IBIMA Publishing
Communications of the IBIMA
http://www.ibimapublishing.com/journals/CIBIMA/cibima.html
Vol. 2010 (2010), Article ID 126850, 7 pages
DOI: 10.5171/2010.126850

A Reflection of Search Engine Strategies

Anushia Inthiran[1], Saadat M. Alhashmi[1], and Pervaiz K. Ahmed[2]

[1]School of Information Technology, Monash University Sunway Campus, Malaysia
[2]School of Business, Monash University Sunway Campus, Malaysia

---

Abstract

Information retrieval and search engines are almost synonymise. Usually search engines are employed to perform the search activity. If search engines merely return search results without much analysis, a user will be overwhelmed with search results. The aim is to return results that are relevant and not cornucopia of search results. Search engines today utilise many methods in order to provide users with relevant search results. Relevant results can only be provided if search engine strategies are able to discern the users' information seeking goal. However the tremendous growth of the Internet and the variety of users using the search engine make it difficult for search engines to satisfy the user's diverse information seeking goal. Today users demonstrate various nuances while searching; parallel searching, multiple information seeking goals in a single search session and the use of multiple browsers for a single search. It has become pressing that search engines take into account these search behaviours in the attempt to provide users with relevant search results. In this paper we discuss three methods: query expansion, user search history and re-ranking- in an attempt to provide searchers results that match their information needs. These methods are common strategies used by search engines. Unfortunately, these techniques are not satisfactory when assessed against providing users with relevant results. We also provide insights to a new direction that search engines have venture into. Rather than just limiting search strategies to technical implementation/aspect, the bigger picture of the search process and the user needs to be looked into.

Keywords: re-ranking, search history, query expansion, personalization

---

Introduction

Search engines provide an interface for users to enter the query. Most search engines like *Google*, *Yahoo* and *AltaVista* provide a free form text box that allows for the user to key in the query. *Google*, now provide users with options, for example, the user can now request for the search to be done on local pages or for pages from the World Wide Web. In addition, search is limited to certain parameters, for example search for video or images. Search engines recognise that users prefer to have options while searching. A single interface is not longer suitable to fit the multitude of users on the Internet. Fu (2007) states the search engine is becoming increasing

 Contact author: Anushia Inthiran e-mail: anushia.inthiran@infotech.monash.edu.my

important as a tool to acquire information and enable self directed learning. In this aspect, Teevan et al.,(2005) states the search engine is now not just seen as a tool to acquire information from the World Wide Web but also to extend the process of searching for information into a lifelong learning process. In this day and age, the search engine plays an important role in supporting this process. This is further accentuated by the fact that Xu (2007) states computers become consumer products; and the Internet a mass medium, searching the web had become a daily activity for everyone from children to scientist. The search engine now has to cater to various people from different background and age group.

The search engines task is to present results to the user based on the query provided. There are generally three methods of how the search engine function, crawler based, directory or human powered and hybrid. In crawler based search engines, listings are created automatically. In human powered directory, a short description about the web site is submitted to the directory, in the hybrid technique, both methods are merged. These are the techniques used that will allow for the search engine to present results to the user. There are two categories of users involved when using a search engine; one is the webmaster and the other the searcher. For the webmaster, listing in major search engines is an important place to be, because they can potentially generate more traffic. Sullivan (2007) state for the searcher, using well known and dependable search engines generally mean more dependable results. In this paper the focus in on the aspect of the user.

Micarelli et al., (2007) states the exponential growth of the Internet, has made the task of the search engine more difficult, users surfing the web in search of resources to satisfy their information need have less and less time and patience to formulate queries, wait for the results and sift through them. The search engine now has to deal a variety of information needs, user behaviour and eccentricities. Besides just looking at time as a factor, other issues like not being able to discern and individual's goal add to the frustration of the searcher. Lv et al., (2006) states some searchers also indicate that results are being ordered by popularity rather than relevance to the user's individual need. If search results deviate from the user's information goal then the searcher will definitely not be satisfied with the search results and have to spend more time 'searching further' in the results page. Additionally, Beaza-Yates et al., (2004) state users also complain that, although the user interface of search engines is user friendly, it is not always easy for searchers to express or formulate their queries. Since the query is the only input that the search engine is provided with in order to present relevant results, the search engine has to now maximise the usage of the query for it to be beneficial to the user.

Three popular techniques used by search engines to provide relevant results to users are query expansion, search history and re-ranking. Generally these techniques used past and present activities/user actions or logs to predict information seeking goals. Once the information seeking goal is obtained, search results are tailored to fit the goal. In the following sections, we discuss how search engines utilise these techniques to provide users with relevant results. Query expansion, search history and re-ranking will be examined in Section 2.0 3.0 and 4.0 respectively. In Section 5.0 an analysis and conclusion will be provided.

Query Expansion

Query expansion is a technique that uses the original query provided by the user. Expansion is performed by the search engine by viewing the user's history log or by using algorithms. The rational for expanding the query is related to being able to provide users with more results. From the search engine perspective, *Google* reports that the average *Google* query is 3 in the year 2007. With just 3 words, it is not possible for the search engine to discern the searchers information need. Query expansion is used to provide more words. The assumption made is that more words results in more information for the search engine to use to provide searchers with relevant results.

Intelligent Client Side Web Search Agent (*UCAIR*) developed by Shen et al., (2005) is a client side browser plug-in that acts as a proxy for web search engines. One of the methods used in *UCAIR* to provide searchers with relevant results is query expansion. In order to perform query expansion, the users' actions such as viewing a document and clicking on the '*Back*' or '*Next*' button is used as indicators. With these responses, the search engine will decide if previous queries are related to the current query and if so, expand the current query with useful terms from the previous query.

This technique requires clear demarcation of session boundary. In the event the adjacent queries are not related to each other, the search engine will need to be able to differentiate the two. If this is not done then the query expansion will be filled with unrelated terms that will cause the search results to deviate from the users information need. Currently in *UCAIR*, textual similarity is used to perform boundary detection. Textual similarity is performed by looking at similarity in terms of concepts.

Teevan et al., (2005) states that web log analysis is also used to perform query expansion. If query expansion is limited to statistical information rather than textual similarity, this limits the expansion technique. This is because related queries do not necessarily share the same words, for example, when a user keys in the query, 'java island' and if the adjacent query was 'travel Indonesia', and if similar words is used, then although the two queries are related, the query expansion mechanism will not be able to relate the two queries. This will lead to poor results list. However if textual similarity is used then, the search engine will be able to determine that both the queries are related.

In general, when query expansion is performed, the search engine will need to determine the number of words to select for expansion. Selecting too few words or too many words will affect the expansion process. The focus on the number of words required for query expansion is an area which has not been dealt with much in literature. Popular research area for query expansion is related to where to obtain information to perform the expansion.

Xu et al., (2007) states query expansion is also explored using an algorithm used to perform correlation of terms. An overlay algorithm is used to perform expansion on individual terms. To date, this is the first research technique that performs expansion on individual words of the query and not the entire query string. However, with more query terms natural language issues will complicate the expansion process. Some form of distinction must be made between the selection of query terms to expand, If two unrelated query terms is expanded this will lead to relevant results being missed out.

Current query expansion techniques look at textual similarity, number of keywords and whether the entire query is expanded

or individual terms expanded in totality. There is also the need to consider if expansion is performed semantically. Currently, conceptual and contextual similarities is used, however since semantic expansion is rarely researched, we are unable to conclude as to which of the three techniques will perform better.

In the next section search history will be examined.

Search History

Usage of search history requires a tool to collect information from the user, in work by Speretta and Gauch (2005), *Google Wrapper* is used. User browsing history is seen as a source of information that can be used to narrow down the users' interest and search goal.

However, merely collecting the search history without organisation and categorization is not beneficial to the overall results presentation. A method is needed to decide what needs to be collected and where to place this collection. Speretta and Gauch (2005) state, semantics and ontology can be used to organise the search history automatically. In work done by Kim and Chan (2003) browsing history is organised into clusters to create user interest hierarch. Web pages collected is organised into clusters, when a user visits a page it is taken as an indication of interest. This technique of measuring interest is error nous. Visiting a page is also an indication of disinterest. Verification method is necessary to quantify if a page is of interest to the user. An extension to this idea is implemented by Chan (1999) where a metric is developed to determine user level interest, for example, the percentage of links visited on a page or *URL's* present in bookmarks.

The implementation of a monitoring tool on the user's machine to perform logging is a cause for concern. Concerns over privacy protection is growing in parallel with the demand for services. Gauch et al., (2007) state that these two trends, privacy and obtaining better search results seem to be in direct opposition to each other, so privacy protection must be a crucial component of every personalized system.

Similarly, if a monitoring tool is implemented on the users' desktop, space necessary to store logs is another cause for concern. Speretta and Gauch (2005) state maintenance is required on the logs to determine which ones is relevant to the users' query. Differentiation must be made between short term and long term search context. Speretta and Gauch (2005) state implicit feedback information collected over a long period of time is unlikely to be very useful, but the immediate search context and feedback information is expected to be much more useful. Unfortunately, current search history technique is not able to differentiate between short term and long term interest.

In a home personal computer, there is a possibility that there can be more than one user to the computer. Speretta and Gauch (2005) states that each user will need to register their email address to obtain a cookie to store and upload their user id on their local machine. This solves the problem of multiple users on a machine, however, this task is painful for the user to perform, especially if the cookie is lost. Users do not want to manually perform these responsibilities in order to obtain relevant results. They require these activities to be performed automatically. Organisation of web logs collected requires further research to determine an effective method of usage.

The technique of re-ranking is examined in the next section.

Re-ranking

Re-ranking is a technique where the original order of corpus of information is changed to suite an individual user. The change in rank is necessary as generic rank provides generic search results. This eliminates the user from browsing through irrelevant search results. Irrelevant information browsing only adds to the cognitive burden and increases the amount of time taken to find information.

In work by Shen et al., (2005), re-ranking is performed based on two cases when the user clicks on the '*Back*' button and '*Next*' on the browser. Any seen and unseen results is re-ranked so that the user will see improved search results immediately. However, this model does not take into account exceptions when the user interaction is a result of an error. Time spent on a link, before clicking on the '*Back*' or '*Next*' button is a good indication of interest to use before re-ranking is automatically performed. However Shen et al., (2005) states monitoring time, is an overhead. Sometimes this technique is not completely reliable. It only brings up a small number of highest re-ranked results to be followed by any originally high ranked results.

Re-ranking based on historical clicks is discussed in Dou et al., (2007) , however this technique fails for a completely new query since there is no data to manipulate for re-ranking to take place. The issue of 'cold start' was first coined in Zigoris and Zhang (2006). Any new user to the system must endure poor initial performance until sufficient feedback from that user is provided.

In re-ranking, the issue of using validity metrics is important to determine the user's interest when clicking on the '*Back*' or '*Next*' button. Bearing in mind that it is not possible to re-rank all results the objective of re-ranking is that users find relevant pages to their information need at the top of the results list. Users are biased to results that are listed at the top of the results list. A rational searcher might be expected to asses each of these page summaries against their information need and click on the one that appears as the most relevant. Keane (2008) state people may not search in such a way. The issue with cold start when using historical clicks to perform re-ranking persists whenever a new query is input. However, this is only an intermittent issue.

In the next section, we will look at the overall analysis of these techniques and a conclusion for future direction in the area of personalized search results.

Analysis and Conclusion

In query expansion, the issue to expand individual terms or to collectively expand terms remains an unsolved issue. Although more query words result with more information, the search engine needs to avoid issues with the natural language. Demarcation of session boundary is necessary for query expansion. If two adjacent queries is not related and the search engine expands the current query with terms of the previous query this results with users having to browse irrelevant pages before actually finding related information to the users' information seeking goal. Using textual similarity contextually and conceptually have been researched. However, keyword matching makes natural language issues more apparent when compared to conceptual and contextual similarity.

In user search history, the main issue is the location of implementation. Privacy issues, organisation and storing of the logs require the need to organise and effectively use the logs. Logs provide abundance of information but at the same time, an effective technique is needed to reap the

full benefit of these logs to provide users with what they want.

Re-ranking bring about issues with cold start and validity of measurements used. The objective of re-ranking is to place relevant results at the top of the search, hence provide users with their information seeking goal quicker. However, it is impossible to perform re-ranking for the entire result page.

Gulli and Signorini (2005) states to a certain extent the review of techniques used provides some solution for users but there is still room for improvement. In an attempt to tailor accurate and appropriate results to users in the result page, a search engine is required to crawl over a massive and ever-growing collection of online content. Searching remains a popular activity on the Internet, in a survey conducted by Pew Internet (2008) , 89% of Americans used a search engine to find information on the Internet. However, according to Marchionini (2006), due to the sheer volume, the techniques mentioned above, search engines only continue to perform 'lookup search", users still need to browse from page to page to locate the information that they need].

The probable solution to this problem that is contributed by volume, natural language issues, limited query words, difficulty in expressing information need in the query and obtaining the users information goals is personalization. This is a method according to Xu et al., (2007) tries to adapt to individual needs and to move away from the one size fits all.

Speretta and Gauch (2005) state personalization is the process of presenting the right information to the right user at the right moment. Let us take an example of why it is important to provide users with the best possible results. In a survey done by Search Engine Watch (2008), 'redbox' appeared in the movies category as a top ten search term in December 2008. However in Singapore, 'redbox' is an online florist, in Malaysia a karaoke outlet. In these scenarios, personalization based on the 'right user' and 'right moment' is not easy to achieve. Moreover, the 'right moment' and 'right user' can change from time to time.

When personalization techniques is employed instead of only focusing on technical details of query expansion, user logs and re-ranking the bigger picture of the user's information need and what the user needs 'right now' is achieved. Personalization is also the solution to 'do what I mean and not what I say'.

References

Baeza-Yates, R. Hurtado, C. & Mendoza, M. (2004). Query Recommendation Using Query logs in Search Engines, Current Trends in Database Technology, *EDBT 2004 Workshop*, 3268, pp. 588-596.

Chan. P.K.,(1999) A non-invasive learning approach to building web user profiles, KDD-99, *Workshop on web usage analysis and user profiling*, pp. 7-12.

Dou, Z., Song, R., & Wen, J.R. (2007), A large scale evaluation and analysis of personalized search strategies, *Proceedings of the World Wide Web Conference*, pp.581-590.

Fu, X. (2007). Evaluating sources of implicit feedback in web searches, *ACM Conference On Recommender Systems Proceedings of the 2007 ACM conference on Recommender Systems*, pp. 191-194.

Gauch, S., Speretta, M., Chandramouli,A. & Micarelli,A. (2007). User Profiles For Personalized Information Access, *Lecture Notes in Computer Science*, The Adaptive Web, Computer Science, pp 54-89.

Gulli, A. and Signorini, A. (2005), The indexable Web is more than 11.5 billion pages. *In proceedings of the 14th International Conference in the World Wide Web, 2005*, USA, pp.902 – 903 .

Huang X., Peng F., An, A., & Schuurmans, D. (2004). Dynamic web log session identification with statistical language models. *Journal of the American Society for Information Science and Technology*, 55(4):1290-1303.

Keane,M.T., O'Brien,M. & Smyth, B. (2008). Are People Biased in Their Use if Search Engines?, *Communications of the ACM*, 51(2), pp. 49 -52.

Kim H.R. & Chan P.K. (2003). Learning implicit user interest hierarchy for context in personalization, *Proceedings of the 8th International conference on Intelligent user interfaces, USA, 2003*, pp.101-108.

Lv, Y., Sun,L., Zhang, J., Nie, J.Y., Chen, W. & Zhang, W. (2006). An Iterative Implicit Feedback Approach to Personalised Search, *21st International Conference on Computational Linguistics*, pp.585-593.

Marchionini G. (2006). Exploratory Search: From Finding to understanding, *Communications of the ACM,* Vol. 49 No. 4.

Micarelli,A., Gauch, S., Speretta, M., & Chandramouli, A. (2007). User Profiles For Personalized Information Access, The Adaptive Web, *Lecture Notes in Computer Science 4321*, pp 54-89.

Pew Internet & American Life Project Tracking Surveys, Internet Activities March 2000 – December 2008 (online) from http://www.pewinternet.org/Data-Tools/Download-Data/Trend-Data.aspx [Accessed 30th June 2009].

Shen, X., Tan, B. & Zhai,C. (2005). Implicit User Modeling for Personalized Search, *Proceedings of the 14th International Conference on Information and Knowledge Management*, pp. 824-831.

Speretta, M. & Gauch,S. (2005). Personalizing Search Based on User Search Histories, *IEEE/WIC/ACM International Conference on Web Intelligence*, pp.622-628

Sullivan D. (2007). (online) Search Engine Watch, Major Search Engines and Directories from Sullivan, D. (2007), from http://searchenginewatch.com/showPage.html?page = 2156221 [Accessed 30th June 2008].

Teevan,J., Dumais,S.T. & Horvitz, E. (2005). Personalizing Search via Automated Analysis of Interests and Activities, *Proceedings of the 29th Annual International ACM Special Interest Group on Information Retrieval,* pp. 449-456.

Top 10 Search Terms in 10 Categories, December 2008, Search Engine Watch, from http://searchenginewatch.com/3632401.print (13th February 2009) .

Xu,J., Zhu,Z., Ren,X., Tian, Y. & Luo,Y. (2007). Personalized Web Search Using User Profile, *Proceedings of the 2007 International Conference on Computational Intelligence and Security*, pp. 222-226.

Zigoris, P. & Zhang,Y.(2006). Bayesian Adaptive User Profiling with Explicit and Implicit Feedback, *Proceedings of the 2006 ACM CIKM International Conference on Information and Knowledge Management*, pp. 397-404.