STUDIA INFORMATICA 2010
Volume 31 Number 2B (90)

Anna KOTULLA
Politechnika Śląska, Instytut Informatyki

ANALIZA ZASOBÓW INTERNETOWYCH NA PODSTAWIE STRUKTURY POŁĄCZEŃ

Streszczenie. Opracowanie omawia możliwości analizy zasobów sieci *World Wide Web* na podstawie struktury połączeń. Przedstawione są dwa najważniejsze podejścia, wyszukiwanie zasobów w całej sieci oraz wyszukiwanie informacji w zależnej od zapytania części sieci. Wskazano nowe obszary zastosowań dla metod analizy struktury połączeń.

Słowa kluczowe: Web Mining, Web Structure Mining, analiza struktury połączeń

ANALYSIS OF INTERNET RESOURCES DUE TO THE LINK STRUCTURE

Summary. This study discusses the possibilities of the analysis of the resources of the *World Wide Web* network due to the link structure. The most important ways, the searching for resources in the entire network and the searching for information in the query-depended part of the network, are presented. New application areas of the link structure analysis are indicated.

Keywords: Web Mining, Web Structure Mining, link analysis

1. Wprowadzenie do eksploracji zasobów sieci WWW

Sieć *WWW* (*World Wide Web*) [1] jest bardzo szybko rosnącym zbiorem informacji – liczba [4] indeksowanych stron sieci *WWW* w styczniu 2010 przekroczyła 21 milliardów stron. Struktura sieci *WWW* jest prosta, sieć ta składa się z węzłów, czyli dokumentów, i połaczeń między nimi, czyli hiperłączy, nazywanych również linkami, odnośnikami czy odsyłaczami. Nieuporządkowany charakter informacji - w przeważającej części dokumentów tekstowych – opublikowanych w sieci *WWW* umożliwia z jednej strony łatwą publikację czy edycję, z drugiej

strony jednak znacznie utrudnia wyszukiwanie, ponieważ nie istnieje żaden dokładny i jednoznaczny indeks dokumentów sieci *WWW*, np. podobny do indeksów baz danych czy bibliotek. Odszukiwanie informacji ułatwiają np. wyszukiwarki internetowe czy katalogi tematyczne, inną możliwością jest zastosowanie odpowiednich technik eksploracji danych.

Do analizy dokumentów sieci *WWW* i eksploracji zawartych w nich danych używane są metody zaliczane do tzw. *Web Mining* [10], czyli 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 zasób 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*) [5]. Zamiennie z terminem *Web Content Mining* używany bywa *Web Text Mining*.

Niniejsza praca omawia możliwości analizy zasobów sieci *WWW* na podstawie struktury połączeń, czyli sposób analizy zaliczany do *Web Structure Mining*. Przedstawione przykłady ilustrują omawianą dziedzinę eksploracji danych. Opisano również dwa najpopularniejsze algorytmy używane do analizowania struktury połączeń, *PageRank* oraz *HITS*.

2. Zastosowanie struktury połączeń do analizy zasobów sieci WWW

Struktury połączeń [6, 12] zasobów sieci *WWW* (ang. *link structure*) umożliwiają niezależną ocenę popularności poszczególnych stron internetowych. Sieć *WWW* rozpatrywana może być jako przykład sieci społecznej, czyli sieci złożonej ze zbioru jednostek (np. osób, organizacji, uczelni, państw) oraz ze zbioru powiązań pomiędzy jednostkami (np. hierarchia, więzy rodzinne, społeczne itp.). Sieci społeczne wizualizowane są zazwyczaj jako grafy, których wierzchołkami są wspomniane jednostki, natomiast krawędzie odzwierciedlają relacje pomiędzy jednostkami. W ten sposób [6, 12] przedstawić można również sieć *WWW* (bądź jej fragment). Formalnie wierzchołki grafu zapisać można jako zbiór

$$V = \{V_1, V_2, ..., V_N\} \tag{1}$$

Krawędzie grafu odzwierciedlające relacje pomiędzy dwoma wierzchołkami grafu V_i oraz V_j zapisane być mogą jako macierz sąsiedztwa A o rozmiarze NxN, dla której:

- $A(V_i, V_j) = 1$, jeżeli dokument V_i wskazuje na V_j, czyli są ze sobą w relacji,
- $A(V_i, V_j) = 0$, jeżeli dokument V_i nie wskazuje na V_j.

Graf ilustrujący sieć *WWW* lub jej fragment to $G=(V,A)$.

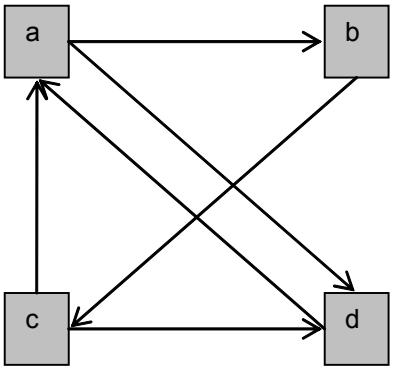
Każdemu węzłowi (dokumentowi) przypisana jest wartość prestiżu [12] (ang. *prestidge*), definiowana jako suma wartości prestiżu dokumentów, ktore na niego wskazują:

$$p(V_i) = \sum_{V_j} A(V_j, V_i)\, p(V_j)\,. \tag{2}$$

Innym spotykanym terminem [6] zamiast prestiżu jest pojęcie centralności węzła w sieci (ang. *centrality*). Wartości prestiżu [6,12] tworzą wektor p i mogą być wyznaczone jako rozwiązanie równania

$$\lambda p = A^T p\,. \tag{3}$$

Aby rozwiązać równanie (3), należy wyznaczyć wektor własny p macierzy A oraz odpowiadającą mu wartość własną λ, dalsze informacje na temat wyznaczania wektorów własnych zawarte są np. w [8]. Ponieważ dla macierzy rozmiaru NxN istnieje N wektorów własnych, rozwiązaniem jest wektor własny [6,12] z największą odpowiadającą mu wartością własną.



Rys. 1. Przykład sieci
Fig. 1. A network example

Przykładowa analiza przeprowadzona zostanie dla grafu przedstawionego na rys. 1, składającego się z wierzchołków

$$V = \{a, b, c, d\}, \tag{4}$$

dla którego pozostają w relacji

$$\{(a,b), (a,d), (b,c), (c,a), (c,d), (d,a)\}. \tag{5}$$

Macierz sąsiedztwa dla powyższego grafu przedstawia się następująco:

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{pmatrix}. \tag{6}$$

Największa wartość własna macierzy A wynosi 1,4655712, natomiast odpowiadający jej wektor własny to

$$p = \begin{pmatrix} 0{,}3702232 \\ 0{,}6228367 \\ 0{,}4249788 \\ 0{,}5425884 \end{pmatrix}. \tag{7}$$

Wartości prestiżu dla poszczególnych węzłów grafu przedstawionego na rys. 1 to 0,3702232 dla *a*, 0,6228367 dla *b*, 0,4249788 dla *c* oraz 0,5425884 dla *d*.

2.1. Algorytm *PageRank*

Algorytm *PageRank* jest najbardziej znaną metodą umożliwiającą określenie dla indeksowanych stron internetowych ich popularności na podstawie odnośników. *PageRank* [3] został stworzony przez Brina i Page'a pod koniec lat dziewięćdziesiątych. W odróżnieniu od wcześniej proponowanych rozwiązań, *PageRank* analizuje strukturę połączeń (bez podziału na linki wewnętrzne i zewnętrzne), porządkując dokumenty niezależnie od ich zawartości. Zamierzeniem Brina i Page'a było odwzorowanie zachowania przypadkowego użytkownika sieci *WWW* (ang. *random surfer modell)*, który przechodzi od strony do strony klikając w odnośniki z pewnym określonym prawdopodobieństwem.

Początkowo *PageRank* [3] przedstawiony został równaniem

$$r(a) = (1 - DF) + DF\left(\frac{r(t_1)}{C(t_1)} + \frac{r(t_2)}{C(t_2)} + \ldots + \frac{r(t_n)}{C(t_n)}\right), \tag{8}$$

gdzie: $r(a)$ jest wartością *PageRank* strony *a*, $r(t_i)$ jest wartością *PageRank* strony t_i, zawierającą odnośnik do strony *a*, $C(t_i)$ jest liczbą wszystkich odnośników zawartych na stronie t_i, DF jest współczynnikiem tłumienia, $0 \le DF \le 1$.

Współczynnik tłumienia określa przekazywaną dalej część wartości *PageRank*. Jak określili Brin i Page, współczynnik DF określa, z jakim prawdopodobieństwem przypadkowy użytkownik sieci *WWW* zniechęci się do przejścia na kolejną stronę korzystając z odnośnika, a zamiast tego wybierze inną, losową stronę. Współczynnikowi tłumienia [3] najczęściej przypisywana jest wartość 0,85.

W praktyce często stosowany jest algorytm PageRank zadany równaniem

$$r(a) = \frac{(1 - DF)}{N} + DF\left(\frac{r(t_1)}{C(t_1)} + \frac{r(t_2)}{C(t_2)} + \ldots + \frac{r(t_n)}{C(t_n)}\right), \tag{9}$$

gdzie *N* jest całkowitą liczbą stron w sieci (węzłów grafu). Równanie (13) w postaci macierzowej przedstawia się następująco:

$$r = \frac{(1 - DF)}{N} + DF \cdot A \cdot r, \tag{10}$$

gdzie A jest macierzą sąsiedztwa, której elementami niezerowymi są

$$A(t_i, t_j) = \frac{1}{C(t_j)}, \text{ jeżeli } t_i \text{ zawiera odnośnik do } t_j. \tag{11}$$

Wartości *PageRank* wyznaczyć można iteracyjnie, na przykład za pomocą algorytmu Jacobiego, opisanego np. w [8].

Page i Brin podają, że wystarczy około 100 iteracji algorytmu do obliczenia wartości *PageRank* dla całej sieci WWW.

Do słabych stron algorytmu *PageRank* zaliczyć trzeba możliwość celowego pozycjonowania stron, poprzez sztuczne utworzenie odpowiednio dużej liczby odnośników, określanego angielskim terminem *spamdexing*[1]. Zupełna niezależność otrzymanego za pomocą *PageRank* rankingu dokumentów od informacji w nich zawartych powoduje, że brak jest gwarancji, iż dokumenty zwrócone jako najbardziej pasujące do wyszukiwania rzeczywiście są najwłaściwsze.

Markov i Larose [8] wymieniają dalsze możliwe zastosowania algorytmu *PageRank*:

- Przewidywanie ruchu w sieci.
- Optymalne przeszukiwanie sieci.
- Nawigacja po stronach internetowych.

Najpopularniersza obecnie przeglądarka, czyli *Google*, korzysta m.in. z algorytmu *PageRank* i jego modyfikacji. Implementacja algorytmu tworzącego rankingi stron, którą posługuje się *Google*, jest pilnie strzeżoną tajemnicą firmy. Wspomnieć należy również, że *Google* świadomie i skutecznie walczy z praktykami zaliczanymi do *spamdexing*.

Tabela 1

Wybrane wartości *PageRank* przeglądarki *Google* (sprawdzono 16.01.2010)

URL	Wartość *PageRank*
http://www.bdas.pl	3
http://www.polsl.pl	7
http://www.ustron.pl	5
http://www.wikipedia.org	9
http://www.google.com	10
http://www.google.pl	7

Możliwe jest sprawdzanie (uproszczonych) wartości *PageRank* dla poszczególnych odwiedzanych stron. W tym celu można skorzystać z oferowanej przez *Google* wtyczki dla przeglądarek internetowych *Internet Explorer* oraz *Mozilla Firefox*, tzw. *Google Toolbar* (wtyczka jednak jednocześnie śledzi zachowanie użytkownika), bądź z dostępnych licznych

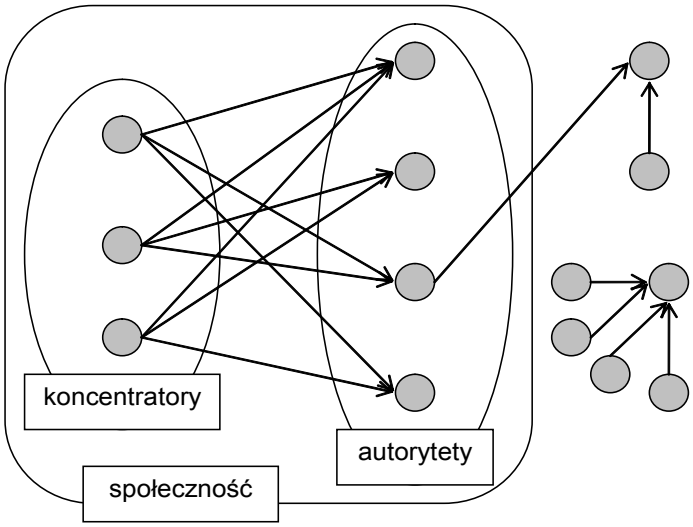
[1] Termin utworzony od angielskich słów *spam* oraz *indexing*.

serwisów (łatwe do odnalezienia poprzez wyszukiwarkę, bazujące na *Google Toolbar*). Wartość *PageRank* dla wybranych stron przedstawiono w tabeli 1.

Przeglądarki uaktualniają ciągle swoje indeksy stron. Dla stron o wysokiej wartości *PageRank* aktualizacje odbywają się częściej, gdyż strony te są również częściej zmieniane i odwiedzanie przez wielu użytkowników.

2.2. Algorytm HITS

W sieciach społecznych wyodrębnić się dają grupy jednostek połączonych między sobą relacjami, przykładowo pracownicy pewnej firmy. Podobnie w sieci *WWW* istnieją pewne skupiska połączonych pomiędzy sobą odnośnikami stron, na przykład dotyczących jednej grupy tematycznej.



Rys. 2. Przykład sieci z koncentratorami i autorytetami
Fig. 2. An example of a network with hubs and authorities

Strony, które wskazują na inne strony, nazywane są koncentratorami (ang. *hub*). Strony, do których prowadzą odnośniki, nazywane są autorytetami (ang. *authority*). Między sobą połączone strony tworzą tzw. społeczność (ang. *community*). Przykładowa sieć z wyszczególnionymi autorytetami i koncentratorami przedstawiona została na rys. 2.

Kleinberg [9] zaproponował algorytm działający na pewnej stosunkowo małej i zależnej od zapytania części grafu reprezentującego sieć. Algorytm ten nazywany jest *HITS*, od *Hyperlink-Induced Topic Search*.

Przed właściwym przyporządkowaniem stronom wartości wyodrębniany jest zbiór istotnych stron, proces ten nosi nazwę wydobywania tematu (ang. *topic distillation*). W ramach tego procesu znajdowany jest zbiór-korzeń (ang. *root set*), który wzbogacany jest o strony połączone odnośnikami ze stronami należącymi do zbioru-korzenia. W ten sposób powstaje zbiór-baza (ang. *base set*).

Każdej stronie v należącej do zbioru bazowego [9, 11] przyporządkowywana jest wartość koncentratora (ang. *hub score*) *h(v)* oraz wartość autorytetu (ang. *authority score*) *a(v)*. Początkowo dla każdej strony przyjmuje się, że $h(v)= a(v)=1$. Niech $v \rightarrow y$ oznacza, że istnieje odnośnik na stronie v wskazujący na stronę y. Kluczową operacją wykonywaną przez iteracyjny algorytm *HITS* jest uaktualnienie wartości koncentratorów oraz autorytetów dla stron, zadanych równaniami (12) oraz (13).

$$h(v) \leftarrow \sum_{v \rightarrow y} a(y) , \tag{12}$$

$$a(v) \leftarrow \sum_{v \rightarrow y} h(y) . \tag{13}$$

Równania (12) oraz (13) dla h oznaczającego wektor wartości koncentratorów oraz a oznaczającego wektor wartości autorytetów w postaci macierzowej przyjmują postać

$$h \leftarrow Aa , \tag{14}$$

$$a \leftarrow A^T h . \tag{15}$$

A jest macierzą sąsiedztwa, dla której:

- $A(v,y) = 1$, jeżeli dokument v wskazuje na y,
- $A(v,y) = 0$, jeżeli dokument v nie wskazuje na y.

Podstawiając do równania (14) wektor a zadany równaniem (15) oraz podstawiając do równania (18) wektor h zadany równaniem (14) otrzymujemy

$$h \leftarrow AA^T h , \tag{16}$$

$$a \leftarrow A^T Aa . \tag{17}$$

Aby rozwiązać równania (16) oraz (17), zapisać je można z użyciem wartości własnych i rozwiązać podobnie jak (3):

$$h = \frac{1}{\lambda_h} AA^T h , \tag{18}$$

$$a = \frac{1}{\lambda_a} A^T Aa . \tag{19}$$

Algorytm *HITS* potrzebuje stosunkowo dużo czasu na ocenę wartości koncentratorów i autorytetów, ponieważ są one zależne od struktury połączeń grafu sieci, nie jest możliwe wcześniejsze wyznaczenie tych wartości.

3. Podsumowanie

Analiza struktury połączeń niesie ze sobą różne problemy. Jednym z podstawowych jest zebranie reprezentacyjnych danych do przeprowadzenia analizy. W przypadku korzystania z wyszukiwarek internetowych, zebrane dane zazwyczaj są niekompletne. Odmienne wyniki otrzymane mogą być np. w zależności od tego, czy adres *URL* wpisany zostanie z kończącym znakiem „/". Zdarza się również, że na stronie *WWW* pewne odnośniki powtarzają się, co wpływa na zwracane przez wyszukiwarki wyniki.

W artykule przedstawiono dwa najważniejsze podejścia do analizy struktury połączeń, tzn. analizowanie sieci jako całości (algorytm *PageRank*) oraz analizowanie zależnej od zapytania części sieci (algorytm *HITS*). Najistotniejsze różnice pomiędzy algorytmami *PageRank* oraz *HITS* przedstawia tabela 2.

Tabela 2

Najważniejsze różnice pomiędzy algorytmami *PageRank* i *HITS*

PageRank	*HITS*
oblicza wartości autorytetów	oblicza wartości autorytetów i koncentratorów
ważny jedynie prestiż konkretnej strony	podąża za połączeniami
działa na całym grafie sieci –wysoka złożoność obliczeniowa	działa na podgrafie sieci
wyliczany niezależnie od wprowadzonego zapytania, stąd krótki czas odpowiedzi	wyliczany dla konkretnego zapytania, stąd dłuższy czas odpowiedzi i lepsze dopasowanie wyników

Algorytm *PageRank* stosowany może być również w innych dziedzinach, jak przykładowo ocena prestiżu artykułów naukowych [13]. Artykuły naukowe w takim przypadku traktowane są jak węzły grafu, natomiast jego krawędziami są w przypadku artykułów naukowych cytowania. Problemem dla określania prestiżu artykułów jest, podobnie jak w przypadku sieci *WWW*, dynamika zmian – ciągle przybywają nowe artykuły i wzrasta liczba cytowań. Zmodyfikowany algorytm *PageRank* może zostać użyty do predykcji liczby cytowań artykułów naukowych. Nowym polem badań może być rozszerzenie badań o sieci zawierające dodatkowe informacje, jak np. konferencje z przypisanymi im artykułami.

Oprócz omówionych algorytmów *PageRank* i *HITS*, znane są inne podejścia, np. zaproponowany przez Bharata i Mihaila algorytm *Hilltrop* [2] czy zaproponowany przez Gyöngyi, Garcia-Molinę i Pedersena algorytm *TrustRank* [7].

BIBLIOGRAFIA

1. Berners-Lee T.: Information Management: A Proposal. CERN, Genewa, 1989-1990, URL: http://www.w3.org/History/1989/proposal.html, (sprawdzono 14.01.2010).

2. Bharat K., Mihaila G. A.: Hilltop: A Search Engine based on Expert Documents. 9th International WWW Conference, 2000.
3. Brin S., Page L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems, 1998.
4. De Kunder M.: WorldWideWebSize.com. Daily estimated size of the World Wide Web, URL: http://www.worldwidewebsize.com, (sprawdzono 14.01.2010).
5. 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, s. 112÷117.
6. Feldman R., Sanger J.: The Text Mining Handbook. Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.
7. Gyöngyi Z., Garcia-Molina H., Pedersen J.: Combating Web Spam with TrustRank. Proceedings of the Thirtieth international conference on Very large data bases – Vol. 30, Toronto 2004.
8. Kiełbasiński A., Schwetlick H.: Numeryczna algebra liniowa. WNT, Warszawa 1992.
9. Kleinberg J.: Authoritative sources in a hyperlink environment. Journal of the ACM, Vol. 46, Issue 5, 1999, s. 604÷632.
10. Kosala R., Blockeel H.: Web Mining Research: A Survey. ACM SIGKDD Explorations Newsletter, Vol. 2, Issue 1, New York 2000, s. 1÷15.
11. Manning Ch. D., Raghavan P., Schütze H.: An introduction to information retrieval. Cambridge University Press, 2008.
12. Markov Z., Larose D. T.: Eksploracja zasobów internetowych. PWN, Warszawa 2009.
13. Sayyadi, H., Getoor, L.: Ranking scientific articles by predicting their future PageRank. Society for Industrial and Applied Mathematics – 9th SIAM International Conference on Data Mining, 2009.

Recenzenci: Dr inż. Michał Kozielski
Dr hab. Tadeusz Pankowski, prof. Uniwersytetu im. Adama Mickiewicza

Wpłynęło do Redakcji 31 stycznia 2010 r.

Abstract

The *World Wide Web* (*WWW*) network is a rapidly growing set of information – the number of indexed pages exceeded 21 billion pages in January 2010. The group

of methods for analysing the documents of the *WWW* network and for exploring the contained data is called *Web Mining Methods* and divides into *Web Content Mining, Web Structure Mining* and *Web Usage Mining*. This study discusses the scope of analyzing the structure of the *WWW* network according to the link structure; it is a part of the *Web Structure Mining*.

The link structure of the *WWW* resources allows an independent evaluation of the popularity of single web pages. The *WWW* network can be investigated as an example of a social network. The social networks are usually illustrated by graphs.

The *PageRank* algorithm is the mostly known method for the calculation of the popularity of indexed web pages. *PageRank* analyses the link structure and orders the documents independent of their content. In the early stages *PageRank* was represented by the equation (8). In practice the (9) representation of *PageRank* is used. Further application areas for the *PageRank* algorithm are e.g. estimating web traffic, optimal crawling, and web page navigation. A modified *PageRank* algorithm can be used e.g. to predict future citations of scientific articles.

In the *WWW* network are commonness of mutual linked pages, e.g. relating to one topic. The pages, which link to other pages, are hubs. Authorities are pages, which are linked from other pages. Mutual connected hubs and authorities establish a community. Algorithm *HITS*, defined by (12) and (13), is running for the relatively small (and depending on the current query) part of the network graph. Algorithm *HITS* needs comparatively a lot of time for the estimation of the hubs and authorities scores, because all steps are done after the query is known.

Adres

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