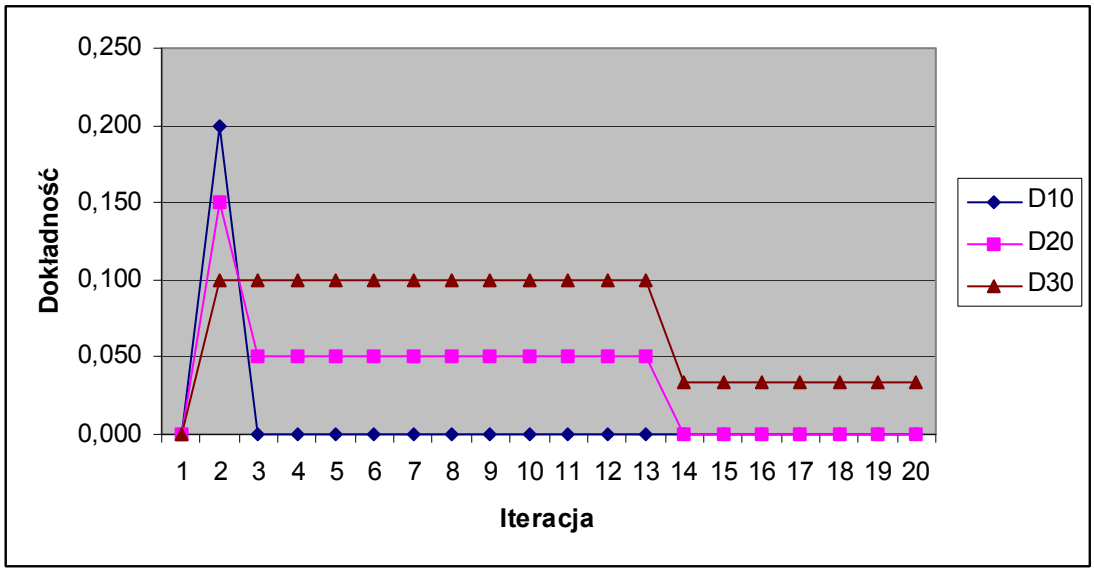
W rozdziale tym zamieszczono wyniki analizy procesu wyszukiwania z modyfikacją pytania. Zestawiono przykłady dla trzech rodzajów zbiorów testowych wykorzystywanych podczas eksperymentów — zbiorów dokumentów: gęstych, rzadkich i mieszanych.

Dla każdego przykładu zamieszczono wykres przedstawiający zmianę dokładności obciętych dla pierwszych 10, 20 i 30 dokumentów odpowiedzi w kolejnych iteracjach wyszukiwania. W każdej iteracji następowała kolejna modyfikacja pytania. Zamieszczono również wykresy przedstawiające procentowy wzrost dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania zadanego przez użytkownika. Trzeci wykres, który zamieszczono dla każdego przykładu pokazuje zależność pomiędzy wzrostem liczby dokumentów relewantnych D_q ,znajdujących się w odpowiedzi na kolejne zmodyfikowane pytanie, a spadkiem ogólnej liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

7.1.1. Przedstawienie szczegółów symulacji dla gęstych zbiorów dokumentów relewantnych — eksperymenty pozytywne

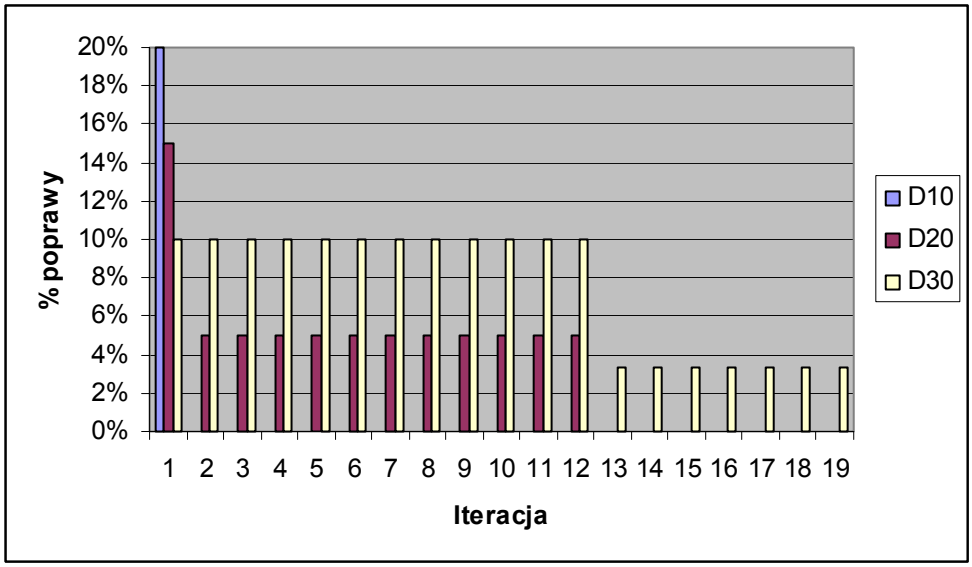
Pytanie q_{28}

1. Dane początkowe
   - zbiór dokumentów testowych: 6 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór: architektura $\wedge$ krajobraz $\wedge$ projekt,
   - pytanie wylosowane q_{28} = slaski $\wedge$ zmienic,
2. Zestawienie dokładności obciętej

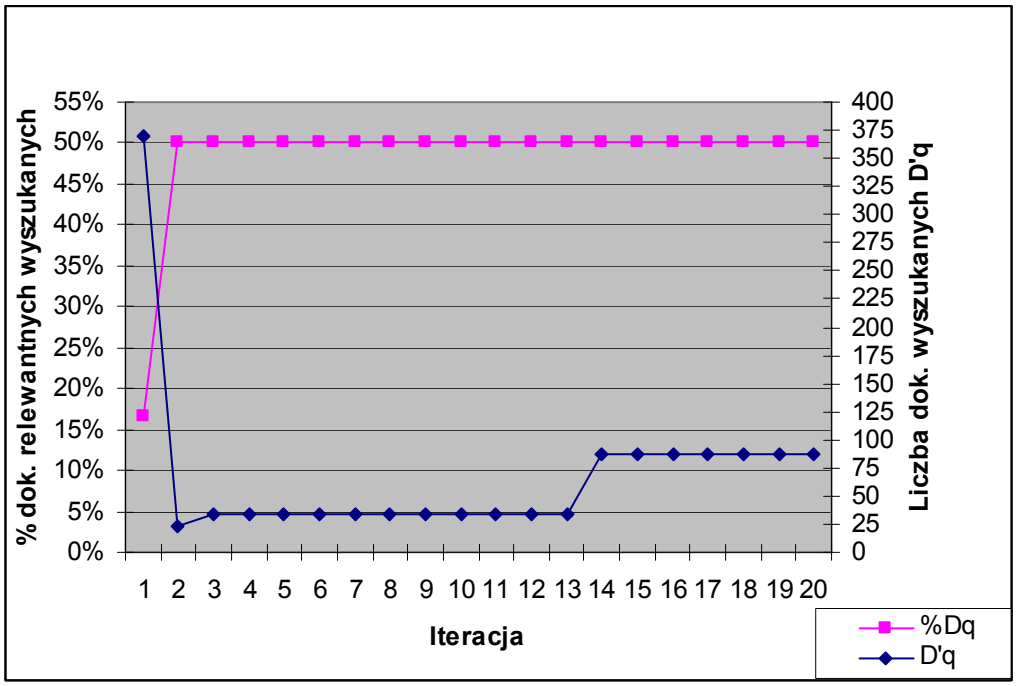


Rysunek 7-1: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{28}).

3. Poprawa efektywności wyszukiwania



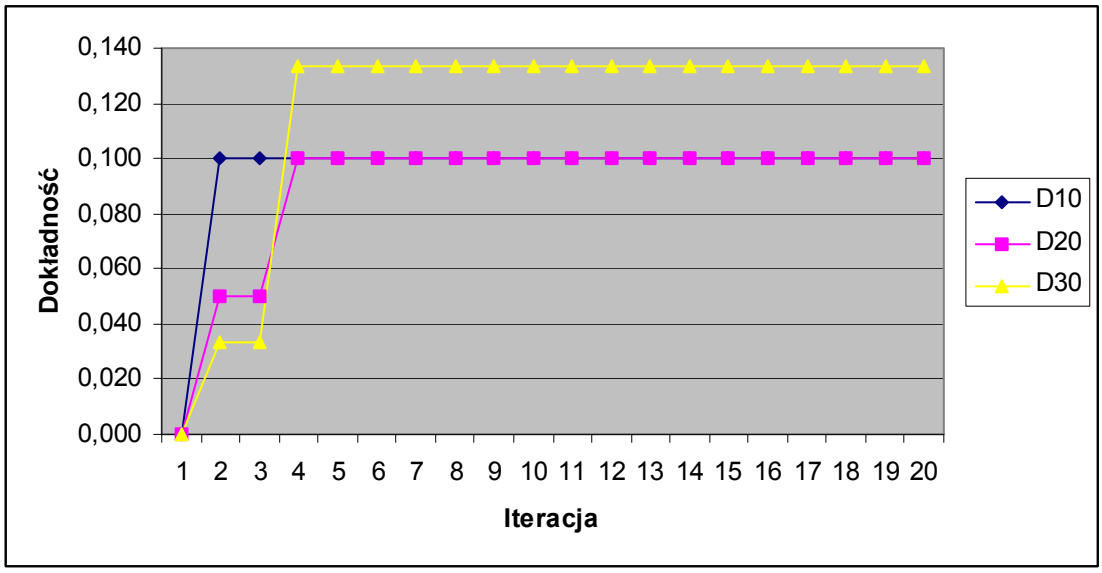
Rysunek 7-2: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{28}.



Rysunek 7-3: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{28} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

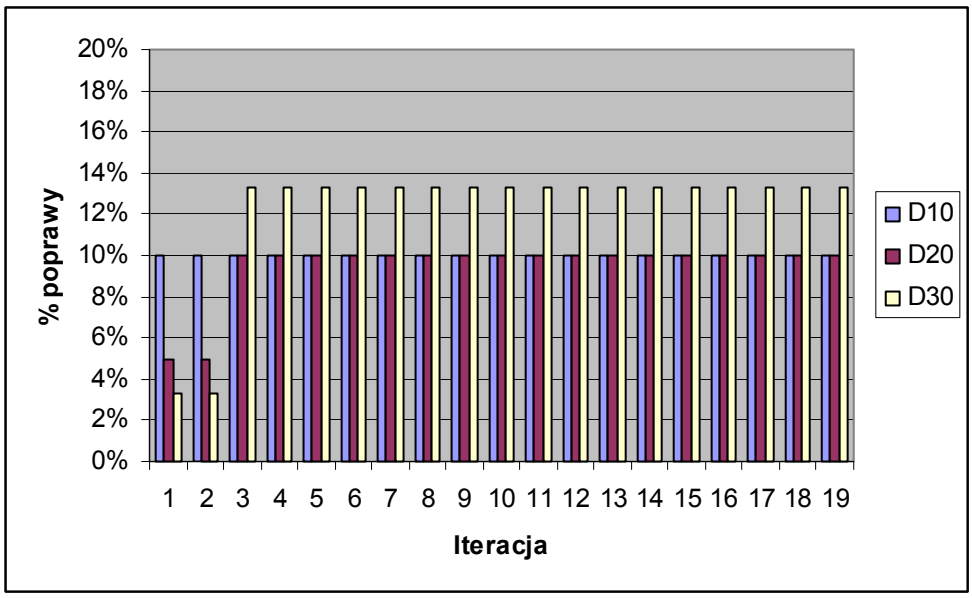
Pytanie q_{41}

1. Dane początkowe
   - zbiór dokumentów testowych: 6 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór: architektura $\wedge$ krajobraz $\wedge$ projekt,
   - pytanie wylosowane q_{41} = swojej,
2. Zestawienie dokładności obciętej

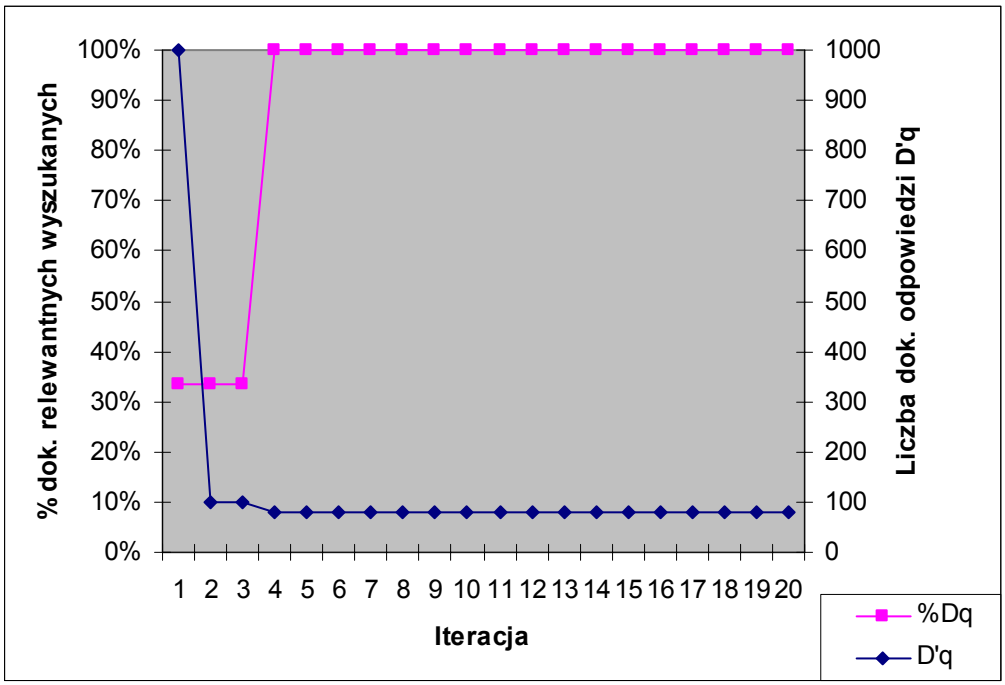


Rysunek 7-4: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{41}).

3. Poprawa efektywności wyszukiwania



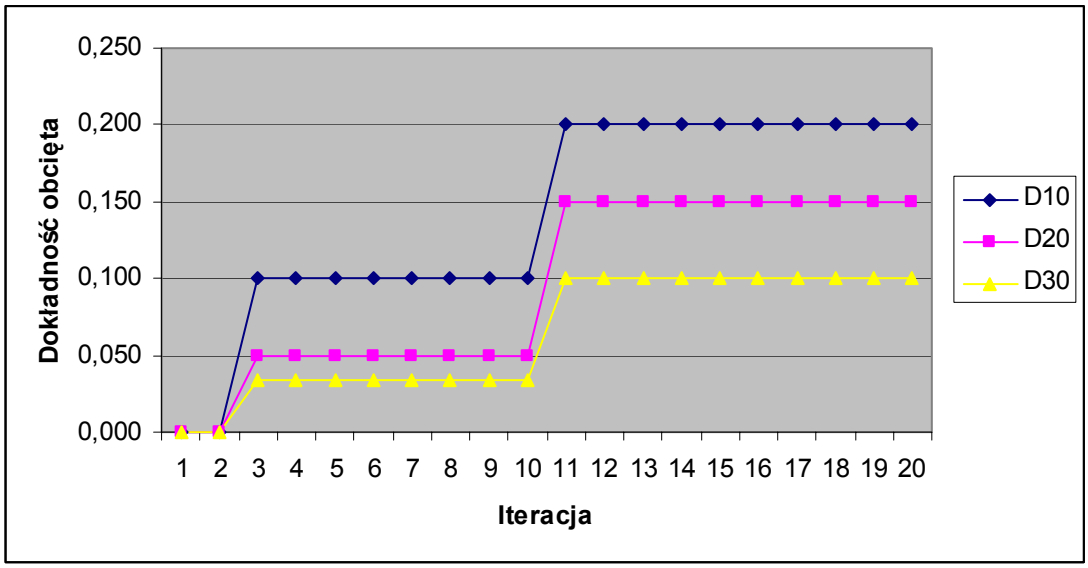
Rysunek 7-5: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{41}.



Rysunek 7-6: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{41} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

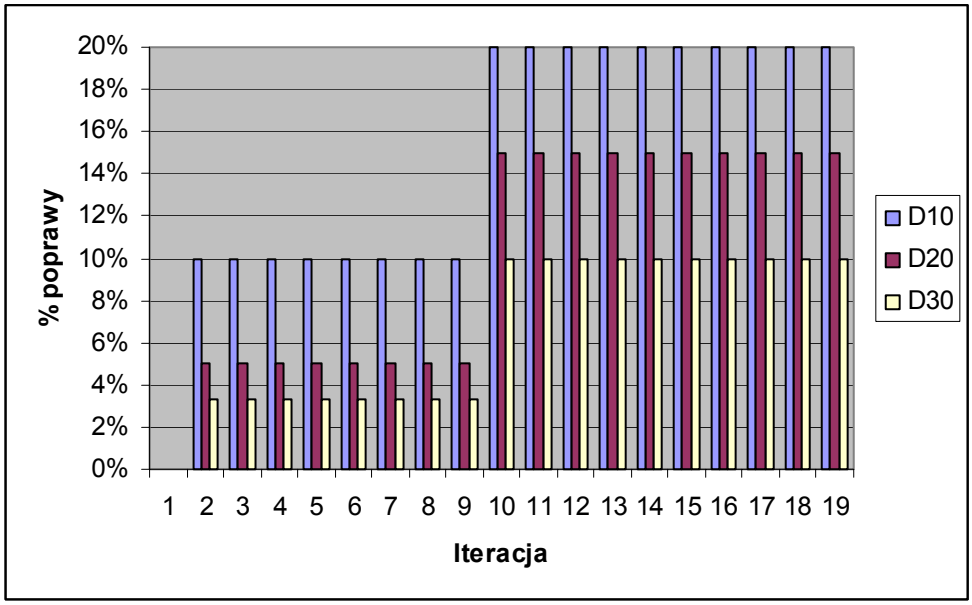
Pytanie q_{47}

1. Dane początkowe
   - zbiór dokumentów testowych: 3 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór: agroturystyka $\wedge$ pojezierze,
   - pytanie wylosowane q_{47} = natura,
2. Zestawienie dokładności obciętej

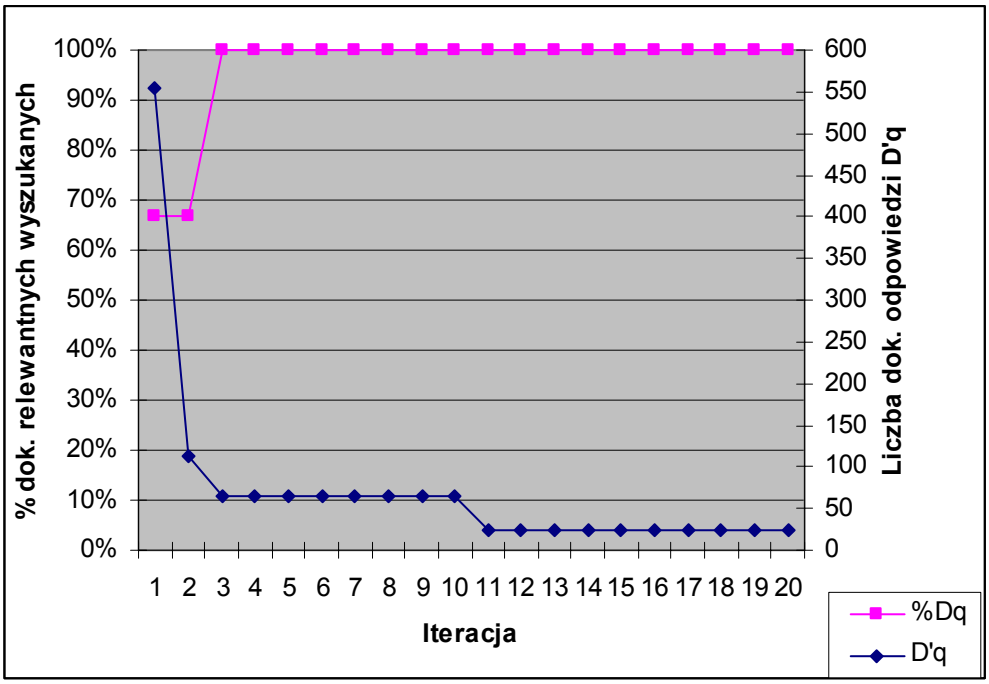


Rysunek 7-7: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{47}).

3. Poprawa efektywności wyszukiwania



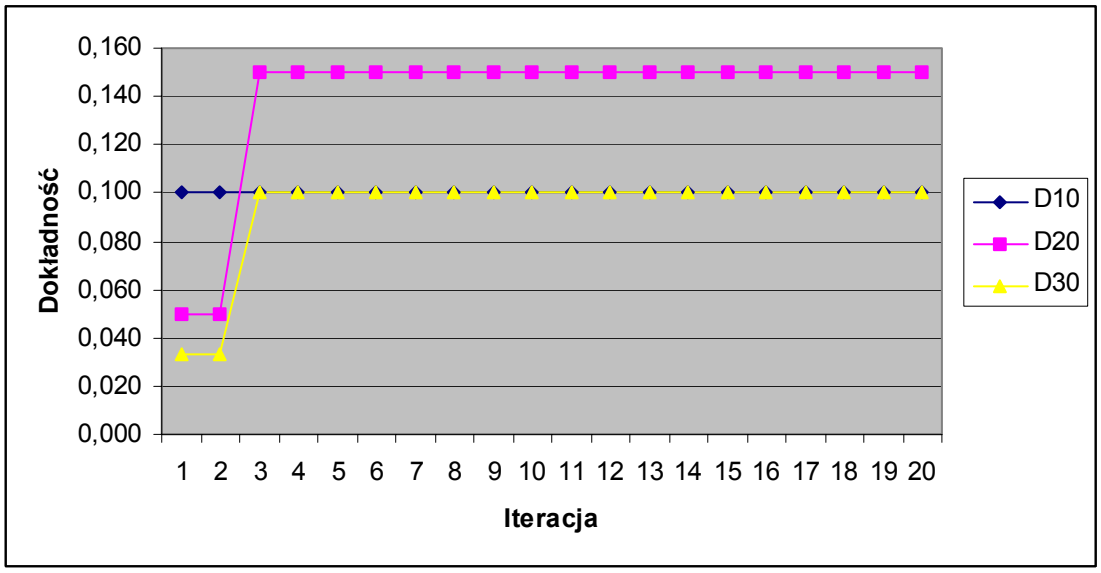
Rysunek 7-8: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{47}.



Rysunek 7-9: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{47} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

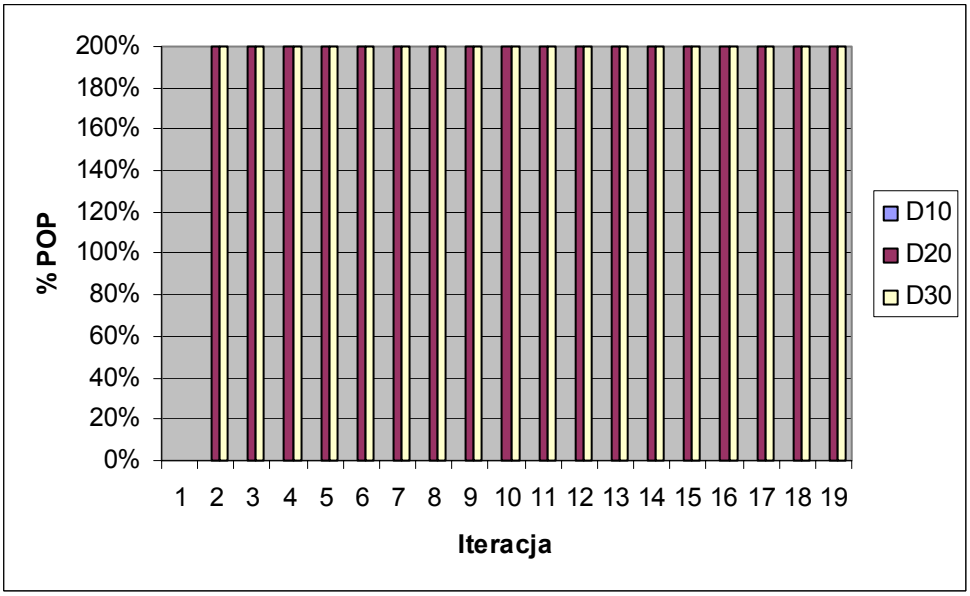
Pytanie q_{49}

1. Dane początkowe
   - zbiór dokumentów testowych: 3 dokumenty,
   - pytanie, na podstawie którego ustalono powyższy zbiór: agroturystyka ∧ pojezierze,
   - pytanie wylosowane q_{49} = lubelski ∧ wiesław,
2. Zestawienie dokładności obciętej

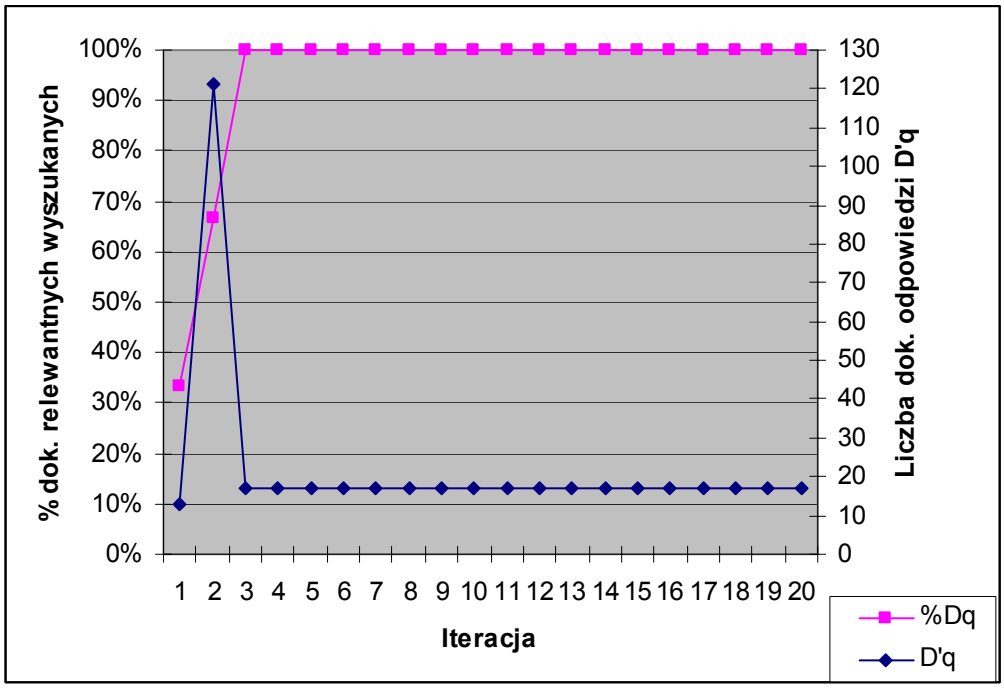


Rysunek 7-10: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{49}).

3. Poprawa efektywności wyszukiwania



Rysunek 7-11: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{49}.

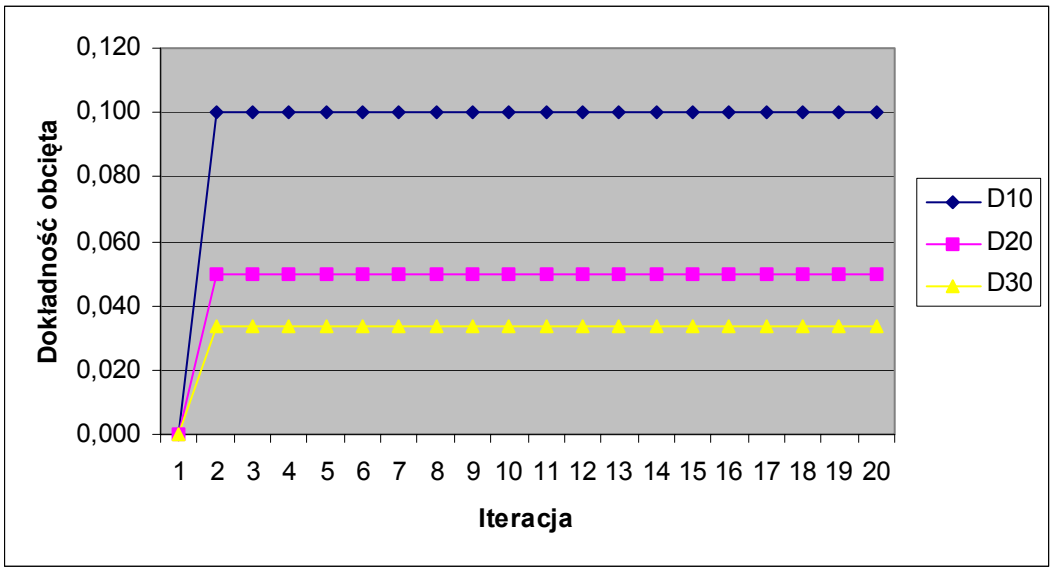


Rysunek 7-12: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{49} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

7.1.2. Przedstawienie szczegółów symulacji dla rzadkiego zbiorów dokumentów — eksperymenty negatywne

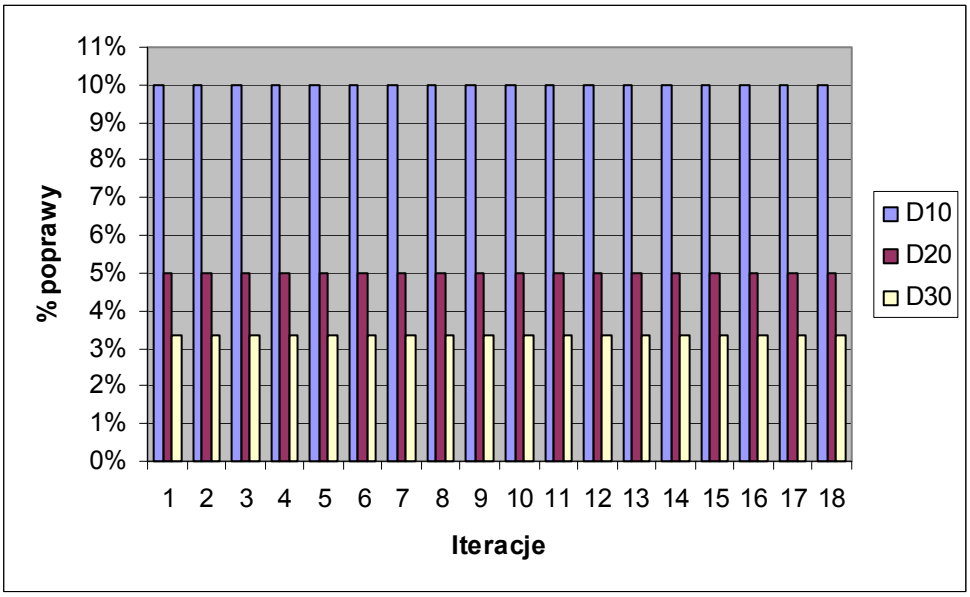
Pytanie q_{53}

1. Dane początkowe
   - zbiór dokumentów testowych: 3 dokumenty,
   - pytanie, na podstawie którego ustalono powyższy zbiór - brak,
   - pytanie wylosowane q_{53} = instytucja $\wedge$ wrzesien,
2. Zestawienie dokładności obciętej.

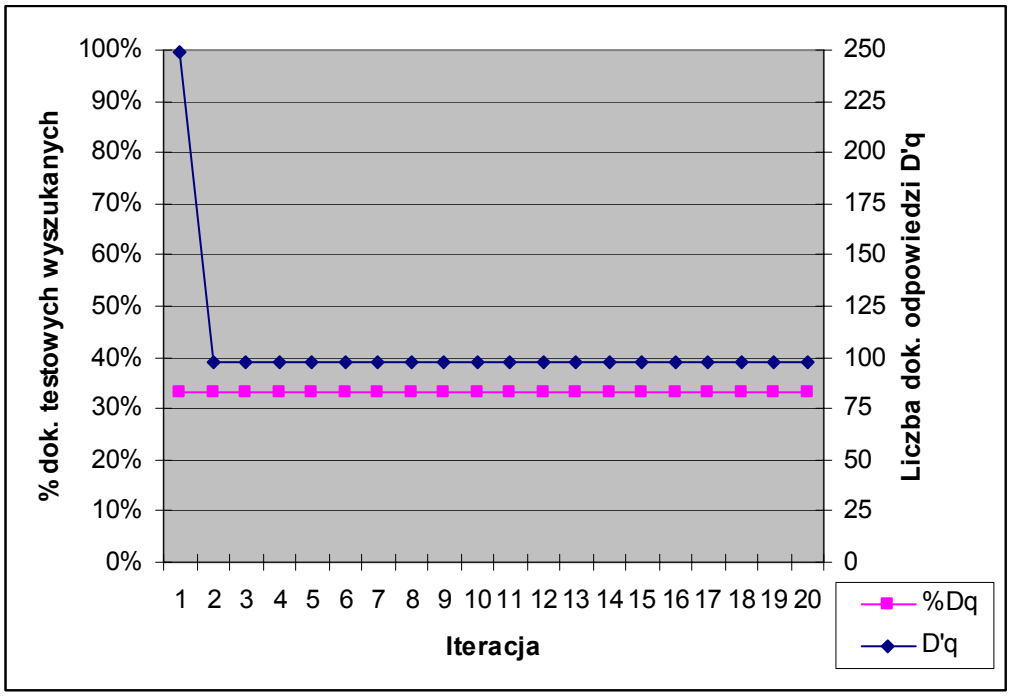


Rysunek 7-13: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{53}).

3. Poprawa efektywności wyszukiwania



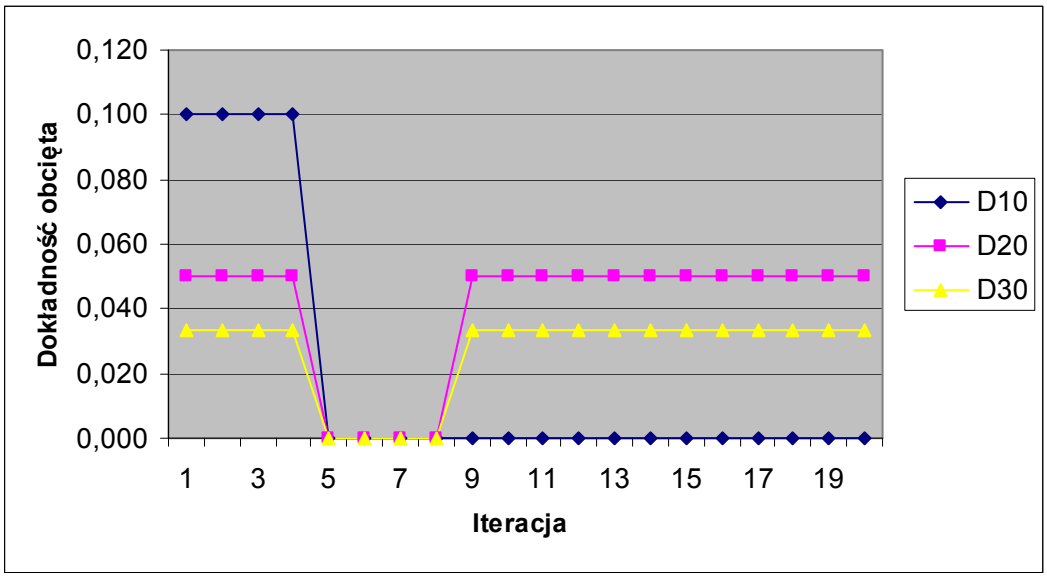
Rysunek 7-14: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{53}.



Rysunek 7-15: Zestawienie liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{53} oraz liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

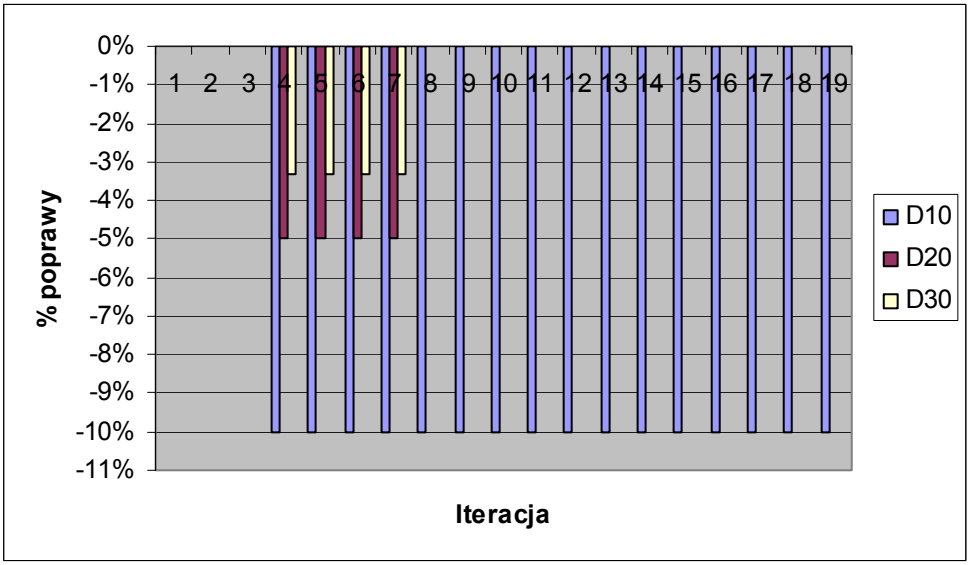
Pytanie q_{73}

1. Dane początkowe
   - zbiór dokumentów testowych: 5 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór - brak,
   - pytanie wylosowane q_{73} = cela,
2. Zestawienie dokładności obciętej.

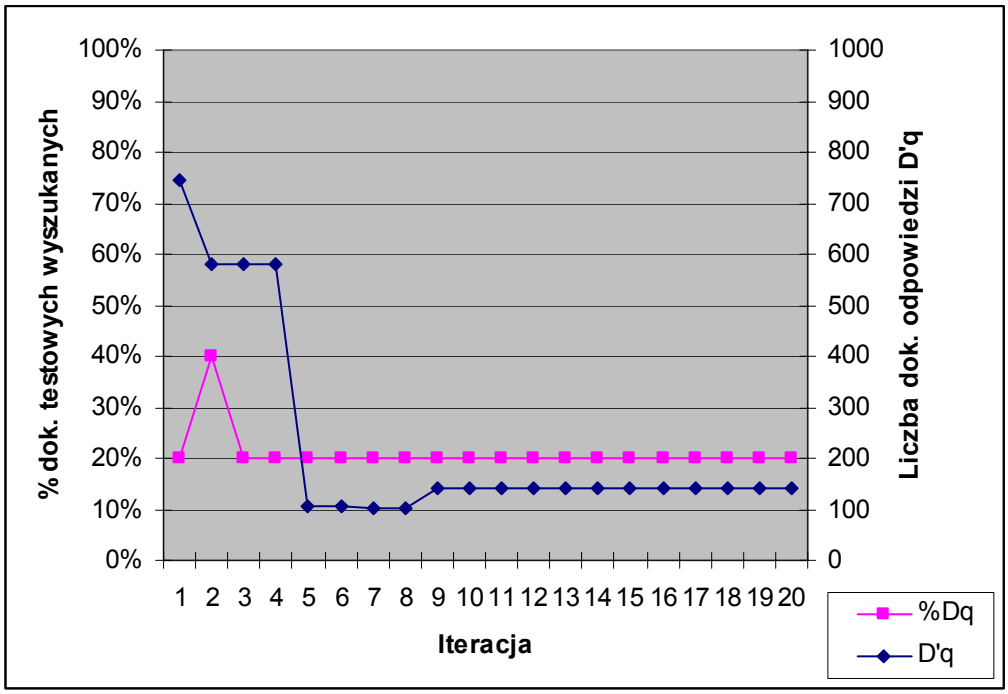


Rysunek 7-16: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{73}).

3. Poprawa efektywności wyszukiwania



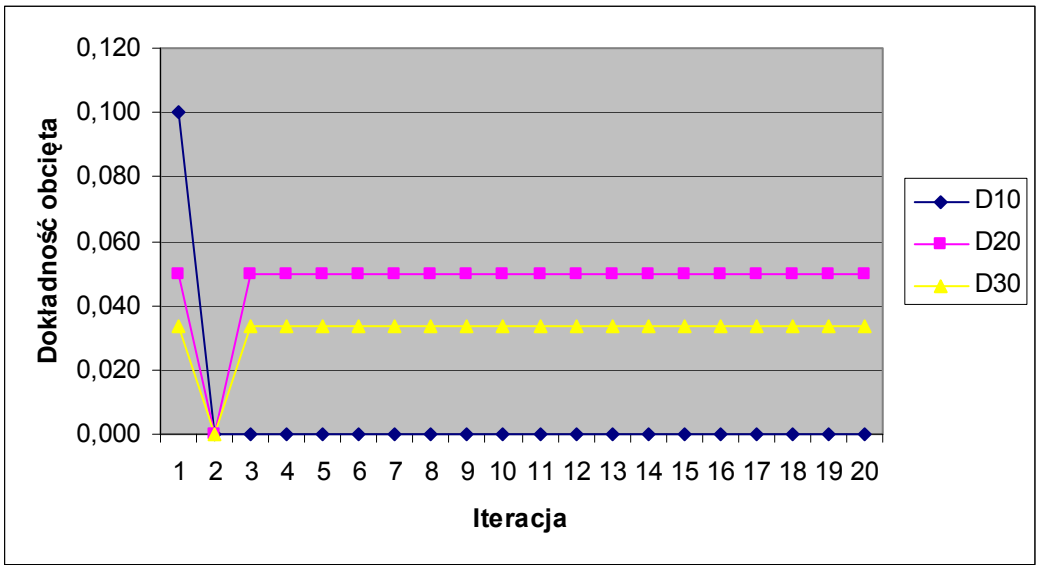
Rysunek 7-17: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{73}.



Rysunek 7-18: Zestawienie liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{73} oraz liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

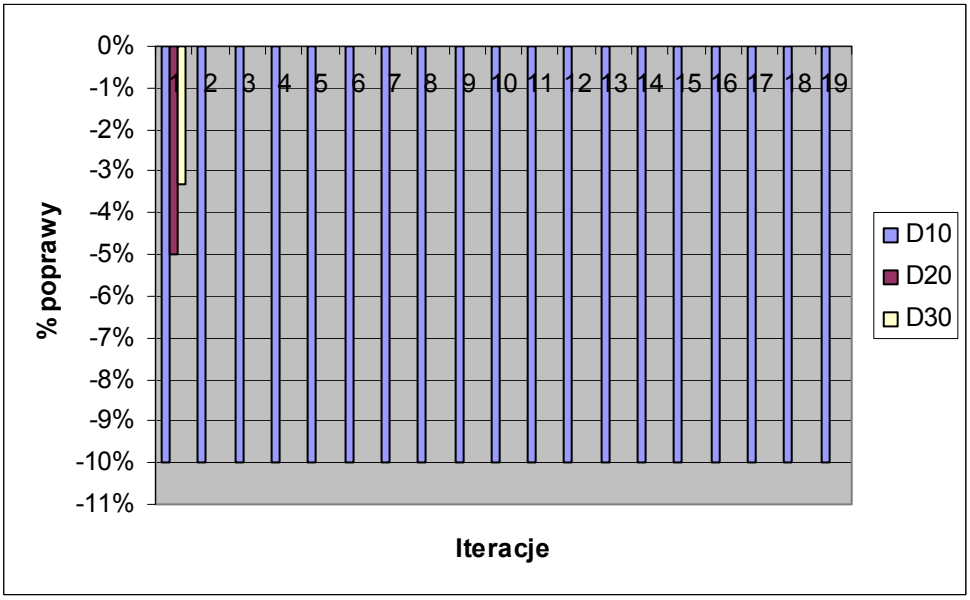
Pytanie q_{74}

1. Dane początkowe
   - zbiór dokumentów testowych: 5 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór - brak,
   - pytanie wylosowane q_{74} = kredyt,
2. Zestawienie dokładności obciętej.

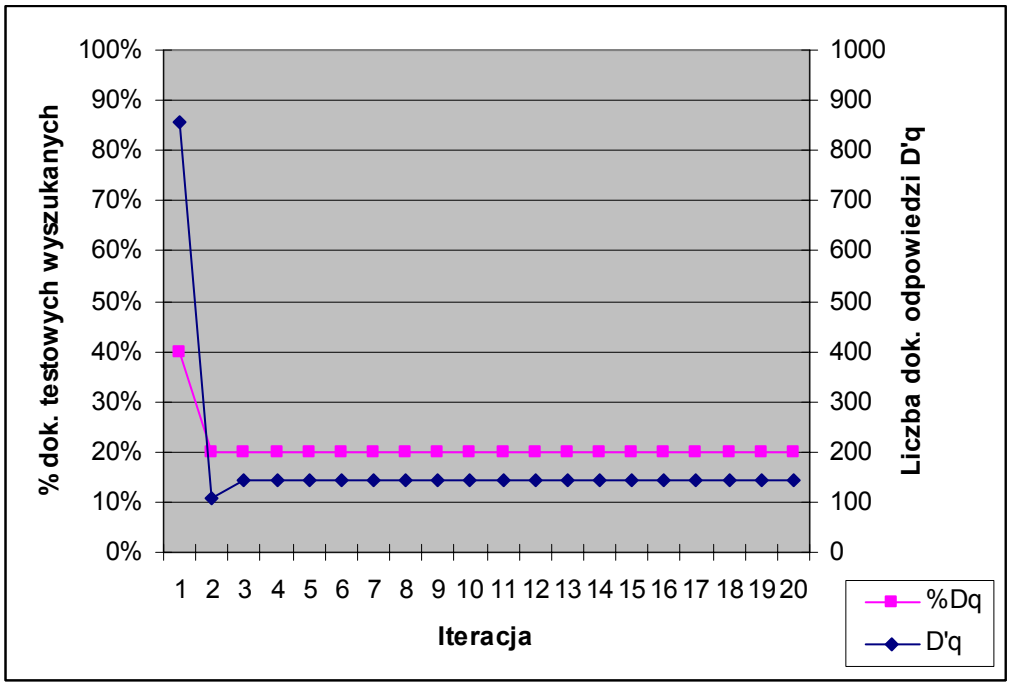


Rysunek 7-19: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{74}).

3. Poprawa efektywności wyszukiwania



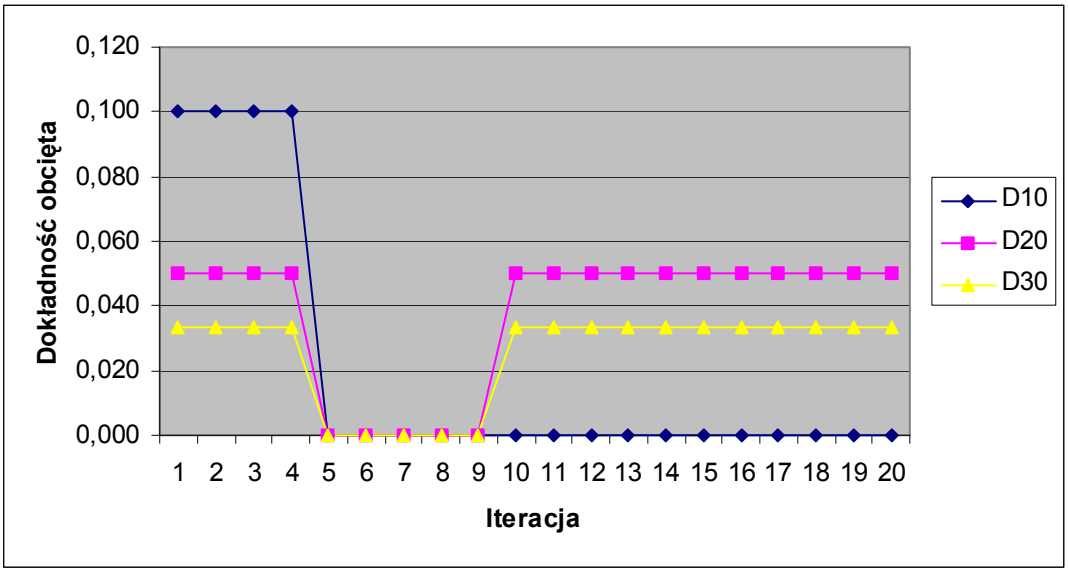
Rysunek 7-20: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{74}.



Rysunek 7-21: Zestawienie liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{74} oraz liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

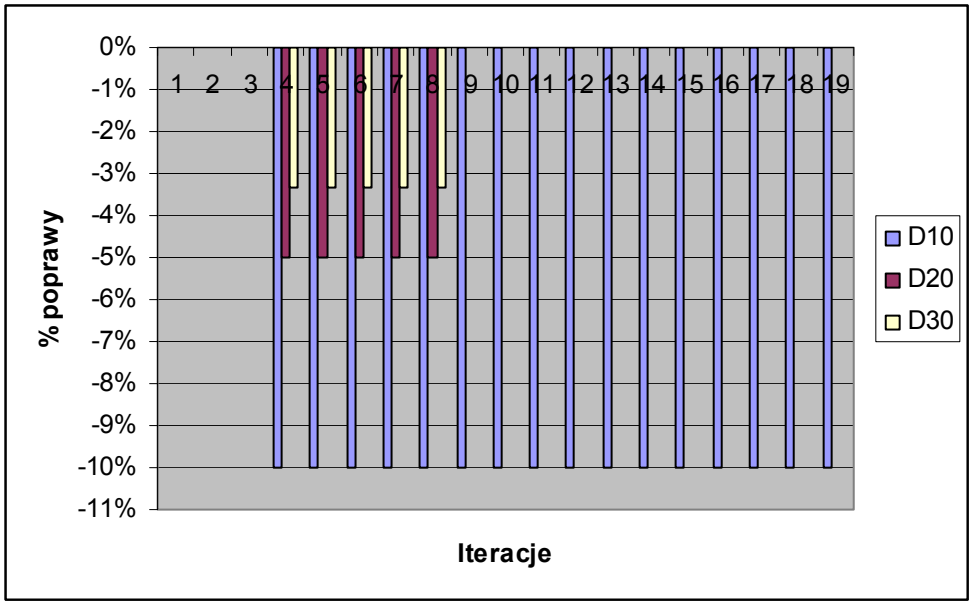
Pytanie q_{75}

1. Dane początkowe
   - zbiór dokumentów testowych: 5 dokumentów,
   - pytanie, na podstawie którego ustalono powyższy zbiór - brak,
   - pytanie wylosowane q_{75} = platny,
2. Zestawienie dokładności obciętej.

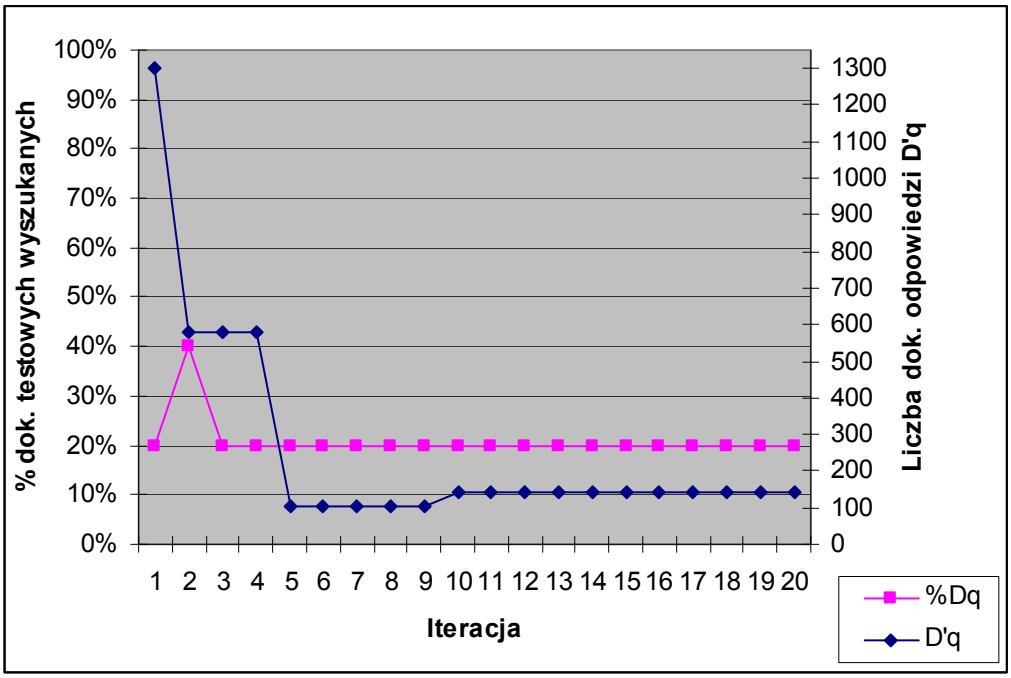


Rysunek 7-22: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{75}).

3. Poprawa efektywności wyszukiwania



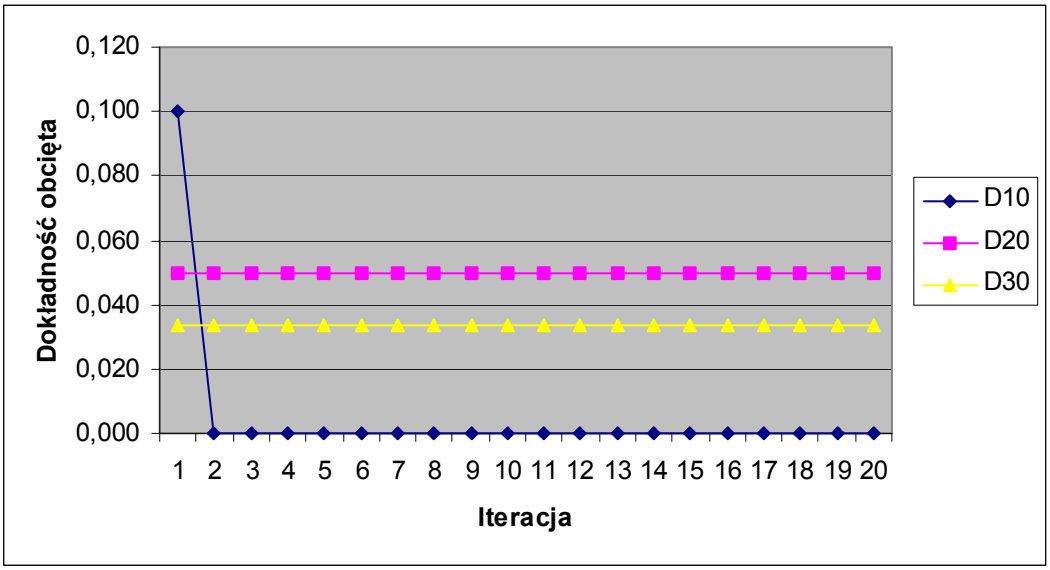
Rysunek 7-23: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{75}.



Rysunek 7-24: Zestawienie liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{75} oraz liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

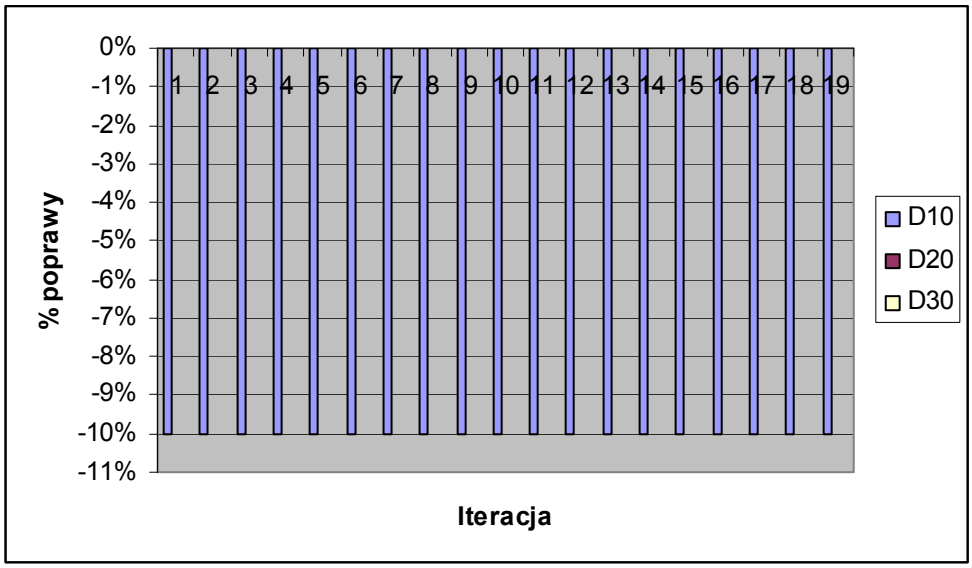
Pytanie q_{80}

1. Dane początkowe
   - zbiór dokumentów testowych: 3 dokumenty,
   - pytanie, na podstawie którego ustalono powyższy zbiór - brak,
   - pytanie wylosowane q_{80} = wroclawskiej $\wedge$ naukowy,
2. Zestawienie dokładności obciętej.

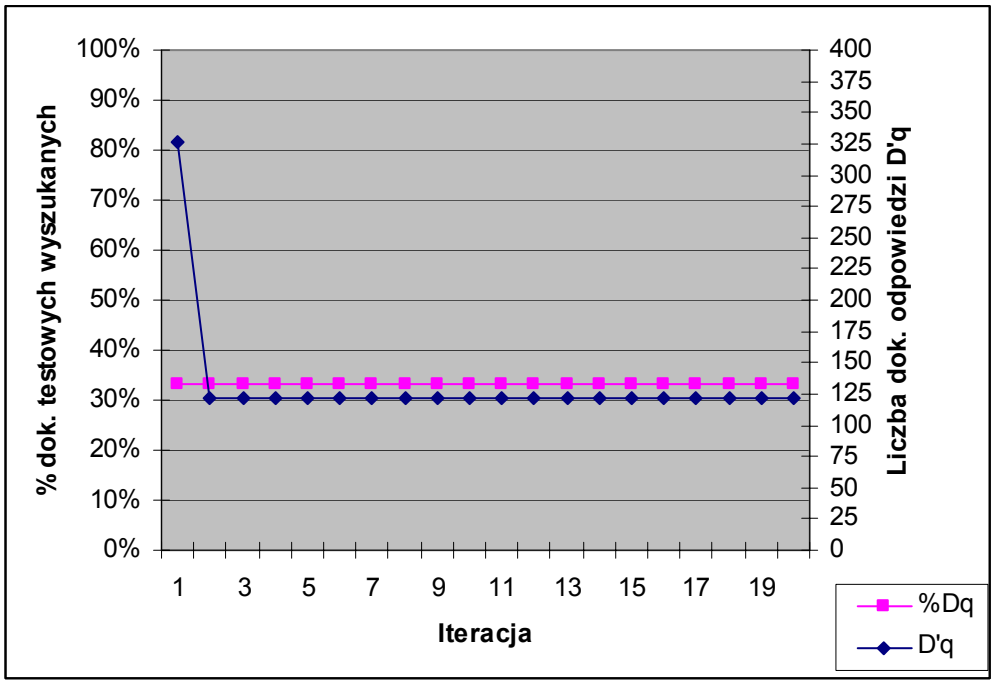


Rysunek 7-25: 13Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{80}).

3. Poprawa efektywności wyszukiwania



Rysunek 7-26: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{80}.

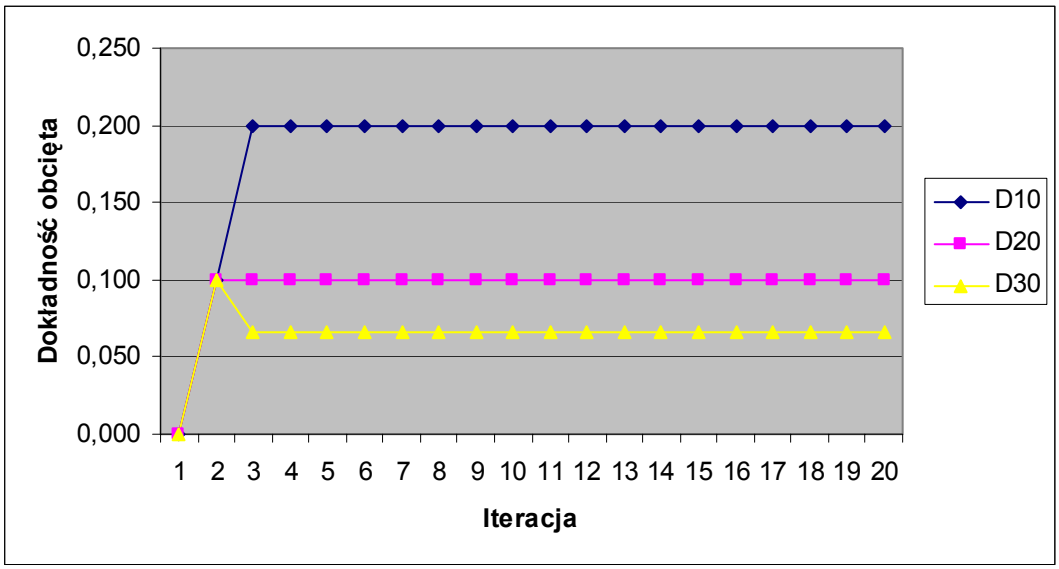


Rysunek 7-27: Zestawienie liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{80} oraz liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

7.1.3. Przedstawienie szczegółów symulacji dla zbioru testowego zawierającego zbiory gęstych dokumentów relewantnych oraz zbiory dokumentów rzadkich — eksperymenty mieszane

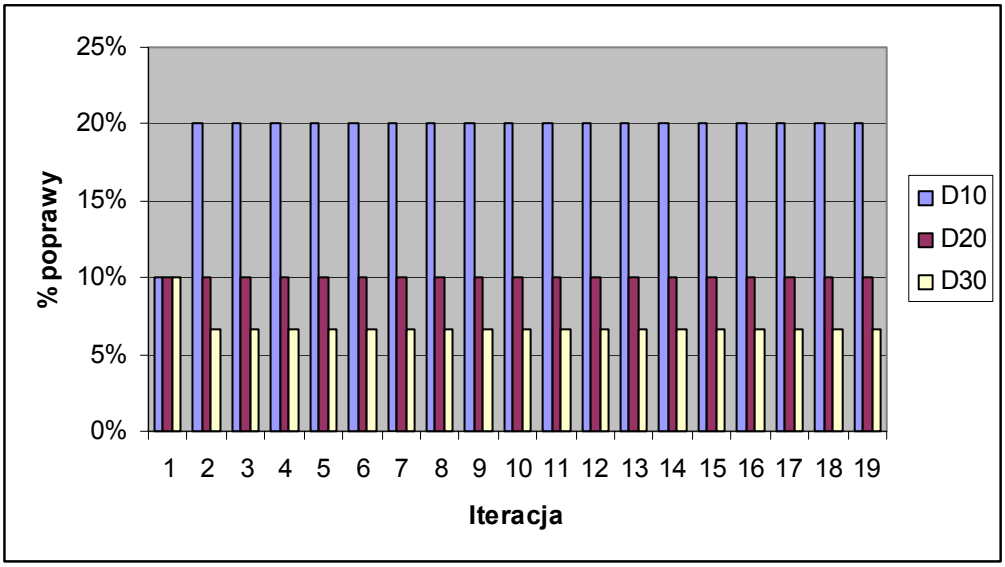
Pytanie q_{89}

1. Dane początkowe
   - zbiór dokumentów testowych: 7 dokumentów tworzących zbiór dokumentów gęstych oraz 3 dokumenty tworzące zbiory dokumentów rzadkich,
   - pytanie, na podstawie którego ustalono powyższy zbiór: dla zbioru dokumentów gęstych — mecz koszykowka sport Wrocław, dla zbioru dokumentów rzadkich — brak,
   - pytanie wylosowane q_{89} = dwie,
2. Zestawienie dokładności obciętej

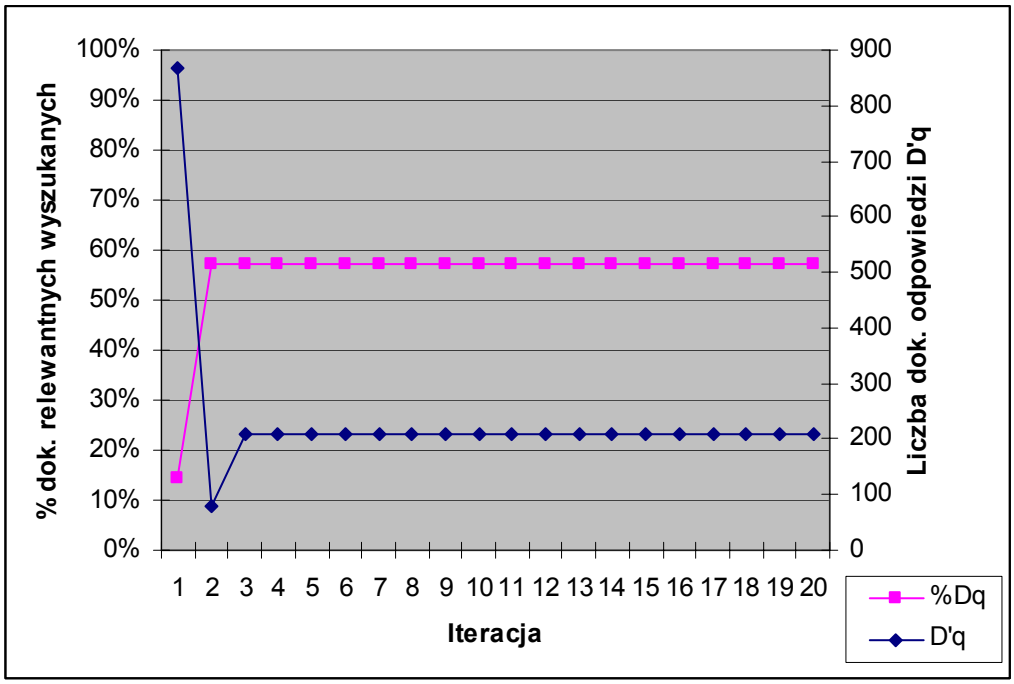


Rysunek 7-28: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{89}).

3. Poprawa efektywności wyszukiwania



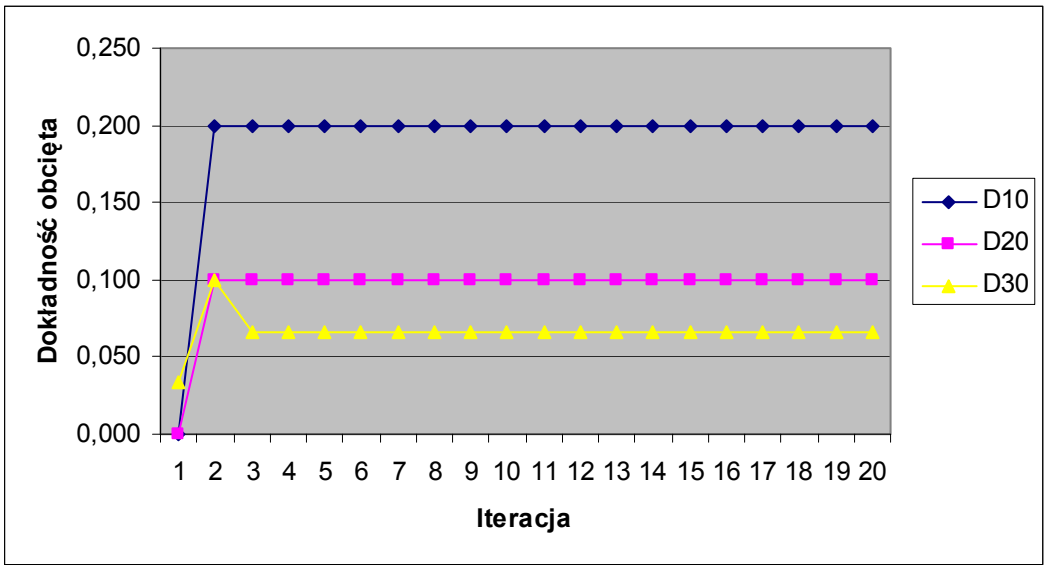
Rysunek 7-29: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{89}.



Rysunek 7-30: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{89} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

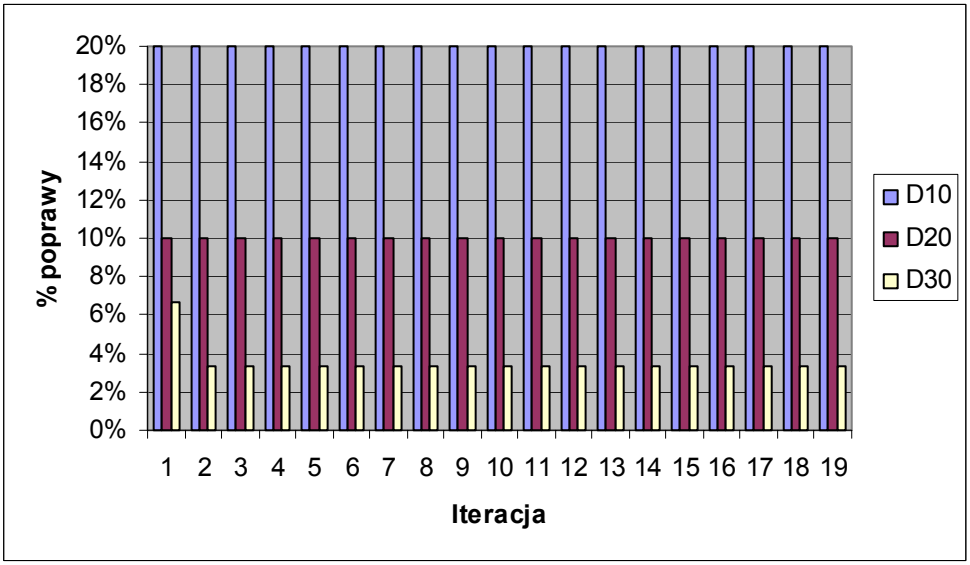
Pytanie q_{93}

1. Dane początkowe
   - zbiór dokumentów testowych: : 4 dokumenty tworzące zbiór dokumentów gęstych oraz 3 dokumenty tworzące zbiory dokumentów rzadkich,
   - pytanie, na podstawie którego ustalono powyższy zbiór: dla zbioru dokumentów gęstych — kino film polski repertuar, dla zbioru dokumentów rzadkich — brak,
   - pytanie wylosowane q_{93} = rola,
2. Zestawienie dokładności obciętej

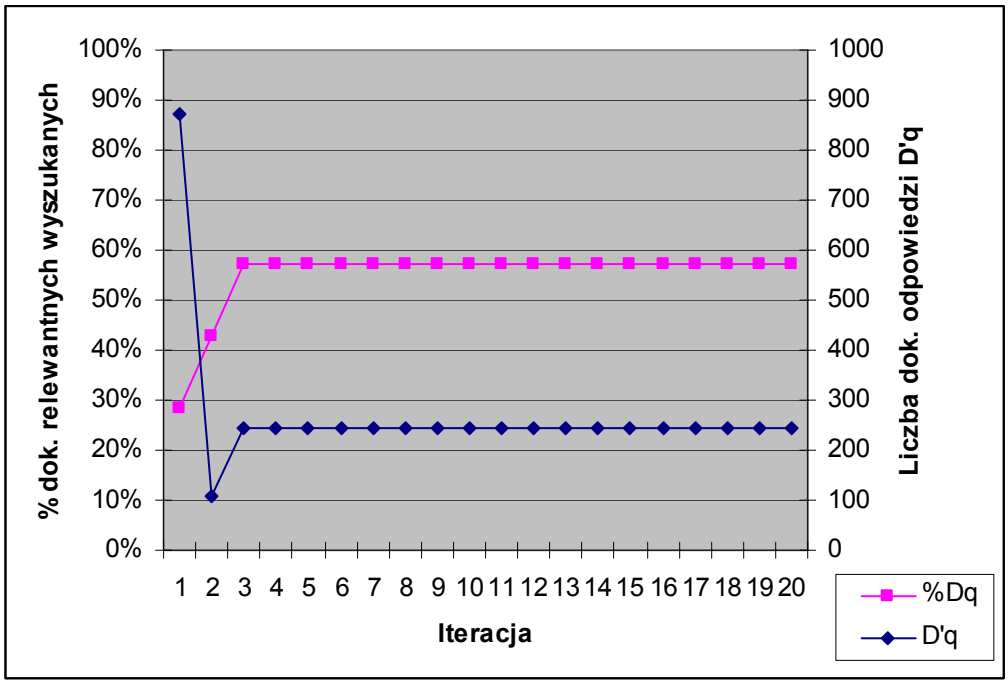


Rysunek 7-31: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q_{93}).

3. Poprawa efektywności wyszukiwania



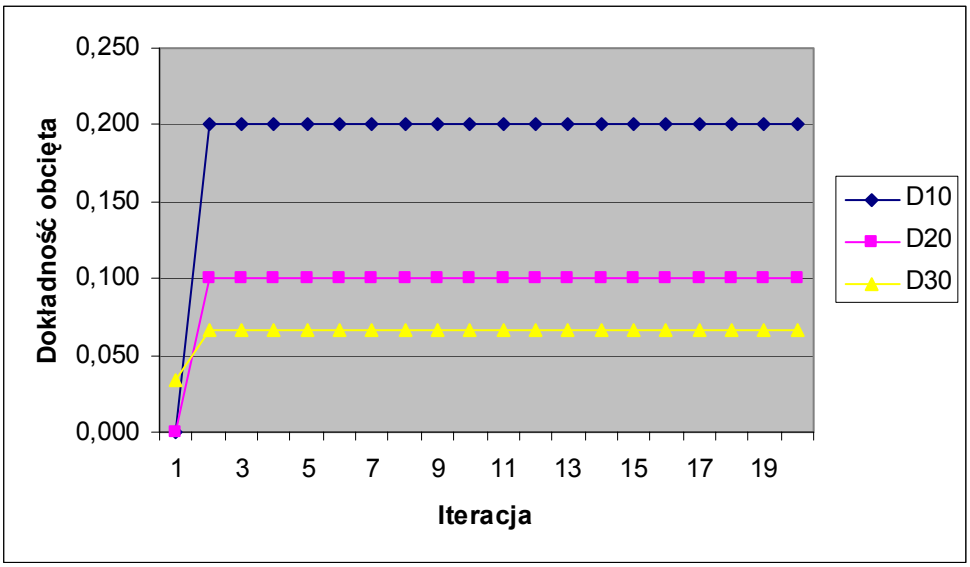
Rysunek 7-32: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{93}.



Rysunek 7-33: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{93} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

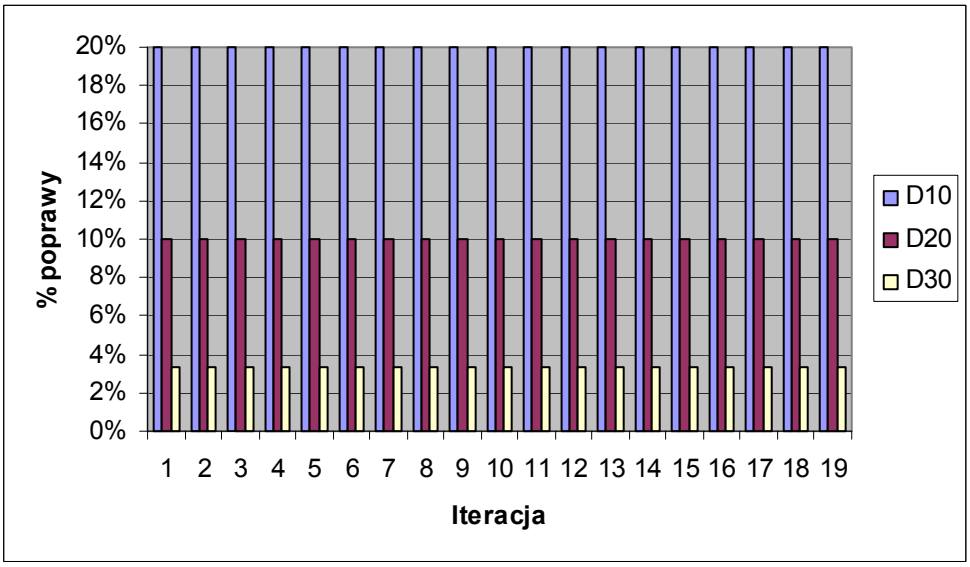
Pytanie q_{98}

1. Dane początkowe
   - zbiór dokumentów testowych: 4 dokumenty tworzące zbiór dokumentów gęstych oraz 3 dokumenty tworzące zbiory dokumentów rzadkich,
   - pytanie, na podstawie którego ustalono powyższy zbiór: dla zbioru dokumentów gęstych — kino film polski repertuar, dla zbioru dokumentów rzadkich — brak,
   - pytanie wylosowane q_{98} = nagroda,
2. Zestawienie dokładności obciętej

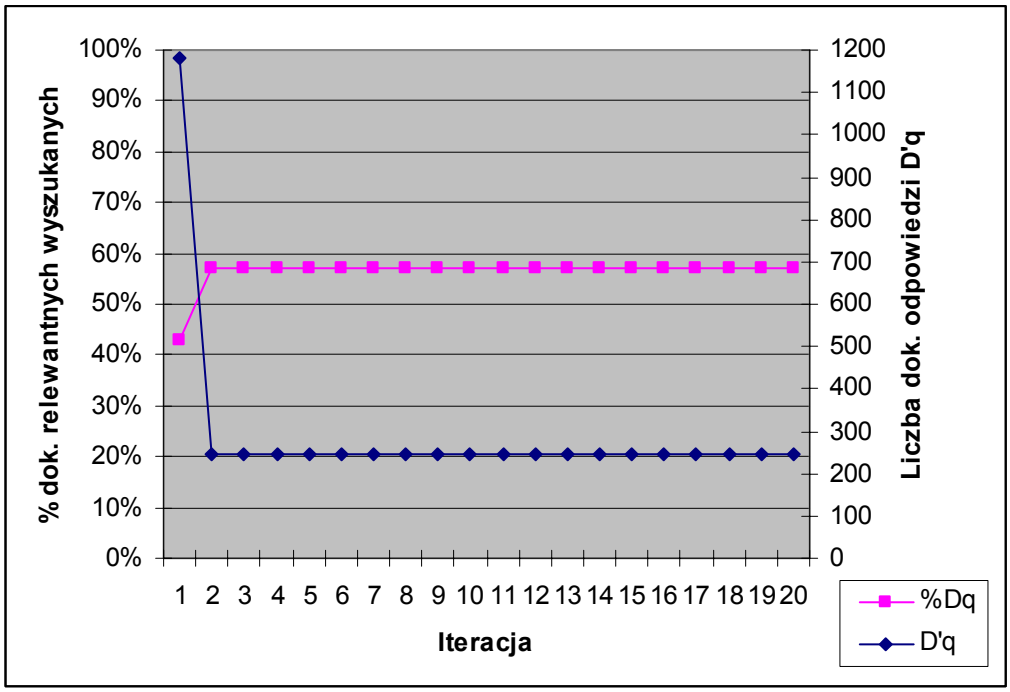


Rysunek 7-34: Dokładność obcięta dla odpowiedzi na kolejne zmodyfikowane pytania (pytanie początkowe q98).

3. Poprawa efektywności wyszukiwania



Rysunek 7-35: Procent wzrostu dokładności wyszukiwania w odpowiedzi na pytania zmodyfikowane w kolejnych iteracjach w stosunku do początkowego pytania q_{98}.



Rysunek 7-36: Zestawienie wzrostu liczby dokumentów relewantnych D_q wśród dokumentów odpowiedzi D'_q na kolejne zmodyfikowane pytanie dla pytania początkowego q_{98} oraz zmniejszenia liczby dokumentów odpowiedzi D'_q w kolejnych iteracjach wyszukiwania z kolejnym zmodyfikowanym pytaniem.

8. Bibliografia

Aas K. (1997): A Survey on Personalised Information Filtering Systems for the World Wide Web, raport techniczny nr 922**,** ISBN 82-539-0442-8 (http://www.nr.no/home/kjersti/SURVEY.ps).

Akoulchina I., Ganascia J. (1997): SATELIT–Agent: An Adaptive Interface Based on Learning Agent Technology, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 21–33, Springer Wien New York.

Albrecht D.W., Zukerman I., Nicholson A.E., Bud A. (1997): Towards a Bayesian Model for Keyhole Plan Recognition in Large Domain, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 21–33, Springer Wien New York.

Ambrosini L., Cirillo V., Micarelli A. (1997): A Hybrid Architecture for User–Adapted Information Filtering on the World Wide Web, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 59–62, Springer Wien New York.

Asnicar F.A., Tasso C., (1997): ifWeb: a Prototype of User Model–Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web, *Proceedings of the workshop "Adaptive System and User Modeling on the World Wide Web" Six International Conference on User Modeling,* Chia Laguna, Sardinia.

Attar R., Fraenkel A. S. (1977): Local Feedback in Full–text Retrieval Systems, *Journal of the ACM*, v. 24, n. 3, str. 397–417.

Baeza-Yates R., Ribeiro-Neto B. (1999), *Modern Information Retrieval,* ACM Press, Adison-Wesley, New York.

Barrett R., Maglio P.P., Kellem D.C. (1997): WBI: A Confederation of Agents that Personalize the Web, *International Conference on Autonomous Agents*, Marina Del Rey, California USA, str. 496–499.

Belkin N.J., Croft W.B. (1992), Information Filtering and Information Retrieval: Two Sides of the Same Coin, *Communications of the ACM*, 35(12), str. 29–38.

Belkin, N. J., Cool, C., Head, J., Jeng, J., Kelly, D., Lin, S., Park, S. Y., Savage-Knepshield, P., Sikora, C. (2000): Relevance feedback versus local context analysis as termsuggestion devices: Rutgers' TREC-8 interactive track experience, *TREC-8, Proceedings of the 8th Text Retrieval Conference* (pod edycją D. Harman, E. Voorhees), Washington, str. 565-574.

Benaki E., Karkaletsis V.A., Spyroupoulus C.D. (1997): Integrating User Modeling Into Information Extraction: the UMIE Prototype, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 55–59, Springer Wien New York.

Benaki E., Karkaletsis V.A., Spyroupoulus C.D. (1997a): User Modeling in WWW: the UMIE Prototype, *Proc. of the Workshop "Adaptive Systems and User Modeling on the World Wide Web", 6th International Conference on User Modeling*, *UM'97*.

Bianchi–Berthouze N., Berthouze L., Kato T.(1997): Understanding Subjectivity: An Interactionist View, *Proc. of the 7th International Conference on User Modeling*, *UM'99*, Banff, Canada, str. 3–12.

Billsus D., Pazzani M. (1999): A Hybrid User Model for News Story Classification, *Proc. of the 7th International Conference on User Modeling*, *UM'99*, Banff, Canada, str. 99–108.

Billsus D., Brunk C., Evans C., Gladish B., Pazzani M. (2002): Adaptive interfaces for ubiquitous web, *Communications of the ACM*, vol. 45, n. 5, str. 34-38.

Brewington B.E., Cybenko G. (2000), How Dynamic is the Web?, *Proc. of 9th International WWW Conference*, Amsterdam, Netherlands.

Brin S., Page L. (1998), Anatomy of a Large–Scale Hypertextual Web Search Engine, *Proc. of 7th Int. WWW Conference*, Brisbarne, Australia, str. 107-117 (http://www–db.stanford.edu/ ~backrub/google. html).

Brusilowski P., Schwarz E. (1997): User as Student: Towards an Adaptive Interface for Advanced Web–Based Applications, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 177–188, Springer Wien New York.

Buckley C., Salton G., Allan J. (1994): The Effect of Adding Relevance Information in a Relevance Feedback Environment, *SIGIR'94, Proceedings of the Seventeenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, Springer–Verlag, str. 292–300.

Bull S. (1997): See Yourself Write: A Simple Student Model to Make Students Think, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 315–326, Springer Wien New York.

Bull S., Smith M. (1997): A Pair of Student Models to Encourage Collaboration, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 339–342, Springer Wien New York.

Carberry S. (2001): Techniques for Plan Recognition, *User Modeling and User Adapted Interaction*, v. 11, n. 1–2, str. 31–48.

Callan J.P., Croft W.B. (1993): An Evaluation of Query Processing Strategies Using the TIPSTER Collection, *SIGIR'93, Proceedings of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA USA, ACM Press, June 27 – July 1, str. 347–355.

Chang Y.K., Cirillo C., Razon K. (1971): Ocena wyszukiwania ze sprzężeniem zwrotnym przy użyciu zmodyfikowanego zamrażania, zbioru pozostałości oraz grup testowych, w Salton G. (pod redakcją) *SMART – Automatyczny system wyszukiwania informacji*, PWN, str. 374–389.

Choroś K. (2002): Efektywność wyszukiwarek internetowych, *III Krajowa Konferencja Multimedialne i Sieciowe Systemy Informacyjne MISSI'02*, Oficyna Wydawnicza PWr., Wrocław 2002, str. 115–123.

Collins J.A., Greer J.E., Kumar V.S., McCalla G.I., Meagher P., Tkatch R. (1997): Inspectable User Models for Just–In Time Workplace Training, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 327–338, Springer Wien New York.

Conati C., Gertner A.S., VanLehn K., Druzdzel M. (1997): On–Line Student Modeling for Coached Problem Solving Using Bayesian Networks, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 231–242, Springer Wien New York.

Corbett A.T., Bhatnagar A. (1997): Student Modeling in the ACT Programming Tutor: Adjusting a Procedural Learning Model with Declarative Knowledge, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 243–254, Springer Wien New York.

Croft W.B., Townsend S.C., Lavrenko V. (2001): Relevance feedback and personalization: A language modeling perspective, *Proc. of the DELOS-NSF Worshop on Personalization and Recommender Systems in Digital Libraries*, str. 49–54.

Crouch C. J., (1990): An Approach to the Automatic Construction of Global Thesauri, *Information Processing and Management*, 26 (5), str. 629–640.

Crouch C. J., Yang B., (1992): Experiments in Automatic Statistical Thesaurus Construction. *Proceedings of the Fifteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Copenhagen, Denmark, str. 77–87.

Daniłowicz Cz. (1982): SDI Systems at the Technical University of Wrocław, *Journal of Information Science*, vol. 5, str. 55–61.

Daniłowicz Cz. (1994), Modelling of user preferences and needs in Boolean retrieval systems, *Information Processing and Management*, 30(5), str. 363–378.

Daniłowicz Cz., (1998): Reprezentacja preferencji użytkownika końcowego w modelach informacyjnych agentów. *I Krajowa Konferencja Multimedialne i Sieciowe Systemy Informacyjne MISSI'98*, Oficyna Wydawnicza PWr., Wrocław, str. 167–172.

Daniłowicz Cz., (1999): Rozpoznawanie zainteresowań i preferencji użytkowników w otwartych systemach informacyjnych, *Materiały I Konferencji Komputerowe Systemy Rozpoznawania KOSYR*, Oficyna Wydawnicza PWr., Wrocław, str. 161–166.

Daniłowicz Cz. (2000), Możliwości i problemy wyszukiwania informacji w otwartym systemie WWW, *Informatyka*, nr 1.

Davies N.J., Weeks R., Revett M.C. (1997): Information Agents for the World Wide Web, *Software Agents and Soft Computing, Towards Enhancing Machine Intelligence, Concepts and Application* pod edycją H.S. Nwana i N. Azarmi, Springer–Verlag Berlin Heidelberg New York, str. 81–99.

Dąbrowski M., Laus_Mączyńska K. (1978): *Metody wyszukiwania i klasyfikacji informacji*, Wydawnictwa Naukowo–Techniczne, Warszawa.

De Carolis B., Pizzutilo S. (1997): From Discourse Plans to User–Adapted Hypermedia, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 37–40, Springer Wien New York.

Doux A.C., Laurent J.P. Nadal J.P. (1997): Symbolic Data Analysis With the K–Means Algorithm for User Profiling, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 359–362, Springer Wien New York.

Efthimiadis E.N. (2000): Interactive Query Expansion: A User–Based Evaluation in Relevance Feedback Environment, *Journal of the American Society for Information Science*, v. 51, n. 11, str. 989–1003.

Eztioni O., Weld D. (1994), A Softbot–Based Interface to the Internet, *Communications of the ACM*, 37(7), str. 72–76.

Faloutsos C., Oard D.W. (1995), A Survey of Information Retrieval and Filtering Methods, *raport techniczny*, Uniwersytet Maryland, Maryland, USA, str.1–24.

Fink J., Kobsa A., Nill A. (1997): Adaptable and Adaptive Information Access for All Users, Including the Disabled and the Elderly, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 171–174, Springer Wien New York.

Frakes W., Baeza–Yates R. (1992), *Information Retrieval: Data Structure and Algorithms*, Prentice–Hall, New Yersey.

Goldberg J. L., (1996): CDM: An Approach to Learning in Text Categorization, *International Journal on Artificial Intelligence Tools*, v. 5(n. 1 i 2), str. 229–253.

Grasso F. (1997): Using Dialectical Argumentation for User Modeling in Decision Support Systems, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 83–86, Springer Wien New York.

Gutkauf B., Thies S., Edwards A. D. N. (1997): A User Adaptive Chart Editing System Based on User Modeling and Critiquing, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 159–170, Springer Wien New York.

Harman D. (1992): Relevance Feedback Revisited, *SIGIR'92, Proceedings of the Fifteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Denmark, ACM Press, str. 1–10.

Hu W.-C., Chen Y., Schmalz M., Ritter G. (2001): An overview of the World Wide Web search technologies, *Proceedings of 5th World Multi-conference on System, Cybernetics and Informatics SCI2001*, Orlando, Florida, July 22-25.

Ide E. (1971): Nowe badania systemów wyszukiwania ze sprzężeniem zwrotnym, w Salton G. (pod redakcją) *SMART – Automatyczny system wyszukiwania informacji*, PWN, str. 353–373.

Indyka-Piasecka A. (2000): Możliwości zastosowania tradycyjnych metod wyszukiwania informacji w sieci WWW, *II Krajowa Konferencja Multimedialne i Sieciowe Systemy Informacyjne MISSI'00*, Oficyna Wydawnicza PWr., Wrocław 2000, str. 269–275.

Indyka-Piasecka A. (2002): Propozycja personalizacji zapytań w internetowym systemie wyszukiwania informacji, *III Krajowa Konferencja Multimedialne i Sieciowe Systemy Informacyjne MISSI'02*, Oficyna Wydawnicza PWr., Wrocław 2002, str. 419–428.

Indyka-Piasecka A., Piasecki M. (2003): Adaptive Translation between User's Vocabulary and Internet Queries, *Proceedings of the International Conference on Intelligent Information Systems*, *New Trends in Intelligent Information Processing and Web Mining: IIPWM'03*, Zakopane, Poland, June 2-5, 2003, Springer, str.149–157.

Indyka-Piasecka A., Daniłowicz Cz. (2004): Dynamic User Profiles Based on Boolean Formulas, R. Orchard et al. (Eds.): *Proceedings of the IEA/AIE 2004 Conference,* LNAI 3029, Springer-Verlag Berlin Heidelberg, str. 779-787.

Kazienko P. (2000): Grupowanie dokumentów hipertekstowych na podstawie drzewa maksymalnych przepływów, praca doktorska, Politechnika Wrocławska, Wrocław.

Kelly D., Belkin N.J. (2002): Modeling Characteristics of the User's Problematic Situation with Information Search Use Behaviours, *Proc. of the JCDL'02*, July 13-17, Portland, Oregon.

Kelly D., Teevan J. (2003): Implicit Feedback for Inferring User Preference: A Bibliography, *SIGIR Forum*, vol. 37, n. 2, str. 18-28.

Kim, J., Oard, D. W., Romanik, K. (2000): Using implicit feedback for user modeling in Internet and Intranet searching, raport techniczny, College of Library and Information Services, University of Maryland at College Park.

Kobayashi M., Takeda K. (2000), Information retrieval on the Web, *ACM Computing Surveys*, vol. 32, n. 2, str. 144-173.

Lau T., Horvitz E. (1999): Patterns of Search: Analysing and Modeling Web Query Refinement, *Proc. of the 7th International Conference on User Modeling*, *UM'99*, Banff, Canada, str. 119–128.

Lawrence S., Giles C.L. (1998), Searching the World Wide Web, *Science*, April, vol. 280, str. 98–100, (www.sciencemag.org).

Lawrence S., Giles C.L. (1998 A), Context and Page Analysis for Improved Web Search, *IEEE Internet Computing*, July–August, str. 38–46.

Lawrence S., Giles C.L. (1999), Accessibility of Information on the Web, *Nature*, vol. 400, str. 107–109.

Levy A.Y., Weld D.S. (2000): Intelligent Internet Systems, *Artificial Intelligence*, v.118, str. 1–14.

Lieberman H. (1995): Letizia: An Agent that Assist Web Browsing, *Proc. of the International Joint Conference on Artificial Intelligence IJCAI–95*, Montreal, Quebec, Canada, str. 924–929.

Luhn H.T. (1957), A statistical approach to mechanised encoding and searching of library information, *IBM Journal of Research and Development*, 1, str.309-313.

Luhn H.T. (1958), The automatic creation of literature abstracts, *IBM Journal of Research and Development*, 2, str. 159-165.

Macskassy S.A. (2003): *New Techniques In Intelligent Information Filtering*, praca doktorska, Department of Computer Science, Rutgers University, New Brunswick, NJ.

Maes P. (1994): Agents that Reduce Work and Information Overload, *Communication of the ACM*, 37 (7), str. 31–40.

Maglio P.P., Barrett R. (1997): How to Build Modeling Agents to Support Web Searchers, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 5–16, Springer Wien New York.

Marchiori M. (1997), The Quest for Correct Information on the Web: Hyper Search Engines, *Proc. of the 6th International WWW Conference*, Santa Clara, California, USA, str. 222–231.

Moukas A. (1996): Amalthea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem, *Proc of the Practical Application of Intelligent Agents and Multi–Agents Technology, PAAM96*, London, UK.

Moukas A. (1997): Amalthea: Information Discovery and Filtering using a Multiagent Evolving Ecosystem, *Proc of the International Conference on Autonomus Agents 97*, ACM Press, str. 394–403.

Moukas A., Maes P. (1998): Amalthea: An Evolving Multiagent Information Filtering and Discovery System for the WWW, *Autonomous Agents and Multi–Agent Systems*, vol. 1.

Müller H., Müller W., Marchand-Maillet S., Pun T., Squire D. (2000): Strategies for positive and negative relevance feedback in image retrieval, *Proc. of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, September 3-8.

Nichols D. (1997): Implicit rating and filtering, *Proc. 5th DELOS Workshop on Filtering and Collaborative Filtering*, Budapest, Hungary, str. 31–36.

Paranagama P., Burstein F. (1996): A Preliminary Study of the Relationship Between the Decision–Makers' Personality and Models of their Preferences, *Proc. (Suplement) of the IFI WG 8.3 Working Conference*, str.19–38.

Paranagama P., Burstein F., Arnott D. (1997): Modelling the Personality of Decision Makers for Active Decision Support, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 79–82, Springer Wien New York.

Pazzani M., Billsus D. (1997): Learning and Revising User Profiles: The Identification of Interesting Web Sites, *Machine Learning*, 27, str. 313–331.

Pazzani M., Muramatsu J., Billsus D. (1996): Syskill and Webert: Identifying Interesting Web Sites, , *Proc. of the 13th National Conference on Artificial Intelligence AAAI-96*, Portland, OR, str. 54–61.

Peat H.J., Willett P. (1991): The Limitation of Term Co-Occurance Data for Query Expansion in Document Retrieval Systems, *Journal of the American Society for Information Science,* 42(5), str. 378–383.

Pinkerton B. (1994): Finding What People Want: Experience with the WebCrawler, *Proc. of the 2nd International WWW Conference*, Chicago, Illinois, USA.

Pretschner A., Gauch S. (1999): Ontology Based Personalized Search, *Proc. of the 11th IEEE International Conf. on Tools with AI* (www4.in.tum.de/pretschn/papers/kuthesis.ps.gz).

Pretschner A., Gauch S. (2000): Personalization on the Web, raport techniczny, Information and Telecommunication Technology Center, Department of Electrical Engineering and Computer Science, The University of Kansas ITTC-FY2000.

Qiu Y., Frei H.P. (1993): Concept Based Query Expansion. *Proceedings of the Sixteenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, Pittsburgh, PA USA, ACM Press, June 27 – July 1, str. 160–169.

Qiu Y., (1996): Automatic Query Expansion Based on a Similarity Thesaurus, praca doktorska, Swiss Federal University of Technology, Zurich, Swiss.

Quinlan J.R. (1986): Introduction of Decision Trees, *Machine Learning*, 1, str, 81–106.

Rao J. S. (1988): An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval ..I.. On effectiveness of full-text retrieval. *Journal of the American Society for Information Science*, 39 (2), str. 73–78.

Rao J. S. (1988a): An evaluation of the applicability of ranking algorithms to improve the effectiveness of full-text retrieval ..II.. On effectiveness of full-text retrieval. *Journal of the American Society for Information Science*, 39 (3), str. 147–160.

Rocchio J.J. (1971): Sprzężenie zwrotne w wyszukiwaniu informacji, w Salton G. (pod redakcją) *SMART – Automatyczny system wyszukiwania informacji*, PWN, str. 328–338.

Salton G. (1971): *SMART – Automatyczny system wyszukiwania informacji*, PWN, 1971.

Salton G. (1971), Recent Studies in Automatic Text Analysis and Document Retrieval, *Journal of ACM*, 20(2), str. 258–278.

Salton G., Yang C.S., Yu C.T. (1975), A Theory of Term Importance in Automatic Text Analysis, *Journal of the American Society for Information Science*, 26(1), str. 33–44.

Salton G., McGill M.J. (1983), *Introduction to Modern Information Retrival*, McGraw–Hill Book Co., New York, str. 52–117.

Salton G., McGill M. (1983a), *Introduction to Modern Information Retrieval*, McGraw-Hill Book Co., New York, str.202-203.

Salton G., Buckley Ch. (1988): Term-Weighting Approaches in Automatic Text Retrieval. *Information Processing & Management*, 24 (5), str. 513–523.

Salton G. (1998): *Automatic Text Processing*, Addison–Wesley.

Salton G., Buckley C. (1990): Improving Retrieval Performance by Relevance Feedback, *Journal of the American Society for Information Science*, v. 41, n. 4, str. 288–297.

Seo Y.W., Zhang B.T. (2000): A Reinforcement Learning Agent for Personalised Information Filtering. *Proceedings of the 2000 International Conference on the Intelligent User Interfaces*, New Orleans, LA USA, ACM Press, January 9 – 12, 2000, str. 248–251.

Spink, A., Jansen, B.J., Ozmultu, H.C. (2000): Use of query reformulation and relevance feedback by Excite users, *Internet Research: Electronic Networking Applications and Policy*, vol. 10, n. 4, str. 317-328.

Staff C. (1997): HyperContext:A Model for Adaptive Hypertext, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 33–36, Springer Wien New York.

Stein A, Gulla J.A., Theil U. (1997): Making Sense of User's Mouse Clicks: Abductive Reasoning and Conversational Dialogue Modeling, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 89–100, Springer Wien, New York.

Strachan L., Anderson J., Sneesby M., Evans M. (1997): Pragmatic User Modeling in A Commercial Software Systems, *Proc. of the 6th International Conference on User Modeling*, *UM'97*, Sardinia, str. 189–200, Springer Wien New York.

Tanudjaja F., Mui L. (2002) Persona: A contextualized and personalized Web search, *Proc. of the 35th Annual Hawaii International Conference on System Sciences HICSS'02*, Big Island, Hawaii, vol. 3, str. 53 (http://citeseer.ist.psu.edu/tanudjaja01persona.html).

van Rijsbergen C.J. (1979), *Information Retrieval*, Butterworths, London, second edition.

Voorhees E.M. (1986): Implementing Agglomerative Hierarchic Clustering Algorithms for Use in Document Retrieval, *Information Processing and Management*, v. 22, n.6, str. 465–476.

Webb G.I., Pazzani M.J., Billsus D. (2001): Machine Learning for User Modeling, *User Modeling and User–Adapted Interaction*, v. 11, n 1–2, str. 19–29.

Weiss S., Kasif S., Brill E. (1996): Text Clasification in USENET A Progress Report, Working Notes of the AAAI Spring Symposium on Machine Learning in Information Access, Stanford, CA.

Weiss S., Kasif S. Brill E. (1997): Text Classification in USENET Newsgroups: A Progress Report, raport techniczny, The Johns Hopkins University, str. 1–15.

Xu J., Croft W.B. (1996): Query Expansion Using Local and Global Document Analysis. *Proceedings of the 19th Annual International ACM–SIGIR Conference on Research and Development in Information Retrieval*, Zurich Switzerland, ACM Press, August 18 – 22, str. 4–11.

Xu J., Croft W.B. (2000): Improving the Effectiveness of Information Retrieval with Local Context Analysis. *ACM Transactions on Information Systems*, 18 (1), January 2000, str. 79–112.

Yang K., Maglaughlin K., Meho L., Sumner R.G. (1999), IRIS at TREC7, w E.M. Voorhees i D.K. Harman (Eds), *Proc. of the 17th Text Retrieval Conference (TREC7)*, Washington DC, US Goverment Printing Office, str. 555–666.

Zhang B., Seo Y. (2001): Personalized web-document filtering using reinforcement learning, *Applied Artificial Intelligence*, v.15, n.7, , str. 665-685.

Zukerman I., Alberecht D.W., Nicholson A.E. (1999): Predicting User's Requests on WWW, *Proc. of the 7th International Conference on User Modeling, UM'99*, Banff, Canada, str. 275–284.

Zukerman I., Alberecht D.W. (2001): Predictive Statistical Models for User Modeling, *User Modeling and User Adapted Interaction*, v. 11, n. 1–2, str. 5–18.