STUDIA INFORMATICA 2010
Volume 31 Number 2B (90)

Anna KOTULLA
Politechnika Śląska, Instytut Informatyki

SYSTEM ZARZĄDZANIA PUBLIKACJAMI NAUKOWYMI

Streszczenie. Artykuł omawia teoretyczne podstawy systemów zarządzania publikacjami naukowymi, opisując m.in. analizę cytowań. Opisano najważniejsze obecnie bazy danych i systemy wyszukiwania artykułów naukowych. Przedstawiony został nowy system zarządzania publikacjami naukowymi.

Słowa kluczowe: bazy wiedzy, bibliografia, analiza cytowań

SCIENTIFIC DOCUMENT RETRIEVAL SYSTEM

Summary. This article presents the theoretical fundamentals of scientific document retrieval system, describing et al. citation analysis. The most important scientific documents databases and search systems for scientific publications are described. A new scientific document retrieval system are introduced.

Keywords: knowledge databases, bibliography, citation analysis

1. Wprowadzenie

Artykuł omawia od strony teoretycznej analizę cytowań, prezentując sposoby przedstawiania dokumentów i cytowań. Opisane są m.in. sposoby wyznaczania podobieństwa pomiędzy dokumentami.

Opracowanie wymienia najpopularniejsze obecnie systemy zarządzające publikacjami, oferujące wyszukiwanie publikacji i realizujące automatyczną analizę cytowań dokumentów naukowych (*CiteSeer*, *Google Scholar*, serwis *DBLP Computer Science Bibliography*, *Scopus*, baza danych *ISI Web of Knowledge*). Omówiona została architektura lokalnych (stosowanych przykładowo przez instytucje naukowe lub gromadzących zasoby konferencyjne) systemów zarządzania publikacjami. Systemy lokalne najczęściej nie uwzględniają cytowań dokumentów.

Systemy zarządzające publikacjami (czyli naukowe bazy danych) są ciągle udoskonalane, pomimo to w dalszym ciągu nie oferują jeszcze pełnego spektrum pożądanych funkcji. Żaden z tych systemów nie oferuje możliwości wyszukiwania w zbiorze wszystkich opublikowanych artykułów naukowych. Brakuje również dopasowania wyników do zapytania w taki sposób, żeby najpierw zwracane były najważniejsze publikacje z szukanej dziedziny.

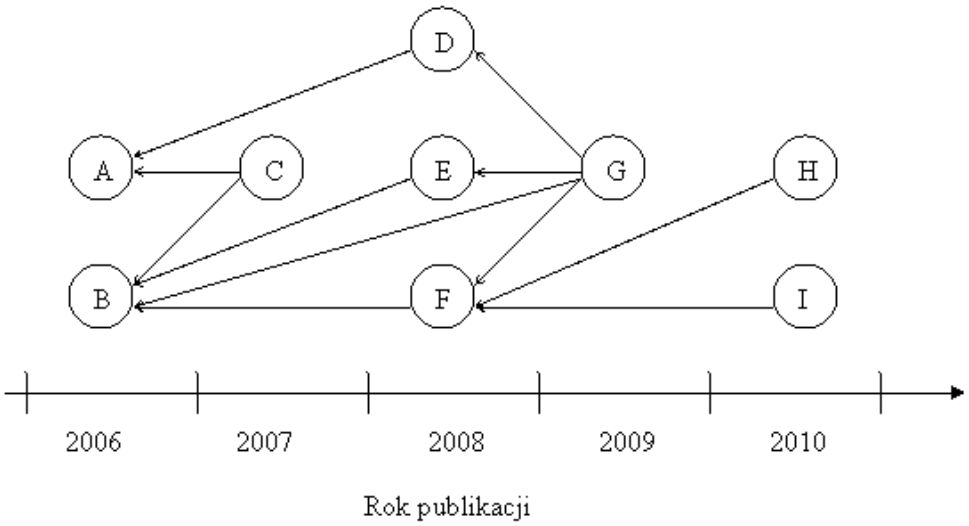
Jaki powinien być system zarządzania i wyszukiwania publikacji? Powinien zgromadzić możliwie kompletny zestaw artykułów naukowych, uaktualniany na bieżąco. Nie jest to dużym problemem w przypadku instytucji naukowych czy wydawnictw, spełnienie tego wymogu staje się jednak trudne w przypadku systemów globalnych. System musiałby wyszukiwać dokumenty dopasowane do zapytań, przy czym szukanie powinno odbywać się nie tylko po metadanych, lecz również po treści artykułu. System powinien być w stanie posortować publikacje pod względem ich prestiżu (możliwe dzięki zastosowanej analizie cytowań). Dodatkowo system musiałby działać szybko, być prosty w obsłudze. Interfejs powinien być udostępniany przez przeglądarkę internetową, dzięki czemu system taki stałby się niezależny od konkretnego systemu operacyjnego. System powinien być zaprojektowany w taki sposób, żeby w poźniejszym czasie możliwe było udostępnienie wyszukiwania semantycznego.

Artykuł przedstawia tworzony obecnie system zarządzania i wyszukiwania publikacji, przeznaczony do lokalnych zastosowań, którego architektura i funkcjonalność stara się sprostać postawionym wymaganiom.

2. Publikacje, powiązania między artykułami naukowymi

2.1. Analiza cytowań

Analiza cytowań [3, 14], będąca dziedziną bibliometrii, zajmuje się badaniem wzorców, struktury i częstotliwości cytowań (referencji) w publikacjach, takich jak artykuły naukowe czy książki. Publikacje i cytowania – jako powiązania między publikacjami – przedstawiane są często w postaci skierowanego grafu cytowań. Wierzchołki grafu reprezentują publikacje, natomiast skierowane krawędzie między wierzchołkami odzwierciedlają cytowania (dokument, od którego wychodzi krawędź, cytuje dokument, do którego skierowana jest krawędź). Rys. 1 przedstawia przykładowy graf cytowań.



Rys. 1. Graf cytowań
Fig. 1. Citation graph

Formalnie graf cytowań zapisuje się jako $G=(V,E)$, gdzie V jest zbiorem wierzchołków (dokumentów), a $E \subset VxV$ jest zbiorem krawędzi (cytowań). Zbiór E przedstawiany jest najczęściej w postaci macierzy, o niezerowych elementach $m_{xy}= 1$, jeżeli dokument x cytuje dokument y. Dla grafu z rys. 1 macierz M przedstawiona jest wzorem (1).

$$M = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix} \quad (1)$$

Najważniejsze rodzaje cytowań (widoczne również na rys. 1) to:

- cytat bezpośredni (ang. *direct citation*) – dokument *D* cytuje dokument *A*,
- współcytowanie (ang. *co-citation*) – dokument *C* cytuje dwa różne dokumenty *A* oraz *B*,
- połączenie bibliograficzne (ang. *bibliographic coupling*) – dwa różne dokumenty *H* oraz *I* cytują ten sam dokument *F*.

Do analizy grafów cytowań używa się najczęściej algorytmu *HITS* (ang. *Hypertext Induced Topic Search*) [4].

2.2. Referencje

Obecnie nie istnieją ogólnie uznane sposoby przedstawiania referencji. W artykułach z dziedziny informatyki podawane są najczęściej według jednego z dwóch popularnych systemów – według *Institute of Electrical and Electronics Engineers* (*IEEE*) [16] lub *Modern Language Assoziation of America* (*MLA*) [17]. Dodatkowo istnieje wiele umownych specyfikacji, np. określonych przez konkretne wydawnictwa.

2.3. Podobieństwo dokumentów

Podobieństwo pomiędzy dwoma dokumentami w najprostszy sposób wyrażone być może poprzez porównanie liczby referencji – biorąc pod uwagę współcytowanie lub połączenie bibliograficzne. Takie podejście faworyzuje dokumenty często cytowane. Wpływ na wyznaczone podobieństwo ma liczba dokumentów, uwzględniana jest ona w systemie (zindeksowana w bazie danych). Wadą podobieństwa wyznaczonego na podstawie liczby referencji jest brak unormowania, a co za tym idzie, trudność przy porównywaniu z wartościami podobieństwa wyznaczonymi w inny sposób.

2.3.1. Metryki jako miary podobieństwa ciągów znaków

Podobieństwo dwóch dokumentów w wiarygodny sposób wyrazić można przy użyciu specjalnych metryk, wyznaczanych dla ciągów znaków.

Cohen, Ravikumar i Fienberg [5] zbadali następujące metryki używane dla porównywania ciągów znaków: odległość *Levenshteina* (przedstawiona jako przykład miary odległości w punkcie 2.3.1.1), odległość *Mongego-Elkana*, odległość *Smitha-Watermana*, odległość *Jaro*, odległość *Jaro-Winklera* oraz miary *TF-IDF* (od ang. *term*[1] *frequency – inverse document frequency*).

Przy porównywaniu wymienionych metryk wspomnieć trzeba, że dzielą się one na dwie duże grupy, w zależności od podejścia do problemu porównywania ciągów znakowych. Miara *TF-IDF* należy do metod wyznaczających statystyczne wagi termów. Miary odległości *Levenshteina*, *Mongego-Elkana*, *Smitha-Watermana*, *Jaro* czy *Jaro-Winklera*, w odróżnieniu od miary *TF-IDF*, wyznaczają liczbę operacji edycji (zamiany, wstawiania i kasowania) potrzebną, by z jednego z porównywanych ciągów znaków uzyskać drugi. Odległość *Jaro-Winklera* [5] okazała się najszybszą z miar porównujących liczbę operacji edycji, choć i tak była znacznie wolniejsza od miary *TF-IDF*.

[1] Jako term rozumie się wyrażenie posiadające sens, które może zawierać zmienne, liczby, symbole działań algebraicznych oraz nawiasy.

2.3.1.1. Odległość Levensteina

Algorytm wyznaczający odległość *Levensteina* [18] posługuje się macierzą *D* o rozmiarze $(n+1)\times(m+1)$, gdzie *n* i *m* są długościami porównywanych ciągów znaków. Wartości macierzy *D* dla $i \geq 1,\ j \geq 1$ wyznaczane są rekurencyjnie:

$$D_{ij} = \min \begin{cases} D_{i-1,j-1} & \textit{ta sama litera} \\ D_{i-1,j-1}+1 & \textit{operacja zamiany} \\ D_{i,j-1}+1 & \textit{operacja wstawiania} \\ D_{i-1,j}+1 & \textit{operacja kasowania} \end{cases} \tag{2}$$

gdzie: $D_{00} = 0,\ D_{io} = i,\ D_{oj} = j$.

Wynikiem jest wartość umieszczona w $D_{n+1,m+1}$.

3. Impact Factor

Rozpowszechnionym i uznawanym wskaźnikiem określającym „jakość" czasopism naukowych jest tzw. *Impact Factor* (*IF*) [15]. Rozważanie *IF* nie jest bezpośrednio przedmiotem tego opracowania, jednak fakt powszechnego uznawania tegoż wskaźnika w świecie naukowym spowodował, że *IF* został krótko opisany. *IF* mierzy częstość cytowań artykułów danego czasopisma w stosunku do ogólnej liczby artykułów opublikowanych w tym czasopiśmie. *IF* wyznaczany jest jako stosunek łącznej liczby cytowań w danym roku kalendarzowym artykułów czasopisma opublikowanych w ciągu dwóch poprzedzających lat (z pominięciem autocytowań) do łącznej liczby artykułów opublikowanych w czasopiśmie w czasie wspomnianych dwóch lat. Bazę do wyznaczania *IF* (obliczanego przez *Institute of Scientific Information* [15]) stanowią bazy danych artykułów *Science Citation Index* (dla nauk technicznych, nauk przyrodniczych i medycyny) oraz *Social Sciences Citation Index* (dla nauk humanistycznych).

4. Systemy zarządzania publikacjami umożliwiające automatyczną analizę cytowań

Jako pierwsze automatyczną analizę cytowań dokumentów naukowych udostępniły niekomercyjne wyszukiwarki *CiteSeer* [10] oraz *Google Scholar* [11], obydwie są w dalszym ciągu dopracowywane, także oficjalnie dostępne są jako beta-wersje. Popularnością cieszy się również serwis *DBLP Computer Science Bibliography* [12]. Dwie najważniejsze komercyjne naukowe bazy danych to *Scopus* [19] oraz baza danych *Thomson Scientific's Institute for Scientific Information* (*ISI*) – *ISI Web of Knowledge* [20].

4.1. CiteSeer

CiteSeer [10] specjalizuje się w dokumentach z dziedziny informatyki. *CiteSeer* wyszukuje w sieci *WWW* dokumenty naukowe w formatach *PDF* oraz *PS*, przeszukuje ich zawartość pod kątem cytowań. Zasoby *CiteSeer* ciągle są aktualizowane. Publikacje są powiązane odnośnikami, które odwzorowują strukturę cytowań. *CiteSeer* wyodrębnia z dokumentów metadane.

4.2. Google Scholar

Google Scholar [11] przy dostępie do dokumentów korzysta z interfejsu i zasobów *CrossRef* [13]. *CrossRef* jest projektem powołanym do życia przez 45 wiodących wydawnictw naukowych, mającym na celu umożliwienie skonstruowania wyszukiwarki artykułów naukowych. *CrossRef* bazuje na systemie *DOI* (ang. *Digital Object Identifier*). Oprócz treści artykułu, interfejs *CrossRef* udostępnia także metadane, np. nazwiska autorów czy rok publikacji. *Google Scholar* wyświetla dla każdej publikacji liczbę jej cytowań. Dla *Google Scholar* nie ma danych o aktualności wyników wyszukiwań, nie wiadomo również, ile artykułów naukowych zostało zindeksowanych.

4.3. DBLP Computer Science Bibliography

DBLP [12] *Computer Science Bibliography* oferuje wyszukiwanie informatycznych publikacji naukowych na podstawie zgromadzonych metadanych. *DBLP* w styczniu 2010 zgromadziło informacje o ponad 1 300 000 artykułach. *DBPL* używa do jednoznacznej identyfikacji autorów specjalnej miary podobieństwa, bazującej na sieci współautorów.

4.4. Scopus

Komercyjna baza publikacji *SCOPUS* [19] jest własnością wydawnictwa *ELSEVIER* i bazuje na bibliotece wirtualnej *ScienceDirect*. Obecnie (źródło [19], stan danych styczeń 2010) *Scopus* oferuje dostęp do prawie 18 000 tytułów wydawanych przez ponad 5 000 wydawnictw. *Scopus* posiada indeks cytowań (dane od 1996), oferuje narzędzia realizujące m.in. identyfikację autorów czy tworzenie spisów literatury.

4.5. *ISI Web of Knowledge*

Komercyjna baza danych *Thomson Scientific's Institute for Scientific Information* (*ISI*) – *ISI Web of Knowledge* [20] oferuje artykuły naukowe z wielu dziedzin, prawdopodobnie udostępnia większą liczbę publikacji niż *SCOPUS*. Dla publikacji udostępnionych przez *ISI Web of Knowledge* wyznaczono indeks cytowań, umieszczono streszczenia (dane od 1991), baza oferuje

też informację o znaczeniu artykułów dla nauki (dane zbliżone do *Impact Factor*). Bibliograficzne dane artykułów bywają niejednolite. *ISI Web of Knowledge* oferuje możliwość zapisywania wyników wyszukiwania.

4.6. Systemy używane lokalnie

Oprócz wymienionych „globalnych systemów" umożliwiających wyszukiwanie publikacji naukowych istnieje wiele „lokalnych systemów", tworzonych np. przez instytucje naukowe. Najczęściej lokalne systemy mają postać architektury trójwarstwowej, z serwerem aplikacji w warstwie pośredniej (tzw. warstwie logiki biznesowej), oraz pracującą w warstwie danych bazą danych. Zazwyczaj elementy, po których możliwe jest wyszukiwanie (np. autor, tytuł, streszczenie), wprowadzane są do pól baz danych. Lokalne systemy oferują sporo możliwości, jednak korzystne byłoby wyszukiwanie również bezpośrednio w artykułach – rozszerzenie lokalnych systemów o tę funkcjonalność nie jest proste.

5. Proponowany system zarządzający publikacjami naukowymi

5.1. Założenia

Celem jest utworzenie systemu zarządzającego publikacjami naukowymi, przeznaczonego w pierwszym rzędzie dla mniejszych zbiorów danych, np. publikacji utworzonych przez poszczególne instytucje naukowe czy powstałych jako rezultat konferencji. Proponowany system będzie w stanie zindeksować dane (lokalne lub już opublikowane w sieci *WWW*), wyodrębnić z publikacji metadane, informacje o cytowaniach oraz treść publikacji. Wyszukiwanie artykułów oferowane będzie poprzez przeglądarkę internetową, co zapewni łatwość użycia oraz możliwość integracji z istniejącą prezentacją internetową instytucji naukowej. System będzie dokonywać analizy cytowań zgromadzonych artykułów.

5.2. Nutch

Proponowany system zarządzający publikacjami naukowymi opiera się na strukturze (ang. *framework*) *Nutch* [6]. 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* [7], do której dodano elementy właściwe dla sieci *WWW*, jak np. robot internetowy (ang. *crawler*), bazę danych zawierającą 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* [8].

Obecnie dostępna jest wersja 1.0 struktury *Nutch*, opublikowana 27.03.2009. Dla struktury *Nutch* zalecane jest używanie serwleta *Apache Tomcat*. W tworzonym przykładowym systemie wykorzystano, obok *Nutch 1.0*, *Apache Tomcat* w wersji 4.1.40 oraz *SUN JAVA* w wersji 1.6.0. Architektura *Nutch* przedstawiona została na rys. 2, sposób działania struktury *Nutch* omówiono poniżej.

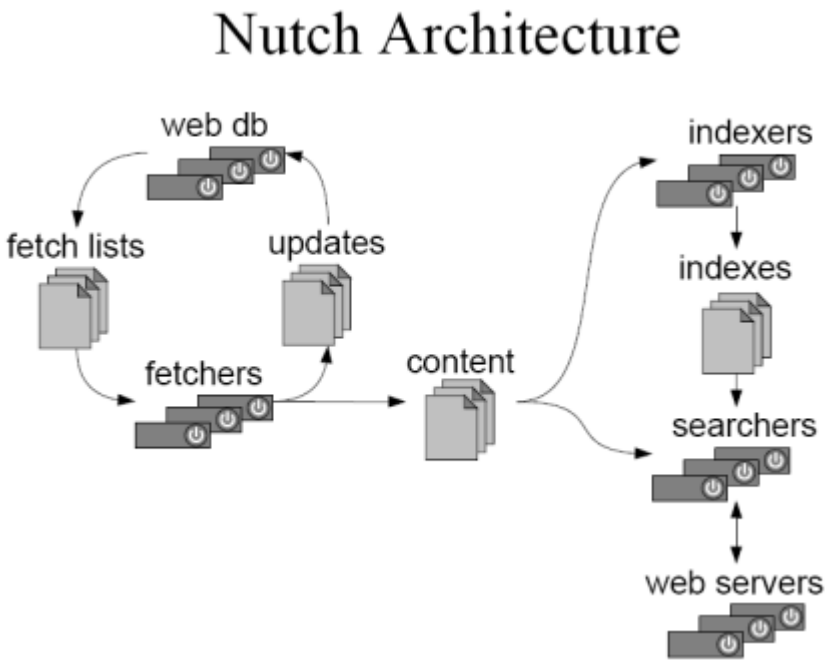
Struktura *Nutch* oferuje dwie metody pobierania stron:

- 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.

W przypadku omawianego systemu, początkowo wystarczające jest zautomatyzowane cykliczne zbieranie danych (ang. *fetching*) za pomocą polecenia *crawl* oraz aktualizowanie (ang. *update*) bazy danych, przechodzenie do kolejnych stron następuje po odnośnikach.

Nutch oferuje, poprzez 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* [2], 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 zbieraniu 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*).



Rys. 2. Architektura struktury Nutch (za [9])
Fig. 2. Architecture of the Nutch framework ([9])

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

5.3. Analiza cytowań

Planowane jest wzbogacenie tworzonego systemu o możliwość analizy cytowań. Pierwszym krokiem w tym kierunku jest wyodrębnienie referencji. Jak wspomniano, brak jest jednolitego formatu cytowań – powoduje to znaczne skomplikowanie procesu wyodrębniania referencji. Dodatkową trudność stanowi fakt, że publikacje tworzone są w różnych językach – początkowo planowane jest ograniczenie do języków polskiego i angielskiego. Konieczne jest również założenie, w jakiej postaci spodziewane są referencje. Bardzo trudne jest takie uogólnienie systemu, żeby w każdym przypadku mógł poradzić sobie z przyporządkowaniem informacji takich, jak: autor (autorzy), tytuł, wydawnictwo, rok publikacji itp. Przed zaimplementowaniem algorytmu wyodrębniającego informacje z referencji konieczne jest dokładne sprawdzenie, jakie formaty cytowań zostały użyte.

Analiza cytowań wymaga, aby dane przedstawione były w postaci grafu skierowanego. Wyodrębnione referencje pomogą utworzyć krawędzie, natomiast wierzchołki grafu zostaną zidentyfikowane poprzez pozyskanie własności dokumentów, takich jak tytuł i autorzy oraz rok publikacji, bezpośrednio z wyodrębnianego tekstu dokumentów. Należy liczyć się z faktem, że uzyskane w opisany sposów informacje o referencjach czy własności dokumentów obarczone są

błędami, stąd potrzebne są metody tworzenia grafu cytowań uwzględniające ewentualne rozbieżności. Podczas dopasowywania wierzchołków i krawędzi konieczne będzie posługiwanie się miarą podobieństwa. Początkowo, ze względu na łatwość implementacji, planowane jest użycie stosunkowo prostej odległości *Levensteina*. Wprowadzony zostanie parametr determinujący procedurę wykorzystaną do obliczenia miary podobieństwa, więc w późniejszych wersjach systemu będzie możliwe posłużenie się dokładniejszymi (i szybciej wyznaczanymi) miarami.

6. Podsumowanie

Artykuł przedstawia teoretyczne podstawy, które powinny zostać uwzględnione w systemie administrowania i wyszukiwania publikacji. Wymienione zostały najważniejsze wymagania, które powinien spełniać taki system.

Omówione zostały najważniejsze obecnie globalne platformy zarządzające publikacjami i umożliwiające wyszukiwanie publikacji.

Zaproponowany został system zarządzania publikacjami oparty na strukturze *Nutch*, przeznaczony w pierwszym rzędzie dla lokalnych zasobów danych o publikacjach. Zaletą proponowanego systemu jest jego skalowalność. Kluczowym w tym przypadku składnikiem jest struktura *Nutch*, która – teoretycznie – gotowa jest przetworzyć zasoby całej sieci *WWW*. Proponowany system poprzez implementację narzędzi analizy cytowań będzie w stanie określać wartość poszczególnych publikacji, optymalizować wyniki wyszukiwania, proponować artykuły naukowe powiązane z dziedziną wyszukiwań. Architektura systemu została tak zaplanowa, aby możliwe było późniejsze wzbogacenie go o semantyczną wyszukiwarkę informacji.

BIBLIOGRAFIA

1. Abiteboul S., Preda M., Cobena G.: Adaptive On-Line Page Importance Computation. 12th international conference on World Wide Web, Budapeszt 2003.
2. Brin S., Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 1998.
3. Havemann F.: Einfühung in die Bibliometrie. Gesellschaft für Wissenschaftsforchung, Berlin 2009.
4. Kleinberg J.: Authoritative sources in a hyperlink environment. Journal of the ACM 46 (5), 1999.
5. Cohen W. W., Ravikumar P., Fienberg S. E.: A Comparison of String Metrics for Matching Names and Records. Workshop on Data Cleaning, Record Linkage and Object Consolidation, KDD 2003.

6. Strona projektu Nutch. URL: http://lucene.apache.org/nutch/ (sprawdzono 30.01.2010).
7. Strona projektu Lucene. URL: http://lucene.apache.org/ (sprawdzono 30.01.2010).
8. Strona Apache Software Fundation. URL: http://www.apache.org/ (sprawdzono 30.01.2010).
9. Cutting D.: Nutch, Open/Source Web Search, 2004, URL: http://nutch.sourceforge.net-/twiki/Main/Presentations/www2004.pdf (sprawdzono 30.01.2010).
10. CiteSeerX. URL: http://citeseerx.ist.psu.edu (sprawdzono 30.01.2010).
11. Google Scholar. URL: http://scholar.google.com/ (sprawdzono 30.01.2010).
12. DBLP Computer Science Bibliography. URL: http://dblp.uni-trier.de/ (sprawdzono 30.01.2010).
13. Strona projektu CrossRef Search Pilot. URL: http://www.crossref.org/crossrefsearch.html (sprawdzono 30.01.2010).
14. Rubin R.: Foundations of Library and Information Science. Wydanie II, Neal/Shuman Publishers, Nowy York 2004.
15. Introducing the Impact Factor. URL: http://thomsonreuters.com/products_services/ science/academic/impact_factor/ (sprawdzono 30.01.2010).
16. Institute of Electrical and Electronics Engineers. URL: http://www.ieee.org/organizations/ pubs/transactions/information.htm (sprawdzono 30.01.2010).
17. Modern Language Association of America. URL: http://www.mla.org/publications/style (sprawdzono 30.01.2010).
18. Levenstein V. I.: Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 1966, t. 10 nr 8 (tłumaczenie artykułu z Doklady Akademii Nauk SSSR, 1965, t. 163, nr 4, s. 845÷848).
19. SCOPUS, URL: http://info.scopus.com (sprawdzono 30.01.2010).
20. ISI Web of Knowledge, URL: http://www.webofknowledge.com (sprawdzono 30.01-.2010).

Recenzenci: Dr inż. Robert Brzeski
Prof. dr hab. inż. Stanisław Wrycza

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

This article gives a theoretical overview of citation analysis and describes the formal representation of documents and citations, as citation graphs. Fig. 1 presents a sample citation graph;

the corresponding citation matrix is given by (1). The methods for measuring documents similarity are described.

This paper lists the most popular scientific databases and document retrieval systems, which offer the search for scientific papers and provide the automatic citation analysis (*CiteSeer*, *Google Scholar*, *DBLP Computer Science Bibliography*, *Scopus*, database *ISI Web of Knowledge*).

The architecture of local scientific documents retrieval systems is described. The local systems are operated by, for instance, scientific institutions. The local systems mostly do not consider the citation analysis of documents.

This paper presents a new scientific document retrieval system, intended particularly for smaller collections of data, for example articles published by a scientific institution or conference. Proposed system will be able to parse the documents into a text format, and then to index the data (stored local or published in the *WWW* network), to extract the meta-data, the information about citations and the content. A web browser will offer the search functionality. The proposed scientific document retrieval system is built on the *Nutch* framework (the *Nutch* architecture is given by fig. 2). The citation analysis will be implemented.

Adres

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