Analiza długiego ogona wizyt z wyszukiwarek internetowych

Artur Strzelecki[1]

Streszczenie: Celem pracy jest zaprezentowanie wyników badań nad popularnością i widocznością witryn internetowych w wyszukiwarkach. Badanie jest oparte o dane pozostawione przez użytkowników odwiedzających witryny w badanym okresie. Omówiono czynniki, które zgodnie z badaniami, mają największy wpływ na popularność badanych witryn. Ze szczególnym uwzględnieniem analizowano ruch generowany przez użytkowników kierowanych z wyszukiwarek internetowych. Ukazano aktualne trendy w długości konstruowanych fraz kluczowych. Artykuł podkreśla znaczenie dywersyfikacji fraz kluczowych opisujących witrynę. Szerokie spektrum fraz kluczowych zapewnia wysoką popularność i oglądalność witryn.

Słowa kluczowe: wyszukiwarka, frazy kluczowe, internet.

1. Wprowadzenie

Każde zapytanie przesłane do wyszukiwarki, zwraca listę z wynikami wyszukiwania. Naturalne wyniki wyszukiwania powstałe w oparciu o algorytm wyszukiwarki, który rankinguje i sortuje zwracane rezultaty, prowadzą do adresów internetowych. Wyszukiwarki potrafią przemierzać usługi WWW, FTP i grupy dyskusyjne. Wybiórczo stosuje się opiniowane redaktorów, których zadaniem jest sprawdzanie wysoko konkurencyjnych słów kluczowych w popularnych branżach. Natomiast inne dokumenty otrzymują ocenę rankingującą jako rezultat wyliczeń relewancji zapytania do znalezionych dokumentów.

Popularność wyszukiwarek internetowych jest silnie uzależniona od dokładności rezultatów, jakie zwracają. Im bardziej dokładne wyniki, tym większą popularność zyskuje wyszukiwarka. Procedura ustalania rankingu jest fundamentalną charakterystyką każdej wyszukiwarki i ma ogromny wpływ na istnienie oraz popularność pozostałych witryn w Internecie. Wysoka pozycja na liście zwracanych wyników dla słowa kluczowego związanego z witryną zazwyczaj przynosi znacznie więcej korzyści, niż bardzo droga kampania reklamowa oparta o banery (patrz Khaki-Sedigh i Roudaki, 2003).

Artykuł przedstawia analizę popularności witryn w wyszukiwarkach. Opiera się o dane pozostawione przez odwiedzających je użytkowników. Koncentruje się na użytkownikach, którzy odnaleźli te witryny w oparciu o mechanizmy wyszukujące w Internecie. Przedstawia istotę długiego ogona w wynikach wyszukiwarek. Długi ogon to ruch w witrynach generowany przez rzadko szukane słowa kluczowe, tzw. słowa z długiego ogona. Podkreśla, że w wielu wypadkach skuteczniejszą metodą

[1] Katedra Informatyki, Akademia Ekonomiczna w Katowicach, Bogucicka 3, 40-226 Katowice
e-mail: strzelecki@ae.katowice.pl

popularyzacji witryny w Internecie jest koncentrowanie się na mniej konkurencyjnych słowach kluczowych, które zebrane razem tworzą wysoki udział w generowaniu ruchu w witrynie.

2. Optymalizowanie wyników w wyszukiwarkach

Celowe działania, które mają doprowadzić do znalezienia się na samym szczycie listy z wynikami wyszukiwania nazywa się pozycjonowaniem witryn internetowych. Pozycjonowanie witryn internetowych jest jednym z narzędzi realizacji strategii promocji przedsiębiorstwa w Internecie. Pozycjonowanie witryn polega na kombinacji działań zwiększających prawdopodobieństwo znalezienia się pozycjonowanej witryny wśród pierwszych wyników wyświetlanych przez wyszukiwarki internetowe. Wyniki mają dotyczyć najbardziej popularnych słów kluczowych związanych z tematyką pozycjonowanej witryny internetowej.

Na podstawie opracowań (patrz Bifet i inni, 2005), (patrz Fortunato i inni, 2006) czynniki, które potencjalnie wpływają na ranking wyszukiwarki internetowej dzielą się na dwie wyraźne kategorie. Pierwsza to czynniki wewnętrzne (ang. query-factors), które są zależne od treści na stronie, jak obecność i częstotliwość słów kluczowych oraz druga to czynniki zewnętrzne (ang. query-independent factors), które są uzależnione od informacji pochodzących z zewnętrznych stron, mających odnośniki do reklamowanej witryny. Jednak obie grupy czynników trudno wyliczyć, ponieważ wyszukiwarki internetowe nie ujawniają, które w szczególności są używane do określenia pozycji w rankingu.

Problem jest złożony w wyniku następujących kwestii (patrz Evans, 2007):

- istnieje ponad 200 różnych czynników używanych m.in. przez Google do wyliczenia rankingu strony,
- ponieważ większość z nich jest nieznana, nie wiadomo także jaki wpływ mają na wynik końcowy,
- waga każdego z użytych czynników do określenia wyników z pierwszej strony rezultatów może być różna od wagi użytej dla pozostałych stron z rezultatami,
- różne zapytania mogą zostać obsłużone przez różne czynniki bądź różne wagi,
- Google posiada wielorakie centra z danymi rozmieszczone po całym globie, nie wszystkie są synchronizowane w tym samym momencie.

To czyni zidentyfikowanie czynników zaangażowanych przez algorytm wyszukiwarki niezwykle trudnym. W konsekwencji, zapotrzebowanie na wiedzę o tych czynnikach doprowadziło do powstania organizacji trudniących się tzw. search engine optimization - SEO, lub w szerszym ujęciu search engine marketing - SEM. Celem tych przedsiębiorstw jest zwiększenie wartości rankingu w wynikach wyszukiwarek dla własnych klientów. Dzięki doświadczeniu i wielu testom są w stanie znacząco zwiększyć wynik witryny. Jednak trzeba pokreślić, że po przeszukaniu wielu for i blogów internetowych prowadzonych przez organizacje SEO wynika, że ich praca ewidentnie opiera się na zgadywaniu, próbach i błędach. Pomimo tego, obroty na rynku SEO/SEM rosną z każdym rokiem.

3. Badanie witryn w internecie

Dotychczas pojawiło się kilka prac, w których autorzy badali różne aspekty związane z ruchem generowanym przez wyszukiwarki. Evans (2007) przedstawił wyniki badań 50 wysoko optymalizowanych witryn internetowych, które zostały stworzone jako część konkursu pozycjonowania witryn internetowych. Praca skupia się na najbardziej popularnych technikach, które zostały wykorzystane do podniesienia rankingu oraz zawiera analizę wpływu PageRank, liczby indeksowanych podstron, liczby odnośników przychodzących, wieku domeny i wykorzystania witryn trzecich jak katalogi i serwisy społecznościowe. Praca prowadzi prosto do istoty procedur, jakie wykorzystują organizacje SEO, aby witryny ich klientów znalazły się wysoko w wynikach wyszukiwania.

Fortunato i inni (2006) wykonali podobny eksperyment, w którym starali się przybliżyć dynamikę algorytmu PageRank poprzez badanie odnośników przychodzących. Mimo że, byli w stanie pokazać, iż liczba odnośników przychodzących jest dobrym przybliżeniem PageRank popularnych witryn, to nie jest to jedyny czynnik używany w tym wypadku przez Google do rankingowania rezultatów.

Bifet i inni (2005) wykorzystali wiele różnych czynników do estymacji funkcji, którą znaleźli do rankingowania w wyszukiwarkach. Użyli tej funkcji do porównania własnych, przewidywanych rezultatów z aktualnymi rankingami w Google. Znaleźli rozmaite czynniki wpływające na rankingi, zależne, na pozór od tematu jaki był szukany. Jednak zbudowana funkcja nie działała tak dobrze jakby autorzy sobie tego życzyli, prowadząc do nieprzekonywujących wyników.

4. Eksperyment i metodologia

Do badania wybrano 21 witryn, które były odwiedzane w okresie od 1 stycznia 2007 do 30 czerwca 2007. Witryny są zróżnicowane pod względem budowy, treści i celu, w jakim zostały stworzone (Tabela 1). Badanie opiera się na statystykach pobranych z systemu analitycznego Awstats zainstalowanego na serwerze z usługą hostingową. Badane witryny zostały arbitralnie nazwane kolejnymi literami alfabetu, aby ułatwić analizę porównawczą.

Możliwość pozyskania danych do badania nie jest łatwa. Właściciele witryn internetowych wszelkie informacje statystyczne traktują jako poufne i nie są one powszechnie dostępne. Dane dotyczące pierwszych siedmiu witryn autor uzyskał poprzez użyczony dostęp do systemu analitycznego. Natomiast pozostałe informacje o 14 witrynach zostały znalezione w sieci Internet. Odpowiednio sformułowane zapytanie do wyszukiwarki zwróciło wiele publicznie dostępnych statystyk. Po zagłębieniu się w nie, okazało się, że prawie wszystkie miały powiązania ze stronami o charakterze pornograficznym, co zniekształcało obraz rzetelności danych statystycznych. Po przejrzeniu około 100 statystyk autor wybrał 14 witryn z wiarygodnymi danymi o różnej tematyce do dalszego badania. Zostały odrzucone witryny z powiązaniami do stron pornograficznych oraz te, które nie miały żadnej treści.

Witryny zostały także ocenione pod kątem następujących kryteriów, Tabela 2:

– liczba stron z witryny zaindeksowanych przez wyszukiwarkę, niektóre witryny są większe niż inne, być może większe oznacza lepsze,

Tabela 1. Podstawowe informacje o badanych witrynach

Witryna	Adres	Technologia	Utworzona
A	www.piotrowice.katowice.pl	xthml	11/2006
B	www.fuji-foto-centrum.com.pl	xthml	11/2005
C	www.info-global.com.pl	xthml	05/2006
D	www.2px.com.pl	flash	05/2005
E	www.logomaker.pl	xthml	02/2003
F	www.www.mysms.pl	xthml	04/2002
G	www.poldruk.com	xthml	01/2004
H	www.immigrantinfo.org	xthml	10/2001
I	www.nikicruz.com	xthml	08/1999
J	www.theatrescene.net	xthml	06/2001
K	www.wavplanet.com	xthml	01/2000
L	www.francisshanahan.com	xthml	11/2002
M	www.glenhelen.com	xthml	01/1996
N	www.ikona.cz	xthml	12/2000
O	www.jiminypeak.com	xthml	06/1996
P	www.longdistancerelationships.com	xthml	09/2001
Q	www.nostradamus.cz	xthml	05/2004
R	www.onlinepot.org	xthml	09/2002
S	www.salonhogar.net	xthml	12/2005
T	www.sieveonline.it	xthml	01/1999
U	www.webdesign101.dk	xthml	12/2000

– wynik PageRank witryny w ocenie algorytmu Google opracowanego przez (patrz Page i Brin, 1998),

– liczba odnośników przychodzących, uzyskana operatorem 'link' w wyszukiwarce

– wiek nazwy domeny witryny internetowej, spekuluje się że starsze domeny są oceniane wyżej niż nowe w przypadku tej samej zawartości,

– obecność witryny w najważniejszych katalogach, edytowanych przez redaktorów, wpisy z katalogów Yahoo i DMOZ (Open Directory Project) zasilają wyniki w wyszukiwarkach Yahoo i Google. Ponieważ witryny znajdujące się w tych katalogach przechodzą kontrolę wysokiej jakości, strony tych katalogów są traktowane jako autorytety, które wyszukiwarki mogą wykorzystać jako jeden z czynników rankingu.

5. Wyniki badań

Do badania wzięto następujące informacje:

Tabela 2. Rozszerzone informacje o badanych witrynach

	Strony w indeksie Google	Strony w indeksie Yahoo	PageRank	Liczba odnośników Google	Liczba odnośników Yahoo	DMoz
A	58	0	1	0	162	tak
B	12	12	3	2	802	tak
C	32	99	1	1	497	tak
D	1	1	3	3	391	tak
E	59500	17700	4	73	55172	nie
F	505	1636	0	35	22718	nie
G	8	13	3	0	866	tak
H	343	2082	5	6	144	nie
I	79	186	4	3	67	nie
J	356	3803	4	31	3322	nie
K	2660	11467	4	7	341	nie
L	673	1839	6	229	3127	nie
M	1550	3422	4	20	1777	tak
N	457	907	3	4	529	nie
O	364	973	5	72	3092	tak
P	379	974	4	10	4219	nie
Q	1110	1129	4	2	651	nie
R	2910	3880	3	8	3492	tak
S	215	438	4	12	135	tak
T	7200	15609	4	47	6040	nie
U	79	557	4	11	1632	nie

1. Oglądalność, czyli
   - Liczba unikalnych gości - użytkownicy, którzy odwiedzili stronę z niepowtarzalnym numerem IP.
   - Ilość wizyt - użytkownicy, którzy odwiedzili stronę, wraz z powracającymi do niej.
2. Źródła połączeń a w tym:
   - Liczba wizyt bezpośrednich lub z Ulubionych/Zakładek.
   - Liczba wizyt z wyszukiwarek.
   - Liczba wizyt z odnośników na innych stronach.
3. Popularność słów kluczowych
   - Poszukiwane słowa kluczowe (frazy).
   - Poszukiwane wyrazy.

5.1. Oglądalność witryn

Oglądalność jest uzależniona od popularności witryny, częstotliwości z jaką użytkownicy odnajdują daną stronę w Internecie oraz samej budowy wewnętrznej witryny. Oczywiście istnieje całe mnóstwo innych stron ze znacznie większą oglądalnością niż te wzięte do analizy. Mimo to, już na uzyskanym poziomie wyświetleń, można wyprowadzić wiele cennych wniosków i porównań. Łączna suma wizyt w badanych witrynach wyniosła ponad 4,5 miliona. Liczba wizyt jest zawsze większa od liczby wizytujących użytkowników, różnica polega na zliczaniu wielokrotnych wizyt jednego, tego samego użytkownika na podstawie adresu IP.

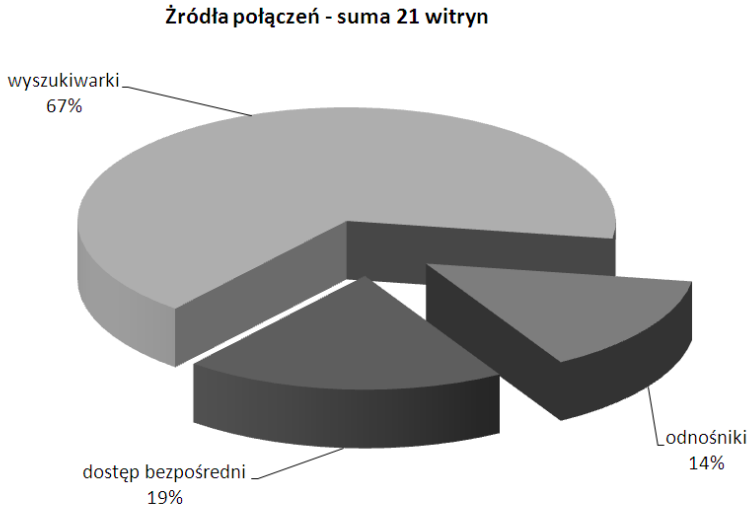
5.2. Źródła połączeń

Do witryny internetowej można dotrzeć trzema sposobami. Pierwszy to wpisanie adresu URL do paska adresu przeglądarki internetowej, drugi to skorzystanie z bezpośredniego odnośnika do witryny i trzeci, znalezienie adresu witryny w zasobach wyszukiwarki.

Do wyszukiwarek na Rysunku 1 należą: Goole, Live Serach, MSN, Dmoz, Yahoo, Alexa, AOL, Altavista i Seznam. W ich skład wchodzą również wyszukiwarki polskie jak: OnetSzukaj, Szukacz i NetSprint. Średnia ważona rozkładu źródeł połączeń wskazuje na kluczową rolę wyszukiwarek w docieraniu użytkowników do witryn internetowych. Ponad 2/3 użytkowników korzysta z wyszukiwarek żeby odnaleźć pożądaną stronę. Dalsze badanie pokaże jak kształtuje się rozkład słów kluczowych, po których te wizyty następują. Dostęp bezpośredni to wejścia na witrynę przez użytkowników, którzy adres do strony mają zapisany w zakładce ulubione swojej przeglądarki internetowej albo zapamiętany adres w historii przeglądanych witryn lub znają adres na pamięć i wpisują go w pasek adresu przeglądarki. Dostęp przez odnośniki to przejście z jednej witryny na drugą z wykorzystaniem odsyłacza hipertekstowego.

Tabela 3. Oglądalność witryn

	Użytk.	Wizyty	Słowa kluczowe	Frazy niepowt.	Frazy powtarzalne	Znalezione wyrazy
A	5374	7197	2513	1743	770	1318
B	4097	5099	734	454	280	318
C	744	1943	182	129	53	160
D	784	1153	65	27	38	43
E	13514	18384	8355	6384	1971	1244
F	16030	19886	8595	5871	2724	2980
G	1793	2312	294	193	101	bd
H	48971	60108	28104	23957	4147	11905
I	5057	12285	380	292	88	298
J	28935	101310	7806	6834	972	6932
K	237332	328112	16981	13681	3300	5042
L	443617	121470	6780	5696	1084	5056
M	54124	101863	6343	4417	1926	4925
N	9915	21185	1879	1409	470	2244
O	187872	488677	10025	6014	4011	3658
P	28903	41470	2850	2040	810	1522
Q	79462	142003	11596	9314	2282	8584
R	111690	141000	35930	26616	9314	11122
S	1946588	2581590	470252	350669	119583	86856
T	265185	356086	139955	103765	36187	18226
U	28531	62961	10977	7420	3557	4049



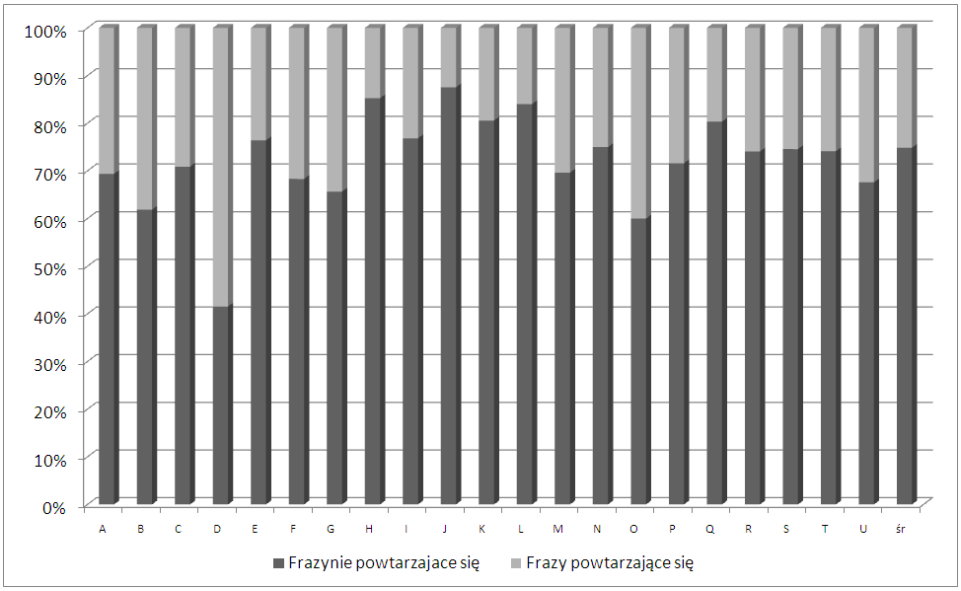
Rysunek 1. Źródła generowanych wizyt w witrynach

5.3. Frazy kluczowe

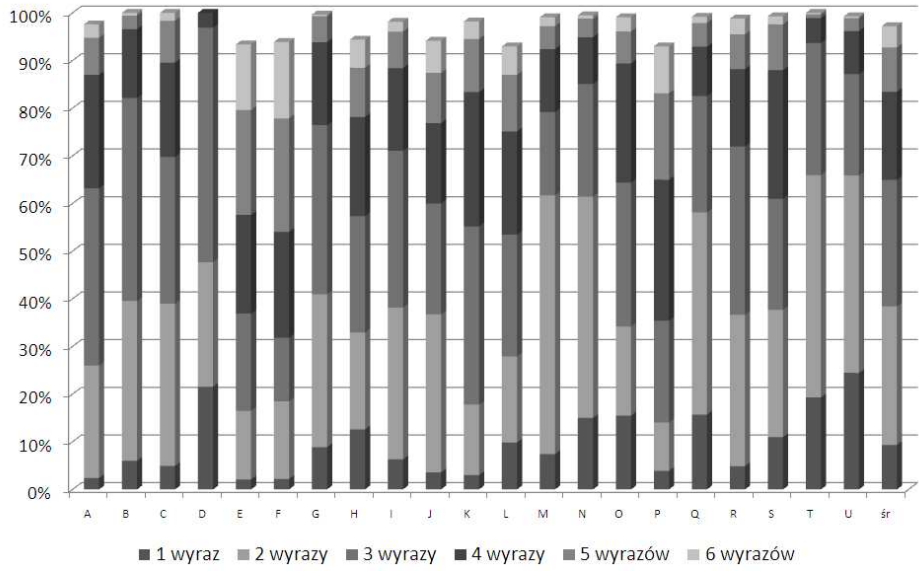
Wyniki z Tabeli 3 pokazują także, jak wielką przewagę przy wyszukiwaniu mają witryny budowane w technologii tekstowej, wypełnione wysokiej jakości pożądanymi treściami, nad witrynami, które nie mają takiej zawartości. Ilość utworzonych słów kluczowych, czyli fraz, dzięki którym witryna została znaleziona w wyszukiwarkach we wszystkich przypadkach jest średnio 4-5 razy większa od całkowitej liczby słów znalezionych na tej stronie. Jasno z tego wynika, że użytkownicy wykorzystują różne kombinacje słów kluczowych. Ilość możliwych kombinacji, jakie jeszcze mogą powstać, rośnie w postępie geometrycznym.

Wyniki na Rysunku 2 prezentują procentowy udział wizyt wygenerowanych przez pierwsze dziesięć najpopularniejszych fraz ze słów kluczowych do wizyt wygenerowanych przez pozostałe frazy. W badanej próbce dwie witryny wykazywały się anomalią i pierwsze słowo kluczowe wygenerowało około 50% (O i P) ruchu z wyszukiwarek dla obu. Średnia ważona pokazuje, że pierwsze dziesięć słów kluczowych daje niecałe 14% ruchu w witrynach. Natomiast zdecydowana większość, ponad 86% ruchu z wyszukiwarek pochodzi ze słów kluczowych spoza pierwszej dziesiątki. Ogólnie wiadomo, że witrynę internetową można opisać za pomocą kilku słów kluczowych, tych najważniejszych, związanych z jej tematyką. Wykres pokazuje, że kilka czy kilkanaście głównych słów kluczowych, generuje niewielki udział w oglądalności witryny. Nie ma tu zastosowania zasada Pareto. Wyniki sugerują zupełną odwrotność tej zasady. Zaistniałą sytuacje nazwano długim ogonem wyszukiwania. Najwięcej jest słów kluczowych, które tworzą niewielki ruch na stronie, ale razem zebrane, decydują o wysokiej widoczności witryny w wyszukiwarkach.

Rysunek 2. Procentowy udział poszczególnych fraz



Rysunek 3. Całkowity udział fraz niepowtarzalnych do powtarzających się



Rysunek 4. Rozkład długości fraz kluczowych wpisywanych do wyszukiwarki

Ostatnim mierzalnym parametrem, charakteryzującym wysoką widoczność w wynikach wyszukiwania, są formułowane zapytania do wyszukiwarki. Można pokusić się o twierdzenie, że ilu użytkowników wyszukiwarek, tyle sposobów na wyszukanie informacji. Autorzy Google podają (patrz Battelle, 2005), że około 50% wpisywanych do niej zapytań każdego dnia jest niepowtarzalnych, nigdy wcześniej niestworzonych.

Na Rysunku 3 w witrynach widać zróżnicowany poziom niepowtarzalnych wpisywanych fraz kluczowych do wyszukiwarki w stosunku do całkowitej liczby wszystkich fraz, bez powtórzeń. Obraz zmienia się prawie odwrotnie, gdyby brać pod uwagę powtórzenia, czyli całkowitą liczbę wizyt pochodzących z wyszukiwarek. Średnia ważona przedstawia łączny udział fraz niepowtarzających się ze wszystkich witryn.

W statystykach odwiedzin najczęściej powtarzają się kombinacje, dwu, trzy i cztero wyrazowe. Rysunek 4 ilustruje rozkład długości wpisywanych fraz do wyszukiwarki. 8 witryn zanotowało kilkukrotne wejścia dla fraz o długościach od 7 do 13 wyrazów, natomiast dwie (J i L) frazy od 14 do 20 słów kluczowych, ale nie zostały one pokazane na wykresie. Rysunek przedstawia ilość różnych długości fraz wprowadzonych do wyszukiwarki, bez uwzględniania częstotliwości ich ponownego wprowadzenia. Opiera się wyłącznie na jednokrotnych wystąpieniach. Średnia ważona przedstawia całkowity rozkład pochodzący ze wszystkich witryn.

6. Wnioski

Artykuł na podstawie przeprowadzonych studiów i analiz wskazuje na korzyści uzyskiwane przez witryny internetowe. Umiejętna budowa stron i wykorzystanie technologii tekstowej powodują znaczne zwiększenie widoczności witryny w wynikach wyszukiwania. Wyraźnie podkreślono jak możliwa jest dywersyfikacja wizyt i ruchu pochodzącego z wyszukiwarek. Ruch nie powinien być wyłącznie oparty o kilka popularnych fraz ze słów kluczowych. Widać to doskonale w witrynach składających się z wielu podstron. Nawet, jeśli okresowo witryna pod pewnymi frazami ze słów kluczowych nie jest na wysokich pozycjach, to pozostają jeszcze możliwości pod innymi frazami. Słabiej na tym tle wypadają witryny z zawartością multimedialną lub małą ilością podstron. Mogą być atrakcyjniejsze w wyglądzie, ale nie ma to absolutnie znaczenia, kiedy zupełnie nie będą odwiedzane z powodu braku widoczności w wyszukiwarkach.

Właściciele wyszukiwarek są obecnie na kluczowej pozycji w wirtualnym świecie. Dzieje się to w wyniku popularności ich usług. Ponad połowa użytkowników odwiedzających witryny internetowe jest do nich kierowana prosto z wyszukiwarki, mniej natomiast poprzez bezpośredni odnośnik. Wyszukiwarki odnotowują miesięcznie ponad 4,5 miliarda zapytań wprowadzanych przez użytkowników. Witryny konkurują ze sobą o to, do której z nich w wyniku wyszukiwania przejdzie użytkownik. Proste zapytanie do wyszukiwarki o dużych zasobach zwraca tysiące, a nawet miliony odpowiedzi. 73% użytkowników wyszukiwarek nie patrzy dalej niż pierwszą stronę wyników (patrz Jansen i Spink, 2006).

Literatura

Battelle, J. (2005) *The Search. How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. Portfolio, Penguin Group.

Bifet, A., Castillo, C., Chirita, A. and Weber, I. (2005) An analysis of factors used in search engine ranking. *1st International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*, Chiba.

Evans, M.P. (2007) Analysing Google rankings through search engine optimization data. *Internet Research* (17), 1, 21–37.

Fortunato, S., Boguna, M., Flammini, A. and Menczer, F. (2006) How to make the top ten: approximating PageRank from In-degree. *Proceedings 14th International World Wide Web Conference*, Edinburgh.

Jansen, B.J. and Spink, A. (2006) How are we searching the world wide web? A comparison of nine search engine transaction logs. *Information Processing and Management*. 42, 248–263.

Khaki-Sedigh, A. and Roudaki, M. (2003) Identification of the dynamics of the Google ranking algorithm. *13th IFAC Symposium On System Identification*, Iran.

Page, L., Brin, S., Motwani, R. and Winograd, T. (1998) The PageRank Citation Ranking: Bringing Order to the Web. Report, Stanford University.

An analysis of long tail data from search engine entries

The aim of the work is to present research's results of the popularity and visibility of sites in search engines. The research is based on data left by users who visits in a given period. Here were discussed factors which in agreement with research hale the biggest influence on the popularity of studied sites. The traffic is generated by users directed by search engines with a specific approach. Current trends are shown in the length of constructed key phrases. The article emphasizes the importance of diversification of key phrases describing the site. The wide spectrum of the key phrases gives high popularity and visibility of the sites.