ARTUR STRZELECKI*

AUTORYTATYWNE I EKSPERCKIE STRONY ŹRÓDŁEM RZETELNYCH WYNIKÓW W WYSZUKIWARKACH INTERNETOWYCH

Eksploracja sieci www (*Web mining*) to odkrywanie nowej, interesującej, potencjalnie użytecznej i dotychczas nieznanej wiedzy ukrytej w strukturze, zawartości i sposobie korzystania z sieci www. Sieć www zawiera, poza użyteczną wiedzą, ogromną ilość szumu informacyjnego i śmieci. Pozyskanie użytecznej i wartościowej wiedzy z gigantycznego repozytorium, jakim jest sieć www, jest zadaniem trudnym. Metody eksploracji połączeń (*Web linkage mining*) analizują strukturę połączeń i odnośników między dokumentami www w celu opracowania rankingu dokumentów. Metody te znajdują zastosowanie przede wszystkim w wyszukiwarkach internetowych., choć nie tylko, można bowiem dzięki nim również definiować przestrzeń wokół organizacji [5]. W artykule omówiono oraz pokazano zastosowanie dwóch znanych algorytmów usprawniających pracę wyszukiwarek: HITS oraz Hilltop.

HITS

HITS (*Hyper Induced Topic Search*) [4] skupia się na analizie struktury odnośników ze zbioru stron relewantnych w szeroko rozumianym temacie poszukiwań (*broad search topic*) i odkryciu najbardziej autorytatywnych stron w tym temacie. Autor Jon Kleinberg jest jednym z pierwszych badaczy, którzy zainteresowali się badaniem struktury dużych sieci i ma spory wkład w ich rozwój. Pierwszy raz algorytmowi przypisano nazwę HITS w drugiej, zespołowej pracy w laboratorium IBM [3].

Na poszukiwanie tematu ma wpływ rodzaj zadanego zapytania. Istnieje więcej niż jeden podstawowy typ zapytania. Uwzględnia się następujące typy zapytań (*query*).

- Zapytania szczegółowe, z wyraźnie określonym celem, np. Czy Mozilla obsługuje najnowszą wtyczkę Adobe Player?

* Wydział Zarządzania, Akademia Ekonomiczna w Katowicach, ul. Bogucicka 3, 40-226 Katowice.

- Zapytanie o szerokiej tematyce, np. Znajdź informacje o języku programowania Java.
- Zapytanie o podobne strony, np. Znajdź strony podobne do *java.sun.com*.

Koncentrując się tylko na dwóch pierwszych typach zapytań, widać różnego rodzaju przeszkody. Trudność w zapytaniu szczegółowym w przybliżeniu koncentruje się wokół problemu niedostatku. Jest bardzo mało stron, które zawierają wymagane informacje i często jest trudno określić tożsamość tych stron.

Dla zapytania z szerokiej tematyki należy się spodziewać znalezienia tysięcy albo setek tysięcy relewantnych stron w sieci. Część wyniku może być generowana przez różne warianty dla zapytań składających z więcej niż jednego wyrazu lub przez bardzo popularne słowa. W tym przypadku główną przeszkodą do poszukiwań jest obfitość, nadmiar wyników. Liczba stron, która w uzasadniony sposób została zwrócona jako relewantne jest za duża dla użytkownika do przeszukiwania. Żeby wprowadzić efektywne metody szukania w tych warunkach, należy odfiltrować spośród ogromnej listy relewantnych stron mały zbiór najbardziej autorytatywnych stron.

W tych warunkach pojawiają się komplikacje, o których należy wspomnieć. Po pierwsze, uważa się, że *www.sgh.pl*, strona domowa Szkoły Głównej Handlowej naturalnie będzie najlepszym autorytetem dla zapytania „sgh". Niestety w sieci jest ponad 47 milionów stron zawierających zwrot „sgh", a *www.sgh.pl* nie jest tą która używa tego zwrotu najczęściej. Nie ma czystej, endogenicznej (wewnętrznej) miary strony, która pozwoliłaby właściwie ocenić autorytet strony. Po drugie, zauważa się problem znalezienia stron, które nie używają w swojej treści słów kluczowych odpowiednich do zapytania. Np. dla zapytania „wyszukiwarki" spodziewamy się znaleźć naturalne autorytety (Google, Yahoo, Ask). Ten problem powtarza się w wielu miejscach. Inny przykład to znalezienie określenia „producent samochodów" na stronach Hondy lub Toyoty.

Analiza struktury odnośników pomiędzy stronami w sieci www daje podstawę do rozwiązania wcześniej wskazanych przeszkód. Odnośniki przenoszą znaczną ilość ukrytego człowieczego zdania, a ten typ opinii jest dokładnie tym, co potrzeba do sformułowania pojęcia autorytetu. Utworzenie odnośnika w sieci www reprezentuje konkretne wskazanie na następujący typ opinii. Autor strony p poprzez umieszczenie odnośnika do strony q, w pewnej mierze dodaje autorytetu do strony q. Odnośniki dają możliwość znalezienia potencjalnych autorytetów poprzez strony, które do nich odsyłają. Pozwala to rozwiązać problem, w którym wiele ważnych stron nie opisuje siebie samych dostatecznie. Na przykład duże korporacje projektują swoje witryny internetowe bardzo ostrożnie, aby przekazać pewną atmosferę, przenieść swój wizerunek. Cel takiego projektu może być zupełnie różny od samego opisu przedsiębiorstwa. Ludzie poza przedsiębiorstwem często tworzą bardziej rozpoznawalne i czasami lepsze opinie niż przedsiębiorstwo same o sobie [2].

Kleinberg opracował oparty na odnośnikach model nadawania autorytetu i pokazał jak prowadzi to do metody, która konsekwentnie identyfikuje zarazem relewantne i autorytatywne strony dla zapytania o szerokiej tematyce. Model bazuje na związku, który istnieje pomiędzy autorytetem w danym temacie a tymi stronami, które odsyłają do wielu powiązanych tematycznie autorytetów. Ten drugi typ stron został nazwany koncentratorami (*hub*). Zaobserwowano, że pomiędzy autoryteta-

mi i koncentratorami istnieje pewna naturalna równowaga w grafie zdefiniowanym przez strukturę odnośników. Wykorzystano to do rozwinięcia algorytmu, który identyfikuje jednocześnie oba typy stron. Algorytm operuje na skupionym podgrafie, który został skonstruowany z listy wyników wyszukiwania tekstowej wyszukiwarki. Technika konstruowania podgrafu jest zaprojektowana do uzyskania małego zbioru stron, który najprawdopodobniej zawiera najbardziej autorytatywne strony dla danego tematu.

Koncentratory są reprezentowane przez różne formy stron, od profesjonalnie przygotowanych list zasobów przez komercyjne strony do stron z odnośnikami na indywidualnych stronach domowych. Koncentratory same nie muszą być znaczącymi stronami lub mieć odnośniki wskazujące do nich. Ich rozpoznawalną cechą jest, że silnie oddziałują na przyznawany autorytet w wybranym temacie.

Nazwano dowolny zbiór V stron połączonych odnośnikami jako ukierunkowany graf $G = (V, E)$, węzły odpowiadają stronom a krawędź pomiędzy $(p, q) \in E$ wskazuje na obecność odnośnika ze strony p do q. Stopień wyjścia (*out-degree*) z węzła p to węzły, do których odsyła, natomiast stopień wejścia (*in-degree*) węzła p to liczba węzłów, które mają odnośniki do p. Z grafu G wyizolowano mały rejon, podgraf, w taki sposób, że $W \subseteq V$ jest podzbiorem stron, w którym wektor $G[W]$ oznacza graf w przestrzeni W taki, że wszystkie węzły są stronami należącymi do W, a jego krawędzie odpowiadają wszystkim odnośnikom na stronach należących do W.

Założono, że dane zapytanie o szerokiej tematyce, jest określone przez łańcuch zapytania σ. Do wyznaczenia autorytatywnych stron, należało wpierw określić podgraf, na którym algorytm będzie operował. Autor chciał skupić się na najbardziej relewantnych stronach, na przykład wziąć do wyliczeń wszystkie strony zawierające dane zapytanie σ. Jednak pamiętał przy tym, że posiada to dwie poważne wady. Po pierwsze, zbiór zawierałby miliony stron, co znacząco zwiększyłoby wymagane obliczenia, po drugie niektóre ze stron autorytatywnych mogłyby się nie znaleźć w tym zbiorze.

Spróbowano zbudować zbiór S_σ stron, który ma następujące właściwości:

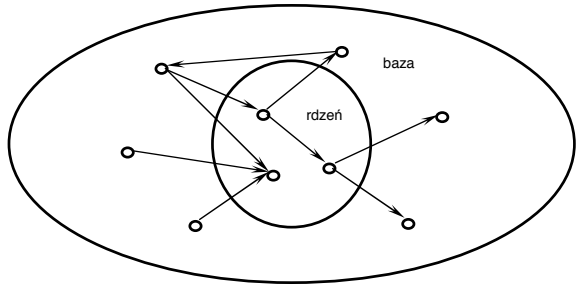
- S_σ jest relatywnie mały,
- S_σ jest bogaty w relewantne strony,
- S_σ zawiera wiele najsilniejszych autorytetów.

Utrzymanie S_σ małym, pozwala na uruchomienie nie trywialnych algorytmów, a pewność, że zawiera relewantne strony ułatwia szybsze znalezienie dobrych autorytetów. Do znalezienia takiego zbioru stron wykorzystano pierwotne wyniki z wyszukiwarki. Zgromadzono $t = 200$ pierwszych wyników dla danego zapytania σ z wyszukiwarki AltaVista lub Hotbot. Zbiór t określono jako rdzeń R_σ.

Zbiór R_σ zaspokaja wymagania postawione w punktach (a) i (b), ale jest daleki od spełnienia założenia (c). Zbiór R_σ nadal jest podzbiorem stron, które zawierają wyłącznie zapytanie σ, co jak wcześniej zauważono nie zawsze powoduje wyszukanie wszystkich autorytetów. Zaobserwowano także, że bardzo często istnieje nadzwyczajnie mało odnośników pomiędzy stronami ze zbioru R_σ, co zasadniczo wpływa na brak struktury w grafie. Na przykład, w eksperymencie zbiór R_σ dla zapytania „java" zawierał tylko 15 odnośników pomiędzy stronami z różnych domen, a zbiór R_σ dla zapytania „censorship" zawierał 28 odnośników pomiędzy stronami

z różnych domen. Te liczby są typowe dla różnych testowanych zapytań. Należy je porównać z 200*199 = 39800 potencjalnymi odnośnikami, które mogłyby istnieć w zbiorze R_σ.

Rysunek 1. Rozszerzenie rdzenia zbioru do podstawowego zbioru



Źródło: J.M. Kleinberg, *Authoritative Sources in a Hyperlinked Environment*, Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms, ACM Press, New York 1998.

Zbiór R_σ trzeba rozszerzyć do zbioru S_σ. Najlepszy autorytet dla danego tematu, może nie znajdować się w pierwszym zbiorze, ale jest bardzo prawdopodobne, że co najmniej jedna ze stron należących do R_σ, posiada odnośnik do tego autorytetu. Dlatego rdzeń grafu rozszerzono o odnośniki wychodzące i wchodzące do niego (rys. 1). Rozszerzenie zbioru polegało na umieszczeniu w nim dodatkowo wszystkich stron, do których wychodziły odnośniki ze stron znajdujących się w zbiorze R_σ, oraz dodanie stron, które same posiadają odnośniki do stron znajdujących się w zbiorze R_σ z limitem, że tych drugich nie może być więcej niż $d = 50$. Zbiór S_σ nazwano zbiorem podstawowym (*base*), którego rozmiar przy pierwotnych założeniach $t = 200$ i $d = 50$, zawierał się w przedziale 1000–5000 stron.

Należy jeszcze wspomnieć o odnośnikach, które służą wyłącznie celom nawigacyjnym. Rozróżniono dwa rodzaje odnośników. Pierwszy to odnośniki poprzeczne (*atransverse*) pomiędzy stronami z różnych domen, drugi to odnośniki wewnętrzne (*intrinsic*) znajdujące się pomiędzy stronami z tej samej domeny. Ta sama domena oznacza nazwę domeny pierwszego poziomu w adresie URL do strony. Ponieważ odnośniki wewnętrzne bardzo często istnieją wyłącznie do nawigacji, przenoszą znacznie mniej informacji niż odnośniki poprzeczne, o autorytecie strony, na którą wskazują, dlatego zostały z całego modelu usunięte. Jest to bardzo prosta heurystyka, ale efektywnie usuwa wszelkie patologie spowodowane traktowaniem odnośników nawigacyjnych w ten sam sposób jak pozostałe.

Wyznaczenie koncentratorów i autorytetów

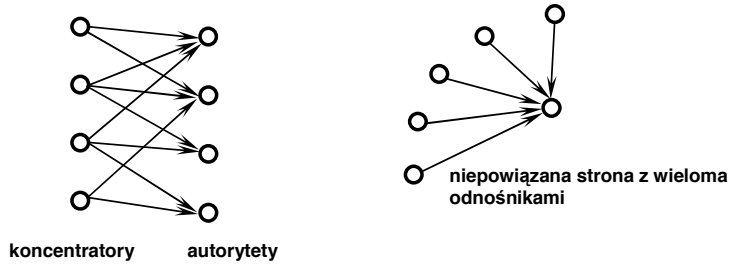
Omówiona metoda wyznacza mały podgraf G_σ, który jest relatywnie skoncentrowany na temacie zapytania, posiada wiele relewantnych stron i silnych autorytetów. Najłatwiejszym podejściem do problemu ekstrakcji tych autorytetów z całego zbioru

stron wyłącznie poprzez analizę struktury odnośników byłoby uporządkowanie stron zgodnie z ich siłą wejścia, czyli ilością odnośników odsyłających do danej strony. Ten pomysł nie sprawdził się w przypadku zbioru zawierających wszystkie strony dotyczące danego zapytania, ponieważ zdarzało się wyznaczyć na autorytet stronę zupełnie niezwiązaną z tematem. Natomiast w wyraźnie skonstruowanym małym zbiorze zawierającym większość autorytetów ta metoda się sprawdzi, dlatego że każdy autorytet należący do G_σ jest silnie powiązany odnośnikami z pozostałymi stronami w G_σ.

Takie podejście czystego zliczenia odnośników zazwyczaj działa dużo lepiej w kontekście G_σ, aniżeli we wcześniejszych ustawieniach. Jednak nadal pojawiały się problemy z właściwymi wynikami, np. dla zapytania „java" strony z największą liczbą odnośników to *www.gamelan.com* oraz *java.sun.com*, razem ze stroną reklamującą wakacje na Karaibach oraz księgarnią Amazon. Ta mieszanina w wyniku powstała na bazie prostego zliczania odnośników, jako że pierwsze dwie strony to dobre odpowiedzi, nie można tego powiedzieć o pozostałych dwóch wynikach, które mają dużo odnośników, ale są słabo związane z tematem.

W trakcie obserwacji zauważono, że autorytatywne strony relewantne do początkowego zapytania powinny mieć nie tylko wysoką liczbę odnośników, ale będąc autorytetami we wspólnym temacie, powinno istnieć znaczne pokrycie w zbiorze stron, które do nich *odsyłają. Dlatego oprócz wyszukania wysoce autorytatywnych stron, spodziewano się znaleźć koncentratory (hub pages*), czyli strony, które mają odnośniki do wielu autorytatywnych stron. To właśnie koncentratory trzymają razem autorytety we wspólnym temacie i pozwalają pozbyć się niepowiązanych stron z dużą liczbą odnośników (rys. 2).

Rysunek 2. Ścisłe powiązanie autorytetów z koncentratorami



Źródło: jak rys. 1.

Koncentratory i autorytety wykazują wzajemny, obopólnie wzmacniający związek (*mutually reinforcing relationship*). Dobry koncentrator to strona, która wskazuje do wielu dobrych autorytetów. Dobry autorytet to strona, która jest wskazywana przez wiele dobrych koncentratorów. Autor zauważył, że rezultaty uzyskane poprzez czystą analizę struktury odsyłaczy dają o wiele lepsze rezultaty, niż wyszukiwarki oparte o przeszukiwanie tekstu. W tym przypadku zastąpiono globalną analizę całej struktury odsyłaczy w WWW bardziej lokalną metodą analizy małego skupionego podgrafu.

Algorytm skutecznie sprawdza się w szerokim zakresie tematów, gdzie najsilniejsze autorytety świadomie nie zawierają do siebie wzajemnych odnośników. Mogą one być połączone pośrednio przez warstwę relatywnie anonimowych koncentratorów, które są skorelowane i odsyłają do tematycznie powiązanych grup autorytetów. Ten dwupoziomowy wzór powiązań odsłania strukturę pośród obu zbiorów, koncentratorów, które mogą wzajemnie o sobie nie wiedzieć i autorytetów, które mogą nie chcieć pogodzić się z istnieniem innych autorytetów.

Dla wielu popularnych tematów w www ilość relewantnych informacji rośnie w zaskakującym tempie, co czyni coraz trudniejszym dla indywidualnych użytkowników przeglądanie i filtrowanie dostępnych zasobów. Celem powyższej metody jest odkrywanie autorytatywnych stron, ale pokazała także bardziej skomplikowany model społecznej organizacji danych w www, gdzie koncentratory zawierają wiele odnośników do zbioru tematycznie powiązanych autorytetów. Równowaga pomiędzy koncentratorami i autorytetami jest fenomenem, który powtarza się w kontekście szerokiej różnorodności tematów w www.

Powyższej metody nie można całkowicie oddzielić od analizy zawartego na stronie tekstu. Wyszukiwanie oparte wyłącznie o odnośniki może w kilku przypadkach mijać się z prawdą [2]. W ściśle skoncentrowanych tematach, np. narciarstwo w Beskidach, HITS najczęściej zwraca dobre źródła dla bardziej ogólnych tematów, jak turystyka w Beskidach. Ponieważ każdy odnośnik w koncentratorze rozprowadza tę samą wagę, HITS czasami odchyla wyniki, gdy koncentrator omawia wiele różnych tematów. Na przykład, strona domowa chemika, może zawierać dobre odnośniki do przemysłu chemicznego oraz do źródeł związanych z jego prywatnymi zainteresowania lub informacje o mieście, w którym żyje. W tym przypadku, HITS nada część autorytetu o chemii stronom, które nie mają z nim nic wspólnego. Żeby zapobiec tym niedogodnością rozszerzono algorytm o analizę treści odnośnika oraz zaczęto dzielić duże koncentratory na mniejsze jednostki, jeśli nie był on wyłącznie skoncentrowany na jednym temacie. Treść odnośnika, który byłby brany pod uwagę przy obliczeniach, powinna zawierać słowo związane z tematem.

Hilltop

Algorytm Hilltop wyszukuje w sieci Internet autorytatywne strony. Jego autorami są K. Bharat i G. Mihaila [1]. Ich praca została napisana i opublikowana w 1999 roku. Na potrzeby nowego algorytmu stworzono prototypową wyszukiwarkę z działającym algorytmem. Autorzy w trakcie prac i badań wielokrotnie odnosili się i porównywali swoją metodę do dwóch wcześniej powstałych algorytmów, PageRank i HITS. Jednak jako pierwsi zwrócili uwagę na termin „spam pages", czyli strony, które są wyłącznie tworzone po to żeby wprowadzić w błąd wyszukiwarki. Algorytm Hilltop bazuje na tych samych założeniach, co inne algorytmy badające strukturę odnośników w www. Liczba i jakość źródeł odsyłających do strony jest dobrą miarą wartości samej strony. Kluczowa różnica polega na braniu pod uwagę wyłącznie źródeł eksperckich.

Strona jest autorytetem w temacie zapytania, jeśli i tylko jeśli, najlepsze ze stron eksperckich posiadają do niej odnośniki. Oczywiście w praktyce strony eksperckie

mogą być ekspertami w szerszym lub zbliżonym temacie poszukiwań. Jeśli tak, to tylko podzbiór odnośników na stronie eksperckiej może być relewantny. W takich wypadkach, odnośniki brane pod uwagę są starannie porównywane i sprawdzane pod względem zgodności treści odnośnika z treścią zapytania. Porównując relewantne odnośniki od wielu ekspertów w jednym temacie, autorzy znajdują strony, które są najwyżej cenione w społeczności stron o tej tematyce. To podstawa wysokiej jakości i trafności w algorytmie Hilltop. Strony autorytatywne muszą mieć, co najmniej dwa odnośniki z niestowarzyszonych stron eksperckich, żeby mogły znaleźć się w ogóle w rankingu. Ranking jest obliczany na podstawie ilości i jakości stron eksperckich.

Strony eksperckie

Strona ekspercka jest stworzona w konkretnym, określonym celu poprowadzenia użytkownika w kierunku dobrego źródła. Strona ekspercka jest skupiona na określonym temacie i posiada odnośniki do wielu niestowarzyszonych stron o tym samym temacie. Dwie strony są niestowarzyszone, jeśli ich autorami są sobie obce, niestowarzyszone organizacje. W eksperymencie tylko podzbiór stron z bazy wyszukiwarki AltaVista został uznany za ekspercki. Wybrano 2,5 miliona stron ze 140 milionów stron jako ekspertów.

Popularne tematy są dobrze reprezentowane w sieci www, tworzy się dla nich wiele ręcznie edytowanych list zasobów. Indywidualnym osobom lub organizacjom zależy żeby tworzyć listy zasobów w określonym temacie. Podnosi to ich popularność i wpływ wewnątrz społeczności zainteresowanej danym tematem. Autorzy list mają motywację do utrzymywania ich obszernych, wyczerpujących i aktualnych. Odnośniki z tych list są rekomendacjami, a strony które je zawierają są ekspertami. Żeby odróżnić obiektywnie strony eksperckie od innych, muszą one być wyraźnie bezstronne i niestowarzyszone.

Strony stowarzyszone

Strona jest stowarzyszona, kiedy jedno lub oba założenia są prawdziwe.

- Dzielą te same pierwsze trzy oktety adresu IP.
- Nazwa domeny drugiego poziomu, pomijając poziom najwyższy jest taka sama.

Domena najwyższego poziomu, potocznie nazywana końcówką, jest uzależniona od rodzaju działalności prowadzonej przez organizację. Na przykład „.com.pl" i „.pl", są domenami występującymi w wielu adresach i dlatego są traktowane jako rodzajowe, gatunkowe. Jeśli po usunięciu domeny najwyższego poziomu, rodzajowej, pozostała nazwa domeny drugiego poziomu jest taka sama, to strony są traktowane jako stowarzyszone. W porównaniu dwóch adresów np. *www.mbank.com.pl* oraz *www.mbank.pl* należy odrzucić końcówki rodzajowe „.com.pl" i „.pl". Tak pozostała domena drugiego poziomu „mbank", która w obu przypadkach jest identyczna. Obie strony są uznane za stowarzyszone.

Wybór ekspertów

Strona ma charakter ekspercki, jeśli posiada, co najmniej 5 odnośników wychodzących do odrębnych, niestowarzyszonych stron. Każda, która spełnia powyższy warunek jest ekspertem. Ekspert jest dopasowywany do zapytania użytkownika na podstawie stworzonej mapy z fraz kluczowych. Fraza kluczowa obejmuje fragment strony, który zalicza jeden lub więcej odnośników do mapy. Każda fraza kluczowa obejmuje różny zakres wewnątrz strony, a odnośniki znajdujące się w zakresie objętym przez frazę kluczową są zaliczane do mapy. Na przykład, znacznik <TITLE>, nagłówki (tekst pomiędzy parą znaczników <Hx></Hx>) oraz treść odnośnika, znacznik <A></A>, na stronie eksperckiej są traktowane jako frazy kluczowe. Znacznik <TITLE> obejmuje swoim zakresem wszystkie odnośniki ze strony. Fraza oparta o nagłówek zalicza odnośniki występujące po nim aż do następnego nagłówka o tym samym lub wyższym poziomie. Treść odnośnika obejmuje swoim zakresem tylko ten odnośnik.

Do listy stron eksperckich zostaje zakwalifikowanych pierwsze 200 stron zwracanych przez wyszukiwarkę. Następnie algorytm sortuje je pod względem relewantności znajdujących się na nich odnośników i wyznacza wynik eksperta. Strona zostanie uznana za ekspercką, jeśli zawiera, chociaż jeden odnośnik, który w treści ma wszystkie słowa kluczowe składające się na zapytanie. Dodatkowo są punktowane elementy w dokumencie, które zawierają słowa kluczowe. Tytuł strony zawierający słowa kluczowe jest 16 razy, a nagłówek 6 razy ważniejszy niż każdy następny odnośnik. Po wyznaczeniu uszeregowanej listy ekspertów, algorytm przystępuje do wyznaczenia autorytetów. Autorytet musi być, co najmniej wskazywany przez dwóch ekspertów, niestowarzyszonych ze sobą ani z autorytetem. Wszystkie autorytety otrzymują ocenę na podstawie ilości i relewantności powołujących się na nich ekspertów.

Autorzy przeprowadzili testy, które potwierdziły wysoką skuteczność ich pomysłu. Zaprezentowane wyniki w oryginalnej pracy pokazują wysoką skuteczność, przewyższającą wyniki pochodzące z innych wyszukiwarek, oprócz Google. Prawie każde zapytanie zwracało wynik trochę gorszy bądź nieznacznie lepszy od Google.

Optymalizacja pod Hilltop

Wiele stron eksperckich takich jak listy zasobów lub katalogi niszowe są aktualizowane okresowo o nowe odnośniki. Natomiast inne to statyczne strony, które nigdy nie zostaną odnowione. Wtedy liczy się „czynnik ludzki". Z czasem witryna będzie rozpoznawana jako wartościowa, autorzy innych stron zaufają jej na tyle, aby do niej odsyłać. Jednak faktem jest, że nowej witrynie jest bardzo trudno stać się autorytetem w oczach wyszukiwarki. Nasuwa to pytanie, jak nowa witryna może szybciej pojawić się w swoim sąsiedztwie? Należy zbadać, jakie powiązania mają eksperci i autorytety. Znalezienie stron eksperckich i autorytetów nie stanowi problemu, ponieważ są to witryny najwyżej rankingowane dla danego zapytania. Za pomocą narzędzi identyfikujących odnośniki można odszukać te strony z odnośnikami. Ze

znalezionych witryn należy pozyskać odnośniki. Warto także spróbować zbudować własną stronę ekspercką, na przykład poprzez stworzenie katalogu branżowego. Oczywiście należy dodać własną stronę do innych katalogów tematycznych.

Zakończenie

Podstawowym zadaniem algorytmów wykorzystywanych w wyszukiwarkach jest ranking stron, czyli ocena względnej ważności, dokumentów www. Problem rankingu jest znany od wielu lat i występuje w wielu dziedzinach zastosowań. Skutecznie działające algorytmy będą coraz szerzej wykorzystywane w wyszukiwarkach z uwagi na wzrastającą częstotliwość korzystania z nich przez użytkowników.

Literatura

1. Bharat K., Mihaila G.A., *Hilltop: A search Engine based on Expert Documents*, Poster of the 9th International World Wide Web Conference, Amsterdam 2000.
2. Chakrabarti S., Dom B.E., Gibson, D., Kleinberg J., Kumar R., Raghavan P. et all, *Mining the Link Structure of the World Wide Web*, IEEE Computer, 1999.
3. Gibson D., Kleinberg J. M., Raghavan P., *Inferring Web communities from link topology*, Proceedinsg 9th ACM Conference on Hypertext and Hypermedia, 1998.
4. Kleinberg, J.M., *Authoritative Sources in a Hyperlinked Environment*, Proceedings 9th ACM-SIAM Symposium on Discrete Algorithms, ACM Press, New York 1998.
5. Pawełoszek-Korek I., *Grupy niejawne w wirtualnym otoczeniu organizacji*, materiały konferencyjne *Technologie Wiedzy w Zarządzaniu Publicznym*, Akademia Ekonomiczna, Katowice 2007.