Zastosowania algorytmu pagerank w wyszukiwaniu

Artur Strzelecki

2008-09-23

Streszczenie

Zaczęło się od poszukiwania ciekawego tematu na pracę doktorską. Lawrence Page skupił uwagę na rodzącej się wówczas sieci World Wide Web [Batt06]. Sieć WWW była interesująca ze względu na pewne cechy matematyczne, każda strona była węzłem i każdy odnośnik na stronie był połączeniem między węzłami. Była to klasyczna struktura grafu. Page teoretyzował, że World Wide Web jest być może największym grafem w historii i rośnie w zawrotnym tempie. Zauważył, że przechodzenie odnośnikami z jednej strony na inną jest niewiarygodnie proste, jednak nie można tego powiedzieć o drodze powrotnej. Oglądając stronę internetową nie wiadomo było, jakie strony do niej odsyłają. Przydałaby się wiedza, kto jest z kim połączony

Taka analiza wywodzi się z bibliometryki publikacji naukowych. W publikacjach, oprócz recenzji, ważna jest lista cytatów. Cytat to odnośnik albo lista kluczowych informacji o publikacji, umożliwiających jej zidentyfikowanie i ponowne znalezienie. Drugim, istotnym dla publikacji naukowych pojęciem są adnotacje. Adnotacje służą do opisywania cytatów. Mogą zawierać krytykę albo komentarz. Adnotacja to osąd cytowanego dokumentu. Ostatnim elementem jest ranga publikacji. Prace są oceniane również na podstawie liczby cytowanych dokumentów, liczby dokumentów, które się potem do nich odwołują oraz domniemanej wagi każdego z cytatów.

Page doszedł do wniosku, że cała sieć WWW opiera się na wykorzystaniu cytatów i adnotacji. Odnośnik jest cytatem, a tekst odnośnika jest adnotacją. Page stworzył BackRub, system, który wykrywał odnośniki w sieci WWW, zapisywał je do analizy, a następnie prezentował adresy sieciowe, w których znajdują się odnośniki do określonej strony. Zainspirowany analizą cytatów, Page zaproponował, że liczba odnośników do danej strony może być wskazówką przy ustalaniu rangi tej strony. Uważał też, że każdy odnośnik wymaga własnego ran-

kingu, opartego na liczbie odnośników z pierwotnej strony. Zatem zliczać trzeba nie tylko odnośniki na poszczególnych stronach, ale również odnośniki załączone do odnośników, co istotnie zwiększa ilość wymaganych obliczeń. Page i Brin, z którym współpracował od początku projektu, opracowali algorytm PageRank, nazwany na cześć pierwszego, który potrafił uwzględniać zarówno liczbę odnośników do określonej strony, jak i liczbę odnośników na każdej z odsyłających stron.

PageRank jak wcześniej wspomniano, to metoda wyliczenia rankingu dla każdej strony na podstawie grafu sieci. Graf sieci składa się z wierzchołków lub węzłów, czyli stron oraz krawędzi, czyli odnośników (ang. links). Każda strona może mieć odnośniki wychodzące (ang. outlinks, outboundlinks) z niej oraz odnośniki przychodzące (ang. backlinks, inboundlinks) Te ostatnie dzielą się na wewnętrzne (ang. internal links) i zewnętrzne (ang. external links). Ranking strony zależy od ilości odnośników, które do niej prowadzą. Generalnie, strony z większą ilością odnośników mają lepszy ranking niż strony, które mają ich mniej. Tak działa proste zliczanie odnośników. Natomiast PageRank wprowadza bardziej wyrafinowaną metodę zliczania odnośników [PBMW98], która spowodowało, że nowatorski algorytm był lepszy od konkurencji w ówczesnym czasie. W wielu przypadkach zwykłe zliczanie odnośników nie koresponduje z rzeczywistą wartością strony, na którą wskazują. Na przykład, jeśli witryna ma odnośnik ze strony głównej znanego uniwersytetu, może mieć tylko ten jeden, ale jest on wartościowy. Ta witryna powinna być wyżej w rankingu niż wiele stron, które mają więcej odnośników, ale z ciemnych miejsc Internetu. Na podstawie powyższych założeń wyciągnięto wniosek, że strona ma wyższy ranking, jeśli suma rankingów stron do niej odsyłających jest wysoka. Obejmuje to dwa przypadki, gdy strona ma wiele odnośników oraz gdy strona ma ich mniej, ale wyżej rankingowanych.

Uproszczony wzór rankingu wygląda tak:

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v}$$

gdzie R(u) jest rankingiem strony u, c współczynnikiem normalizującym (częściej nazywanym d, współczynnik tłumienia – na początku autorzy przyjęli d=0,85), takim, że całkowity ranking wszystkich stron jest stały, Bu to grupa stron zawierających odnośniki do u, Nv to liczba odnośników wychodzących ze strony u. R(v) jest rankingiem strony v, która zawiera odnośnik do u. Całe równanie jest rekurencyjne. Oryginalna hipoteza autorów algorytmu zakłada, że średnia z wszystkich wartości PageRank [BrPa98] wszystkich witryn wynosi 1.

Wynika z tego, że dopóki każda strona posiada odnośnik wychodzący to suma wszystkich wartości PageRank jest równa liczbie stron w systemie.

W pierwszych testach autorzy do wyliczenia PageRank z 322 milionów odnośników potrzebowali 52 iteracji, natomiast połowa próby, czyli 161 milionów odnośników zabrała 45 iteracji. Ilustruje to wzrastającą wydajność PageRank w miarę rozrostu grafu sieci a wykres iteracji jest odwzorowany logarytmicznie. W czasie testów napotkano na problem funkcji rankingującej. Zakładamy, że dwie strony wzajemnie do siebie odsyłają, ale już nigdzie indziej. Przypuszczamy także, że jest gdzieś strona, która ma odnośnik do jednej z tych dwóch. W trakcie iteracji PageRank będzie akumulowany w pętli, natomiast nie będzie dystrybuowany dalej, ponieważ brakuje wyjścia tej pętli. Tak powstała sytuacja została określona jako rank sink.

Do obejścia tej pułapki, zastosowano model losowego surfera (ang. random surfer model). Nawiązuje on do losowego poruszania się po grafie. Losowy surfer po prostu sukcesywnie klika w losowo wybrane odnośniki. Jeśli prawdziwy użytkownik sieci dostałby się do małej pętli pomiędzy stronami, jest nieprawdopodobne, aby kontynuował poruszanie się w niej bez końca. Zamiast tego, użytkownik przeskoczy do innej strony. Takie zachowanie określa się jako okresowe znudzenie i przeskoczenie do innej strony.

Drugim zagrożeniem prawidłowego obliczenia modelu PageRank są odnośniki prowadzące do stron bez żadnego odnośnika wychodzącego (np. polityka prywatności). W trakcie badań okazało się, że w bazie znajduje się spora liczba takich odnośników. Ponieważ nie ma możliwości dystrybuowania ich rankingu do innych stron, podjęto decyzję o usunięciu z systemu wszystkich takich odnośników zanim PageRank został wyliczony. Autorzy podkreślili, że ma to wpływ na końcowe wartości, ale nie zmienia go znacząco.

Zastosowanie w wyszukiwaniu

Głównym zastosowaniem PageRank jest wyszukiwanie. Na potrzeby badań zostały zbudowane dwie wyszukiwarki używające PageRank. Pierwsza z nich to prosta wyszukiwarka oparta o tytuły stron. Druga z nich to pełno tekstowa wyszukiwarka Google. Ówcześnie Google łączyło już w sobie kilka czynników do rankingowania i szeregowania rezultatów wyszukiwania takich jak, standardowe miary pozyskiwania danych (ang. standard information retrieval measures) odległość pomiędzy słowami kluczowymi w treści (ang. proximity), treść odnośnika (ang. anchor text) oraz PageRank.

Treść odnośnika została (BrPa98) specjalnie potraktowana w tej wyszukiwarce. Zazwyczaj inne wyszukiwarki wiązały treść odnośnika ze stroną, na której się znajdował, natomiast

Google dodatkowo przywiązało tę treść do strony, na którą odnośnik wskazywał. Rozszerzenie treści odnośnika na stronę, do której odsyła najwcześniej pojawiło się w pierwszej, najstarszej wyszukiwarce McBryan's Word Wide Web Worm [McBr94]. Miało to kilka zalet, po pierwsze taki odnośnik często dostarczał ściślejszych informacji o stronie niż ona sama o sobie. Po drugie można było dotrzeć do dokumentów, które nie były wykryte przez wyszukiwarki tekstowe, takich jak obrazy, programy i bazy danych. Był to sposób na wykrywanie stron, które wcześniej nie znalazły się w indeksie.

Najlepsze rezultaty PageRank prezentuje dla sprecyzowanych zapytań. Zwracając rezultat z wynikami zapytania, testowa wyszukiwarka oparta tylko o tytuły stron, wynajdywała wszystkie strony, których tytuł zawierał każde słowo składające się na zapytanie. Następnie rezultat był sortowany według PageRank. Te dwa elementy sprawiły, że wyszukiwarka tak dobrze działała. Dopasowanie tytułów odpowiadało za precyzyjne wyniki, natomiast PageRank zapewnił wysoką jakość.

PageRank potrafi także reprezentować pewną miarę autorytetu lub zaufania. Np. mała hobbistyczna witryna może być często odwiedzana przez użytkowników, ponieważ odnośnik do niej znajduje się na stronie głównej poczytnej gazety. Taka mała witryna otrzyma wysoki PageRank, ponieważ została zauważona przez bardzo ważną stronę. Reprezentuje to pewien obraz zaufania, gdzie wspomniana witryna zyskuje na swoim autorytecie lub zaufaniu, dzięki odnośnikowi z autorytatywnego źródła.

PageRank niestety jest narażony na manipulację celem uzyskania komercyjnych korzyści. Żeby poprawić PageRank witryny trzeba pozyskać odnośniki z wysoko rankingowanych stron lub bardzo dużo odnośników ze stron o nikłej wartości. Zdaniem autorów algorytmu najgorszą formą manipulacji jest kupowanie reklamowych odnośników na popularnych stronach z wysokim PageRank. Ta słabość na komercyjne manipulacje rankingiem ma daleko idące skutki, w efekcie pogarsza jakość działania samej wyszukiwarki.

Wykorzystanie wiedzy o PageRank

Niektórzy autorzy witryn po przestudiowaniu algorytmu PageRank mieli pomysł wygenerowania milionów stron, które mogłyby wyprodukować PageRank i poprawić ranking ich własnych witryn. Teoretycznie powinno to działać przy właściwej strukturze odnośników, jednak praktycznie nie działa wcale. Google zmieniło swój algorytm lata temu. Jedna ze zmian zapobiega takiemu działaniu, inna zmienia wartość współczynnika tłumienia. Zakłada się, że usprawnienia w algorytmie są wprowadzane średnio raz na kwartał.

Inny aspekt nasuwa pytanie, które odnośniki są zliczane do obliczenia PageRank. Jeśli strona A dwukrotnie odsyła do strony B i raz do strony C to istnieją różne możliwości jak PageRank zostanie podzielony pomiędzy strony B i C. Obecna wersja algorytmu PageRank ignoruje wielokrotne linki z tych samych stron. Od 18 stycznia 2005 r., Google razem z innymi wyszukiwarkami rozpoznaje nowy atrybut w znaczniku anchor [Prcs05]. To atrybut „rel" i jest zapisywany jako: treść odnośnika . Atrybut mówi wyszukiwarce żeby kompletnie zignorowała odnośnik. Poprzez ten odnośnik nie zostanie przekazany PageRank do wskazywanej strony i nie pomoże jej rankingowi, jakby odnośnik wcale nie istniał.

Aktualnie istnieją dwa sposoby pozyskania niezależnej informacji o wartości PageRank: poprzez pasek narzędzi (ang. toolbar) Google oraz katalog Google. Oczywiście istnieje wiele innych narzędzi, które oferują wyświetlanie wartości PageRank w przeglądarce lub w kodzie HTML, jednak wykorzystują one informacje pochodzące z usługi paska narzędzi.

Pasek narzędzi Google (Toolbar Google) używa logarytmicznej skali w przedziale od 0 do 10. Przyrost wartości na logarytmicznej skali w pasku narzędzi odpowiada przyrostowi rzeczywistego PageRank logarytmowanego przy podstawie b, ustalanej na nowo przy każdorazowej aktualizacji rankingu. Najbardziej prawdopodobne jest rozwiązanie, że Google normalizuje skale tak, aby zawsze strona z najwyższym PageRank miała wartość nie większą niż 10. Skorzystanie ze skali logarytmicznej, bądź podobnej oznacza, że potrzeba o wiele więcej dodatkowych odnośników, które przeniosą wysoki PageRank, żeby przenieść stronę z obecnego poziomu PageRank na następny.

Katalog Google używa logarytmicznej skali z przedziału od 0 do 7. Dwie różne skale mogą dostarczyć dodatkowych, bardziej ścisłych informacji o PageRank. Strony, które mają jeden wynik w pasku narzędzi mogą mieć inny wynik w katalogu. Oczywiście takie informacje można wyświetlić tylko dla stron, które znajdują się w ODP (Open Directory Project, www.dmoz.org), ponieważ katalog Google wyświetla jego zawartość. Jednakże mogą wyniknąć nieścisłości, gdy w różnym czasie zostaną wykonane aktualizacje wartości w pasku narzędzi i w katalogu. Wydaję się, że katalog jest częściej aktualizowany.

W ostatnim czasie zaobserwowano wiele prób manipulowania wynikiem PageRank pokazywanym w pasku narzędzi. Można to wykonać przez przekierowanie jednej strony na drugą z wysokim PageRank. Google łączy te strony i pokazuje jeden PageRank dla strony docelowej, jaki i dla przekierowanej. Jeśli przekierowanie zostanie zastąpione nową treścią, pasek narzędzi pokazuje dla niej przez jakiś czas fałszywy PageRank. Czasem strony są zamasko-

wane (ang. cloaked) i przekierowanie widoczne jest jedynie dla wyszukiwarki, podczas gdy użytkownicy widzą normalną treść. Fałszywy PageRank można wykryć poprzez sprawdzenie zarchiwizowanej wersji strony w wyszukiwarce, przeprowadzając inspekcję odnośników, sprawdzając czy PageRank jest przekierowany na inną stronę oraz używając operatora zaawansowanego wyszukiwania ‘info’.

W internecie istnieje kilka witryn utrzymujących, że potrafią przewidzieć przyszłą wartość PageRank danej strony. Oczywiście nikt nie posiada tej samej struktury grafu sieci, jaką ma Google do obliczenia PageRank. Dlatego takie przewidywanie opiera się na innego rodzaju danych. Na podstawie odnośników przychodzących, co jest bardzo niedokładne lub na odnośnikach przychodzących i odpowiadającej danej stronie wartości PageRank z paska narzędzi. Nawet w drugim przypadku istnieje wiele niewiadomych jak:

- Lista odnośników zwracanych przez Google nie jest wyświetlana w całości, pokazuje niewielką ilość. Gdyby korzystać z odnośników z innej wyszukiwarki, nie wiadomo czy Google akurat te wszystkie zlicza.
- Pasek narzędzi pokazuje tylko liczbę całkowitą na logarytmicznej skali. PageRank równy 5 może mieć wartość 5,0 lub 5,99.
- Aktualnie pokazywana wartość jest brana jako dana wejściowa. Jednakże, właśnie ta wartość zmieni się przy następnej aktualizacji całego indeksu.
- Podstawa logarytmiczna jest częścią obliczania prognozy. Jednak większość osób używa wyłącznie wartości z przedziału 5 do 8.
- Liczba odnośników wychodzących musi zostać zastąpiona średnią ilością odnośników w trakcie danego okresu.
- Skala wyświetlana na pasku narzędzi nie jest stała.

Wszystkie powyższe niewiadome kumulują się, zatem rozsądne przewidywanie jest niemożliwe. Najbardziej prawdopodobne trafienie nie jest lepsze niż proste przewidzenie, że „PageRank się nie zmieni’ lub ustalenie, że będzie wynosić maksymalny PageRank ze strony odsyłającej minus jeden.

W pierwotnym założeniu PageRank miał być systemem do sortowania odnośników. Jeśli strona miała wiele odnośników to najlepsze z nich wyświetlane były najwyżej. Taki rozkład odnośników jest również interesujący w przypadku analizy stron konkurencji. Podsumowując PageRank jest systemem, który spłaszcza każdą stronę w sieci Internet do jednej liczby. PageRank to globalny ranking wszystkich stron, niezależny od ich zawartości, bazujący wyłącznie na ich położeniu w strukturze grafu sieci.

Literatura

1. Battelle, J. (2006). Szukaj: jak Google i konkurencja wywołali biznesową i kulturową rewolucję. Wydawnictwo Naukowe PWN.
2. Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.
3. Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, *30*(1-7), 107-117.
4. McBryan, O. A. (1994, May). GENVL and WWWW: Tools for taming the web. In *Proceedings of the first international world wide web conference* (Vol. 341).

Informacje o autorze

Mgr Artur Strzelecki
Katedra Informatyki, Akademia Ekonomiczna
ul. Bogucicka 3 40-226 Katowice – Polska
e-mail: strzelecki@ae.katowice.pl