STUDIA INFORMATICA 2011
Volume 32 Number 2A (96)

Jakub CIEŚLEWICZ, Adam PELIKANT
Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych

ZASTOSOWANIE SIECI JĘZYKOWYCH W REPREZENTACJI DOKUMENTÓW TEKSTOWYCH

Streszczenie. Artykuł opisuje zastosowanie sieci językowych w reprezentacji dokumentów tekstowych. Przedstawia dwa modele reprezentacji: statystyczny oraz z wykorzystaniem sieci językowych. Opiera się na przeprowadzonej analizie literaturowej, której celem było poszukiwanie wydajnej metody reprezentacji dokumentów, mającej służyć do dalszych badań w dziedzinie wyszukiwania dokumentów tekstowych na podstawie rzeczywistych treści.

Słowa kluczowe: zgłębianie tekstu, sieci językowe, reprezentacja dokumentu

USING LANGUAGE NETWORK FOR TEXT DOCUMENT REPRESENTATION

Summary. The article propels the problem of building the model of continuous texts representation. It presents mechanisms of weights assignment to the individual document features based on statistical analysis and text networks. The review on document representation is the first step to investigation into searching documents.

Keywords: text mining, text network, document representation

1. Wstęp

Szeroko rozumiana informatyzacja wraz z dynamicznym rozwojem nowych technologii i Internetu powoduje niewyobrażalny przyrost informacji w zbiorach danych. W obecnych czasach składowanie ogromnych baz danych nie stanowi już kłopotu. Problemem jest brak wydajnych narzędzi do wydobywania z nich wiedzy czy wyszukiwania istotnych dla użytkownika treści, przy akceptowalnym czasie oczekiwania i trafności wyników. Jedną z najbardziej naturalnych dla człowieka form przekazywania i zapisywania wiedzy jest posługi-

wanie się opisowym językiem naturalnym w formie najróżniejszych dokumentów tekstowych. Artykuł powstał jako podsumowanie przeprowadzonej analizy literaturowej, której celem było poszukiwanie wydajnych struktur reprezentacji dokumentów tekstowych, nadających się do dalszych badań w temacie wyszukiwania dokumentów tekstowych, bazującym na rzeczywistej treści.

Przetwarzanie ciągłych tekstów z wykorzystaniem komputerów jest bardzo skomplikowane w porównaniu do operowania na danych prostych. Jednym z najistotniejszych problemów analizy dokumentów jest opracowanie odwzorowania, które przy stosunkowo prostej strukturze zawierać będzie możliwie największą ilość relewantnych informacji, uzyskanych w wyniku procesu ekstrakcji cech dokumentu. Jakość wyselekcjonowanych cech ma wpływ na trafność uzyskiwanych rezultatów na dalszym etapie wydobywania danych.

Celem niniejszego artykułu jest przedstawienie metod przygotowania reprezentacji dokumentów, które znajdują zastosowanie w zaawansowanych przetwarzaniach. Krótko zostanie scharakteryzowana najpowszechniej stosowana metoda analizy statystycznej, która zastosowana została w pierwszym etapie badań nad metodami wyszukiwania dokumentów tekstowych. Druga część wprowadzi w tematykę sieci językowych, dzięki którym możliwe jest zachowanie w modelu informacji na temat wybranych relacji pomiędzy wyselekcjonowanymi cechami.

2. Statystyczne analizy zbiorów dokumentów

Istnieje wiele prostych metod, które pozawalają na przygotowanie dokumentu do analizy w kontekście całego zbioru, przy stosunkowo niewielkim nakładzie mocy obliczeniowej. Najbardziej popularnym podejściem jest zastosowanie reprezentacji wektorowej, w której współrzędne wektorów odpowiadają wyekstrahowanym cechom (słowom) i określają jej istotność w dokumencie albo nawet w całym zbiorze. Przed wyznaczeniem wag, niezależnie od przyjętej metody, tekst taki należy poddać wstępnemu przetwarzaniu. W skład działań wchodzących w ten etap są między innymi: identyfikacja typu dokumentu i wersji językowej, ujednolicenie wielkości liter, separacja wyrazów, n-gramami słów, a nawet całych zdań w zależności od przyjętych założeń. Niektóre metody wymagają również eliminacji słów, których obecność podyktowana jest zasadami gramatycznymi i koniecznością zachowania logicznej formy przekazu. Wyrazy te stanowią duży procent wszystkich słów w tekście i przez to przodują w statystykach. Obecność w każdym dokumencie powoduje, że ich siła dyskryminacyjna jest zerowa. Wyrazy te zebrane są na tzw. stop-listach i zaliczają się do nich np. spójniki, przyimki, zaimki.

Jak już zostało wspomniane w modelu wektorowym, współrzędna odpowiada istotności cechy w obrębie dokumentu albo w całym zbiorze i jest nazywana częstością (ilość). Jedną z najprostszych metod, która wykorzystuje analizę statystyczną jest reprezentacja Boolowska, która dla cechy, skojarzonej ze współrzędną wektora, występującej w dokumencie ustala wagę równą 1, a dla cechy obecnej w całym zbiorze, niezwiązaną z analizowaną treścią, przypisuje wartość 0. W konsekwencji otrzymujemy reprezentację na podstawie, której możemy mylnie wywnioskować o tematyce tekstu, gdyż obecność słowa nie musi świadczyć o faktycznej treści, szczególnie, jeśli dane słowo użyte jest w konstrukcji z zaprzeczeniem.

Uszczegółowieniem poprzedniej metody jest metoda ilościowa, która w ustalaniu wagi uwzględnia częstość słowa (*Term Frequency* – TF). Metoda ilościowa pozwala na wyznaczenie wagi bezwzględnej, określającej tylko liczbę wystąpień słowa w dokumencie, lub względnej wyrażonej jako stosunek liczby wystąpień wyrazu w tekście, do liczby wszystkich relewantnych cech. Dzięki metodzie względnej reprezentacje dokumentów krótkich i długich są podobne. W przeciwnym wypadku należy zastosować normalizację bądź wykorzystać metrykę, która nie jest czuła na bezwzględne wartości współrzędnych.

Najdokładniejsze wartości współrzędnych, w kontekście wyszukiwania dokumentów w zbiorze, otrzymuje się stosując odwrotną częstość dokumentu TF-IDF (*Term Frequency – Inverse Document frequency*):

$$TFIDF(i) = TF(i) \cdot IDF(i) = \frac{N_i}{N} \cdot \ln \frac{D}{D_i}, \tag{1}$$

gdzie N określa łączną liczbę cech danego rodzaju (słów, n-gramów), N_i informuje o liczbie wystąpień konkretnej cechy w danym dokumencie, D to liczba wszystkich dokumentów w zbiorze, a D_i określa liczbę dokumentów zawierających cechę i.

Stosowanie wag TFIDF eliminuje efekt przeszacowania. Oprócz częstości cechy w dokumencie d_i pod uwagę brana jest również obecność tej cechy w innych dokumentach. Z własności logarytmu naturalnego wynika, że siła dyskryminacyjna dla wyrazu występującego w niewielkiej liczbie dokumentów jest większa, niż dla słowa występującego w znacznej części zbioru tekstów. Współczynnik TFIDF spełnia rolę podobną do entropii w zagadnieniach zgłębiania danych.

Posiadanie wektorowej reprezentacji poszczególnych dokumentów umożliwia przygotowanie reprezentacji zbioru N dokumentów tekstowych w postaci macierzy nazywanej TFM (*Term Frequency Matrix*), której elementy $TFM[d_i,n_i]$ opisują wagę cechy n_i w dokumencie d_i. Reprezentacje wektorowe poszczególnych tekstów są wektorami o różnej długości, stąd dla cech występujących w dokumencie d_i, nieobecnych w dokumencie d_{i+1}, ustala się wagę 0. W związku z tym macierz TMF jest macierzą rzadką.

Tak przygotowana reprezentacja stanowi podstawę do przeprowadzenia automatycznego grupowania, klasyfikowania czy wyszukiwania dokumentów w zbiorze, bazując na jego rzeczywistej treści, a nie wąskiej grupie sztucznie określonych słów kluczowych.

3. Sieci językowe

Rozwój dziedziny *text miningu* stawia coraz większe wymagania przed mechanizmami wydobywania informacji ze zbioru tekstowego. Przedstawiona powyżej krótka charakterystyka reprezentacji wektorowej, opartej na metodzie zliczania, jest stosunkowo prosta w realizacji, nie wymusza tworzenia skomplikowanych parserów i analizatorów treści, a wyznaczenie istotności cechy w zbiorze nie wymaga przeprowadzania trudnych i długotrwałych obliczeń. Jednak ogólnoświatowy trend w poszukiwaniu inteligentnych algorytmów zgłębiania wiedzy powoduje, iż znaczenie powyższej metody maleje ze względu na brak zachowania logicznych, a nawet semantycznych relacji w modelu reprezentacji. Wyodrębnienie poszczególnych cech dokumentu i traktowanie ich jako osobnych jednostek, nieposiadających wzajemnych relacji między sobą powoduje, iż tracone są istotne informacje, wynikające z zasad i reguł rządzących językami naturalnymi.

Język naturalny jest pewnego rodzaju systemem ewoluującym, w którym obecne struktury są determinowane przez wcześniejszą ewolucję [4]. System ten jest bardzo złożony i można traktować go jako rozrastającą się sieć słów, będących w interakcjach. W mowie słowa nie występują pojedynczo, zawsze pozostają w relacji z innymi słowami tak, aby wypowiedź zawierała precyzyjne informacje. Sama ewolucja sieci językowej polega zaś na pojawianiu się w niej nowych wyrazów, które zaczynają współistnieć z już obecnymi elementami sieci lub polega na pojawianiu się nowych powiązań pomiędzy istniejącymi elementami sieci.

Bazując na teorii ewoluującego systemu przygotowanie reprezentacji dokumentu, w której zachowane są relacje pomiędzy poszczególnymi cechami dokumentu zrealizować można za pomocą mechanizmu tzw. sieci językowych, których idea opiera się na teorii grafów. Sieci te, podobnie jak fizyczne (np. komputerowe,) są systemem komponentów, zwanymi węzłami (*ang. node*), pomiędzy którymi występują połączenia (ang. *link*). W przypadku analizy dokumentów nie mamy do czynienia z sieciami, w których da się wskazać fizyczne połączenie między wierzchołkami. W sieciach językowych połączenia są abstrakcyjnymi relacjami, opisującymi zależności pomiędzy węzłami, za które przyjmuje się wyrazy, natomiast relacje odzwierciedlają zależności między nimi, wynikające czysto z zasad i reguł językowych (np. składnia, semantyka). W zależności od relacji pomiędzy wyrazami możliwe jest zbudowanie kilku rodzajów sieci.

Sieci językowe tak, jak inne sieci złożone charakteryzują się podobnymi własnościami co sieci małych światów (*small-word*) oraz sieci dowolnie skalowane (*scale-free*) [4]. *Small-world* jest typem grafu, w którym większość węzłów bezpośrednio ze sobą nie sąsiaduje, ale można do nich przejść z dowolnego punktu w niewielkiej liczbie kroków. Przyjmuje się, że odległość (liczba przejść przez pośrednie węzły) pomiędzy dowolnie wybranymi wierzchołkami wzrasta proporcjonalnie do logarytmu liczby węzłów w sieci [5].

W sieci *scale-free* występują węzły z ogromną liczbą połączeń, a także węzły, które mają krawędzie z niewielką liczbą wierzchołków. Węzły, które wskazują na bardzo dużo innych węzłów, a także te, które są tak samo licznie wskazywane przez inne wierzchołki nazwane są *koncentratorami* (ang. *hub*).

W zależności od przyjętej metody wyznaczania relacji pomiędzy wyrazami, w sieciach językowych koncentratory mogą pełnić różne funkcje. Czasem może zdarzyć się tak, że ze względu na małą istotność koncentratory muszą zostać usunięte. W literaturze najczęściej spotykanymi typami sieci językowych są sieci współwystępowania, syntaktyczne lub regułowe oraz semantyczne, a ich nazwy wynikają z typu interakcji między wyrazami.

3.1. Sieci współwystępowania

Sieci współwystępowania są najprostszą formą odwzorowania. Idea budowy zależności pomiędzy wyrazami opiera się na występowaniu dwóch wyrazów w tym samym przedziale bądź oknie. W przypadku języka mówionego naturalnym zakresem, w którym badane jest sąsiedztwo wyrazów jest zdanie rozumiane jako ciąg wyrazów zakończonych odpowiednim separatorem. Jeśli dwa wyrazy występują w bezpośrednim sąsiedztwie (w analizowanej jednostce tekstu) to znaczy, że są ze sobą połączone. Wyróżniane są dwa rodzaje relacji: niezorientowana i zorientowana. W zależnościach niezorientowanych informacja o kolejności słów w zadaniu jest traktowana jako nadmiarowa, a przyjęcie takiego postępowania sprawia, że ten typ analizy charakteryzuje się tymi samymi słabymi stronami, co analiza statystyczna. Dlatego w odniesieniu do języka naturalnego bardziej realistyczne jest posługiwanie się zależnościami zorientowanymi, ze względu na możliwość zachowania istotnych informacji, wnikających z utrzymania kolejności słów w wypowiedzi. Ponadto, posługiwanie się kierunkowymi zależnościami między wyrazami pozwala na zidentyfikowanie wyrazów o niskim znaczeniu semantycznym, ale o bardzo istotnej funkcji gramatycznej. W przypadku sieci współwystępowania koncentratory te nie są słowami kluczowymi opisującymi dokument. Ze względu, że pełnią one funkcje tylko gramatyczne, podobnie jak w opisanej wcześniej analizie statystycznej, wyrazy takie są usuwane (mechanizm stop-listy).

3.2. Sieci syntaktyczne

Mechanizm odwzorowania dokumentów w sieciach syntaktycznych opiera się na analizie składniowej. Wyrazy również stanowią tutaj podstawową jednostkę i są węzłami w sieci. Natomiast krawędzie wyznaczają poprawność informacji pod względem ich prawidłowej budowy – składni.

W momencie budowania sieci zależności, zdania w pierwszej kolejności są parsowane pod kątem relacji gramatycznych. W ogólnym przypadku budowa zdania zależy od orzeczenia i to od niego zależą pozostałe wyrazy w zdaniach. W tego rodzaju sieciach koncentratorami zwykle są czasowniki, które w małym stopniu nadają się jako słowa kluczowe, opisujące dokument. W rzeczywistości zawartość dokumentu opisana jest przez grupę semantycznie współzależnych słów.

3.3. Sieci semantyczne

W języku naturalnym wypowiedzi budowane są w sposób logiczny, tak by przekaz był jednoznaczny i zrozumiały. W sieciach semantycznych wyrazy lub złożone konstrukcje znaczeniowe są węzłami, natomiast najróżniejsze semantyczne relacje stanowią połączenia między nimi. Definicja sieci językowej może odbyć się na wiele sposobów. Relacje pomiędzy obiektami mogą bazować na *ISA-relation* (hiponimia: zwierzę – pies – bokser), *part-whole relation* (meronimia – *X* jest meronimem *Y*, jeśli *X*(-y) są częściami lub członami *Y*(-ów)) czy nawet antonimii (góra – dół).

Analiza semantyczna jest jedną z najtrudniejszych analiz, gdyż wymaga weryfikacji relacji pomiędzy informacjami analizując ich zawartość. W przypadku sieci semantycznych koncentratorami okazują się być polisemiczne zbiory pojedynczych wyrazów, a także zwrotów o bardzo dużym znaczeniu semantycznym dla dokumentu.

4. Algorytm rankingu - TextRank

Samo zbudowanie sieci nie jest jednoznaczne z posiadaniem struktury, która pozwala na wykonywanie wnioskowań na temat treści dokumentu. Konieczne jest ustalenie wagi dla poszczególnych węzłów w grafie. W tym celu można wykorzystać algorytm rankingu. Sieć językowa w swojej strukturze przypomina organizację stron internetowych, gdzie strony będące analogią dla wyrazów połączone są za pomocą fizycznych odnośników, linków. W literaturze [2][5] można odnaleźć wiele przykładów na potwierdzenie możliwości zastosowania algorytmów rankingu w sieciach językowych.

Dziedziną, w której najpowszechniej wykorzystywane są algorytmy rankingu są wyszukiwarki internetowe (np. wyszukiwarka Google). Idea działania opiera się na analizie odwołań i wskazań na siebie stron internetowych, tworzących pewnego rodzaju sieć zależności. Na podstawie wzajemnych głosowań algorytm PageRank ocenia wartość strony, a swoje zastosowanie znajduje między innymi w silniku wyszukiwarki Google. Algorytmy rankingu stosowane są również do analizy cytowań dla publikacji naukowych. W przypadku analizy sieci tekstowych algorytm został zmodyfikowany tak, by ustalanie istotności każdego węzła sieci było wyznaczane rekurencyjnie i bazowało na budowie całej sieci oddziaływań między wyrazami [2].

Przyjmijmy, że $G = (V,E)$ jest skierowanym grafem zależności ze zbiorem wierzchołków V i zbiorem krawędzi E, gdzie E jest podzbiorem VxV. Dla dowolnego wierzchołka V_i, $In(V_i)$ określa zbiór wierzchołków z wychodzącymi krawędziami w kierunku tego wierzchołka – poprzednika, $Out(V_i)$ jest zbiorem następników. Wartość wierzchołka V_i definiowana jest jako [2]:

$$S(V_i) = (1-d) + d \cdot \sum_{j \in IN(V_i)} \frac{1}{|Out(V_j)|} S(V_j), \tag{2}$$

gdzie d jest współczynnikiem tłumienia, przyjmującym wartości z przedziału 0 i 1. Do modelu wprowadza on prawdopodobieństwo przejścia z danego wierzchołka do innego, dowolnego wierzchołka w grafie. W kontekście wyszukiwania w sieci Internet oznacza to wybranie przez użytkownika dowolnego linku z prawdopodobieństwem d i przejście do zupełnie nowej strony z prawdopodobieństwem $(1-d)$. Zwykle przyjmuje się wartość 0,85 [1].

Dla każdego punktu grafu z dowolnie przyporządkowaną wartością, algorytm iteracyjnie wyznacza nową wartość, aż do uzyskania oczekiwanej granicy błędu. Po zakończeniu działania algorytmu wyznaczone wartości zostają przypisane odpowiednio do węzłów i oznaczają ich istotność w grafie. Wybór wartości początkowych dla węzłów nie ma wpływu na końcowe wyniki, uzyskane w wyniku działania algorytmu rankingu testowego. Mają wpływ tylko i wyłącznie na liczbę iteracji, po których uzyskiwana jest oczekiwana zbieżność. Zbieżność zostaje osiągnięta, jeśli stopa błędu dla wszystkich wierzchołków jest mniejsza od założonej wartości. Stopa błędu dla wierzchołka V_i jest definiowana jako różnica pomiędzy oczekiwaną wartością $S(V_i)$ oraz wartością wyznaczoną w k-tej iteracji. Do czasu, gdy wartość oczekiwana nie jest znana, różnica definiowana jest jako:

$$S(V_i) = S^{k+1}(V_i) - S^k(V_i) \tag{3}$$

Oprócz powyższego algorytmu istnieją również inne (np. HITS ang.*Hyperlink Induced Topic Serach),* które mogą zostać wykorzystane w modelu rankingu tekstowego [2].

Algorytm TextRank może również być z powodzeniem stosowany w sieciach niezorientowanych. W tym przypadku $In(V_i)$ oraz $Out(V_i)$ są sobie równe. W strukturach grafowych

luźno połączonych, w których liczba krawędzi jest porównywalna z liczbą wierzchołków wymagana jest większa liczba iteracji, w celu osiągnięcia zadowalających wyników.

Operując algorytmem rankingu, w kontekście poruszania się w Internecie, nie jest konieczne uwzględnianie wag dla wskazań dla kolejnych wierzchołków, gdyż zwykle strony nie wskazują na siebie wiele razy. Często elementy sieci zbudowane z zależności wynikających z reguł języka naturalnego wielokrotnie na siebie wskazują. Właściwość ta jest wykorzystywana do wyznaczenia wagi w_{ij} połączenia pomiędzy dwoma wierzchołkami V_i i V_j. Waga ta kojarzona jest z krawędzią łączącą jednostki, będące we wzajemnej relacji. Po uwzględnieniu wagi krawędzi wzór na istotność wierzchołka w grafie przyjmuje postać:

$$S(V_i) = (1-d) + d \cdot \sum_{j \in IN(V_i)} \frac{w_{ij}}{|\,Out(V_j)\,|} S(V_j) \tag{4}$$

Podsumowując, zastosowanie algorytmu TextRank w pierwszym kroku wymaga przeprowadzenia ekstrakcji zbioru relewantnych cech, który różni się w zależności od wybranej, jednej z powyżej opisanych sieci. W dalszej kolejności wybór sieci pociąga za sobą typ relacji pomiędzy wierzchołkami grafu. Krawędzie te mogą być skierowane bądź nieskierowane, z wagą lub bez. W celu określenia istotności w sieci poszczególnych wierzchołków algorytm rankingowy wykonywany jest iteracyjnie do osiągnięcia przyjętej dokładności. Wyznaczone wartości z algorytmu z powodzeniem mogą zostać wykorzystane jako współrzędne reprezentacji wektorów.

5. Plan dalszych prac

Przedstawione w artykule formy odwzorowania ciągłych treści mają istotne znaczenie dla prowadzonych badań, które poświęcone są metodom wyszukiwania zadanego wzorca w rzeczywistej treści dokumentów tekstowych. W pierwszym etapie prac nad algorytmami wyszukiwania zastosowano metodę statystyczną, która wraz ze wzrostem wymagań okazała się być zbyt prosta. Z tego powodu pojawiła się potrzeba odszukania formy, która oprócz wyekstrahowanych cech zachowywać będzie również relacje pomiędzy wyrazami występujące w języku naturalnym.

Dalsze prace badawcze skierowane będą na opracowanie algorytmów, które z wykorzystaniem przedstawionych w artykule reprezentacji w sieciach językowych, pozwolą na wyszukiwanie i określanie podobieństwa z uwzględnieniem relacji semantycznych, określanych w językoznawstwie terminami: hiponimia, polisemia i meronimia.

6. Podsumowanie

Zastosowanie sieci językowych do budowy reprezentacji wektorowych dokumentów pozwala na znaczne poprawienie wyników analiz tekstów w stosunku do metod statystycznych. Analiza literaturowa wykazała, iż zastosowanie modelu sieci semantycznej do automatycznej ekstrakcji słów kluczowych dało o 20% lepsze wyniki dla poziom ufności i 10% w przypadku precyzji, w porównaniu do metody statystycznej *Term Frequency* [5]. W przypadku poziomu ufności (ang. c*overage*) autorzy przyjęli liczbę poprawnych słów kluczowych podzieloną przez liczbę słów kluczowych, przypisaną przez autora tekstu, natomiast precyzja (ang. *precision*) wyznaczona jest jako iloraz poprawnych cech w 10 najistotniejszych cechach, wybranych przez algorytm.

Uzyskanie lepszych wyników w wyznaczaniu istotności cech algorytmem TextRank poprawia wyniki, ponieważ algorytm nie bazuje tylko na lokalnym kontekście wierzchołka w sieci, ale jego istotność w tekście wyznaczana jest rekursywnie, na podstawie budowy całego grafu [7]. Dodatkowa zaleta algorytmu to możliwość jego zastosowania w każdego typu sieci, gdyż jego implementacja nie wymaga zastosowania wiedzy lingwistycznej. Doskonałość metody polega również na tym, iż uwzględniając w budowaniu reprezentacji, czy to reguły gramatyczne czy semantyczne często odkrywane są nowe, istotne dla opisu, dokumenty cechy, które w przypadku analiz częstotliwościowych, wrzucone do „worka słów" (ang. *bag-of-words*), mogą zniknąć ze względu na małą liczność, pomimo iż odgrywają znaczącą rolę w oddawaniu charakteru treści.

BIBLIOGRAFIA

1. Cieślewicz J., Pelikant A.: Reprezentacja i wyszukiwanie dokumentów tekstowych w bazach danych. BDAS'09.
2. Brin S., Page L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 1998, s. 1÷7.
3. Mihalcea R., Tarau P.: TextRank: Bringing Order into Texts. Conference on Empirical Methods in Natural Language Processing, 2004.
4. Sole R. V., Murtra B. C., Valverde S., Steels L.: Language Networks: their structure, function and evolution. Complexity, Vol. 15, Issue 6, 2010, s. 20÷26.
5. Dorogovtsev S. N., Mendes J. F. F.: Language as an evolving word Web. The Royal Society, 2001, s. 2603÷2606.
6. Watts D. J., Strogatz S. H.: Collective dynamics of 'small-world' networks. Nature 393, 1998, s. 440÷442.

7. Liu J., Wang J.: Keyword Extraction Using Language Network. Natural Language Proccessing and Knowledge Engineering, 2007, s. 129÷134.
8. Jin W., Srihari R. K.: Graph-based Text Representation and Knowledge Discovery. SAC'07, 2007, s. 808÷811.

Recenzenci: Dr hab. inż. Krzystof Goczyła, prof. Pol. Gdańskiej
Dr inż. Sławomir Niedbała

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

Keyword extraction is an important application in the area of information technology. Automatic keyword extraction can help know what is the article primary talking about. The review describe statistical analyses and introduce a network language as a way to building document representation. Firstly, the content in a single documents is associated with text network. The text network can based on three different relations: co-occurrence, semantic or syntax. Then TextRank algorithm is using on this network to assign the rank for each feature. Finally, ranked words can be used as vector representation of documents to future investigation into searching documents.

Adresy

Jakub CIEŚLEWICZ: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego18/22, 90-924 Łódź, Polska, jakub.cieslewicz@p.lodz.pl.
Adam PELIKANT: Politechnika Łódzka, Instytut Mechatroniki i Systemów Informatycznych, ul. Stefanowskiego18/22, 90-924 Łódź, Polska, apelikan@p.lodz.pl.