PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU
Nr 7 (1207) 2008
TAKSONOMIA 15
Klasyfikacja i analiza danych – teoria i zastosowania

Paweł Lula
Uniwersytet Ekonomiczny w Krakowie

EKSPLORACYJNA ANALIZA POLSKOJĘZYCZNYCH OFERT ZATRUDNIENIA SPECJALISTÓW Z ZAKRESU INFORMATYKI

1. Wstęp

Niemal w każdej dziedzinie życia człowieka zauważa się dynamiczny rozwój dostępnych zasobów informacyjnych. Zjawisko to można analizować w aspekcie zarówno ilościowym (biorąc pod uwagę ilość udostępnianych zasobów), jak i w jakościowym (uwzględniając jakość informacji, formę jej reprezentacji lub sposób dostarczenia). Rozważania przedstawione w niniejszej pracy koncentrują się na zagadnieniu wspomagania człowieka w procesie pozyskiwania informacji z zasobów tekstowych. Potrzeba automatyzacji tego typu zadań wynika m.in. ze stałego zwiększania się ilości i znaczenia zasobów informacyjnych o charakterze tekstowym (raporty, oceny, ekspertyzy, doniesienia agencyjne, korespondencja, zasoby internetowe) oraz wysokiego kosztu angażowania człowieka–eksperta w pozyskiwanie i analizę informacji pozyskiwanych z zasobów tekstowych.

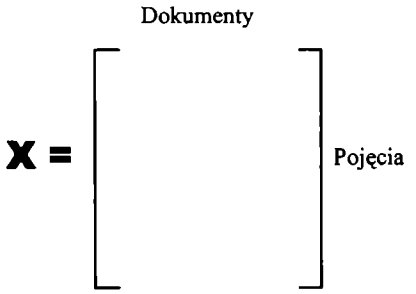
Praca zawiera przede wszystkim rozważania dotyczące metod przydatnych do pozyskiwania informacji z polskojęzycznych dokumentów tekstowych. Przedstawiono metody bazujące na dwóch odmiennych podejściach do zagadnienia reprezentacji informacji dostępnych w dokumentach, a mianowicie na reprezentacji w postaci listy słów oraz reprezentacji ontologicznej. Za podstawę do sformułowania wniosków dotyczących ocen stosowanych metod analizy przyjęto zaprezentowane w pracy wyniki badań dotyczące zautomatyzowanej analizy polskojęzycznych ofert pracy dla specjalistów z zakresu informatyki[1].

[1] Należy podkreślić, że przeprowadzone badania ukierunkowane były na dokonanie oceny metod analizy, a wykorzystane teksty ofert traktowane były jedynie jako przykładowe dokumenty dotyczące jednej wyodrębnionej dziedziny aktywności człowieka. Z uwagi na subiektywny dobór źródeł ofert nie należy w żaden sposób odnosić uzyskanych rezultatów do sytuacji na polskim rynku pracy.

Struktura pracy przedstawia się następująco. Pierwszy punkt pracy poświęcony jest metodzie reprezentacji informacji wykorzystującej listę słów. W kolejnym punkcie zaprezentowano podstawy reprezentacji ontologicznej. W trzecim punkcie przedstawiono charakterystykę materiału wykorzystywanego w trakcie badań. Punkt czwarty zawiera opis przeprowadzonych eksperymentów i prezentuje uzyskane wyniki. Pracę kończą wnioski.

2. Reprezentacja informacji bazująca na liście słów

Celem operacji przedstawionych w bieżącym punkcie artykułu jest przekształcenie kolekcji przetwarzanych dokumentów do postaci macierzy częstości (rys. 1). Kolumny macierzy reprezentują poszczególne dokumenty wchodzące w skład przetwarzanej kolekcji, zaś wiersze odpowiadają pojęciom występującym w dokumentach.



Rys. 1. Macierz częstości

Źródło: opracowanie własne.

Elementami macierzy częstości są wartości x_{ij} określające liczbę wystąpień i-tego pojęcia w j-tym dokumencie. Kluczowym etapem w procedurze wyznaczania macierzy częstości jest określenie listy uwzględnianych pojęć. Zwykle problem ten jest rozwiązywany poprzez realizację następujących kroków [Lula 2005]:

- przygotowanie zbioru dokumentów,
- podział dokumentów na wyrazy,
- wstępne przetworzenie list wyrazów obejmujące:
  - sprowadzenie do formy podstawowej (redukcja do rdzenia),
  - usunięcie wyrazów nieistotnych (wchodzących w skład stop-listy),
- skonstruowana w ten sposób lista wyrazów stanowi listę pojęć wykorzystywanych w kolejnych krokach do opisu zawartości dokumentów.

Wyznaczona macierz częstości stanowi numeryczną reprezentację poddawanych analizie tekstów. Większość badaczy uważa jednak, że powinna ona stanowić punkt wyjścia do kolejnych przekształceń mających na celu poprawę prawidłowo-

ści reprezentowania informacji zawartych w dokumentach. Wspomniane metody dalszego przekształcania macierzy częstości polegają zazwyczaj na:

- zastosowaniu reprezentacji binarnej – w tym celu niezerowe elementy macierzy częstości zastępowane są jedynkami, natomiast elementy zerowe pozostają niezmienione;
- zastosowaniu reprezentacji logarytmicznej – decydując się na tę metodę reprezentacji, należy zlogarytmować elementy niezerowe;
- zastosowaniu ważonej reprezentacji logarytmicznej, w której wartości zlogarytmowane są ważone w sposób preferujący słowa występujące w niewielkiej liczbie dokumentów;
- zastosowaniu dekompozycji macierzy według wartości osobliwych (dekompozycja SVD). Operacja ta może dotyczyć oryginalnej macierzy częstości lub macierzy uzyskanej w wyniku zastosowania jednej z przedstawionych powyżej formuł. Zastosowanie dekompozycji SVD pozwala na skonstruowanie nowego zbioru pojęć wykorzystanego do opisu dokumentów. Ze względu na zagregowany charakter nowych pojęć obserwuje się redukcję wymiaru przestrzeni, w której opisywane są dokumenty.

Próbując dokonać oceny przedstawionego powyżej sposobu reprezentacji informacji, należy wskazać na pozytywne jego cechy, do których zalicza się:

- uzyskanie reprezentacji w postaci macierzy wartości numerycznych, a więc w postaci dogodnej do dalszej analizy za pomocą klasycznych metod analizy danych,
- stosunkowo prosty sposób realizacji obliczeń zmierzających do skonstruowania macierzowej reprezentacji kolekcji dokumentów.

Nie należy jednak zapominać o wadach reprezentacji w postaci oryginalnej lub przekształconej macierzy częstości:

- reprezentacja macierzowa nie uwzględnia w żaden sposób informacji o kolejności pojęć w dokumentach źródłowych,
- nie jest też w stanie oddać różnic w relatywnej ważności zidentyfikowanych pojęć.

Reprezentacja wykorzystująca macierz częstości jest obecnie powszechnie wykorzystywana przez twórców pakietów obliczeniowych, co z kolei powoduje, że odgrywa ona dominującą rolę w zastosowaniach praktycznych.

3. Ontologiczna reprezentacja informacji

W zagadnieniach pozyskiwania, przechowywania i przetwarzania informacji pojęcie ontologii rozumiane jest jako sformalizowany opis wyodrębnionego fragmentu rzeczywistości. Dziedzinowy charakter ontologii wart jest podkreślenia, gdyż zawsze jest ona modelem fragmentu świata i nie pretenduje do roli narzędzia opisu całej otaczającej nas rzeczywistości. Ontologiczny model rozpatrywanej części świata składa się z obiektów, charakteryzujących je cech oraz relacji opisujących związki między obiektami. Poszukiwania struktury danych właściwej do przechowywania informacji zawartych w dokumentach jednoznacznie wskazują na

potrzebę zastosowania grafów, w których węzły byłyby odzwierciedleniem obiektów, krawędzie zaś wskazywałyby rodzaj powiązania istniejącego między połączonymi przez krawędź węzłami. Zarówno w rozważaniach teoretycznych, jak i praktycznych dość często spotyka się szczególny przypadek grafu, jakim jest struktura drzewiasta. Drzewo jest doskonałym narzędziem opisu zależności między obiektami w sytuacji, w której wszystkie istniejące zależności mają charakter hierarchiczny. Są one odzwierciedlone poprzez relacje między węzłami drzewa. Zastosowanie ontologicznej reprezentacji informacji zawartych pierwotnie w kolekcji dokumentów wymaga:

- zdefiniowania ontologii właściwej do opisu zagadnień charakterystycznych dla dziedziny, której dotyczą dokumenty. W tym celu należy zidentyfikować kluczowe obiekty, ich cechy oraz relacje zachodzące między nimi;
- opisania ontologii w przyjętym języku opisu ontologii – w chwili obecnej wiele narzędzi tego typu definiowanych jest w postaci aplikacji języka XML;
- identyfikacji w przetwarzanej kolekcji dokumentów obiektów, cech i relacji charakterystycznych dla stworzonej ontologii;
- opracowania metod przetwarzania tak reprezentowanych informacji; w większości przypadków zastosowanie klasycznych metod analizy danych wymaga wprowadzenia modyfikacji w zakresie sposobów wyznaczania odległości lub podobieństwa uwzględniających fakty wynikające bezpośrednio z przyjętej ontologii.

Już wstępna analiza podejścia ontologicznego wskazuje na znacznie większą jego złożoność w porównaniu z podejściem bazującym na liście słów. Z tego powodu za konieczne należy uznać przeprowadzenie badań mających służyć ocenie, czy ze zwiększonej złożoności wynika poprawa jakości uzyskiwanych rezultatów.

4. Charakterystyka materiału badawczego

Przetwarzane były polskojęzyczne oferty dotyczące zatrudnienia informatyków pochodzące z portalu ***Praca*** z serwisu www.gazeta.pl. Analizie poddano fragmenty ofert zawierające charakterystykę wymagań (nie przetwarzano charakterystyki pracodawcy). Przetwarzany zbiór liczył 94 oferty, wśród których występowały ogłoszenia zawierające dokładny opis wymagań. Przykładem może być następujący tekst:

Utrzymanie infrastruktury teleinformatycznej w CEMEX Polska. Zarządzanie zespołem utrzymania infrastruktury i zespołem wsparcia dla użytkowników. Wdrażanie lokalnych i globalnych projektów informatycznych w zakresie infrastruktury. Określanie kierunków rozwoju, planowanie, budżetowanie i realizacja. Współpraca z dostawcami i firmami zewnętrznymi. Od Kandydatów oczekujemy: Kilkuletniego doświadczenia w pracy na podobnym stanowisku w działach IT, w tym: pięcioletniego doświadczenia w kierowaniu podległym zespołem pracowników. Minimum dwuletniego doświadczenia w zarządzaniu projektami informatycznymi. Dobrej znajomości technologii informatycznych, środowisk sieciowych, aspektów

telekomunikacyjnych i bezpieczeństwa. Wyższego wykształcenia na kierunku informatyka lub pokrewnym. Znajomość ITIL w praktyce oraz doświadczenie w pracy w koncernie międzynarodowym mile widziane. Bardzo dobrej znajomości języka angielskiego w mowie i piśmie. Umiejętności pracy w stresie oraz pod presją czasu.

W kolekcji dokumentów znalazły się również znacznie krótsze opisy wymagań stawianych poszukiwanemu pracownikowi, czego przykładem mogą być przytoczone poniżej trzy oferty:
Administrator IT.
Serwisanta sieci LAN. Montera instalacji teletechnicznych.
Grafika komputerowego ze znajomością programów graficznych Adobe Ilustrator Photoshop.

W trakcie wstępnego przygotowania dokumentów wszystkie oferty przekształcono do unikodowego formatu tekstowego.

5. Realizacja obliczeń

Pierwszy etap obliczeń wykorzystywał reprezentację bazującą na liście słów. Słowa występujące w przetwarzanej kolekcji dokumentów przekształcono do formy podstawowej, a następnie zastosowano stop-listę w celu wyeliminowania wyrazów o niewielkiej wartości informacyjnej. Uzyskana w ten sposób lista wyrazów liczyła 118 elementów (informacje dotyczące występowania tych właśnie wyrazów stały się narzędziem reprezentacji informacji zawartych w dokumentach). Następnie wyznaczono macierz częstości, którą w kolejnych krokach poddawano dalszym transformacjom. Opisywany eksperyment powtórzono siedmiokrotnie, stosując następujące sposoby reprezentacji:

- macierz częstości w postaci niezmodyfikowanej,
- reprezentację binarną,
- ważoną reprezentację logarytmiczną,
- macierze uzyskane w wyniku zastosowania redukcji opartej na dekompozycji według wartości osobliwych (uwzględniono cztery wersje wykorzystujące 3, 7, 25 oraz 49 składowych).

W każdym przypadku uzyskane dane poddano klasyfikacji realizowanej za pomocą algorytmu Warda. Po uzyskaniu pełnego dendrogramu wybrano podział optymalny (wskazywany przez najdłuższą odległość w dendrogramie pomiędzy dwoma kolejnymi krokami podziału). Wyniki uzyskane dla rozpatrywanych kolejno metod reprezentacji informacji należy uznać za zbliżone. Duże różnice dało się zauważyć tylko w przypadku wykorzystującym jedynie trzy składowe SVD (wydaje się, że odmienne wyniki klasyfikacji mogą być spowodowane zbyt małą liczbą uwzględnionych składowych). We wszystkich pozostałych przypadkach za podział optymalny uznany został podział na dwie klasy. Po przeanalizowaniu treści ofert przypisanych do każdej z grup w kilkunastu przypadkach trudno jest w sposób

logiczny uzasadnić prawidłowość klasyfikacji. Wydaje się, że na nieuzyskanie w pełni prawidłowych wyników duży wpływ miała długość tekstu oferty (w jednej grupie przeważają oferty długie, w drugiej krótkie). Można też zauważyć, że większość pomyłek dotyczyła zatrudnienia grafików – zostały one przypisane do obu grup. Wyniki wskazują na potrzebę wypracowania innej metody klasyfikacji ofert dotyczących zatrudnienia.

W kolejnym etapie badań zastosowano podejście bazujące na ontologii. Prace rozpoczęto od zdefiniowania formalnego modelu pracownika. Głównym celem modelu było opracowanie zestawu cech potencjalnego pracownika istotnych z punktu widzenia rekrutacji pracowników na stanowiska informatyczne. Przyjęto, że rozpatrywany układ cech ma strukturę hierarchiczną.

Pracownik

WymaganiaInne

CechyCharakteru:	*Komunikatywnosc, Kreatywnosc, OdpornoscNaStres, Samodzielnosc, Zaangazowanie*
Doswiadczenie	
JezykiObce:	*Angielski, Niemiecki*
PrawoJazdy	

WymaganiaIT

BazyDanych:	*Bazy, MySQL, Oracle, SQL*
Grafika:	*Adobe, Corel, PhotoShop*
Internet:	*Flash, HTML, PHP, XHTML, XML*
Programowanie:	*J2EE, Java, JavaScript, NET, Programista, Visual*
RealizacjaProjektow:	*UML, Wdrozenia*
SieciKomputerowe:	*LAN, Serwer, WAN*
Sprzet:	*Elektronika, Telekomunikacja*
SystemyOperacyjne:	*Linux, Unix, Windows*

Drzewo prezentujące zaproponowaną ontologię liczy 50 węzłów. Pełna definicja zapisana została w języku OWL. Następnie wyznaczono miary podobieństwa między zdefiniowanymi pojęciami. Zastosowano w tym celu miernik podobieństwa zaproponowany przez D. Lina [1998]. Uzyskana macierz podobieństwa miała wymiary 50×50 elementów.

Następnie dla każdego pojęcia zdefiniowano warunki, dzięki którym można było sprawdzić, czy w tekście poddawanym analizie jest mowa o rozpatrywanym pojęciu. Warunki te zapisywane były w postaci układu instrukcji sprawdzających fakt wystąpienia lub niewystąpienia odpowiednio dobranych słów kluczowych. Efektem wykonanych operacji było wyznaczenie dla każdego dokumentu wektora binarnego o długości równej liczbie rozpatrywanych pojęć (czyli 50). Wystąpienie jedynki na i-tej pozycji wskazuje na wystąpienie w dokumencie i-tego pojęcia. Następnie wyznaczono podobieństwo między dokumentami, obliczając podobień-

stwo między reprezentującymi je wektorami binarnymi. Podobieństwo to wyznaczono jako uśrednioną wartość podobieństwa liczoną między każdym pojęciem występującym w jednym dokumencie a pojęciami występującymi w drugim tekście (do wyrażenia podobieństw cząstkowych zastosowano wyznaczone wcześniej współczynniki podobieństwa Lina między pojęciami). Macierz podobieństw między dokumentami stała się punktem wyjścia do przeprowadzenia klasyfikacji za pomocą metody Warda. Podejście ontologiczne doprowadziło też do wyodrębnienia dwóch grup dokumentów. Uzyskano znacznie lepsze wyniki niż w przypadku analiz wykorzystujących macierz częstości. W sposób prawidłowy zaklasyfikowane zostały oferty dotyczące zatrudnienia grafików (wszystkie znalazły się w jednej klasie). Natomiast druga klasa zawierała przede wszystkim oferty dotyczące zatrudnienia programistów i specjalistów od sieci komputerowych.

6. Wnioski końcowe

Przeprowadzone eksperymenty jednoznacznie wskazują, że klasyfikacja bazująca na ontologii daje lepsze wyniki. Nie należy jednak zapominać, że jest ona o wiele bardziej kosztowna (konieczność przygotowania ontologii, zdefiniowania sposobu określenia podobieństwa między klasami, określenia sposobu identyfikacji w dokumentach informacji reprezentowanych przez klasy ontologii). Pozytywnym aspektem korzystania z ontologii jest natomiast możliwość ich wielokrotnego użycia, co może pozwolić na wypracowanie jednolitych formuł wyrażania podobieństwa między hierarchicznymi układami pojęć.

Literatura

Lin D. (1998), *An Information-theoretic Definition of Similarity*, Department of Computer Science University of Manitoba, Winnipeg, Manitoba, Canada R3T 2N2, http://www.cs.ualberta.ca/~lind ek/papers/sim.pdf.

Lula P. (2005), *Klasyfikacja dokumentów tekstowych sporządzonych w języku polskim*, Taksonomia 12, red. K. Jajuga, M. Walesiak, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 1076, AE, Wrocław.

TEXT MINING ANALYSIS OF POLISH JOB OFFERS FOR IT SPECIALISTS

Summary

The problem of document clustering is the main topic of the paper. Two approaches to text clustering are studied. The first one is based on the bag-of-words representation. And the second uses a user-defined ontology. The results show that ontology-based approach is more complex than methods based on the bag-of-words representation but it is far superior to them.