Optimization and Effectiveness of Search Engine Results

Trevor Sibuyi and Johnson O. Dehinbo

Abstract—**The web is expanding day by day and people generally rely on web search engines to explore the web. Web search engines are so popular that they are, together with e-mail, the most commonly used services on the Internet. Millions of web search queries are posed daily and in such a scenario it is the duty of the service provider to provide relevant, accurate, and quality information to the internet user. However, web search engine results are filled with unauthorized, irrelevant, misleading content, and advertisements that seems legit to the user but their intent is to mislead and raise their position in web search engine rankings. As a result, users spend a lot of time and bandwidth trying to locate the information they are searching for. In this paper we investigate user search results preferences and incorporate them to significantly improve the relevance, accuracy, and filtering of search results**.

Index Terms—**Optimization, Effectiveness of search results, Web search engines, Search queries, Personalized search**

I. INTRODUCTION

THE Web has now become a major source of information for many people worldwide. The role of World Wide Web as a major information publishing and retrieving mechanism on the Internet is now predominant and continues to grow extremely fast. The amount of information on the Web has long since become too large for manually browsing through any significant portion of its hypertext structure. As a consequence, a number of Web search engines have been developed. Web search engines are used for a wide variety of research purposes, and they are often the first place to go when searching for information and internet users quickly feel comfortable with the act of searching. Millions of web search queries are posed daily and in such a scenario it is the duty of the service provider to provide relevant, accurate, and quality information to the internet user from trusted sources or domains against their query submitted to the web search engine. Typically, a user submits a search query (A set of keywords) to a web search engine, which then returns a list of links to pages that are most relevant to this query. To determine the most-relevant pages, a web search engine selects a set of candidate pages that contain some or all of

the query terms and calculates a page score for each page. Finally, a list of pages sorted by their score is returned to the user [5].

Web search engines are so popular that they are, together with e-mail, the most commonly used services on the Internet [10]. It is a challenge for service providers (Web search engines) to provide proper, relevant, and quality information to the internet user by using the web page contents and hyperlink between the web pages. It's astonishing to what degree users trust Web search engines [11] and they rely on Web search engines to display the most credible search results first. Users select search results and then read Web pages based on their decisions about whether the information presented by the search engine is of value to them.

However, web search engines are faced with difficult problems or challenges in maintaining and enhancing the quality of their performance. The search results page is filled with noisy features (unauthorized, irrelevant, and misleading content) and advertisements that seems legit to the user but the sole intent is to mislead and raise their position in web search engine rankings. Since a better position in the rankings directly and positively affects the number of visits to a site, attackers use different techniques to boost their pages to higher page ranks. In the best case, web spam pages are like a form irritations that provide undeserved advertisement revenues to the page owners [5]. On the other hand, the noise presented on the search results page leaves a negative reputation about the service provider and negative repercussions on the user.

In this paper we investigate user search results preferences and incorporate them to significantly improve ordering of top search results and eliminate noisy features presented in web search results page.

A. The Problem Statement

Web search engines are services that help their users find information on the web with ease. However, web search engines are faced with a number of difficult problems in maintaining or enhancing the quality of their performance such as spam, content quality, page ranking, and duplicate hosts. The amount of information on the web has become too large for manually browsing through any significant portion of its hypertext structure. As a consequence, web search results are filled with noisy features and irrelevant advertisements that provide undeserved advertisement revenues to the page owners. On the other hand, the noise presented on the search results page leaves a negative reputation about the service provider and negative repercussions on the user.

Manuscript received July 6, 2016; revised July 30, 2016. T. Sibuyi is a B.Tech. student with the Department of Computer Science, Tshwane University of Technology, Soshanguve, North of Pretoria, South Africa (Phone: +27123829261; e-mail: sibuyitk@gmail.com).

J.O. Dehinbo is with the Department of Computer Science, Tshwane University of Technology, Soshanguve, North of Pretoria, South Africa (Phone: +27123829219; e-mail: jdehinbo@yahoo.com).

B. Research Questions and Objectives

The main research question is: How can we optimize web search engine results? The sub-questions are: How to retrieve appropriate search results from user query? How to evaluate the quality of the results retrieved? How can we filter search results based on user preferences?

The main research objective is to optimize web search engine results. The sub-objectives are:
- To retrieve appropriate search results from the user search query;
- To evaluate the quality of content retrieved;
- To filter search results based on user preferences.

II. LITERATURE REVIEW

A. Background and Related Works

Though many Internet-enabled services are available today, one of the primary applications is for information retrieval. With the advancement of Web page development tools in both functionality and usability, individuals can publish information on almost any topic imaginable. It is certain that with such diverse content and the enormous volume of information on the Internet, retrieving relevant or needed information is far from assured. In other words, seeking resources on the World Wide Web is a significant task because there is such a vast amount of information available, but the growth of information and the increasing number of users requiring simultaneous access to it all add to its complexity [9]. Thus, it is fair to say that Web information retrieval would collapse if search engines were not available. Essentially, search engines offer four main facilities: first, they gather Web pages from which an individual can retrieve information; second, they cluster Web pages into hierarchical directories; third, they provide hyperlinks to connect Web pages; and fourth, they allow individuals to issue queries, employing information retrieval algorithms to find for them, the most relevant Web pages [9].

In general, search engines are essential tools for finding resources on the World Wide Web; thus, the effective use of search engines for information Retrieval (IR) is a crucial challenge for any Internet user.

B. Challenges in Web Search Engines

Spams

Henzinger et al. [12] observes that for commercially-oriented web sites, whose income depends on their traffic, it is in their interest to be ranked within the top 10 results for the query relevant to the content of the website. To achieve this goal, some web authors try to deliberately manipulate their placement in ranking order of various Web search engines. The resulting pages are called spam. There are three broad spam categories: Text spam, Link spam and Cloaking.

Text spam: Text spam techniques are used to modify the text in such a way that the search engine rates the page as being particularly relevant, even though the modifications do not increase perceived relevance to a human reader of a document.

Link spam: A common approach is for an author to put a link farm at the bottom of every page in a website, where a farm is a collection of links that points to every other page in that website, or to any website the author controls. The goal is to manipulate systems that use raw counts of incoming link to determine a web page's importance. A problem with link farms is that they distract the reader because they are on pages that also have legitimate content.

Cloaking: *Cloaking* involves serving entirely different content to a Web search engine crawler than to other users. As a result, the search engine is deceived as to the content of the page and scores the page in ways that, to a human observer seem rather arbitrary. Typically, cloaking is used to deceive Web search engines, allowing authors to achieve the benefits of link and text spam without inconveniencing human readers of the web page.

Content Quality: It is a mistake to think that in search engines, credibility does not play a role in ranking. The web is replete with text that intentionally or not intentionally misleads its human readers. Other sites contain information that was once correct but is now out of date. The issue of document quality or accuracy has not received much attention in web search or information retrieval. One interesting aspect of the problem of document quality is specific to hypertext quantities such as the web, evaluating the quality of anchor text.

Duplicate Hosts: Web search engines try to avoid crawling and indexing duplicate web pages, since such pages increase the time to crawl and do not distribute new information to the search results. While mirror detection and individual page detection try to provide a complete solution to the problem of duplicate pages, a simpler variant called duplicate host detection can reap most of the benefits while requiring less computational resources.

C. Relevance Measurement

Relevance Measurement is crucial to web search and to information retrieval in general. Evaluating user preferences of web search results is crucial for search engine development, deployment, and maintenance. Traditionally, search relevance is measured by using human assessors to judge the relevance of query-document pairs. However, explicit human ratings are expensive and difficult to obtain. At the same time, millions of people interact daily with web search engines and a significant distinction is that Web search is not controlled.

Ranking search results is a fundamental problem in information retrieval. The most common approaches in the context of the web use both the similarity of the query to the page content, and the overall quality of a page. A state-of-the-art search engine may use hundreds of features to describe a candidate page, employing sophisticated algorithms to rank pages based on these features. Current search engines are commonly tuned on human relevance judgments. Human annotators rate a set of pages for a query according to perceived relevance, creating the "gold standard" against which different ranking algorithms can be evaluated. Reducing the dependence on explicit human judgments by using implicit relevance feedback has been an active topic of research (Agichtein et al., 2006).

Several research groups have evaluated the relationship between implicit measures and user interest. In these studies, both reading time and explicit ratings of interest are

collected. Morita et al. [1994] studied the amount of time that users spend reading news articles and found that reading time could predict a user's interest levels. Konstan et al. [16] showed that reading time was a strong predictor of user interest in their Group Lens system. Oard et. al. [1998] studied whether implicit feedback could substitute for explicit ratings in recommender systems and more recently presented a framework for characterizing observable user behaviours using two dimensions, the underlying purpose of the observed behaviour and the scope of the item being acted upon.

Goecks et al. [8] approximated human labels by collecting a set of page activity measures while users browsed the World Wide Web. The authors hypothesized correlations between a high degree of page activity and a user's interest. While the results were promising, the sample size was small and the implicit measures were not tested against explicit judgments of user interest. Claypool et al. [4] studied how several implicit measures related to the interests of the user. They developed a custom browser called the Curious Browser to gather data, in a computer lab, about implicit interest indicators and to probe for explicit judgments of Web pages visited. They found that the time spent on a page, the amount of scrolling on a page, and the combination of time and scrolling have a strong positive relationship with explicit interest, while individual scrolling methods and mouse-clicks were not correlated with explicit interest. Fox et al. [6] explored the relationship between implicit and explicit measures in Web search. They built an instrumented browser to collect data and then developed Bayesian models to relate implicit measures and explicit relevance judgments for both individual queries and search sessions. They found that click through was the most important individual variable but that predictive accuracy could be improved by using additional variables, notably dwell time on a page.

More recently, Joachims et. al. [13] presented an empirical evaluation of interpreting click through evidence. By performing eye tracking studies and correlating predictions of their strategies with explicit ratings, the authors showed that it is possible to accurately interpret click through events in a controlled, laboratory setting. Unfortunately, the extent to which existing research applies to real-world web search results is unclear. In this paper, we build on previous research to investigate user search results preferences and incorporate them to significantly improve ordering of top search results and eliminate noisy features presented in web search results page.

D. *Personalized Search Based on Users' Search History*

Personalization is the process of presenting the right information to the right user at the right moment [22]. However, Search engines order their results based on small amount of information available in the user's queries and by web site popularity, rather than individual user interests. Thus, all users see the same results for the same query, even if they have wildly different interests and backgrounds. To address this issue, interest in personalized search had grown in the last several years, and user profile construction is an important component of any personalization system.

Systems can learn about user's interests collecting personal information, analyzing the information, and storing the results in a user profile. Information can be captured from users in two ways: explicitly, for example asking for feedback such as preferences or ratings; and explicitly, for example observing user behaviors such as the time spent reading an online document. Explicit construction of user profiles has several drawbacks. The user provides inconsistent or incorrect information, the profile created is static whereas the user's interests may change over time, and the construction of the profile places a burden on the user that she/he may not wish to accept. Thus, many research efforts are underway to implicitly create accurate user profiles.

To achieve effective personalization, profiles should distinguish between long-term interests and include a model of the user's context, i.e. the task in which the user is currently engaged and the environment in which they situated [18]. User browsing histories are the most frequently used source of information about user interests. User profiles are created by classifying the collected Web pages with respect to a reference ontology. Kim and Chan [15] also build user profiles from browsing histories, however they use clustering to create a user interest hierarchy. The collected Web pages are then assigned to the appropriate cluster. The fact that a user has visited a page is an indication of user interest in that page's content.

Several systems have attempted to provide personalized search that is based upon user profiles. The Personal Search Assistant [14] acts as an independent agent that collects and organizes information on behalf of its user. Similarly, the Competitive Intelligent Spider and Meta Spider [3] autonomously gather information for a user based on their preferences. Collected documents are then analyzed and noun phrases are extracted to create a personal dictionary for the user to guide future searches.

In contrast to the above systems, the OBIWAN project [7] focuses on interactive, personalized search rather than background processes. Another difference is that the user profiles are implicitly created based on browsing histories rather than explicitly created from user input. Search results from a conventional search engine are then classified with respect to the same concept hierarchy used to represent the user profiles. Documents then are re-ranked based upon how well their concepts match those that appear highly weighted in the user profile.

In order to capture information about users, [2] implemented GoogleWrapper to anonymously monitor the search activities of a set of volunteers. Two different types of information was collected per individual user:
- Queries submitted through GoogleWrapper and for which at least one result was visited.
- Snippets of results in the list selected by the user.

Each piece of information collected about a user was classified into a concept hierarchy based upon the Open Directory Project hierarchy. Once the user profiles are built, they are used to provide personalized search.

E. Typical System Architecture

The architecture of the system consists of two main modules:

- GoogleWrapper, a wrapper for Google responsible for collecting information from users. When queries are submitted by users, GoogleWrapper stores in a session variable the query and the userID and then forwards the query to the Google search engine.

- The category from KeyConcept, a conceptual search engine is used to classify each query and snippet into a list of weighted concepts from the reference concept hierarchy.

A set of scripts was also implemented to conduct the experimental analysis of the effectiveness of using search-history based user profiles for personalized search, comparing Google's original rank with our conceptual rank (Speretta and Gauch, 2005).

User Profiles

User profiles are presented with a weighed concept hierarchy. The reference concept hierarchy contains 1,869 categories in the top 3 levels of the Open Directory Project, and the weights represent the amount of user interest in the specific category. The classifier was trained using 30 documents listed for each category that were collected by a spider. The user profile concept eights are assigned by classifying textual content collected from the user into the appropriate categories. This process produces a list of concepts with associated weight that can be accumulated over the queries, or snippets is compared with the vocabulary for each category's set of training documents and the classifier reports back a similarity value.

Personalized Search

During the evaluation phase, each search result is classified to create a document profile in the same format as the user profile. The document profile is then compared to the user profile in order to calculate the conceptual similarity between each document and the user's interests. The conceptual match between the document profile and user profile is calculated using the cosine similarity function.

F. Page Ranking

Search engines generally return a large number of pages in response to user queries. To assist the users to navigate in the result list, ranking methods are applied on the search results. Most of the ranking algorithms proposed in the literature are either link or content oriented, which do not consider user usage trends. In order to measure the relative importance of web pages, (Page et al., 1999) proposed a basic ranking algorithm from Google i.e. PageRank, a method for computing a ranking for every web page based on the graph of the web. To assist the users to navigate in the result list, ranking methods are applied on the search results. Most of the ranking algorithms proposed in the literature are either link or content oriented, which do not consider user usage trends.

III. RESEARCH METHODOLOGY

In general, both interpretive and positivist approaches are used in this study. The interpretive research design used involves elements of descriptive studies. The positivism research design used involves methodologies including prototyping used in the development phase.

A. Design Science Encapsulation of the Methodology

The basic methodology of the research is Design Science methodology. A suitable methodology for research involving software development is Design Science. In line with [17], a typical design science research effort as illustrated as follows.

Awareness of the Problem

In this phase, we identify gaps or problems that can come from multiple sources such as new developments in a field of interest in specific discipline or in the industry. Reading widely is critical in providing opportunity for the awareness of a problem that can be researched. The result of this phase is a proposal, formal or informal, for a new research effort.

Suggestion

The resulting output of the suggestion phase is a tentative design in which new functionality is envisioned. This could likely include the performance of a prototype based on the design. In the absence of an output of this phase, circumscription involves looping back into the problem awareness phase, else the proposal will be set aside.

Prototyping

This phase involves further elaboration, creative development and implementation of the tentative design. Depending on the artifact to be constructed, this could involve using various techniques including algorithm construction, expert system development using a high-level package or tool, etc. On errors or the absence of an output of this phase, circumscription involves looping back into the suggestion phase. The development stage of the design science approach will use prototyping. According to [17], prototyping refers to a simplified program or system that serves as a guide or example for the complete program or system. Therefore, based on the user experience and their perceptions, a prototype that tend to meet all their needs and expectation is developed. As the goal of the project is to provide a web search engine that eliminates noisy results and returns relevant and useful information to the user.

This approach helps to eliminate uncertainties and ensure that the system will do its purpose. This is a good method to identify problems that users encounters when using the system so that an improvement can be made. The application's working principle is similar to Google and other professional web search engines.

Evaluations

After development, the artifact is evaluated according to criteria possibly stated in the awareness phase. Here deviations from expectations, either quantitative or qualitative, are carefully noted and tentatively explained. The evaluation results and additional information gained in the construction and implementation of the software or artifact are fed back to another round of "suggestion" through circumscription. The overall functionality of the system and the performance of the system is tested and the content quality is evaluated according to the user specifications.

B. Research sub-questions Mapped to Methodologies

Table 1 below shows a mapping of the research objectives to respective methodologies and their deliverables.

TABLE 1: Mapping of the research objectives to respective methodologies

Objective	Methodology	Deliverable
To optimize web search engine results.	Prototyping	Optimized web search engine results.
To retrieve appropriate search results based on the user search query.	Prototyping	Accurate search results based on the user search query.
To evaluate the quality of content retrieved.	Evaluation	Quality content.
To filter search results based on user preferences.	Survey	Filtered results based on user preferences

IV. RESULTS

A. Figures and Tables

The application consists of a back-end and a front-end setup. The back-end is only accessible to system administrator and the front-end is what users see when they visit http://kaymasukela.co.za/searchengine.html.

Backend
URL:
https://cse.google.co.za/cse/setup/basic?cx=007115424577962053366:3zpnco-u82o

Adding Trusted Backend
The search allows for users/administrators to define their trusted sources, their frequently visited (Figure 1) sites in order to emphasize and prioritize (Figure 2) these sites on the search results.

Removing Unauthorized Advertisement
Removing unauthorized advertisements or 'adds' is one of the key objectives of this paper. The application allows administrators to enable/unable promotions that appear on the search result page.
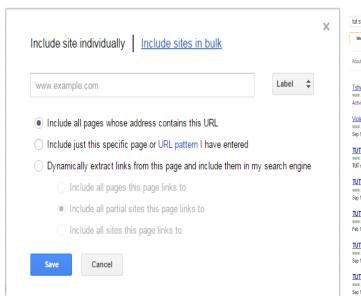

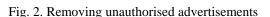Fig. 1. Including trusted sources in the search engine results

Fig. 2. Removing unauthorised advertisements

Fig. 3. Filtering of Search results

Filtering of Search Results
The search results can either by filtered or sorted by either the date (publishing date of the page/image) or by relevance (Relevance of the page/image).

Final Search Results Based on Users' Preferences
Figure 4 demonstrates proper, accurate, relevant, and clean search results. The results shown are from the trusted sources and are ranked higher than the ones from other sources.
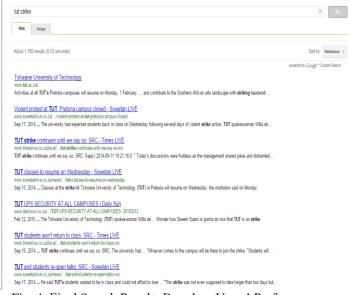

Fig. 4. Final Search Results Based on Users' Preferences

B. Evaluation Survey Results

An online survey consisting of eight questions was conducted and received fifteen responses from the users. Multiple choice questions with only one field that require user input were formulated. The questions are based on user experiences with web search engines. The online survey is accessible from URL http://goo.gl/forms/LJ5pekQ22Q.

All respondents indicate they frequently use a search engine to find information and all mostly use google. Twenty percent of them say they are occasionally likely are you to click on a Pay-Per-Click (PPC) link provided by a search engine (i.e. Google, Yahoo, Bing) while 80% rarely do. 66.7% of them can you distinguish between Pay-Per-Click (PPC) results and Organic (SEO) Results while 26.7% can't. One third of the respondents say their type of search (i.e. shopping or research) influence their likelihood of using or not using a PPC link. 46.7% of them trust the results they get from a search engine while 53.3% only sometimes do. The websites often visited by respondents in a week include academic, sports, new and current affairs, entertainment, scientific/discovery and others.

The search results are preferred by the respondents to be in a presentable manner listed in order of relevancy; to give accurate results; to be quick, fast with recommendations. Some prefer known sites or authors as not everything on the internet is accurate and reliable. Some say that as long the results can be returned better than they are now, they are okay. A respondent wants keywords ordered and used websites first, then blogs, sites with youtube and rss feeds later, not on top. Most respondents prefer results to be accurate as long as they are relevant search results. Most prefer the most important things on the sites to be listed on page one of search results.

V. CONCLUSION

Most users have a routine way of finding information on the web. This means that they often visit the same websites over and over. Since users know which websites to visit for news, research, or whatever interest they maybe have, we can include all the websites and emphasize them in the search results to be the ones that appear first if they make a search query. This study highlighted the importance of identifying Spam or misleading content and unauthorized advertisements on the results page, which will help users differentiate between a Pay-Per-Click link and an Organic link when they choose they're search results.

With this application, users can actually decide what information they want to see, how they want to see it, and where to find it. All the unnecessary noisy presented on the result page which misleads the user, affects performance, and uses a lot of bandwidth is eliminated using this system.

Service providers must provide user training and awareness/tips on web search engines. The impact of clicking on a misleading link can result in cost incurred, viruses, and a waste of time on an irrelevant website. Since most service providers are businesses with profit their main target, it is the responsibility of the user to understand how pages are returned against a search query and how most internet marketers manipulate the search results to rank their pages high on the list. Users must differentiate between an organic link and a pay-per-click or advertisements provided by marketers for profit purposes.

REFERENCES

[1] E. Agichtein, E. Brill, S. Dumais & R. Ragno. Learning user interaction models for predicting web search result preferences. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006. ACM, 3-10.

[2] M. K. Bergman. 2001. White paper: the deep web: surfacing hidden value. Journal of electronic publishing, 7.

[3] M. Chau, D. Zeng & H. Chein. Personalized spiders for web search and analysis. Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries, 2001. ACM, 79-87.

[4] M. Claypool, P. Le, M. Wased & D. Brown. Implicit interest indicators. Proceedings of the 6th international conference on Intelligent user interfaces, 2001. ACM, 33-40.

[5] M. Egele, P. Wurzinger, C. Kruegel & E. Kirda. 2009. Defending browsers against drive-by downloads: Mitigating heap-spraying code injection attacks. Detection of Intrusions and Malware, and Vulnerability Assessment. Springer.

[6] S. Fox, K. Karnawat, M. Mydland, S. Dumais & T. White, 2005. Evaluating implicit measures to improve web search. ACM Transactions on Information Systems (TOIS), 23, 147-168.

[7] S. Gauch, J. Chaffee & A. Pretschner. 2003. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems, 1, 219-234.

[8] J. Goecks & J. Shavlik. Automatically labeling web pages based on normal user actions. Proceedings of the IJCAI Workshop on Machine Learning for Information Filtering, 1999.

[9] M. Gordon & P. Pathak. 1999. Finding information on the World Wide Web: the retrieval effectiveness of search engines. Information processing & management, 35, 141-180.

[10] K. Hampton, L.S. Goulet, L.Rainie & K. Purcell. 2011. Social networking sites and our lives. Retrieved July 12, 2011.

[11] E. Hargittal, L. Fullerton, E. Menchen-Trevino & K.Y. Thomas. 2010. Trust online: Young adults' evaluation of web content. International Journal of Communication, 4, 27.

[12] M.R. Henzinger, R. Motwani & C. Silverstein. Challenges in web search engines. ACM SIGIR Forum, 2002. ACM, 11-22.

[13] T. Joachims. Optimizing search engines using clickthrough data. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002. ACM, 133-142.

[14] D.K. Kaushik & K.N. Murthy. Personal search assistant: A configurable personal meta search engine. AUSWEB99: Proceedings of the 5th Australian World Wide Web Conference. Southern Cross University, 1999.

[15] H.R. Kim & P.K. Chan. Learning implicit user interest hierarchy for context in personalization. Proceedings of the 8th international conference on Intelligent user interfaces, 2003. ACM, 101-108.

[16] J.A. Konstan, B.N. Miller, D. Maltz, J.L. Herlocker, L.R. Gordon & J. Riedl. 1997. GroupLens: applying collaborative filtering to Usenet news. Communications of the ACM, 40, 77-87.

[17] B. Kuechler & V. Vaishnavi. 2008. On theory development in design science research: anatomy of a research project. European Journal of Information Systems, 17, 489-504.

[18] S. Mizzaro & C. Tasso. 2002. Ephemeral and persistent personalization in adaptive information access to scholarly publications on the web. Adaptive Hypermedia and Adaptive Web-Based Systems, 2002. Springer, 306-316.

[19] M. Morita & Y. Shinoda. 1994. Information filtering based on user behavior analysis and best match text retrieval. Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, Springer-Verlag New York, Inc., 272-281.

[20] D.W. Oard & J. Kim. 1998. Implicit feedback for recommender systems. Proceedings of the AI workshop on recommender systems, 81-83.

[21] L. Page, S. Brin, R. Motwani & T. Winograd. 1999. The PageRank citation ranking: bringing order to the Web.

[22] M. Speretta & S. Gauch. Personalized search based on user search histories. Web Intelligence, 2005. Proceedings. The 2005 IEEE/WIC/ACM International Conference on, 2005. IEEE, 622-628.