STUDIA INFORMATICA 2011
Volume 32 Number 2B (97)

Anna KOTULLA
Politechnika Śląska, Instytut Informatyki

WYSZUKIWANIE INFORMACJI Z UWZGLĘDNIENIEM DANYCH DOTYCZĄCYCH LOKALIZACJI

Streszczenie. Znacząca część zapytań realizowanych przez wyszukiwarki internetowe dotyczy wyszukiwania lokalnego. W artykule omówiona została problematyka wyszukiwania informacji uwzględniających lokalizację. Zaproponowano system dla zasobów internetowych w języku polskim, umożliwiający pozyskiwanie informacji uwzględniających lokalizację.

Słowa kluczowe: wyszukiwanie informacji, web mining, szukanie lokalne

INFORMATION RETRIEVAL CONSIDERING LOCALIZATION DETAILS

Summary. A significant part of search queries parsed by web search engines refers to local resources. The problem of searching for information considering localization details is described in this paper. A new local search system is introduced, for web resources in polish language.

Keywords: information retrieval, Web Mining, local search

1. Wprowadzenie

Wyszukiwanie danych powiązanych pod względem lokalizacji geograficznej, często nazywane też wyszukiwaniem lokalnym (ang. *local search*), jest złożonym zagadnieniem. Jednocześnie wiadomo, że sporo (za [14]: 20%–25%) zapytań dotyczy danych związanych z pewną określoną lokalizacją geograficzną – największe wyszukiwarki uruchomiły serwisy ułatwiające wyszukiwanie takich treści, np. *Google Maps* (wcześniej znanego pod nazwą *Google Local,* URL: *http://local.google.pl/*) czy serwisu udostępnionego przez *Yahoo!*,

URL: *http://local.yahoo.com/*. Szczegóły algorytmów używanych przez największe z wyszukiwarek są nieznane.

Analiza dokumentów sieci WWW i techniki eksploracji zawartych w nich danych bazują na metodach zaliczanych do tzw. *Web Mining* [13]. Jako *Web Mining* rozumiane są te metody eksploracji danych, które przydatne są do odkrywania wzorców w sieci *WWW*. *Web Mining* dzieli się na:

- *Web Content Mining* – odszukiwanie informacji w zawartości zasobów sieci *WWW* (np. treść stron, zamieszczone informacje graficzne czy multimedialne),
- *Web Structure Mining* – rozpoznawanie struktur stron bądź domen na podstawie hiperłączy,
- *Web Usage Mining* – odszukiwanie wzorców w przypadku użytkowania stron czy zasobów.

Zasoby sieci *WWW* składają się w przeważającej części z różnego rodzaju informacji tekstowych, w związku z czym, w przypadku *Web Content Mining* stosowane są metody przypisane do *Text Mining* (ang. *Knowledge Discovery from Text*) [8]. Zamiennie z terminem *Web Content Mining* używany bywa *Web Text Mining*.

Artykuł omawia problematykę analizowania zawartości zasobów sieci *WWW* w taki sposób, aby wyodrębnić informacje geograficzne (nazwy miejscowości itp.) Jako przedmiot badania obrane zostały zasoby internetowe dotyczące Polski.

2. Wyszukiwanie w polskich zasobach internetowych

2.1. Wyszukiwarki używane w Polsce

Według danych udostępnianych przez wiodące serwisy statystyczne[1], rynek wyszukiwarek w Polsce zdominowany jest przez *Google*.

Testy wykonane zostały za pomocą wymienionych powyżej, najpopularniejszych dla stron polskojęzycznych wyszukiwarek: *Google, bing, Yahoo!* i *NetSprint*. Dla najczęściej używanych zapytań – jak wyszukiwanie hoteli w pewnej miejscowości, np. „hotel Ustroń“ – przeglądarki zwracają wyniki bardzo dobrze odpowiadające szukanej frazie. Dużo gorzej wygląda sprawa w przypadku zapytań bardziej specyficznych, np. wyszukania miejsc (sklepów) w mieście Opole, w których kupić można foteliki samochodowe, za pomocą zapytania „fotelik samochodowy sklep Opole“ czy „fotelik samochodowy Opole“. Wśród pierwszych

[1] Dane na podstawie serwisu *StatCounter Global Stats*, URL: *http://gs.statcounter.com/* oraz *NetApplications*, URL: *http://www.netapplications.com/Default.aspx*

10 wyników pojawiają się wyłącznie portale porównujące ceny, odnośniki do sklepów internetowych czy ogloszeń drobnych.

StatCounter Global Stats podaje, że w 2010 roku w Polsce używane były następujące wyszukiwarki:

- *Google* (URL: *http://www.google.pl*): 97,88%
- *Onet PL* (URL: *http://www.onet.pl*) bazująca na *Googl*e: 0,76%
- *bing* (URL: *http://www.bing.com/?cc=pl*), wyszukiwarka udostępniona przez firmę Microsoft: 0,7%
- *Yahoo!*(URL: *http://www.yahoo.com/*): 0,35%
- Wirtualna Polska (URL: *http://www.wp.pl*), bazująca na wyszukiwarce internetowej *NetSprint* (URL: *http://www.netsprint.pl*): 0,11%
- inne: 0,21%

2.2. Żółte strony

Wśród zasobów internetowych (nie tylko polskich, ale również tych w innych językach), oferujących dane dotyczące lokalizacji geograficznej, najpowszechniejsze i najbardziej uniwersalne są zasoby udostępniane w forme tzw. żółtych stron (ang. *yellow pages*). Wersja internetowa w gruncie rzeczy niewiele różni się od książek telefonicznych, w których umieszczone są informacje dotyczące firm, względnie instytucji. W przypadku wersji elektronicznej, wyszukiwanie jest możliwe po podaniu słowa kluczowego (np. nazwy firmy, branży) oraz danych dotyczących lokalizacji (np. miejscowość, powiat, województwo). Największą przewagą elektronicznej wersji żółtych stron nad wersją tradycyjną, drukowaną, jest aktualność zawartych w nich danych. Do największych polskich serwisów tego typu zaliczają się *http://www.pkt.pl* oraz *http://www.ditel.pl* (obydwa należące do spółki pkt.pl Polskie Książki Telefoniczne) czy *http://www.yellowpages.pl* (należący do Yellow Pages Sp. z o.o.). Serwisy typu *yellow pages* są serwisami komercyjnymi, wysoką pozycję na liście wyników wyszukiwania można po prostu nabyć – takie posortowanie wyników nie zawsze jest jednak zgodne z intencją szukającego. Dodać jednak trzeba, że większość serwisów specjalnie oznacza sponsorowane odnośniki.

3. Przegląd dotychczasowych badań

Wyszukiwanie danych pod względem kryterium lokalizacji było już wcześniej przedmiotem badań naukowych. Zaproponowane rozwiązania mają jednak w dalszym ciągu charakter prototypowy, autorzy w swoich badaniach zwracali uwagę, że nadal istnieją pewne ograni-

czenia. Dotychczasowe badania przyporządkować można do następujących, omówionych poniżej, grup:

- geokodowanie,
- semantyczny internet ukierunkowany geograficzne,
- wyszukiwarki geograficzne.

3.1. Geokodowanie

Jako geokodowanie [6] (ang. *geocoding, Geo Coding*) rozumiany jest zautomatyzowany proces przyporządkowania współrzędnych geograficznych (kodów, np. długości i szerokości geograficznej) na podstawie danych zawartych w opisie (np. ulica, miejscowość, kod pocztowy, kierunkowy numer telefonu).

McCurley [16], obecnie pracujący dla Google Inc., opublikował wyniki swoich badań w zakresie geokodowania. Opisał różne przydatne dla geokodowania informacje, jak nazwy miast, kody pocztowe itd. Problematyka pozyskiwania tych informacji nie jest omówiona w jego pracy.

Jako odwrotne geokodawanie (ang. *reverse geocoding*) [5] rozumiane jest przyporządkowanie pewnemu miejscu, którego współrzędne geograficzne są znane, danych określających (np. ulica, miejscowość, kod pocztowy).

3.2. Semantyczny internet ukierunkowany geograficzne

Koncepcja semantycznego Internetu zaproponowana została przez T. Berners-Lee [4]. M.J. Egenhofer [7] zaproponował stworzenie narzędzi uwzględniających dane geograficzne. Wymagane było stworzenie specjalnych ontologii, np. dla danych przestrzennych. Różne aspekty semantycznego Internetu ukierunkowanego geograficznie omówione zostały przykładowo w [9] czy [17].

Pomimo ogromnych możliwości, powszechne wykorzystanie semantycznego Internetu jest na dzień dzisiejszy trudne. Wyszukiwarki aktualnie nie interpretują w wystarczający sposób informacji semantycznych – w związku z tym informacje takie nie są załączane. Nie można również wykluczyć, że nastąpią próby wykorzystania semantycznego Internetu do manipulowania wyników wyszukiwania, co ma miejsce również w przypadku wykorzystywanych obecnie sposobów indeksowania i wyszukiwania.

3.3. Wyszukiwarki geograficzne

Problematyka wyszukiwarek geograficznych omówiona została np. w [15]. Koncepcja wyszukiwarek geograficznych zakłada, że użytkownicy często zainteresowani są informacja-

mi lokalnymi, związanymi z danym regionem. Wyszukiwarki geograficzne porządkują wyniki pod względem dopasowania do zapytania oraz pod względem odległości od zadanego miejsca.

Spośród dostępnych wyszukiwarek dobrym przykładem wyszukiwarki porządkującej informacje pod względem regionów jest szwajcarska wyszukiwarka *http://www.search.ch/ /index.en.html*. Trwają badania nad doskonaleniem standardów dla wyszukiwarek geograficzanych, przykładowy schemat przedstawiony został w [10].

3.4. Niejednoznaczność

W [2] omówiony został problem wieloznaczności nazw. Określone zostały dwa typy wieloznaczności, tzw. *geo/non-geo* oraz *geo/geo*.

Wieloznaczność *geo/geo* zachodzi, kiedy jednej nazwie przyporządkowanych jest więcej miejsc geograficznych, np. dla nazwy „Węgry" prawidłowe lokalizacje geograficzne[2] to:

- państwo w Europie,
- następujące miejscowości w Polsce:
  - wieś w woj. dolnośląskim, w pow. wrocławskim, w gminie Żórawina,
  - wieś w woj. opolskim, w pow. opolskim, w gminie Turawa,
  - wieś w woj. pomorskim, w pow. sztumskim, w gminie Sztum,
  - wieś w woj. wielkopolskim, w pow. ostrowskim, w gminie Nowe Skalmierzyce.

Wieloznaczność *geo/non-geo* zachodzi, kiedy jakiś termin, oprócz nazwy geograficznej, posiada jeszcze jakieś inne znaczenie, przykładem może tu być góra Rysy.

W [2] stwierdzono, że najwięcej problemów przysparza wieloznaczność *geo/non-geo*. Testy wykonane zostały dla zasobów w języku angielskim, w przypadku zasobów w innym języku, a także stosowanych tu innych algorytmów, wyniki mogą się różnić.

4. Przyporządkowywanie informacji geograficznych wyodrębnionych w zasobach internetowych

4.1. Podstawowe informacje

Oficjalne spisy nazw geograficznych pielęgnowane są przez dane państwo, a zmiany komunikowane są w wydanych oficjalnie aktach prawnych.

[2] Za Wikipedią, *http://pl.wikipedia.org/wiki/W%C4%99gry_%28ujednoznacznienie%29*

Naturalnie nasuwającą się podstawą dla przyporządkowywania informacji geograficznych są:

- kody pocztowe (pocztowe numery adresowe) wraz z nazwami miejscowości,
- symbol tzw. strefy numeracyjnej (dawniej numer kierunkowy) dla podanych telefonów stacjonarnych.

Kody pocztowe, oficjalnie nazywane Pocztowymi Numerami Adresowymi (PNA), dla terenu Polski określane są w rozporządzeniach Ministra Infrastruktury, natomiast zarządzane są przez Pocztę Polską (URL: *http://www.poczta-polska.pl/spispna/*). Na podstawie kombinacji PNA oraz miejscowości, możliwe jest określenie dodatkowo gminy, powiatu i województwa. Często zdarza się, że jeden PNA odpowiada paru miejscowościom. Z kolei w miejscowości (dotyczy szczególnie większych miast) występować może więcej niż jeden PNA. W Polsce istnieje ponad 22 000 kodów pocztowych. Przykładowo wymienione zostały nazwy miejscowości wraz z gminą (wszystkie w powiecie krapkowickim, województwo opolskie) dla kodu pocztowego 47-330: Dalnie (gm. Zdzieszowice), Dąbrówka (gm. Gogolin), Januszkowice (gm. Zdzieszowice), Jasiona (gm. Zdzieszowice), Krępna (gm. Zdzieszowice), Oleszka (gm. Zdzieszowice), Rozwadza (gm. Zdzieszowice), Zakrzów (gm. Gogolin), Zdzieszowice (gm. Zdzieszowice), Żyrowa (gm. Zdzieszowice).

W przypadku numerów telefonów stacjonarnych, w Polsce obowiązuje obecnie tak zwany zamknięty plan numeracji krajowej, charakteryzujący się stałą długością numerów telefonicznych. Początkowe dwie cyfry numeru telefonu stacjonarnego stanowią jednak symbole tzw. stref numeracyjnych, regulowane przez rozporządzenia Ministra Infrastruktury i zarządzane przez Urząd Komunikacji Elektronicznej (URL: *http://www.uke.gov.pl*).

Istnieje również komercyjny serwis hoga.pl (URL: *http://bazy.hoga.pl*), należący do spółki giełdowej WASKO SA / HOGA SA, oferujący możliwość weryfikowania dla Polski danych adresowych i oferujący różnego rodzaju słowniki. Serwis udostępnia również współrzędne geograficzne dla miejscowości / kodów.

4.2. Utworzone słowniki

Dla potrzeb omawianego systemu używane będą dwa słowniki, kodów pocztowych / miejscowości oraz dodatkowo (uzupełniająco) słownik symboli strefy numeracyjnej. Słownik kodów pocztowych / miejscowości zawiera również informację, czy miejscowość jest miastem oraz, w przypadku miast, informację o liczbie ludności. W dalszej części przedstawiona zostanie ulepszona struktura słownika, w której powiązane zostaną informacje dotyczące kodów pocztowych / miejscowości oraz symboli strefy numeracyjnej. Początkowo, dla celów testowych, informacje te zostaną skorelowane dla miast powyżej 20000 mieszkańcow (według danych Głównego Urzędu Statystycznego z 01.01.2011, w Polsce takich miast jest 108).

4.3. Proponowany algorytm

Wyszukiwanie informacji geograficznych na stronach tekstowych odbywa się w następujących krokach:

- skanowanie strony, porównywanie ze słownikiem informacji geograficznej
  Dane tekstowe zostają najpierw oczyszczone, np. ze zbędnych spacji, potencjalne numery telefonów konwertowane zostają do dziewięcioznakowych ciągów cyfr. Kody pocztowe i nazwy miejscowości porównywane są bezpośrednio. Spośród rozpoznanych numerów telefonu zapamiętywane są takie, które zawierają na pierwszym miejscu symbol strefy numeracyjnej.
- określanie istotności dla znalezionych fraz
  Każdej ze znalezionych fraz przyporządkowywana jest wartość, tzw. waga, z zakresu od 0 (mała istotność) do 1 (duża istotność). Wartości bliskie 1 są przydzielane, kiedy oprócz właściwej kombinacji kod pocztowy / miejscowość znaleziono odpowiedni numer telefonu stacjonarnego dla jednoznacznej nazwy (tzn. nie występuje wieloznaczność *geo/geo*). W przypadku wystąpienia wieloznaczności *geo/geo*, przy założeniu, że znaleziona nazwa określa miasto (tylko dla miast w słowniku znajdują się informacje o liczbie ludności), znaleziona nazwa przyporządkowywana jest miastu o danej nazwie i najwyższej liczbie ludności, natomiast przyporządkowana waga to 0,5. Bywają jeszcze inne przypadki, np. czasem na podstawie zestawu nazw występujących na stronie internetowej możliwe jest dokładniejsze określenie rejonu, a co za tym idzie tej miejscowości, która wydaje się być najbardziej prawdopodna. Znajduje to również wyraz w wysokości przyporządkowanej wagi. Zgodnie z wyznaczonymi wartościami wag, strona zostaje oznaczona jako przynależna do jednej (lub wielu) lokalizacji geograficznej.

5. Zarys systemu testowego

Celem jest utworzenie systemu, który będzie w stanie zindeksować dane (lokalne lub już opublikowane w sieci *WWW*), wyodrębnić ze stron dane zawierające informacje geograficzne i uporządkować pozyskane informacje.

Wyszukiwanie oferowane będzie przez przeglądarkę internetową.

Proponowany system opiera się na strukturze (ang. *framework*) Nutch [18]. Jest to jest obecnie najbardziej wszechstronna struktura, umożliwiająca implementację wyszukiwarek oferowaną jako otwarte oprogramowanie (ang. *open source*). Struktura Nutch, napisana w języku JAVA, bazuje na wyszukiwarce pełnotekstowej Lucene [19], do której dodano elementy właściwe dla sieci *WWW*, jak np. robot internetowy (ang. *crawler*), bazę danych zawie-

rającą informacje o strukturze odnośników (linków) czy analizator składniowy (ang. *parser*) do HTML i innych formatów (między innymi PDF, Microsoft Word, Microsoft Excel, Microsoft PowerPoint, pliki *.SWF Adobe Flash). Zarówno Nutch, jak i Lucene tworzone są przez programistów Apache Software Fundation [20].

Obecnie dostępna jest wersja 1.2 struktury Nutch, opublikowana 24.09.2010. Dla struktury Nutch zalecane jest używanie serwleta Apache Tomcat. W tworzonym przykładowym systemie wykorzystano, obok Nutch 1.2, Apache Tomcat w wersji 4.1.40 oraz SUN JAVA w wersji 1.6.0.

Struktura Nutch oferuje dwie metody pobierania stron:

- przy użyciu polecenia *crawl* – zalecane, gdy będzie pobieranych do miliona stron z niewielu serwerów,
- przy użyciu poleceń niższego poziomu *inject, generate, fetch, updatedb* – zalecane, gdy pobieranych będzie więcej danych niż poprzednio.

Nutch oferuje, przez tzw. analizator linków (ang. linkanalyser), metodę prioryzacji popularnych stron internetowych. Analizator linków korzysta z informacji o odnośnikach zgromadzonych w bazie danych, używając algorytmu OPIC [1] (ang. Online Page Importance Computation). Algorytm OPIC jest modyfikacją algorytmu PageRank [3], który wyznacza dla każdej zindeksowanej strony internetowej wartość prestiżu. PageRank wyznacza wartości prestiżu dla całej sieci, np. sieci WWW, przedstawionej w postaci grafu. Algorytm OPIC wprowadza tzw. wirtualny węzeł, który połączony jest z każdym innym węzłem grafu przedstawiającego sieć. Algorytm PageRank potrzebuje przed startem kompletnej, zindeksowanej sieci, natomiast algorytm OPIC może wyznaczać wartości prestiżu już w czasie zbierania informacji o dokumentach.

Po zebraniu i przetworzeniu (wyczytaniu zawartości przez analizatory składniowe) następuje tworzenie bądź modyfikacja indeksu. Przed użyciem indeksu do wyszukiwania wskazane jest usunięcie duplikatów stron (za pomocą polecenia *dedup*).

Po wykonaniu wymienionych kroków system jest gotowy do wyszukiwania. Nutch oferuje do celów testowych wyszukiwanie lokalne z linii poleceń. Wyszukiwanie w zasobach przez przeglądarkę internetową możliwe jest po zainstalowaniu odpowiedniego archiwum WAR w serwlecie Apache Tomcat.

6. Podsumowanie

W artykule przedstawiona została problematyka wyszukiwania informacji. Jednym z ważnych kryteriów, na podstawie których zwracane są wyniki, jest zawarta w zasobach informacja o lokalizacjach geograficznych.

Zapytania analizowane są na podstawie słownika, by stwierdzić, czy zapytanie dotyczy określonej lokalizacji geograficznej. Jeżeli nie, zwracane są wyniki według prioryzacji ustalonej przez wbudowane w strukturę Nutch mechanizmy. Jeżeli zapytanie dotyczy lokalizacji geograficznej zawartej w słowniku, prioryzacja wyników opiera się na metodzie zaproponowanej w niniejszym opracowaniu.

Podstawę ulepszania przedstawionego w artykule systemu stanowi zmodyfikowany słownik geograficzny, przebudowany i wzbogacony o dalsze treści. Brakuje jednak ogólnodostępnych, wiarygodnych (zweryfikowanych) i kompletnych źródeł. Potencjalnie podstawę stanowić mogą wydawane przez Ministra Spraw Wewnętrznych i Administracji „Wykazy ustalonych nazw miejscowości załączniki", zawierające nazwy miejscowości przyporządkowane do województw, powiatów i gmin, wraz z drugim przypadkiem deklinacji oraz formą przymiotnika. Jako struktura planowany jest słownik geograficzny zbudowany w postaci drzewa, gdyż jego przeszukiwanie będzie najefektywniejsze. Innym możliwym podziałem wydaje się być podział – przyjmuje się, że chodzi o strony polskie – pod względem aktualnych jednostek terytorialnych (województwa, powiaty, gminy, miejscowości). W węzłach zamieszczone będą również informacje o używanych skrótach (np. Strzelce Opolskie, Strzelce Op.) oraz formy fleksyjne nazw. Dodatkowo umieszczone będą w węzłach również właściwe kody pocztowe oraz symbole telefonicznej strefy numeracyjnej. Tak zbudowany słownik może służyć również do polecenia dodatkowych wyników wyszukiwania z miejscowości znajdujących się blisko miejscowości docelowej.

Wzbogacony słownik geograficzny pozwoli lepiej zweryfikować przynależność strony do określonej lokalizacji geograficznej i ulepszyć tym samym przedstawiony system.

BIBLIOGRAFIA

1. Abiteboul S., Preda M., Cobena G.: Adaptive On-Line Page Importance Computation. 12th International Conference on World Wide Web, Budapeszt 2003.
2. Amitay E., Har'El N., Sivan R., Soffer A.: Web-a-where: geotagging web content. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, New York 2004, p. 273÷280.
3. Brin S., Page L.: The Anatomy of a Large-Scale Hupertextual Web Searcg Engine. Computer Networks and ISDN Systems, 1998.
4. Berners-Lee T., Hendler J., Lassila O.: The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. Scientific American, No. 284 (5), 2001, p. 34÷43.

5. Brossier V.: Developing Android Applications with Adobe AIR. O'Reilly Media, Inc., 2011.
6. Christen P., Willmore A., Churches T.: A Probabilistic Geocoding System Utilising a Parcel Based Address File. Advances in Data Mining: Theory, Methodology, Techniques, and Applications. State-of-the-Art Lecture Notes in Artificial Intelligence, Vol. 3755, Springer-Verlag, 2006.
7. Egenhofer M.: Toward the semantic geospatial web. Proceedings of the 10th ACM International Symposium on Advances in Geographic Information Systems, 2002, s. 1÷4.
8. Feldman R., Dagan I.: Knowledge Discovery in Textual Databases (KDT). Proceedings of the First International Conference on Knowledge Discovery and Data Mining (KDD-95), Montreal 1995, Consolidation, KDD 2003, s. 112÷117.
9. Fonseca F., Rodriguez A.: From Geo-Pragmatics to Derivation Ontologies: new Directions for the GeoSpatial Semantic Web. Transactions in GIS, 2007.
10. Fonts O., Huerta J., Díaz L., Granell C.: OpenSearch-geo: The simple standard for geographic web search engines. IV Jornadas de SIG Libre, Girona 2010.
11. Gotlib D., Iwaniak A., Olszewski R.: GIS. Obszary zastosowań. PWN, 2007.
12. Harmon J.E., Anderson S.J.: The design and implementation of geographic information systems. John Wiley & Sons, Inc., Hoboken, New Jersey 2003.
13. Kosala R., Blockeel H.: Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter, Vol. 2, Issue 1, New York 2000, s. 1÷15.
14. Lee H.Ch., Liu H., Miller R.J.: Geographically-Sensitive Link Analysis. Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence. Washington DC 2007.
15. Markowetz A., Chen Y.-Y., Suel T., Long X., Seeger B.: Design and Implementation of a Geographic Search Engine. 8 International Workshop on the Web and Databases (WebDB), Baltimore 2005.
16. McCurley K.: Geospatial mapping and navigation of the web. Proceedings of the 10th World Wide Web Conference, 2001, s. 221÷229.
17. Silveira Chaves M., Silva M.J., Martins B.: A Geographic Knowledge Base for Semantic Web Applications, [in:] Proceedings of SBBD-05, the 20th Brazilian Symposium on Databases, 2005.
18. Strona projektu Nutch. URL: http://lucene.apache.org/nutch/ (sprawdzono 30.01.2011).
19. Strona projektu Lucene. URL: http://lucene.apache.org/ (sprawdzono 30.01.2011).
20. Strona Apache Software Fundation. URL: http://www.apache.org/ (sprawdzono 30.01.2011).

Recenzenci: Prof. dr hab. inż. Andrzej Grzywak
Dr inż. Hafed Zghidi

Wpłynęło do Redakcji 31 stycznia 2011 r.

Abstract

This article gives a theoretical overview of local search problem. The features offered by popular search engines are listed.

This paper describes the common methods groups for search considering localization information, like Geocoding, Geographic Semantic Web and Geographic Search Engines.

The basic for a gazetteer, like the polish list of zip codes / places and dialling codes is introduced.

The geographically sensitive parsing of web resources, data indexing and extracting of geographically relevant information is presented. One of the more important tasks to solve is the ambiguity of geographical names.

A more detailed and effective gazetteer is described. This gazetteer will be the basis for the improved version of the system.

The proposed local search system is built on the Nutch framework. A web browser will offer the search functionality.

Adres

Anna KOTULLA: Politechnika Śląska, Instytut Informatyki, ul. Akademicka 16,
44-100 Gliwice, Polska, anna.kotulla@polsl.pl.