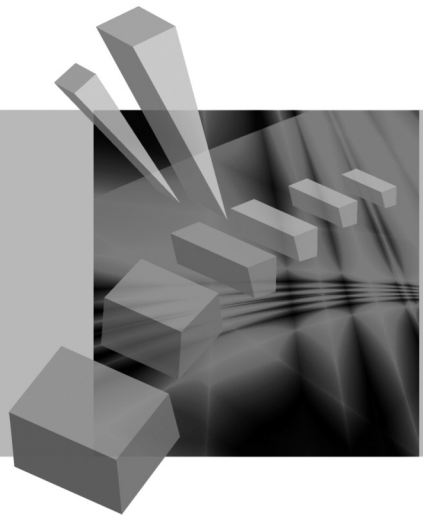
PRACE NAUKOWE
Uniwersytetu Ekonomicznego we Wrocławiu
RESEARCH PAPERS
of Wrocław University of Economics

242

Taksonomia 19.

Klasyfikacja i analiza danych – teoria i zastosowania



Redaktorzy naukowi
Krzysztof Jajuga
Marek Walesiak

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2012

Recenzenci: Eugeniusz Gatnar, Elżbieta Gołata, Tadeusz Kufel, Józef Pociecha,
Mirosław Szreder, Feliks Wysocki

Redaktor Wydawnictwa: Aleksandra Śliwka

Redaktor techniczny: Barbara Łopusiewicz

Korektor: Barbara Cibis

Łamanie: Małgorzata Czupryńska

Projekt okładki: Beata Dębska

Tytuł sfinansowano ze środków Sekcji Klasyfikacji i Analizy Danych PTS
i Uniwersytetu Ekonomicznego we Wrocławiu

Publikacja jest dostępna na stronie www.ibuk.pl

Streszczenia opublikowanych artykułów są dostępne w międzynarodowej bazie danych
The Central European Journal of Social Sciences and Humanities http://cejsh.icm.edu.pl
oraz w The Central and Eastern European Online Library www.ceeol.com,
a także w adnotowanej bibliografii zagadnień ekonomicznych BazEkon http://kangur.uek.krakow.pl/
bazy_ae/bazekon/nowy/index.php

Informacje o naborze artykułów i zasadach recenzowania znajdują się
na stronie internetowej Wydawnictwa
www.wydawnictwo.ue.wroc.pl

ISSN 1899-3192 (Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu)
ISSN 1505-9332 (Taksonomia)

Wersja pierwotna: publikacja drukowana

Druk: Drukarnia TOTEM
Nakład: 320 egz.

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 242
RESEARCH PAPERS OF WROCŁAW UNIVERSITY OF ECONOMICS 2012

Taksonomia 19. Klasyfikacja i analiza danych – teoria i zastosowania ISSN 1899-3192

Paweł Lula
Uniwersytet Ekonomiczny w Krakowie

UCZĄCE SIĘ SYSTEMY POZYSKIWANIA INFORMACJI Z DOKUMENTÓW TEKSTOWYCH

Streszczenie: Zasadniczym celem pracy jest prezentacja, klasyfikacja i ocena systemów pozyskiwania informacji z dokumentów tekstowych konstruowanych przy udziale mechanizmów uczących. W początkowej części pracy zdefiniowano pojęcie systemu uczącego się oraz zagadnienie pozyskiwania informacji. Następnie zaprezentowano strukturę oraz sposób funkcjonowania uczącego się systemu pozyskiwania informacji. Kluczowym elementem tego typu rozwiązań jest model zawartości informacyjnej. Jego charakterystyka i rodzaje są zasadniczym tematem kolejnego punktu pracy. Następnie przedstawiono rolę wiedzy zewnętrznej i metod uczenia maszynowego w poszczególnych rozwiązaniach. W kolejnej części artykułu zaprezentowano rozważania dotyczące lokalnego lub globalnego charakteru poszczególnych rozwiązań.

Słowa kluczowe: pozyskiwanie informacji z dokumentów tekstowych, analiza tekstu, text mining.

1. Wstęp

Występujący praktycznie w każdej dziedzinie dynamiczny wzrost liczby dokumentów sprawia, że koniecznością staje się automatyzowanie procesów pozyskiwania informacji z dokumentów tekstowych. W wielu przypadkach konstrukcja tego typu rozwiązań jest zadaniem praco- i czasochłonnym. Dlatego też pozytywnie należy ocenić podejmowane próby zastąpienia zaangażowania człowieka w budowę systemu przez mechanizmy pozwalające na doskonalenie systemu dzięki zastosowaniu algorytmów uczenia maszynowego.

Zasadniczym celem pracy jest prezentacja, klasyfikacja i ocena systemów pozyskiwania informacji z dokumentów tekstowych konstruowanych przy udziale mechanizmów uczących.

2. System uczący się

System uczący się charakteryzuje się zdolnością do poprawy sposobu swojego funkcjonowania na podstawie wniosków płynących z przeprowadzonej przez niego ana-

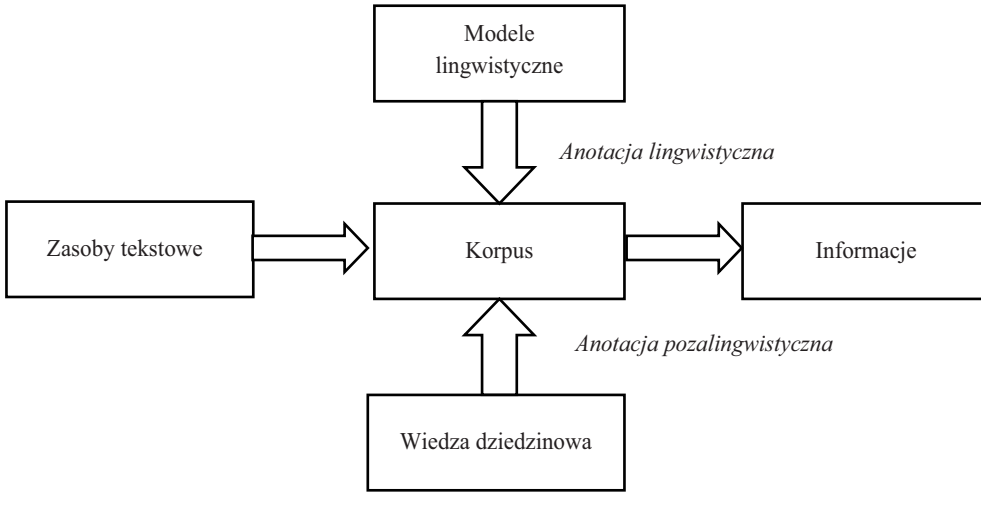
lizy dostarczonych danych. W przeważającej liczbie przypadków systemy uczące się przyjmują postać algorytmów komputerowych służących do rozwiązywania ściśle określonej klasy zadań. Niech *P* będzie przyjętą miarą jakości sposobu realizacji rozpatrywanego typu zadania, zaś *D* – zbiorem danych opisującym rzeczywisty przebieg procesów zmierzających do ich rozwiązania. Algorytmowi *A* można przypisać zdolność do *uczenia się*, jeśli został on przystosowany do polepszania swojego działania w sensie przyjętej miary *P* poprzez analizę danych zawartych w zbiorze *D*.

3. Pozyskiwanie informacji

Pozyskiwanie informacji z zasobów tekstowych będziemy rozumieć jako obszar badań naukowych oraz zastosowań wypracowanych metod i narzędzi mających na celu wydobycie istotnych informacji z zasobów przyjmujących postać nieustrukturyzowanego tekstu (np. dokumentów w postaci elektronicznej, zawartości stron WWW, blogów, poczty elektronicznej). System pozyskiwania informacji wykonuje swoje zadanie jako odpowiedź na zapytanie sformułowane przez użytkownika.

4. Struktura i etapy budowy uczącego się systemu pozyskiwania informacji z zasobów tekstowych

Struktura systemu pozyskiwania informacji z zasobów tekstowych przedstawiona została na rys. 1.



Rys. 1. Struktura systemu pozyskiwania informacji

Źródło: opracowanie własne.

Punktem wyjścia w procesie budowy uczącego się systemu pozyskiwania informacji z dokumentów tekstowych jest przekształcenie zestawu dokumentów źródłowych w jednolity korpus. Operacja ta obejmuje przekształcenie dokumentów do postaci plików tekstowych, usunięcie wszelkich informacji formatujących oraz ujednolicenie sposobu kodowania.

Kolejnym etapem jest anotacja (*annotation*) korpusu. Pojęcie anotacji określane jest również mianem znakowania. Polega ona na dodaniu do tekstu źródłowego dodatkowych informacji w postaci znaczników obejmujących swoim zasięgiem fragmenty tekstu. Znaczniki odgrywają zwykle podwójną rolę: określają charakter fragmentu tekstu przez zidentyfikowanie jego charakteru i przypisanie mu właściwej etykiety oraz pozwalają na przypisanie pewnych dodatkowych informacji do fragmentów tekstu źródłowego. Jako przykład ilustrujący wspomniane dwie funkcje posłużyć może zapis:

Jacek<*CZASOWNIK bezokolicznik=”kolekcjonować”*>**kolekcjonuje**</*CZASOWNIK*>**znaczki pocztowe**.

Dodane znaczniki wskazują, że słowo „kolekcjonuje” jest czasownikiem i jednocześnie informują o postaci bezokolicznika.

Anotacja może dotyczyć informacji lingwistycznej lub pozalingwistycznej. Znaczniki pozalingwistyczne mają zwykle na celu wzbogacenie warstwy znaczeniowej tekstu, dlatego też określane są mianem anotacji semantycznej.

Najbardziej upowszechnioną formą anotocji lingwistycznej jest identyfikacja części mowy (*POS tagging* lub *POST – part-of-speech tagging*). W trakcie realizacji tego zadania poszczególnym wyrazom przypisywane są informacje określające m.in. właściwą dla niego część mowy, postać podstawową, formę występującą w dokumencie. Identyfikacja części mowy może być realizowana za pomocą słowników lub odpowiednich algorytmów.

Natomiast powszechnie stosowanym elementem anotacji pozalingwistycznej jest identyfikacja znaczeniowa elementów tekstu (NER – *Named Entity Recognition*). Zadanie polega na wyodrębnieniu takich elementów, jak: imiona i nazwiska osób, nazwy geograficzne, nazwy organizacji i instytucji, określenia czasu, jednostki pieniężne, adresy poczty elektronicznej, numery telefonów i wiele innych. Elementy te są następnie znakowane.

Korpus dokumentów wzbogacony o dodatkowe informacje dodane w postaci anotacji określany jest mianem *korpusu anotowanego*. Jest on punktem wyjścia do budowy *modelu zawartości informacyjnej* korpusu. Podstawowym zadaniem modelu jest reprezentowanie informacji występujących w tekście, zapewnienie do nich dostępu oraz dostarczenie narzędzi pozwalających na ich przetwarzanie w sposób pożądany przez użytkownika systemu.

5. Modele zawartości informacyjnej dokumentów

Prezentowane w niniejszej pracy podejście zakłada, że system pozyskiwania informacji nie operuje bezpośrednio na dokumentach źródłowych, lecz na modelu reprezentującym informacje pochodzące z dokumentów. Przyjęty sposób modelowego ujęcia informacji ma istotny wpływ na możliwości i efektywność systemu. W sposób znaczący determinuje również koszty związane z budową systemu. W niniejszej pracy wyróżnione zostały cztery typy modeli opisujących zawartość dokumentów. Każdy z nich może zostać wykorzystany w systemach pozyskiwania informacji. Jednakże funkcje, jakie mogą spełniać poszczególne rozwiązania, są bardzo zróżnicowane.

6. Modele oparte na koncepcji przestrzeni wektorowej

W przedstawianym podejściu konstruowana jest przestrzeń, której poszczególne wymiary odpowiadają terminom indeksującym przyjętym dla danego zestawu dokumentów. W najprostszym przypadku funkcję terminów indeksujących spełniać mogą wyrazy (ewentualnie po sprowadzeniu do formy podstawowej). W bardziej zaawansowanych podejściach rolę terminów indeksujących mogą odgrywać frazy lub identyfikatory przypisane do zidentyfikowanych w tekście faktów. Dokument jest więc kombinacją liniową wektorów odpowiadających występującym w nim terminów indeksujących. Bardzo ważnym problemem jest sposób ustalania wag odpowiadających poszczególnym wymiarom. Najczęściej spotykaną metodą jest technika *tf-idf* zakładająca, że waga jest proporcjonalna do częstości występowania terminu indeksującego w dokumencie i odwrotnie proporcjonalna do częstości dokumentowej (liczby dokumentów zawierających ten termin). Model przestrzeni wektorowej zaproponowany został w [Salton i in. 1975].

Realizacja zapytań polega na potraktowaniu zapytania jako pseudodokumentu i wyznaczeniu jego reprezentacji w tej samej przestrzeni, w której ulokowane zostały pozostałe elementy korpusu. Takie rozwiązanie pozwala na potraktowanie zarówno dokumentów, jak i samego zapytania jako punkty ulokowane w pewnej przestrzeni, co pozwala na wyznaczenie odległości pomiędzy zapytaniem i dokumentami i wybór tych, dla których tak określona miara przyjmuje wartość najmniejszą.

Przy stosowaniu przedstawionego podejścia metody uczenia maszynowego znajdują szczególne zastosowanie na etapie analizy zawartości informacyjnej dokumentów. Reprezentacja wyrazów w postaci punktów w wielowymiarowej przestrzeni, dokumentów zaś jako ich zbiorów pozwala na stosowanie wszystkich metod opartych na miarach odległości lub podobieństwa.

7. Modele probabilistyczne

Modele probabilistyczne zakładają sekwencyjny charakter tekstu. W zależności od celu badań tekst może być rozumiany jako ciąg liter, wyrazów czy sylab. Związki

pomiędzy elementami sekwencji opisywane są za pomocą pojęć rachunku prawdopodobieństwa. Tego typu podejście w badaniach lingwistycznych zostało zapoczątkowane przez Andrieja Markowa, twórcę koncepcji procesów Markowa, w których rozważana jest sekwencja stanów pewnego systemu i prawdopodobieństwo wystąpienia stanu uzależnione jest od stanu bezpośrednio go poprzedzającego. W 1913 r. tego rodzaju model Markow wykorzystał do opisu sekwencji liter w tekstach rosyjskich [Manning, Schütze 1999].

Na bazie klasycznego modelu Markowa stworzona została koncepcja *ukrytego procesu Markowa* (HMM – *Hidden Markov Model*), w którym sekwencja kolejnych stanów systemu nie jest bezpośrednio obserwowalna, lecz dostępne są jedynie sekwencje wartości pewnej funkcji losowej przekształcającej ukryty dla obserwatora stan systemu w wartość obserwowalną.

Parametry klasycznego oraz ukrytego modelu Markowa szacowane są na podstawie zbioru uczącego. Ukryty model Markowa wykorzystywany jest z powodzeniem w zadaniach identyfikacji części mowy, gdzie ma szczególne znaczenie przy podejmowaniu decyzji dotyczących wyrazów wieloznacznych. Natomiast w językach azjatyckich modele tego typu stosowane są do wyznaczania granic słów.

8. Modele oparte na gramatykach formalnych

Gramatyka formalna jest opisem zasad budowy poprawnych wypowiedzi na bazie przyjętego słownika elementów składowych. Gramatyka formalna operuje dwiema grupami pojęć: pojęciami terminalnymi oraz pojęciami nieterminalnymi. Pojęcia terminalne to te, które mogą wystąpić w wypowiedzi. Pojęcia nieterminalne mają charakter abstrakcyjny, nie pojawiają się w wypowiedzi, lecz są zastępowane przez odpowiadające im symbole terminalne. Szczególne znaczenie ma symbol startowy, mający charakter nieterminalny i reprezentujący całą wypowiedź. Definicja gramatyki obejmuje również reguły pozwalające na sprawdzenie, czy rozpatrywany ciąg symboli terminalnych jest prawidłowy. Reguły te określane są mianem produkcji. Sprawdzając poprawność wypowiedzi, należy znaleźć ciąg produkcji pozwalających na jej wyprowadzenie z symbolu startowego. Identyfikacja produkcji pozwalających na wyprowadzenie analizowanego zdania pozwala również na dokonanie interpretacji jego elementów. Autorem koncepcji gramatyk formalnych jest Noam Chomsky. Wśród wielu publikacji prezentujących zagadnienia gramatyk formalnych i ich zastosowań warto wymienić prace: [Révész 1983; Clark i in. 2010].

Gramatyki formalne są narzędziem opisu i analizy zarówno języków sztucznych (np. języków programowania), jak i języków naturalnych. Szczególnie przydatną klasą gramatyk są gramatyki bezkontekstowe, w których produkcje mają postać:

symbol nieterminalny → ciąg symboli terminalnych i/lub nieterminalnych,

czyli sposób traktowania symboli nieterminalnych nie zależy od ich kontekstu.

Zastosowanie reguł zdefiniowanych w ramach gramatyki bezkontekstowej pozwala na dokonanie rozbioru wypowiedzi na poszczególne elementy składowe. Sposób dokonywania rozbioru może zostać przedstawiony w postaci drzewa rozbioru.

O ile gramatyki bezkontekstowe dobrze radzą sobie z analizą języków sztucznych, o tyle języki naturalne, ze względu na ich nieregularność i dopuszczalność wielu alternatywnych sposobów konstrukcji wypowiedzi, stanowiły wyzwanie dla podejmowanych prób ich opisu. Koncepcją pozwalającą na zastosowanie produkcji gramatyk bezkontekstowych do opisu języków naturalnych są probabilistyczne gramatyki bezkontekstowe (określane również jako stochastyczne gramatyki bezkontekstowe). W rozwiązaniach tych z każdą produkcją powiązane jest prawdopodobieństwo jej wystąpienia. Prawdopodobieństwa te przypisywane są w taki sposób, aby suma prawdopodobieństw przypisanych do produkcji o tym samym poprzedniku równała się jedności. W gramatykach probabilistycznych istnieje wiele alternatywnych sposobów rozbioru analizowanej wypowiedzi. Jednakże biorąc pod uwagę prawdopodobieństwa produkcji, można wyznaczyć prawdopodobieństwo wystąpienia każdego drzewa rozbioru (i wybrać najbardziej prawdopodobny sposób interpretacji) i wybrać to, które jest najbardziej prawdopodobne. Budowa zestawu produkcji i przypisywanie im prawdopodobieństw może zostać przeprowadzone przez uczenie maszynowe, głównie przez zastosowanie algorytmów genetycznych.

Do znajdowania najbardziej prawdopodobnego drzewa rozkładu stosuje się zwykle algorytm inside-outside lub algorytm A^* [Manning, Schütze 1999]. Algorytm inside-outside pozwala na dokonanie wyboru zbioru produkcji maksymalizującego prawdopodobieństwo uzyskania zdań wchodzących w skład zbioru uczącego. Algorytm pracuje w trybie bez nauczyciela. Jego wadą jest skłonność do zatrzymywania się w minimach lokalnych optymalizowanej funkcji. Natomiast algorytm A^* należy do metod przeszukiwania grafów. W lingwistyce metoda ta jest stosowana do identyfikacji najbardziej prawdopodobnego drzewa rozkładu analizowanego zdania przy założeniu, że produkcje gramatyki typu PCFG są znane. Autorami koncepcji zastosowania algorytmu A^* do parsowania zdań w języku naturalnym są Klein oraz Manning [2003].

9. Modele w postaci sieci semantycznej

Sieć semantyczna w niniejszej pracy rozumiana będzie jako narzędzie opisu wyodrębnionego fragmentu rzeczywistości pozwalające na wyodrębnienie klas obiektów, definiowanie schematów ich opisów i przedstawienie relacji pomiędzy klasami oraz reprezentowanie rzeczywistych instancji obiektów i relacji. Definiowanie schematu opisu obiektu polega na przyjęciu zestawu własności charakteryzujących obiekt. Dla każdej własności może zostać ustalony typ przypisanych jej wartości oraz mogą zostać nałożone warunki precyzujące zakres dopuszczalnych wartości. Relacje opisywać mogą hierarchię klas oraz związki zachodzące pomiędzy klasami i obiektami.

Przyjęty za obowiązujący, akceptowany przez ogół użytkowników i pozwalający na jednoznaczny opis wymienianych informacji sposób opisu wyodrębnionej dziedziny określany jest terminem ontologii.

Zastosowanie ontologii znacznie podnosi jakość wyników eksploracyjnej analizy tekstów. Podejście wykorzystujące wiedzę dziedzinową w postaci ontologii nie wyklucza również stosowania metod uczenia maszynowego. Należy zauważyć, że algorytmy uczące się mogą być stosowane w trzech całkowicie różnych obszarach. Po pierwsze, mogą być stosowane na etapie konstrukcji ontologii. Zastosowanie metod eksploracyjnych pozwala na identyfikację klas oraz pozwala określić ich hierarchię. Analiza danych jest również przydatnym narzędziem identyfikacji związków pomiędzy klasami [Buitelaar i in. 2003]. Drugim zastosowaniem jest uczenie zasad transformacji tekstu do postaci sieci semantycznej. Tworzone na tym etapie rozwiązanie ma na celu analizę fragmentu tekstu i utworzenie instancji klasy, przypisanie właściwych wartości własnościom obiektu lub rozpoznanie relacji pomiędzy klasami. Trzecim zastosowaniem algorytmów uczenia maszynowego jest analiza danych tworzących sieć semantyczną.

Ontologiczny model zawartości informacyjnej dokumentu w sposób precyzyjny i jednoznaczny reprezentuje treść tekstu. Ważną zaletą tego podejścia jest możliwość wykorzystania modelu zarówno przez człowieka, jak i przez system komputerowy. Reprezentacja ontologiczna pozwala na realizację zapytań. Zapytania kierowane do sieci semantycznej formułowane są za pomocą przyjętej formalnej notacji. Przykładami tego typu rozwiązań jest język SPARQL (*Simple Protocol and RDF Query Language*) oraz OWL-QL (*Web Ontology Language – Query Language*).

10. Rola wiedzy dziedzinowej i uczenia maszynowego w procesie konstrukcji modeli zawartości informacyjnej

Cechą wspólną wszystkich przedstawionych powyżej podejść jest możliwość budowy modelu na bazie danych w postaci korpusu dokumentów. Jednakże wszystkie przedstawione powyżej rozwiązania wymagają również dostarczenia wiedzy zewnętrznej (lingwistycznej lub dziedzinowej). Aspekt ten jest szczególnie istotny z punktu widzenia czasochłonności i kosztów budowy modeli. Zgromadzenie odpowiednich zasobów wiedzy oraz opracowanie właściwego sposobu jej reprezentacji jest szczególnie kosztownym elementem tworzonych systemów (w wymiarze zarówno czasowym, jak i finansowym). Z drugiej jednak strony próby szukania oszczędności na tym polu prowadzić mogą do istotnego spadku jakości konstruowanych rozwiązań.

Warto prześledzić rolę, jaką odgrywa wiedza zewnętrzna i uczenie maszynowe w poszczególnych klasach modeli. Zestawienie przedstawiano w tab. 1.

Tabela 1. Rola uczenia maszynowego i wiedzy zewnętrznej w modelach zawartości informacyjnej dokumentów tekstowych

Model	Uczenie maszynowe	Wiedza zewnętrzna
Model przestrzeni wektorowej	• konstrukcja modelu na podstawie macierzy częstości • redukcja wymiaru modelu • analiza informacji zawartych w dokumentach i reprezentowanych przez punkty w przestrzeni	• listy terminów indeksujących • listy terminów nieistotnych (stop-lista) • synonimy • frazy • forma podstawowa wyrazów
Model probabilistyczny	• konstrukcja modelu • szacowanie prawdopodobieństwa wystąpienia wypowiedzi	• anotacja lingwistyczna wyrazów
Model gramatyczny	• identyfikacja reguł gramatycznych, • oszacowanie prawdopodobieństw dla poszczególnych reguł • parsowanie tekstu i wybór najbardziej prawdopodobnego drzewa rozbioru	• anotacja lingwistyczna i semantyczna wyrazów • definicja produkcji
Model w postaci sieci semantycznej	• wspomaganie procesu tworzenia ontologii • tworzenie reguł przekształcających fragmenty tekstu w elementy sieci semantycznej • analiza danych w postaci sieci semantycznej	• wiedza dziedzinowa w postaci ontologii • anotacja morfologiczna i semantyczna wyrazów • identyfikacja nazw własnych • reguły interpretacji fragmentów tekstu

Źródło: opracowanie własne.

11. Lokalny i globalny charakter poszczególnych modeli zawartości informacyjnej

Porównując zawartość informacyjną korpusu dokumentów oraz zasięg obszaru opisywanego przez model zawartości informacyjnej, za celowe należy uznać wyróżnienie modeli o charakterze:

- globalnym – w których zakres obszaru opisywanego przez model pokrywa się z zakresem całego dokumentu,
- lokalnym – które opisują jedynie zawartość pewnego, zwykle niewielkiego, fragmentu dokumentu.

Wśród przedstawionych powyżej koncepcji modele oparte na koncepcji przestrzeni wektorowej oraz wykorzystujące sieci semantyczne mają charakter globalny. Natomiast modele probabilistyczne oraz wykorzystujące gramatyki formalne przydatne są raczej do opisu fragmentu dokumentu – można je więc określić jako lokalne.

Lokalny bądź globalny charakter modelu nie przesądza o ocenie danego rozwiązania. Dla każdego rodzaju modelu przewidziany jest odmienny zakres zadań.

Szczególnie interesujące jest porównanie dwóch podejść pozwalających na reprezentację całości informacji zawartych w korpusie. Model oparty na przestrzeni wektorowej oraz sieć semantyczna różnią się praktycznie wszystkimi cechami: sposobem reprezentacji informacji, czasem i kosztem budowy czy sposobem analizy reprezentowanych informacji. Z tego powodu dokonanie właściwego wyboru jest w tym przypadku szczególnie ważne.

Podejścia stosowane w modelach lokalnych są znacznie bardziej do siebie zbliżone pod względem możliwości i kosztów konstrukcji.

Należy również zwrócić uwagę na możliwości łącznego wykorzystania modeli. Szczególnie przydatne może być wykorzystanie modeli lokalnych przy konstrukcji modeli globalnych. Podejście takie może być stosowane w szerokim zakresie przy budowie modeli w postaci sieci semantycznej. Natomiast przydatność modeli lokalnych w trakcie tworzenia modelu opartego na przestrzeni wektorowej jest ograniczona (do identyfikacji fraz i przekształcenia słów do postaci podstawowej).

12. Podsumowanie

Głównym celem niniejszej pracy było przedstawienie roli uczenia maszynowego w procesie pozyskiwania informacji z dokumentów tekstowych. Założono, że zadanie to realizowane jest za pośrednictwem modelu zawartości informacyjnej dokumentu. Przedstawione charakterystyki poszczególnych rozwiązań wskazują na bardzo duże zróżnicowanie poszczególnych koncepcji.

W przypadku podejść opartych na koncepcji przestrzeni wektorowej uczenie maszynowe odgrywa szczególnie istotną rolę na etapie zastosowania modelu. W modelach probabilistycznych oraz opartych na gramatykach formalnych i na sieciach semantycznych uczenie maszynowe jest bardzo przydatne już na etapie ich tworzenia. Pozyskiwanie informacji z dokumentów tekstowych pozwala na wykorzystanie zarówno nadzorowanych (uczenie z nauczycielem), jak i nienadzorowanych (uczenie bez nauczyciela) metod uczenia maszynowego.

Wydaje się, że podstawową przesłanką przemawiającą za korzystaniem z uczenia maszynowego w zadaniach pozyskiwania informacji z zasobów tekstowych jest brak możliwości formalnego ujęcia wszystkich poprawnych wypowiedzi formułowanych w języku naturalnym za pomocą precyzyjnych reguł. Ze względu na złożony charakter wypowiedzi oraz dużą ilość zasobów tekstowych rozwój zastosowań metod uczenia maszynowego wymaga rozwiązań charakteryzujących się dużą mocą obliczeniową i dużymi możliwościami w zakresie przechowywania danych.

Literatura

Buitelaar P., Cimiano Ph., Magnini B., *Ontology Learning from Text: An Overview*, 2003, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.70.3041.

Clark A., Fox C., Lappin S. (red.), *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell Publishing Ltd, 2010.

Farkas R., *Machine learning techniques for applied information extraction*, Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and the University of Szeged, June 2009, https://docs.google.com/viewer?url=http%3A%2F%2Fwww.sci.u-szeged.hu%2Ffokozatok%2FPDF%2FFarkas_Richard%2Fthesis.pdf.

Klein D., Manning C., *A* parsing: fast exact viterbi parse selection*, Proceedings of the Human Language Technology Conference and the North American Association for Computational Linguistics (HLT-NAACL 2003), Main Papers, Edmonton, May-June 2003.

Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA: May 1999.

Révész G., *Introduction to Formal Languages*, McGraw-Hill Book Company, 1983.

Salton G., Wong A., Yang C.S., *A vector space model for automatic indexing*, „Communications of the ACM" 1975, vol. 18, no 11.

LEARNING-BASED SYSTEMS OF INFORMATION EXTRACTION FROM TEXTUAL RESOURCES

Summary: The main aim of this work is the presentation, classification and evaluation learning-based systems of information extraction from textual resources. In the initial part of the paper the concept of the learning-based system and the problem of information extraction are presented. The next part of the article presents the structure and the functioning of the information extraction solutions. The model of information content is a key element of such systems. Its characteristics and types are the principal subject of the next point. The following point presents the role of external knowledge and machine learning approach in various solutions. In the next part of the article some remarks concerning local or global character of the individual solutions are presented.

Keywords: information extraction from textual resources, text mining.