AUTOMATYKA • 2008 • Tom 12 • Zeszyt 2

Krzysztof Dorosz*

Ekstrakcja spójnych tekstów z Internetu na potrzeby algorytmów lingwistycznych

1. Spójność tekstów a przetwarzanie języka naturalnego

Statystyczna lingwistyka komputerowa posiada zarówno swoich zwolenników, jak i przeciwników, natomiast faktem niezaprzeczalnym jest jej popularność i stosunkowa łatwość implementacyjna algorytmów [1]. Przykładowe wyszukiwanie tekstów po słowach kluczowych w zależności od ilości wystąpień tychże słów w tekście może przy nieodpowiednim dobraniu korpusów tekstów nie działać prawidłowo. Rozważmy przykład, w którym posiadamy dwa odrębne teksty:

Kapitan Joe wydał rozkaz odpalenia wszystkich armat.

Statek pasażerski przewozi 300 osób.

Informacji o kapitanie statku pasażerskiego można by wyszukać, korzystając z następujących słów kluczowych: *kapitan*, *statek*, *pasażerski*. Ponieważ żadne ze zdań nie zawiera jednocześnie wszystkich słów kluczowych, żaden z tekstów nie zostałby dopasowany do wyszukania. Gdyby jednak w korpusie tekstów pojawił się jeden tekst będący złączeniem obu przypadkowych tekstów:

Kapitan Joe wydał rozkaz odpalenia wszystkich armat. Statek pasażerski przewozi 300 osób.

wtedy proste mechanizmy wyszukiwania słów kluczowych wybrałyby ten tekst jako jeden z pasujących do wzorca (posiada wszystkie słowa kluczowe). Natomiast faktem jest, iż tekst ten nie zawiera żadnej informacji o kapitanie statku pasażerskiego.

Już tak elementarne zastosowanie statystyki do przetwarzania tekstów obrazuje nam, jak bardzo mechanizmy takie nie są odporne na teksty wielotematyczne. Problematyka ta przenosi się na wszystkie inne zastosowania, jak np. wykrywanie tematu tekstu, kategoryzowanie tekstu, rozpoznawanie języka tekstu metodą n-gramów, rozpoznawanie autorstwa tekstów itp.

* Katedra Informatyki, Akademia Górniczo-Hutnicza w Krakowie

Dodatkowo można problem ten rozpatrywać także w kontekście wieloznaczności wyrazów (form). W przypadku wyszukiwania po słowach kluczowych wspomnianych wyżej, mogą pojawić się *false positives* związane z tym, że formy wieloznaczne znajdą się w pobliżu innych wyrazów, które mogą występować z daną formą wieloznaczną, ale w odniesieniu do innego pojęcia. Dla przykładu podany jest tekst skonkatenowany z dwóch niezależnych zdań (np. nagłówków prasowych):

Hrabia odbuduje swój zamek. Błyskawiczny ruch szachisty doprowadził do zwycięstwa turnieju.

W powyższym przykładzie wieloznaczna forma wyrazu *zamek* znalazła się w bliskości formy *błyskawiczny*. Proste mechanizmy (np. wyszukiwanie słów kluczowych) mogłyby wyszukać ten tekst w kontekście zamka do ubrania, natomiast takie pojęcie nie wystąpiło w żadnym z tych tekstów. Gdyby teksty były rozpatrywane osobno, można byłoby uniknąć tego typu skojarzeń form wieloznacznych i innych konsekwencji z tego wynikających.

Przykładem kolejnego zaburzenia poprawnego funkcjonowania algorytmów lingwistycznych przez źle zebrane korpusy tekstów to np. budowanie list skojarzeniowych. Metody te analizują dla każdego słowa kontekst innych słów, w jakim znajdują się w tekstach. Przetwarzanie tego typu tekstu mogłoby doprowadzić do skojarzenia słowa *zamek* ze słowem *błyskawiczny* ale także, ze słowami *hrabia* oraz *szachista*. Problem jaki tu występuje, to z jednej strony tworzenie list skojarzeniowych na podstawie słów dotyczących różnych pojęć, a z drugiej strony niepotrzebne podnoszenie poziomu tak zwanego „szumu" przy tworzeniu list. Słowo *szachista* najprawdopodobniej pozostanie słabo skojarzone ze słowem *zamek* po przejrzeniu innych tekstów, ale powiększy wielkość listy o zupełnie przypadkowe pozycje.

2. Podstawy ekstrakcji tekstów z Internetu

Podstawowym i największym źródłem tekstów w języku naturalnym w Internecie są strony WWW. Jednakże ich ekstrakcja wraz z pojawieniem się nowoczesnych technologii webowych, takich jak Java Script (Ajax), CSS, Flash oraz ogólnie szeroko pojętego zjawiska Web 2.0 (czyli połączenie nowoczesnych technologii webowych z *User Generated Conent*) [5], staje się coraz trudniejsza technologicznie. Same trudności techniczne kodowania stron nie są jednak tak istotne jak forma i dobór materiałów prezentacji treści, które również ewoluowały wraz z nowoczesnymi technologiami. Proste monotematyczne strony zastępuje się portalami o skomplikowanych strukturach, grupującymi w poszczególnych elementach wyglądu wiele różnorodnych treści, m.in.:

- paski nawigacyjne – menu,
- skróty (zajawki) innych artykułów,
- linki nawigujące do innych artykułów w portalu,
- dodatkowe elementy informacyjne (np. kalendarz, pogoda, itp.),
- treści reklamowe,

– treści związane z prawami własności oraz politykami prywatności,
– treści komentarzy,
– treść właściwą dla danej strony.

Mechanizmy do pozyskiwania treści z Internetu tworzone są na bazie robotów internetowych zwanych crawlerami, które trawersują po linkach URL znalezionych na uprzednio odwiedzonych stronach przeglądając systematycznie wycinek sieci. Crawlery są zwykłym oprogramowaniem sieciowym wykorzystującym najczęściej architekturę rozproszoną w celu zwielokrotnienia szybkości działania. Do przetwarzania treści stron HTML wykorzystuje się najczęściej różnorodne warianty metody *HTML Stripping*, która w dużym skrócie polega na umiejętnym wycinaniu tagów języka HTML z treści strony, aby pozostała w niej jedynie treść tekstowa w języku naturalnym [3]. Niezależnie od stosowanego wariantu crawler dla jednej strony zasadniczo dostarcza pojedynczy tekst będący konkatenacją wszystkich tekstów na niej występujących. O ile w przypadku wspomnianych wcześniej monotematycznych stron, taka taktyka okazuje się być wystarczająca, ponieważ w dużej mierze dostajemy dobrze sformatowane spójne tematycznie teksty, o tyle traktowanie w ten sposób nowoczesnych portali WWW (rys. 1) zaowocuje uzyskaniem na wyjściu tekstów dalece niespójnych tematycznie. Teksty takie zawierać będą niektóre lub wszystkie ze wcześniej wymienionych typów napisów, często niestanowiących prawidłowo zakończonych zdań (brak znaków interpunkcyjnych na końcu zdań, itp.).

Na rysunku 1 przedstawiono przykład strony WWW o skomplikowanej strukturze informacyjnej. Przykład ten doskonale obrazuje wszelkie problemy występujące przy automatycznej ekstrakcji tekstu. Przede wszystkim, trudno stwierdzić, co stanowi zasadniczą treść strony, ponieważ stanowią je zbiory przypadkowych napisów i tytułów artykułów będące linkami do kolejnych stron WWW. Niektóre spośród zebranych tytułów (w tym przykładzie *Polskie kino nie jest złe*) posiadają krótkie streszczenie (brak kropki na końcu zdania!), które jest już tematycznie powiązane z tym tytułem, inne występują natomiast samotnie bez powiązania tematycznego z żadnym sąsiednim tekstem. Dodatkowo na stronie widocznych jest wiele elementów nawigacyjnych (np. pasek górny *WWW*, *Katalog*, *Zdjęcia*, *Zakupy*, *Lokalizator* itp.). Na stronie obok innych tekstów znalazły się także treści generowane przez użytkowników (np. *Forum: Marcin: A ja nawet nie mam komu dać róży*, *jestem sam…*) oraz treści reklamowe (np. *Poznaj sekret doskonałej skóry*). Odnaleźć można także informacje dodatkowe, mianowicie datę, imiona solenizantów, informacje o pogodzie.

Wszystkie teksty znalezione na powyższej stronie sklejone w jeden sformatowany tekst nie nadają się do żadnej analizy statystycznej (poza ogólną statystyką charakterystyki słów języka polskiego). Wraz z rosnącą popularnością nowych technologii webowych należy taką sytuację uznać raczej za normalną niż wyjątkową, ponieważ zjawisko to będzie się tylko nasilać. Największą korzyść może przynieść tylko skonstruowanie nowych metod, które umiejętnie rozdzielą poszczególne spójne fragmenty na osobne teksty (np. na poszczególne tytuły, tytuły wraz z odpowiadającymi im skrótami itp., treści artykułów wydzielone od innych elementów strony WWW).

Rys. 1. Przykład portalu internetowego o skomplikowanej strukturze trudnej do automatycznej ekstrakcji tekstu
Źródło: [6]

3. Analiza struktury kodu HTML w celu ekstrakcji spójnych tekstów

Aby przystąpić do ekstrakcji spójnych tekstów, należy najpierw bliżej zdefiniować pojęcie spójności. W przypadku idealnym, wynikiem pracy crawlera powinien być zbiór tekstów, które są wewnętrznie spójne w sensie semantycznym. Nie możemy jednak zakła-

dać tak dalece idącej spójności, ponieważ semantyka tekstu wynika już z intencji jego autora. Poza tym faktem istnieje jeszcze problem swoistego zakleszczenia. Algorytmy lingwistyczne testuje się na tekstach obecnie po to, by między innymi wytworzyć mechanizmy mogące rozpoznać ich semantykę, a jednocześnie do pozyskania tych tekstów potrzebne są algorytmy potrafiące radzić sobie z semantyką, aby odpowiednio podzielić zbierane teksty. Wniosek – należy inaczej podejść do zagadnienia spójności, metodami niebazującymi na semantyce tekstu.

W związku z powyższym, definicję spójności należałoby potraktować trochę ogólniej, jako tekst stanowiący jedną całość w odniesieniu do:

– sposobu prezentacji (np. układ na stronie – tekst w jednej ramce),
– budowy tekstu (tytuł, wprowadzenie, treść).

Spójność tekstu w treści strony WWW zdefiniowana została na potrzeby tej pracy poprzez wyszukiwanie charakterystycznych oznaczeń początku i końca tekstu, co jest pewną alternatywą w stosunku do definicji opierających się na semantyce. Taka definicja spójności tekstu umożliwia pełną automatyzację ekstrakcji za pomocą dostępnej w kodzie HTML hierarchii dokumentu.

Sam format HTML ewoluował w czasie, choć ewolucja ta dotyczyła raczej prób standaryzacji pewnych konwencji kodowania, co ma odzwierciedlenie np. w HTML 4.01 i późniejszych, które są zgodne ze standardem SGML (*Standard Generalized Markup Language*) [4]. Generalną zasadą, jaka towarzyszy kodowaniu treści strony w kodzie HTML, jest duża systematyczność i hierarchiczność bloków kodu, w której umieszcza się treści. Dla przykładu, każdy dokument HTML powinien zawierać sekcję główną <*html*> </*html*>, w której tytuł strony znajdziemy w sekcji <*head*><*title*></*title*></*head*>, a treść zawarta będzie w sekcji <*body*></*body*>. W sekcji treści znajdować się będą kolejno zagnieżdżane elementy, tabele (<*table*>), linki (<*a*>), grafiki (<*img*>), elementy grupujące (<*span*>, <*div*>) i wiele innych. Ostatecznie treść tekstowa znajduje się zagnieżdżona w odpowiednich elementach HTML, od elementu <*body*> począwszy. Elementy treści mogą przeplatać się swobodnie z innymi blokami HTML.

Do reprezentacji hierarchicznej struktury HTML z dużym powodzeniem stosuje się standard DOM (*Domain Object Model*), który jest niezależnym standardem obiektowego modelowania. Strona HTML przedstawiona w modelu DOM jest drzewem obiektów, po którym można swobodnie przemieszczać się w dowolnym kierunku. Każdy węzeł drzewa jest obiektem reprezentującym dany blok HTML. Liśćmi drzewa są tak zwane węzły tekstowe (*Text Node*), które reprezentują (zawierają w sobie) ostateczną treść tekstową strony.

Najważniejszym założeniem, jakie przyjęto przy konstrukcji algorytmów ekstrakcji spójnych tekstów przy wykorzystaniu struktury DOM strony HTML, jest teza, iż istnieje zawsze jeden taki węzeł drzewa DOM, dla którego konkatenacja tekstów z całego poddrzewa jest tekstem spójnym.

Jeśli przyjrzeć się kształtowi obecnych portali WWW, a także sięgnąć do założeń samego języka HTML (szczególnie w połączeniu z CSS), który niejako wymusza dobrą praktykę grupowania treści i nadawania im odpowiednich klas styli CSS, to założenie takie wydaje się co najmniej rozsądne i zgodne z rzeczywistością. Oczywiście znajdą się strony WWW będące wyjątkami od tej reguły, natomiast trzeba mieć na uwadze fakt, iż będzie to

rodzaj niechlujności (przejaw braku sprawności technicznej) twórcy strony, na który nie mamy wpływu tak samo, jak na brak sprawności lingwistycznej (np. umieszczenie w jednym akapicie dwóch zdań o różnej tematyce).

Zgodnie z powyższym założeniem, proces wyszukiwania tekstów spójnych ze strony HTML opierać się będzie na przedstawieniu jej w postaci drzewa DOM, a następnie wyszukiwanie węzłów tego drzewa uznanych jako węzły tekstów spójnych (zwanymi dalej węzłami spójnymi). Węzły te nie powinny należeć do innych poddrzew pozostałych węzłów spójnych (tzn. węzły spójne nie mogą zawierać się wzajemnie w swoich poddrzewach). Dodatkowo powinna zajść własność: dla każdego węzła tekstowego *T* osiągalnego z węzła <*body*> istnieje taki węzeł spójny *S*, że *T* jest osiągalne z *S*. Dzięki temu założeniu proces ekstrakcji spójnych tekstów nie pominie żadnego tekstu na stronie WWW.

3.1. Metoda energii węzłów

Ta autorska metoda przedstawia jedno z podejść do zagadnienia wyszukania węzłów spójnych. Bazuje ona na obserwacji, że nowoczesne portale WWW posiadają bardzo skomplikowane i głębokie struktury drzew (co ma swoje uzasadnienie w sferze prezentacyjnej, gdzie do pozycjonowania treści i formatowania stylami CSS używa się wielokrotnie zagnieżdżonych elementów typu <*div*> i <*span*>). W związku z tym poszczególne jednostki wizualne na stronie (np. ramki z tekstem, akapity, paski nawigacyjne) na poziomie węzłów tekstowych będą oddalone od siebie długościami ścieżek wielokrotnie dłuższymi niż ścieżki znajdujące się w obrębie jednego elementu wizualnego. Dodatkowo na odległość tekstów między sobą wpływa ma także ich ilość. Długie teksty mogą mieć więcej poziomów zagłębienia niż krótkie notatki. Metoda energii węzłów zakłada, iż każdy węzeł tekstowy posiada energię adekwatną do swojej miary informacji (np. do ilości znaków, ilości słów, ilości słów rozpoznanych w słowniku itp.). Energia ta może zostać łączona w węzłach nadrzędnych, część energii przekazywanej wyżej ulega rozproszeniu proporcjonalnie do odległości od węzłów nadrzędnych. Szybkość rozpraszania się energii jest odwrotnie proporcjonalna do względnej wysokości całego drzewa DOM. Oznacza to, że w stosunkowo niewysokich drzewach DOM stosunkowo duża energia jest rozpraszana przy przejściu z węzła na węzeł, a przy wysokich drzewach DOM, energia rozpraszana jest stosunkowo niewielka. Pozwala to na adaptację zarówno do bardzo skomplikowanych, jak i bardzo prostych stron WWW.

Okazuje się, że pewne węzły stosując tę zasadę, skupiają więcej energii od innych (rodzicielskich i potomnych) i osiągają maksima energetyczne na ścieżkach w drzewie. Te właśnie węzły stanowić będą potencjalne węzły spójne. Aby dokonać ostatecznego podziału, należy podążać od korzenia DOM w dół po węzłach potomnych, z każdym razem sprawdzając, czy któryś z węzłów potomnych posiada większą energię od danego węzła. Jeśli nie, dany węzeł uzyskał jako pierwszy maksimum energii na ścieżce od korzenia aż do jego poddrzewa i uznaje się go za węzeł spójny. Jeśli natomiast istnieje jakiś węzeł potomny posiadający większą energię, należy rozdzielić przechodzenie po drzewie na równoległe procesy, zaczynając od każdego węzła potomnego. W szczególności węzłem spójnym może

być pojedynczy węzeł tekstowy. W ten sposób przeszukane zostanie całe drzewo DOM, a wszystkie znalezione węzły spójne będą rozłączne w sensie zawierania się w swoich poddrzewach oraz będą zawierać wspólnie wszystkie węzły tekstowe.

Metodę energii węzłów można opisać za pomocą skróconego formalizmu. Niech DOM będzie drzewem (V, G, R), gdzie V – zbiór wszystkich węzłów, R – zbiór krawędzi (v_1, v_2), gdzie v_1, v_2 należą do V, R – węzeł początkowy (korzeń); R należy do V. Niech *parent*(v) oznacza węzeł nadrzędny dla v oraz *childs*(v) niech oznacza zbiór węzłów potomnych dla v. Niech $m(v)$ oznacza energię tekstu zawartego w v przyjętą zgodnie z dowolną metryką. Niech k oznacza wysokość drzewa DOM, czyli najdłuższą ścieżkę w drzewie DOM o początku w R i końcu w dowolnym liściu g ($g \in V$ i $childs(g) = \varnothing$). Energię węzła v określa się wzorem rekurencyjnym:

$$E(v) = \begin{cases} m(v), & \text{gdy} \quad childs(v) = \varnothing \\ \sum_{w \in childs(v)} \frac{p}{k} E(w), & \text{gdy} \quad childs(v) \neq \varnothing \end{cases} \tag{1}$$

gdzie p jest stałą szybkości rozpraszania energii ustaloną *a priori*.

Węzeł v_i jest węzłem spójnym, począwszy od węzła v_0 wtedy i tylko wtedy, gdy istnieje taki ciąg (v_0, v_1, v_2, ... , v_i, v_{i+1}), że $v_n \in child\ (v_{n-1})$ dla $n \in \langle 1; i + 1\rangle$ oraz $E(v_0) = E(v_1) = E(v_2) = ... = E(v_i)$, oraz $E(v_i) > E(v_{i+1})$, a także $E(v_n) = \max\{E(childs(v_{n-1}))\}$ dla $n \in \langle 1; i\rangle$, czyli $E(v_n)$ jest wartością maksymalną wśród wartości energii wszystkich dzieci rodzica węzła v_{n-1}.

Węzeł v jest węzłem spójnym wtedy i tylko wtedy, gdy:

a) v jest węzłem spójnym, począwszy od R,
b) v jest węzłem spójnym, począwszy od v', takiego że istnieje v'' będące węzłem spójnym należącym do *childs*(*parent*(v')).

3.2. Predykcja spójnych fragmentów tekstu z położenia odsyłaczy URL

Druga autorska metoda nie jest metodą tak sformalizowaną jak metoda energii węzłów, natomiast wykorzystuje do działania zbiór reguł wynikłych z obserwacji budowy stron WWW oraz położenia w nich kluczowego elementu, jakim są odsyłacze URL.

Za regułę generalną można przyjąć, iż każdy węzeł <*a*> jest węzłem spójnym, natomiast pozostałe węzły powstałe z wydzielenia węzłów <*a*> ze struktury DOM stają się węzłami spójnymi agregującymi w sobie pozostałe węzły *Text Node*. Założenie takie jest w pewnym sensie poprawne z punktu widzenia informacyjnej struktury dokumentu HTML (odsyłacz URL prowadzi do odrębnej treści zawartej na innej stronie HTML, więc sama treść linku jest odrębna w stosunku do jego kontekstu na zadany temat). Jest to jednak zbyt daleko idące uproszczenie i okazuje się niepoprawne w wielu przypadkach. Należy rozważyć dokładniej to zagadnienie.

Struktura każdego dokumentu HTML pod względem roli tekstu w niej zawartego służy jednoznacznie dwóm celom: przekazania treści w postaci akapitów tekstu lub przekazania informacji o innych tekstach zawartych na stronach pod podanymi adresami WWW w postaci linków WWW.

W przypadku pierwszym, mamy do czynienia w większości z dużymi fragmentami tekstów podzielonymi na akapity (dużymi w odniesieniu do wielkości tekstów w przypadku drugim), które zazwyczaj nie posiadają żadnych linków. Jest to spowodowane rolą danego tekstu na stronie. Jeśli tekst ten jest tekstem wiodącym danej strony, nie będzie on już linkować do innych miejsc w sieci, ponieważ jest treścią główną strony. Oczywiście zdarzają się tutaj wyjątki, gdzie poszczególne frazy w dużym akapicie tekstu mogą być jednocześnie fragmentem zdań, a także linkami do innych stron np. z definicją tej frazy, natomiast takie wyjątki można z łatwością wykryć. Ich charakterystyczną cechą jest występowanie w dużych partiach tekstu najczęściej w postaci węzłów <*a*> bezpośrednio obok węzłów tekstowych. Z powyższych faktów można wyodrębnić regułę, która mówi że wszelkie linki URL w wyższych węzłach drzewa DOM w stosunku do dużych fragmentów tekstów można traktować jako niezwiązane z tym tekstem (np. paski nawigacyjne, itp...) i stanowią one górne ograniczenie spójności węzła potomnego zawierający tekst zasadniczy.

W przypadku drugim mamy do czynienia z kilkoma wariantami linków. Najczęściej linki występują w zagnieżdżonych strukturach nieotoczone dodatkowymi tekstami – występują samodzielnie. Takie poddrzewo DOM spłaszczone do zwykłej listy elementów (biorąc pod uwagę jedynie elementy <*a*> oraz *Text Node*) zawierać będzie same elementy typu <*a*>. Taki charakterystyczny układ linków wskazuje jednoznacznie na nawigacyjny charakter linków na stronie, niezależnie, czy linki zawierają w sobie pojedyncze słowa kluczowe czy całe frazy będące np. tytułami kolejnych artykułów. Należy odnieść się tutaj do wiedzy o strukturze informacyjnej strony, z której można założyć, że każdy taki link odnosi się do odrębnej informacji (strony WWW), przez co teksty zawarte w linkach należy traktować odrębnie. Często zdarza się także, że linki takie zawierają krótkie skróty tekstowe mające zachęcić do kliknięcia w dany link. Skróty te nie są linkami, więc zaburzają czystą strukturę jednorodnego drzewa linków, a dodatkowo są powiązane tematycznie z linkiem, w obrębie którego występują. W takiej sytuacji potrzebna jest dodatkowa analiza wzajemnego położenia linku i skrótu. Ze względu na kwestie prezentacyjne w HTML-u, skrót wraz z linkiem będą położone w obrębie jednego spójnego węzła w stosunku do węzłów z innymi linkami. Dlatego też jako kolejną regułę można przyjąć, iż węzłem spójnym jest także węzeł grupujący w swoim poddrzewie węzeł <*a*> i *Text Node* pod warunkiem, że węzeł spójny występuje w większej strukturze linków.

Te proste reguły badania hierarchii węzłów tekstowych, ich relatywnych wielkości oraz ich położenia względem linków WWW dają bardzo dobre rezultaty, które są dużo bardziej odporne na zaburzenia w strukturze HTML-a niż metoda energii węzłów. Jest to natomiast metoda trudno formalizowana, ze względu na swoją regałową naturę i szukania prawidłowości, przy wykorzystaniu wiedzy o naturze prezentacji informacji i zasadach łączenia informacji linkami URL.

4. Wnioski

W dobie świetnie opanowanych mechanizmów sieciowych stosowanych do budowania crawlerów główny nacisk kładziony będzie na możliwości lingwistyczne robotów internetowych tak, by polepszyły one jakość i szybkość dostarczania rezultatów. Obie zaprezentowane autorskie metody pozyskiwania spójnych tekstów ze stron WWW z powodzeniem przyczyniły się do uzyskania znaczącego wzrostu jakości produkowanych zbiorów tekstowych przez narzędzia crawlingowe i umożliwiły lepszą adaptację do potrzeb przetwarzania języków naturalnych. Obie metody mają swoje mocne i słabe strony. Z całą pewnością można lepiej wykorzystać zalety różnych podejść tworząc hybrydowe rozwiązania lub zupełnie nowe rozwiązania bazujące na wiedzy o praktycznej strukturze stron HTML pozyskanej z testowania tych metod. Opisane metody pozostają w fazie dalszego rozwoju.

Literatura

[1] Manning C., Schütze H., *Foundations of Statistical Natural Language Processing*. Cambridge, MIT Press 1999.
[2] Understanding Web 2.0
[3] Dorosz K., *System automatycznej ekstrakcji tekstów z Internetu*. Kraków, Akademia Górniczo-Hutnicza 2006 (praca magisterska).
[4] ISO8879 Standard, *Information Processing – Text and Office Systems – Standard Generalized Markup Language (SGML)*.
[5] Murugesan S., *Understanding Web 2.0*. IT Professional Volume 9, Issue 4, July–Aug. 2007, 34–41, IEEE.
[6] onet.pl z dnia 11.02.2008 r.