MECHANIKA
CZASOPISMO TECHNICZNE
TECHNICAL TRANSACTIONS
MECHANICS
WYDAWNICTWO POLITECHNIKI KRAKOWSKIEJ
4-M/2011
ZESZYT 7
ROK 108
ISSUE 7
YEAR 108

ANDRZEJ OPALIŃSKI[*], WOJCIECH TUREK[**], MIROSŁAW GŁOWACKI[***], MARCIN HOJNY[*]

DETEKTORY ZASOBÓW INFORMACJI W CRAWLINGU POLSKIEGO INTERNETU NA PRZYKŁADZIE PRZEMYSŁU TŁOCZNICZEGO

INFORMATION DETECTION IN POLISH WEB RESOURCES CRAWLING BASED ON STAMPING INDUSTRY EXAMPLE

Streszczenie

W artykule zaprezentowano koncepcję stworzenia narzędzia wspomagającego wyszukiwanie informacji zgromadzonych w zasobach polskiego Internetu. Działa ono opierając się na systemie zbierającym i indeksującym dane oraz dedykowane gramatyki wyszukiwania, pozwalając efektywniej odnajdywać wartościowe informacje w sieci. Zaprezentowano przewagę prezentowanej koncepcji w porównaniu z rezultatami otrzymanymi przy użyciu wyszukiwarki Google dla przykładu z przemysłu tłoczniczego. Zaprezentowano także możliwości adaptacji systemu do innych gałęzi przemysłu oraz ewolucję jego wersji podstawowej.

Słowa kluczowe: *wyszukiwanie informacji, crawling internetu, gramatyki*

Abstract

The paper presents the idea of an information extraction and search support system based on polish Web resources. System consist web crawling, data indexing and dedicated grammar syntax modules, which results with results quality improvement. As an usage example, it is presented stamp industry use case, compared to Google search results. Possible usage domains, improvement and evolution directions are shown in conclusion.

Keywords: *data mining, information extraction, web crawling, grammar*

[*] Mgr inż. Andrzej Opaliński, dr inż. Marcin Hojny, Katedra Informatyki Stosowanej i Modelowania, Wydział Inżynierii Metali i Informatyki Przemysłowej, Akademia Górniczo-Hutnicza w Krakowie.

[**] Dr inż. Wojciech Turek, Katedra Informatyki, Wydział Elektrotechniki, Automatyki, Informatyki i Elektroniki, Akademia Górniczo-Hutnicza w Krakowie.

[***] Prof. dr hab. inż. Mirosław Głowacki, Wydział Inżynierii Metali i Informatyki Przemysłowej, Akademia Górniczo-Hutnicza w Krakowie, ZI-IF, Uniwersytet Humanistyczno-Przyrodniczy w Kielcach.

1. Wstęp

Wyszukiwanie danych w sieci Internet jest nieodzowną częścią pracy współczesnego inżyniera. Aby na bieżąco śledzić nowości i wydarzenia w danej branży, nie wystarczy już jedynie przeglądanie czasopism branżowych i udział w konferencjach i sympozjach z danej dziedziny. Należy także śledzić najbardziej dynamiczne medium przekazu informacji, jakim jest sieć WWW. Ilość danych w sieci Internet lawinowo przyrasta i w lutym 2011 roku szacowana jest na około 25 miliardów stron [1].

Do wyszukiwania informacji z sieci używać można różnego rodzaju narzędzi: wyszukiwarek uniwersalnych, wyszukiwarek profilowanych oraz dziedzinowych katalogów stron. Wyszukiwarki uniwersalne są najbardziej popularnymi z używanych narzędzi ze względu na ich szeroką gamę zastosowań, dużą popularność oraz zdecydowanie największe bazy indeksowanych zasobów.

Aktualnie w rankingu najpopularniejszych wyszukiwarek liczą się 4 produkty: Google (www.google.com), Yahoo! (www.yahoo.com), Bing (www.bing.com) oraz Baidu (www.baidu.com). W aktualnym zestawieniu wyszukiwarek na świecie [2] zdecydowanie prowadzi Google z 85% fragmentem rynku. Kolejna wyszukiwarka to Yahoo z 6% oraz Bing, produkt Microsoftu – następca MSN i Live – z wynikiem 4%. Ostatnią wyszukiwarką, która przekracza 1% próg pokrycia rynku, jest używana praktycznie wyłącznie w Chinach przeglądarka Baidu z 3%. Wszystkie pozostałe wyszukiwarki pokrywają w sumie mniej niż 2% całości światowego rynku. Podobnie w Polsce przewaga Google jest miażdżąca [3]. Prowadzi on z 97% pokrycia rynku. Za nim jest polski NetSprint pokrywający 1,5% oraz Bing z wynikiem 1%. Pozostałe wyszukiwarki to zaledwie 0,5% pokrycia rynku. Popularność Google jest już tak duża, że stał się on już synonimem wyszukiwarki internetowej, a w języku potocznym powszechne stały się zwroty typu „wygooglaj".

Cechą charakterystyczną wyszukiwarek uniwersalnych jest indeksowanie przez nie zasobów z wszystkich dziedzin, bez priorytetowego traktowania żadnej konkretnej. Jest to zarówno zaletą, jak i wadą, w zależności od sposobu ich wykorzystania. Już w 2004 roku w odniesieniu do Google pojawiło się określenie „infobesity" jako zlepek angielskich słów „informacja" i „otyłość", wskazujące na ilość i jakość informacji dostarczanych przez popularne silniki wyszukiwania [4]. Wskazuje to jednoznacznie na rodzaj danych, jakie możemy uzyskać – ich ogromną ilość, przy jednoczesnej niskiej jakości.

W sieci Web dostępne są także wyszukiwarki dedykowane, indeksujące jedynie materiały z konkretnych dziedzin. Jako przykład podać można dedykowane wersje Google [5], takie jak Google Scholar indeksujący cytowania w materiałach naukowych, lub Google Books przechowujący fragmenty książek w wersji elektronicznej. Porównanie wyszukiwarek cytowań „Google Scholar" i „Web of science" wykazało niedoskonałości w ich działaniu oraz niekompletne pokrycie dziedziny, sięgające aż do 60% dokumentów w przypadku narzędzia Google [6], jednak już sam zakres dziedzinowy tych narzędzi dość mocno zawęża możliwości ich potencjalnego zastosowania w wypadku wyszukiwania informacji z różnych dziedzin przemysłu.

Dodatkowo w sieci możemy znaleźć jeszcze katalogi dziedzinowe stron, takie jak Open Directory Project[7] zbierające w jednym miejscu informacje dotyczące firm i przedsiębiorstw z wybranej branży. Dostarczają one jednak jedynie podstawowych informacji na temat ich działalności, oferując dodatkowo odnośniki do stron źródłowych, a nie indeksując zasobów tam się znajdujących.

Z tego powodu w większości przypadków, gdy zachodzi potrzeba wyszukania materiałów dotyczących konkretnej tematyki, brakuje wyspecyfikowanych konkretnych firm czy też organizacji, konieczne jest użycie wyszukiwarki uniwersalnej, indeksującej wszystkie zasoby, co w praktyce kieruje nas do dominującego w Polsce i świecie Google. Wyszukiwanie w Google informacji z wyspecjalizowanych dziedzin przemysłu nastręcza nierzadko sporych problemów, dlatego powstała koncepcja stworzenia systemu, który mógłby usprawnić wyszukiwanie takich danych. Oparta jest ona na projekcie [8], w którym powstał indeks polskiego Internetu oraz architektura pozwalająca na implementacje koncepcji pozwalającej zwiększyć efektywność wyszukiwania informacji. W prezentowanym artykule posłużono się przykładem z przemysłu tłoczniczego.

2. Wyszukiwanie uniwersalne – wady i zalety

Ze względu na zdecydowaną dominację wyszukiwarki Google zarówno w Polsce, jak i na świecie, w większości przypadków to tam rozpoczyna się poszukiwanie informacji w sieci Web. Zaawansowane opcje zarówno w Google, jak i w innych wyszukiwarkach uniwersalnych, są bardzo podobne i pozwalają doprecyzować charakter zapytania.

Na przykładzie Google można zaprezentować kilka opcji zaawansowanych, które pozwalają zwiększyć precyzję zapytania [9]. Pierwszą jest wybór źródeł danych, np. grupy dyskusyjne, obrazy czy sieć Web, która jest opcją standardową. Kolejne to zastosowanie cudzysłowów, które skutkuje dokładnym dopasowaniem wyników do podanej frazy. Operator „+" uwzględnia w zapytaniu słowa popularne – *stopwords* – które standardowo są pomijane przy wyszukiwaniu. Operator „–„ powoduje wykluczenie wyrazów, które nie powinny pojawić się w rezultatach. Operator „~" zwraca rezultaty także dla synonimów słowa, przy którym został użyty. Operator „*" wskazuje miejsce, w którym mogą wystąpić jedno lub więcej słów. Dodatkowo istnieje też możliwość wyspecyfikowania elementów strony, w których powinny pojawić się wyszukiwane słowa, to jest: „title" – tytuł strony, „url" – adres, „text" – zawartość strony, „link" – linki do tej strony

Wszystkie z powyższych elementów pozwalają po odpowiednim ich użyciu poprawić jakość wyników wyszukiwania w porównaniu do podstawowej wersji zapytania. Niestety w dalszym ciągu jakość zwracanych rezultatów pozostawia często wiele do życzenia. Pierwszym minusem charakteryzującym rezultaty wyników wyszukiwania Google w odniesieniu do poszukiwania informacji ze specjalistycznych gałęzi przemysłu jest sposób sortowania wyników, jakie zostały zwrócone dla użytkownika. W Google opiera się on na algorytmie PageRank [10], w którym dużą rolę odgrywa ilość odnośników do danej strony, co sprawdza się w wypadku wyszukiwania informacji z dziedzin rozrywki, ale niekoniecznie musi być podstawowym atrybutem przy wyszukiwaniu danych ze specjalistycznych gałęzi przemysłu. Dodatkowo część pozostałych czynników wpływających na kolejność sortowania wyników jest nieujawniona przez Google, co skutkuje tym, że rezultaty wyszukiwania często są ustawione w zupełnie innej kolejności niż ta, jakiej oczekiwałby wyszukujący. To wszystko wiąże się z koniecznością spędzenia sporej ilości czasu, poświęconego na ściągnięcie i przeglądnięcie kolejnych stron rezultatów wyszukiwania, zanim natrafimy na te, które są dla nas faktycznie interesujące.

Kolejny ważny element to ilość wyników, jakie zwraca Google. Niezależnie od całkowitej ilości odnalezionych stron, przedstawionej na pierwszej stronie wyszukiwania,

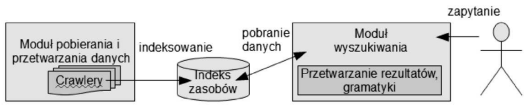
do przeglądnięcia dla użytkownika dostępna jest jedynie niewielka ich część. Aktualnie jest to około 600 wyników i nie ma możliwości, aby uzyskać kolejne. Takie ograniczenie zapewnia efektywność działania systemu przy tak dużej jego popularności, obciążeniu i ilości indeksowanych danych, ale znacznie ogranicza dostęp do całości informacji dostępnych w sieci Web.

Następną kwestią jest ilość danych z poszczególnych stron, które Google przechowuje w swoich indeksach. Z przeprowadzonych testów wynika [11], że crawlery przetwarzające dane dla tej wyszukiwarki indeksują jedynie od 100 do 500 pierwszych kilobajtów tekstu z danej strony. Jest to ilość wystarczająca dla zastosowań uniwersalnych, natomiast mogą pojawić się sytuacje, gdzie jednak będzie to wartość zbyt niska.

Powyższe elementy powodują, że użycie uniwersalnej wyszukiwarki, takiej jak Google, do odnalezienia informacji ze specjalistycznych gałęzi przemysłu może być niewygodne lub nawet niewystarczające. Z tego powodu zaprezentowano koncepcję systemu, który ma za zadanie usprawnić wyszukiwanie informacji z polskojęzycznych zasobów wielu dziedzin przemysłu opierając się na aktualnych dostępnych rozwiązaniach.

3. Opis koncepcji systemu – architektura i działanie

Prezentowana koncepcja oferująca rozwiązanie powyższego problemu opiera się na istniejącym systemie zawierającym indeks zasobów polskiego Internetu, moduły pobierające i indeksujące dane oraz moduł wyszukujący i prezentujący wyniki [8]. Dodatkowym elementem, jaki należałoby dodać do istniejącego już systemu, jest moduł przetwarzania rezultatów, zawierający tzw. „gramatyki". Uproszczony system przetwarzania wygląda tak jak na rysunku 1.



Rys. 1. Uproszczony schemat systemu
Fig. 1. Simplified system's schema

Moduł pobierający dane działa w sposób ciągły, na klastrze 10 komputerów PC, indeksując w swoich bazach danych wszystkie zasoby w języku polskim, jakie zostaną napotkane w trakcie pracy systemu. Za pobieranie kolejnych zasobów odpowiedzialne są tzw. crawlery – komponenty przetwarzające kolejne strony oraz zapisujące znalezione na nich adresy wychodzące, które w przyszłości także zostaną przetworzone. Jednym z elementów procesu przetwarzania zasobów przez crawler jest jego pobranie oraz późniejsze przetworzenie jego struktury i zapisanie w bazie danych (tzw. indeksie zasobów) informacji o zawartości treści strony oraz jej strukturze. Te informacje są podstawą do późniejszego przetwarzania przez system wyszukujący.

Podstawowymi zaletami prezentowanej koncepcji systemu, w porównaniu z uniwersalną wyszukiwarką, są:

– szybszy dostęp do zasobów – indeksowane i przechowywane w bazie danych systemu informacje (treść strony) eliminują konieczność połączenia i pobrania ich z serwerów źródłowych, co wydatnie skraca czas w porównaniu z przypadkiem przetwarzania online,

– możliwość prezentacji dowolnej liczby wyników – w przeciwieństwie do ograniczonej liczby wyników prezentowanej przez wyszukiwarki,
– możliwość definiowania własnego algorytmu sortującego wyniki – w przeciwieństwie do niejawnych algorytmów sortowania wyników oraz reklam dodawanych do rezultatów wyszukiwania w prezentowanej koncepcji systemu, sposób sortowania jest możliwy do zdefiniowania przez użytkownika lub administratora systemu,
– możliwość definiowania zaawansowanych kryteriów wyszukiwania, opierających się na danych o zawartości i strukturze strony. Ten element nazywany jest „gramatyką" i definiuje zasady, na podstawie których wyszukiwane zostaną elementy na stronie. W wyszukiwarkach uniwersalnych nie ma takiej możliwości, a w wypadku proponowanej koncepcji gramatyki mogą być dowolnie zaawansowane, w zależności od potrzeb i wymagań użytkownika. Zapytanie odbywa się w dwóch etapach. W pierwszym etapie wyszukiwane są wybrane słowa znajdujące się w gramatyce, a w drugim strona przetwarzana jest pod kątem zgodności z zasadami danej gramatyki.

Powyższe cechy zapewniają przewagę prezentowanej koncepcji w porównaniu z uniwersalnymi wyszukiwarkami, nawet przy zastosowaniu w nich zaawansowanych formuł zapytań. Gramatyki, które są kluczowymi elementami systemu, mogą operować na:

– liczba wystąpień danego słowa na stronie (liczbowo lub procentowo) lub jego odmiany,
– całkowitej liczba słów na stronie lub rozmiarze strony (w bajtach),
– liczba linków wychodzących z danej strony (obcych/własnych),
– liczba obrazków na stronie,
– liczba wystąpień szukanych słów w poszczególnych segmentach (akapitach, zdaniach),
– zaawansowanych wyrażeniach regularnych stosowanych na tekstowych elementach strony (np. ciągu = słowo1, od 0 do 3 słów, słowo2, znak interpunkcyjny, słowo3).

W zależności od potrzeb użytkownik może utworzyć mniej lub bardziej skomplikowaną gramatykę opisującą jego zapytanie i według tak zdefiniowanych reguł otrzyma zwrócone rezultaty zapytania.

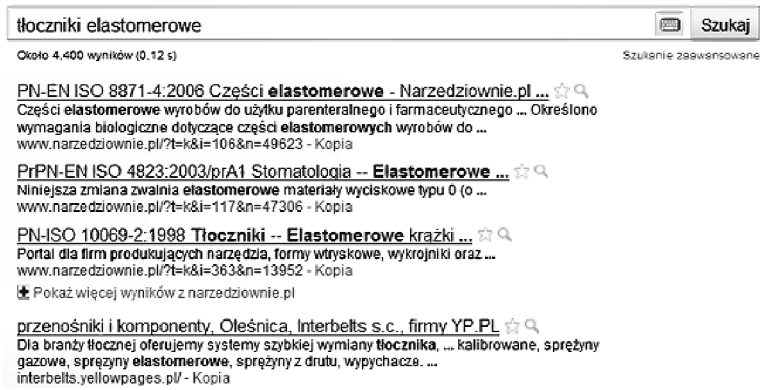
4. Przykładowy problem, propozycja rozwiązania

Jako przykładowy przypadek użycia systemu w dziedzinie przemysłu tłoczniczego zaprezentowano wyszukiwanie firm produkujących tłoczniki elastomerowe, publikujących na swoich stronach ich pełne charakterystyki. Zapytaniem wysłanym do Google będzie w tym wypadku fraza „tłoczniki elastomerowe" (bez cudzysłowów), skierowane standardowo do zasobów sieci Web. Wyszukiwarka zwraca informację, że szukane słowa znalezione zostały na 4400 stronach, co widać na rysunku 2, natomiast dla użytkownika zwrócone jest jedynie 94, co widać na rysunku 3. Przeglądając wyniki zwracane przez przeglądarkę, na pierwszej stronie możemy stwierdzić, że są to dane zupełnie nieprzydatne: katalogi norm, adresy firm, katalogi firm zajmujących się tłocznictwem. Żadne z tych adresów nie zawierają informacji przydatnych, a zwrócone są na szczycie listy wyników.

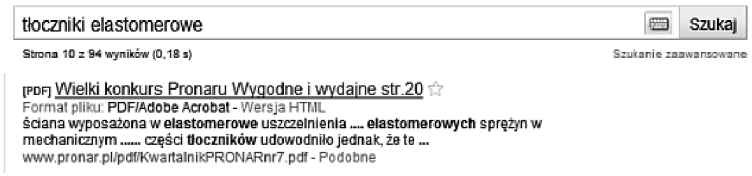
Strona zawierająca interesujące nas dane znajduje się dopiero na 41. miejscu na liście wyników wyszukiwania, a jej zawartość przedstawiona jest na rysunku 4.

Aby uzyskać pożądany, w prezentowanym przypadku, rezultat na wysokiej pozycji w liście wyników, należałoby w naszym systemie utworzyć odpowiednią gramatykę, którą moduł wyszukiwania użyłby do dopasowania i sortowania wyników na danych ze stron

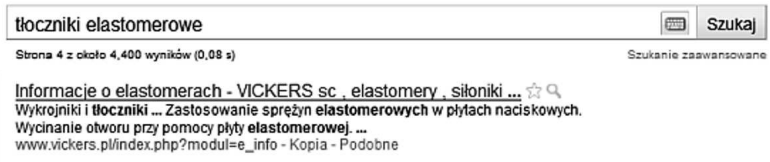
pobranych z bazy. Specyfikacja przykładowej gramatyki zapewniającej wysoką pozycję wartościowych stron, wyglądałaby następująco:
– minimum 7-krotne wystąpienie na stronie słowa „elastomer” lub jego odmiany,
– minimum 2 rysunki na stronie,
– nie więcej niż 2 odsyłacze do stron spoza danej domeny,
– co najmniej 20% akapitów zawiera frazę „tłoczenie”.



Rys. 2. Pierwsza strona wyników wyszukiwania w Google
Fig. 2. First page of search results

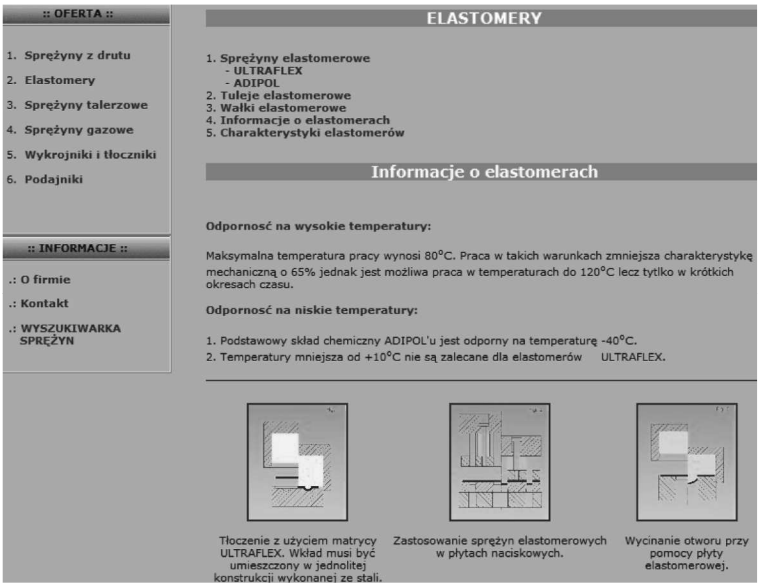


Rys. 3. Całkowita ilość wyników wyszukiwania w Google
Fig. 3. Total number of search results



Rys. 4. Strona z przydatnymi informacjami na 41. pozycji
Fig. 4. 41 position of relevant page

Strony, które w Google zajmują pierwsze kilkadziesiąt pozycji, zostałyby w takim przypadku uszeregowane poniżej właściwej strony, ponieważ prezentowałyby niższą zgodność ze zdefiniowaną gramatyką. Skutkowałoby to niższym czasem potrzebnym na odnalezienie wartościowych informacji. Na tym prostym przykładzie można stwierdzić, że już nieskomplikowana gramatyka oferuje znaczną poprawę wydajności wyszukiwania. Widać także, że istnieje możliwość definiowania bardziej złożonych gramatyk, służących bardziej precyzyjnemu wyszukiwaniu poszukiwanych informacji.



Rys. 5. Strona zawierająca wartościowe informacje (Vickers s.c.)
Fig. 5. Content of web page with valuable data (Vickers s.c.)

5. Wnioski

Wydaje się, że koncepcja prezentowanego systemu jest słuszna w zastosowaniu w wielu dziedzinach przemysłu, gdzie ilość informacji zgromadzonych w sieci jest ogromna, a ich jakość nie idzie w parze z ilością. Konstruowanie gramatyk operujących na zawartości strony, służących bardziej precyzyjnemu wyszukiwaniu informacji oraz korzystanie z indeksu danych przechowywanego w bazie, znacznie przyspiesza proces uzyskania wartościowych danych, w porównaniu z użyciem uniwersalnych wyszukiwarek. Daje to także możliwość definiowania dedykowanych gramatyk dla poszczególnych dziedzin przemysłu, ukierunkowanych na konkretne charakterystyki danej branży. Po odpowiedniej rozbudowie sprzętowej systemu możliwe jest też rozszerzenie dziedziny działania systemu na dodatkowe języki, co jeszcze bardziej zwiększyłoby jego perspektywy wykorzystania.

Praca wykonana w ramach działalności statutowej AGH – umowa nr: 11.11.110.011.

Literatura

[1] De Kunder M., *Daily estimated size of the World Wide Web*, luty 2011 (www.worldwidewebsize.com).
[2] Net Applications, *Search Engine Market Share*, luty 2011 (marketshare.hitslink.com/search-engine-market-share.aspx?qprid=4).

[3] GemiusSA, *Ranking silników wyszukiwarek w Polsce*, luty 2011 (www.ranking.pl/pl/rankings/search-engines.html.
[4] Bell S., *The infodiet: how libraries can offer an appetizing alternative to Google*, The Chronicle of Higher Education, Vol. 50, No. 24, 2004, B15.
[5] Jacso P., *Amazon Google Book Search and Google Scholar*, Online, Wilton, CT, ETATS-UNIS, Vol. 32, No. 2, 2008, 51-54.
[6] Celoch H., *Google Scholar alternatywą dla Web of Science?*, III Ogólnopolska Konferencja Naukowa „Zarządzanie informacją w nauce", Katowice, 15–16 grudnia 2010.
[7] The Open Directory Project. *About the open directory project*, luty 2011 (dmoz.org/about.html).
[8] Opaliński A., Turek W., *Wyszukiwanie informacji i analiza tożsamości*, [w:] *Metody sztucznej inteligencji w działaniach na rzecz bezpieczeństwa publicznego*, Wydawnictwa AGH, Kraków 2009.
[9] Google, *Advanced search options – Google*, http://www.google.com/pdf/Google SearchGuide-back.pdf.
[10] Page L., Brin S., Motwani R., Winograd T., *The PageRank Citation Ranking: Bringing Order to the Web*, 1999.
[11] Price G., *Google Scholar documentation and large PDF files*, 2004, http://blog.searchenginewatch.com/blog/041201-105511.