PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 107
Taksonomia 17 2010

Paweł Lula
Uniwersytet Ekonomiczny w Krakowie

WYKORZYSTANIE INFORMACJI TEKSTOWEJ W MODELOWANIU I PREDYKCJI EKONOMICZNYCH SZEREGÓW CZASOWYCH

Streszczenie: Zasadniczym celem pracy jest przedstawienie i ocena metod pozyskiwania istotnych informacji z dokumentów tekstowych i ich uwzględnienia w statystycznych modelach o charakterze predykcyjnym. W pracy rozpatrywane są trzy podejścia do procesu pozyskiwania i reprezentacji informacji tekstowej: oparte na wyrazach, oparte na wzorcach oraz oparte na modelu ontologicznym. Zrealizowane badania empiryczne dotyczyły metod uwzględniania informacji zawartych w komunikatach spółek giełdowych przy prognozowaniu stóp zwrotu cen akcji.

1. Wstęp

Zgodnie z szacunkami firmy IBM 80% informacji generowanej przez sferę biznesu ma postać nieustrukturyzowaną [Internet 1], a więc głównie tekstową (podana wartość obejmuje również treści tekstowe publikowane na stronach WWW). Ten sposób prezentacji informacji jest dogodny dla człowieka, jednakże automatyczne przetwarzanie przez komputery jest znacznie utrudnione. Ogromne tempo zwiększania się liczby dokumentów i występująca jednocześnie konieczność ich szybkiego przetworzenia sprawia, że w czasach obecnych nie ma alternatywy dla rozwoju metod i narzędzi będących w stanie automatyzować przetwarzanie informacji tekstowej.

Celem niniejszego opracowania jest wskazanie metod pozwalających na uwzględnienie informacji tekstowej przy budowie statystycznych modeli prognostycznych. Praca składa się z dwóch zasadniczych części. Pierwsza z nich zawiera prezentację metod pozyskania informacji z dokumentów i jej uwzględnienia w modelu statystycznym. Druga ukazuje wyniki przeprowadzonych prac badawczych. Wnioski płynące z badań zamieszczone zostały w podsumowaniu.

2. Metody uwzględniania informacji tekstowej w modelach predyktywnych

Istotność informacji tekstowej w pełni uzasadnia podejmowanie badań nad metodami jej uwzględniania w modelach statystycznych. Ani tekst jako całość, ani też jego poszczególne fragmenty nie mogą zostać wprowadzone na wejściu modelu.

Konieczne jest zastosowanie właściwej metody numerycznej reprezentacji informacji zawartych w tekście. W stosunku do proponowanych różnych metod ilościowej reprezentacji informacji pozyskanych z dokumentów sformułować można następujące postulaty:

- powinny one zapewniać prawidłową reprezentację informacji zawartych w dokumencie,
- uzyskane rezultaty muszą mieć postać numeryczną,
- oczekuje się, aby metody uzyskiwania numerycznej reprezentacji były podatne na automatyzację.

Można wyróżnić trzy podstawowe metody uwzględnienia informacji tekstowej w modelach statystycznych, takie jak:

- metoda bazująca na częstości wystąpień poszczególnych wyrazów w dokumentach,
- metoda bazująca na wzorcach definiujących istotne informacje zawarte w dokumencie,
- metoda bazująca na przyjętym modelu rozpatrywanego fragmentu rzeczywistości.

2.1. Metoda bazująca na częstości wystąpień poszczególnych wyrazów w dokumentach

Za najważniejsze etapy tej metody należy uznać [Lula 2005]:

- podział dokumentów na wyrazy,
- usunięcie wyrazów nieistotnych (zawartych na tzw. stop-liście),
- przekształcenie wyrazów do formy podstawowej (redukcja do rdzenia),
- utworzenie macierzy częstości,
- przekształcenie macierzy częstości.

Wiersze utworzonej macierzy częstości reprezentują poszczególne wyrazy pochodzące z przetwarzanego zestawu dokumentów. Kolumny odpowiadają poszczególnym dokumentom. Element macierzy o indeksach (i, j) mówi, ile razy i-ty wyraz występuje w j-tym dokumencie. Uzyskane w ten sposób wartości poddawane są dalszym przekształceniom. Do najważniejszych z nich należy zaliczyć:

- przekształcenie elementów macierzy częstości do ważonej postaci logarytmicznej,
- redukcję wymiaru przestrzeni przez wyznaczenie zmiennych ukrytych za pomocą dekompozycji według wartości osobliwych.

2.2. Metoda bazująca na wzorcach definiujących istotne informacje zawarte w dokumencie

Podstawowym założeniem tego podejścia jest zastosowanie wzorców pozwalających na identyfikację istotnych informacji umieszczonych w tekście. Najistotniejszą zaletą tej metody jest możliwość precyzyjnej interpretacji poszczególnych

fragmentów tekstu w sposób opisany we wzorcu. Należy jednak podkreślić, że koszt wprowadzenia tej możliwości jest stosunkowo wysoki, gdyż metoda bazująca na wzorcach w stosunku do wcześniej omówionej metody bazującej na wyrazach:

- jest znacznie bardziej czasochłonna – na ten fakt wpływa przede wszystkim czas przygotowania wzorców,
- ma charakter dziedzinowy, a nie uniwersalny – należy oczekiwać, że będzie prawidłowo identyfikować informacje jedynie z zakresu dziedzinowego, zgodnego z tematyką skonstruowanych wzorców,
- w wielu implementacjach wymaga przyjęcia miary podobieństwa semantycznego pomiędzy tekstami (lub ich fragmentami).

Definiowanie wzorców może być oparte na:

- podejściu słownikowym – w trakcie analizy dokumentu wyszukiwane są w nim słowa lub frazy pochodzące z uwzględnianych słowników. Sposób interpretacji zidentyfikowanych w ten sposób elementów opisany jest bezpośrednio w słowniku. Tego typu podejście jest stosowane przede wszystkim przy wyszukiwaniu i właściwym zrozumieniu nazw własnych,
- przyjętej notacji formalnej pozwalającej na opisanie wzorców – notacja ta stanowi „język opisu wzorców". Wydaje się, że przy konstrukcji formalizmu pozwalającego na opis wzorców szczególnie przydatny może być mechanizm wyrażeń regularnych. Przykładem tego typu rozwiązania jest język JAPE[1].

Próbując podsumować rozważania dotyczące metody identyfikacji informacji na podstawie wzorców, należy z jednej strony podkreślić czasochłonność metody, z drugiej zaś warto wskazać na możliwość wielokrotnego wykorzystania tych samych wzorców. Można oczekiwać, że tego typu podejście sprawdzi się w przypadku analizy tekstów o jednorodnej tematyce i w miarę uporządkowanej strukturze.

2.3. Metoda bazująca na przyjętym modelu rozpatrywanego fragmentu rzeczywistości

Zaprezentowana w punkcie 1.2 metoda pozyskiwania informacji z dokumentów tekstowych oparta na wzorcach nie definiuje wzajemnych relacji pomiędzy poszczególnymi, zidentyfikowanymi w tekście faktami czy pojęciami. Realizacja tego zadania wymaga przyjęcia modelu opisującego dziedzinę, której dotyczą przekazy tekstowe. Wśród wielu metod pozwalających na budowę modeli dziedzinowych na szczególną uwagę zasługuje technologia sieci semantycznych [Davies, Studer, Warren 2006; Allemang, Hendler 2007]. Do jej podstawowych zalet należy zaliczyć:

- elastyczność,
- mocną podbudowę teoretyczną (teoria grafów),

[1] *Java Annotation Patterns Engine* – oparty na wyrażeniach regularnych język pozwalający na automatyczne definiowanie reguł znakowania istotnych elementów w tekście. Zaimplementowany m.in. w pakiecie GATE.

- dostępność implementacji pozwalających na bezproblemowe przetwarzanie stworzonych modeli przez systemy komputerowe (są to przede wszystkim implementacje oparte na języku XML),
- łatwość interpretacji zapisów opisujących poszczególne modele.

Sieć semantyczna jest modelem, w którym obiekty reprezentowane są przez wierzchołki grafu, występujące zaś pomiędzy obiektami relacje odwzorowują krawędzie grafu. Jednym z najważniejszych elementów składowych technologii sieci semantycznych są ontologie będące hierarchicznym modelem wszystkich klas rozumianych jako wzorce obiektów występujące w konkretnej sieci semantycznej. Ontologia może być więc traktowana jako struktura drzewiasta tworząca hierarchię pojęć dotyczących rozpatrywanego fragmentu rzeczywistości. Jedną z najistotniejszych cech sieci semantycznych i ontologii jest możliwość wyznaczania odległości (lub podobieństwa) pomiędzy elementami składowymi. Z punktu widzenia obliczeniowego problem ten sprowadza się do wyznaczenia odpowiedniej miary w drzewie lub grafie.

W przypadku próby wykorzystania technologii sieci semantycznych w zagadnieniach automatycznego pozyskiwania informacji z dokumentów tekstowych niezbędne jest stworzenie ontologii opisującej hierarchię zdarzeń mogących wystąpić w tekście. Następnie należałoby zdefiniować wzorce pozwalające na automatyczną identyfikację tych fragmentów tekstu, które się do nich odnoszą, i ich interpretację. Uwzględnienie opisanej przez ontologię wiedzy o hierarchii klas pozwala na określenie podobieństwa pomiędzy elementami, które zostały znalezione w dokumencie. Dogodną formą prezentacji numerycznego odwzorowania występujących pomiędzy nimi relacji jest macierz podobieństwa. Jednakże bezpośrednie wykorzystanie macierzy podobieństwa w charakterze danych wejściowych dla modelu predykcyjnego jest trudne do uzasadnienia. Z tego powodu konieczne jest przekształcenie wartości tworzących macierz podobieństwa w wektory mogące być wykorzystane jako dane wejściowe przy opisie kolejnych elementów szeregu czasowego. Za dogodną metodę realizacji tego typu przekształcenia należy uznać skalowanie wielowymiarowe. Realizując procedurę skalowania wielowymiarowego, podać należy wymiar przestrzeni, w której tworzona jest konfiguracja punktów pozostających we wzajemnych relacjach, w maksymalnym stopniu zgodny z relacjami opisanymi przez wejściową macierz podobieństwa. Współrzędne tak wyznaczonych punktów są dogodnymi kandydatami na numerycznych reprezentantów zidentyfikowanych informacji tekstowych.

3. Próba uwzględnienia informacji zawartych w komunikatach spółek notowanych na GPW w modelowaniu stóp zwrotu

Podstawowym celem przeprowadzonych badań empirycznych było opracowanie i przeprowadzenie oceny różnych metod uwzględniania informacji tekstowej w modelowaniu ekonomicznych szeregów czasowych. Analizie poddano stopy zwrotu obliczone na podstawie cen akcji notowanych na Giełdzie Papierów Wartościowych w Warszawie.

3.1. Zakres analizy i charakterystyka danych

W trakcie prac wykorzystano dane dotyczące notowań na GWP w Warszawie akcji przedsiębiorstw branży budowlanej. Analizie poddano dane z okresu od września 2008 r. do sierpnia roku 2009. Dane liczbowe pobrano z serwisu Parkiet.com. Komunikaty tekstowe dotyczące rozpatrywanego zbioru spółek pozyskano z portalu Gazeta.pl. Po pobrania komunikatów wykorzystano autorski program przygotowany w języku Java. Do analizy dokumentów HTML zawierających komunikaty wykorzystano parser Jericho-HTML. Dokumenty HTML przekształcono do postaci tekstowej. W tracie analizy wykorzystano jedynie treści zawarte w nagłówkach wiadomości. Wykorzystując dane dotyczące cen zamknięcia, wyznaczono dla każdego waloru prostą i logarytmiczną stopę zwrotu.

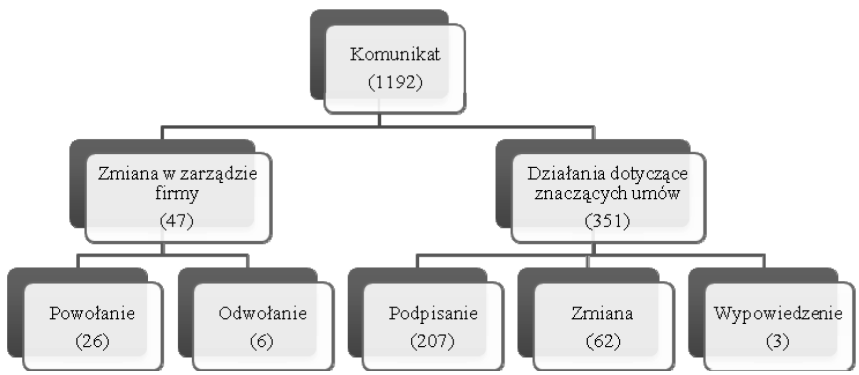
3.2. Analiza oparta na częstości występowania wyrazów w dokumentach

Obliczenia zrealizowano w pakiecie *Text Miner* programu STATISTICA. Dokonano podziału dokumentów na wyrazy. Uwzględniono tylko te słowa, które występowały przynajmniej w jednym procencie wszystkich dokumentów (nagłówków komunikatów). Utworzono macierz częstości wykorzystującą ważoną reprezentację logarytmiczną. Następnie wyznaczono składowe ukryte, stosując dekompozycję według wartości osobliwych. Wzrokowa analiza wykresu ukazującego informacyjność poszczególnych składowych była podstawą do uwzględnienia w dalszych obliczeniach pięciu kolejnych składowych. W charakterze danych wejściowych uwzględnionych przy wyznaczaniu danej wartości szeregu wykorzystano wartości wspomnianych powyżej składowych oraz składowych reprezentujących komunikat z dnia poprzedniego. W celu oceny zastosowanego sposobu reprezentacji informacji tekstowej wyznaczono współczynniki korelacji liniowej pomiędzy stopami zwrotu (prostą i logarytmiczną) a wartościami numerycznymi reprezentującymi informacje zawarte w komunikatach. Statystycznie istotne okazały się jedynie wartości pierwszych dwóch składowych reprezentujących komunikat(-y) opublikowany po zamknięciu sesji w dniu poprzedzającym aktualną sesję.

Zaletą prezentowanej metody jest możliwość wyznaczenia wartości informacyjnej poszczególnych wyrazów. W charakterze miary określającej ważność poszczególnych słów przyjęto odległość pomiędzy punktem opisującym położenie wyrazu w przestrzeni wyznaczonej przez zastosowanie SVD a początkiem układu współrzędnych (im bardziej rozpatrywany wyraz jest oddalony od początku układu współrzędnych, tym większe jest jego znaczenie). Do najistotniejszych wyrazów należy zaliczyć (w kolejności zgodnej z malejącym znaczeniem): *umowy, raportu, przez, zawarcie, akcji, formularz, znaczącej, emitenta, roku, spółki, skonsolidowanego, informacja*, *zależną.*

3.3. Analiza oparta na przyjętym modelu

Analiza treści poszczególnych komunikatów może stanowić podstawę do konstrukcji formalnego modelu opisującego analizowaną dziedzinę. Na potrzeby niniejszej pracy skonstruowano prosty model przedstawiający relacje pomiędzy wybranymi zdarzeniami opisywanymi w komunikatach spółek. Na rysunku 1 przedstawiono model w postaci graficznej.



Rys. 1. Przyjęta ontologia pojęć występujących w komunikatach ze spółek

Źródło: opracowanie własne.

Liczby umieszczone w nawiasach wskazują na liczbę komunikatów, w których zidentyfikowano dane zdarzenie. Liczba wszystkich analizowanych komunikatów wynosiła 1192. Wartości te stanowiły podstawę do wyznaczenia podobieństw pomiędzy zdarzeniami zgodnie z formułą zaproponowaną przez Lina:

$$sim(C_1,C_2)=\frac{2\times\log\left(P\left(C_0\right)\right)}{\log\left(P\left(C_1\right)\right)+\log\left(P\left(C_2\right)\right)}, \tag{1}$$

gdzie: C_1, C_2 – porównywane klasy,
C_0 – najbliższy wspólny przodek porównywanych klas,
$P(C_i)$ – prawdopodobieństwo pojawienia się komunikatu o wystąpieniu zdarzenia klasy C_i.

Uzyskaną macierz podobieństwa przedstawia tab. 1.

Następnie, stosując metodę skalowania wielowymiarowego, skonstruowano zbiór punktów w przestrzeni R^2 w taki sposób, aby podobieństwo pomiędzy nimi było w maksymalnym stopniu zgodne z podobieństwem pomiędzy zdarzeniami. Uzyskane w ten sposób punkty mogą reprezentować poszczególne zdarzenia. Współrzędne punktów przedstawia tab. 2.

Wzajemne relacje pomiędzy punktami (będącymi odpowiednikami poszczególnych zdarzeń) przedstawia rys. 2.

Tabela 1. Macierz podobieństwa pomiędzy rozpatrywanymi zdarzeniami

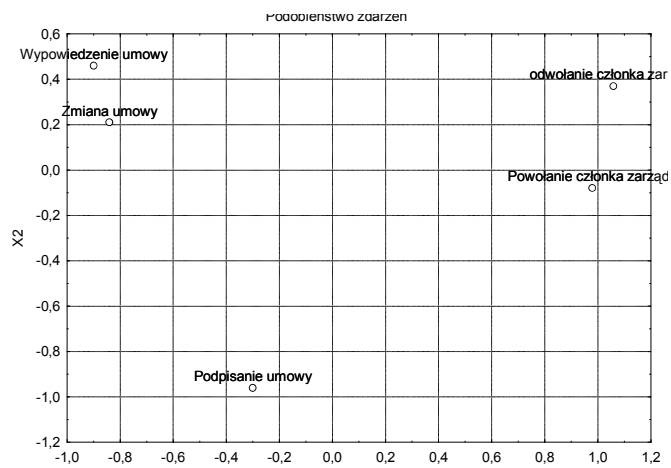
	Podpisanie umowy	Zmiana umowy	Wypowiedzenie umowy	Powołanie członka zarządu	Odwołanie członka zarządu
Podpisanie umowy	1,00	0,72	0,52	0,37	0,32
Zmiana umowy	0,72	1,00	0,80	0,33	0,29
Wypowiedzenie umowy	0,52	0,80	1,00	0,25	0,23
Powołanie członka zarządu	0,37	0,33	0,25	1,00	0,79
Odwołanie członka zarządu	0,32	0,29	0,23	0,79	1,00

Źródło: obliczenia własne.

Tabela 2. Współrzędne zdarzeń w przestrzeni wyznaczonej przez zastosowanie skalowania wielowymiarowego

	X_1	X_2
Podpisanie umowy	–0,30	–0,96
Zmiana umowy	–0,84	0,21
Wypowiedzenie umowy	–0,90	0,46
Powołanie członka zarządu	0,98	–0,08
Odwołanie członka zarządu	1,06	0,37

Źródło: obliczenia własne.



Rys. 2. Podobieństwo zdarzeń występujących w komunikatach

Źródło: obliczenia własne.

Współrzędne punktów mogą służyć jako numeryczna reprezentacja informacji pochodzących z komunikatów dotyczących funkcjonowania spółek giełdowych. Również w tym przypadku stwierdzono słabą, ale statystycznie istotną korelację pomiędzy wartościami stóp zwrotu a wartościami liczbowymi przyjętymi w rozpatrywanej metodzie reprezentacji informacji tekstowej.

4. Podsumowanie

Duży stopień upowszechnienia informacji tekstowej w pełni uzasadnia próby jej wykorzystania w modelowaniu ekonomicznych szeregów czasowych. Trudno jest podać jedną, ogólną metodę uwzględniania informacji tekstowych. Z całą pewnością zależy to od charakteru przetwarzanego zestawu dokumentów.

Metoda oparta na częstości występowania wyrazów jest łatwa w zastosowaniu, nie wymaga definiowania wzorców wypowiedzi, nie wymaga budowy bazy przykładowych dokumentów, jest szybka, może być stosowana do tekstów zróżnicowanych tematycznie (przy czym znajomość tematyki tekstu w chwili przystępowania do realizacji obliczeń nie jest konieczna). Niestety jej wadą są częste problemy z pozyskaniem istotnych informacji z dokumentów.

Metody oparte na modelach rozpatrywanej dziedziny wydają się lepsze. Ale ich stosowanie jest znacznie bardziej czasochłonne (definiowanie wzorców, przygotowanie przykładów), są przystosowane do tekstów o charakterze dziedzinowym (przy czym konieczna jest wcześniejsza znajomość dziedziny problemu). Właściwie przygotowane mogą być wielokrotnie stosowane.

Należy podkreślić, że przedstawione w pracy wnioski nie mają z pewnością charakteru ostatecznego, natomiast wskazane problemy wymagają dalszych badań.

Literatura

Allemang D., Hendler J., *Semantic Web for the Working Ontologist*, Morgan Kaufmann Publisher, 2007.

Davies J., Studer R., Warren P., *Semantic Web Technologies. Trends and Research in Ontology-based Systems*, John Wiley & Sons, 2006.

Lula P., *Text Mining jako narzędzie pozyskiwania informacji z dokumentów tekstowych*, Statsoft, http://www.statsoft.pl/czytelnia/8_2007/Lula05.pdf, 2005.

Źródło internetowe

[1] http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.knowledgeRush.html.

INFORMATION RETRIEVAL FROM TEXT DOCUMENTS FOR STATISTICAL PREDICTIVE MODELS

Summary: The paper discusses the problem of information retrieval from text documents for statistical predictive models. Three approaches are discussed: word-based, pattern-based and ontology-based. During empirical research the problem of information retrieval from announcements published by companies listed on the Warsaw Stock Exchange was studied.