Search Results Re-ranking using User Goals

Kota Vamsee Krishna[#1], Smita Deshmukh[#2]
#Information Technology, Terna Engineering College, Mumbai University
Sector-22, Nerul, Navi Mumbai, Maharashtra, India

***Abstract*— Before submitting any query to a search engine, every user has a specific goal in mind but that goal is not known to the search engine. Based on the query related matching information which is currently available in the database, search engine will display the results and user will have to scan through them to find out the website of his interest. From user experience perspective, this is a very time consuming activity. User would be happy to see the displayed search results to be categorized into various groups (such as Business, Technology, Sports etc..) and the results in each group to be arrangedstarting from the individual user's most visited to least visited website. This project is about a desktop-based application which will re-rank the display of search results, for end users, using user goals. Users will have to register themselves and create user id and password. Considering one login/logout as a single session, usage logs will be captured, feedback sessions and pseudo-documents will be generated to re-rank and optimize display of search results. The performance of this appliation will be evaluated using Classified Average Precision (CAP) factor.**

***Keywords*— User search goals, feedback sessions, pseudo-documents, re-ranking search results, classified average precision.**

I. INTRODUCTION

In today's world, if anyone needs any information then the first thing that they do is to search for it in any of the popular search engines using Internet. Some of the popular search engines are Google, Microsoft Bing etc.. Let us call the information that user is looking for as a 'Query'. Before entering the query in search engine, every user has a goal (i.e. specific information) in mind that theywant to achieve. But the search engine does not know what exactly the user is looking for, so it will display all the available URLs belonging to various domains without any categorization. The user has to scan through all the displayed URLs one-by-one until the desired URL is identified.

Practically every user wants the required/relevant/wanted/needed information to be quickly displayed to save time and this would be possible only if there is an improvement in the display of search results. This gives rise to the need for developing a desktop based application which will analyse usage logs, produce feedback sessions, map the feedback sessions to pseudo-documents, cluster the pseudo-documents, infer user search goals, re-rank the search results whenever user submits the same/similar query next time, evaluate the performance of the re-ranked results using Classified Average Precision(CAP) and further optimize it.

II. RELATED WORK

The authors of reference paper [1],propose a unique approach to conclude user search goals by adeep analysis of search engine query logs. The work in reference paper [2],talks about text classification algorithms used to automatically classify random search results into an existing category structure as a continuous on-going activity. Organizing search results will allow a user to concentrate on URLs in the user required categories.Further, the reference paper [3] discusses about organization ofthe web search results into a cluster and thusfacilitate a user in quick browsing of the search results. The reference paper [4] highlights regarding effective organizing of search results as a critical activityso as to improvise the utility of any search engine. To navigate through relevant documents easily, clustered search results prove to be efficient.Displaying the relevant documents in the order starting from most relevant to least relevant is done by a good information retrieval system and this is explained inreference paper [5]. The goal is to develop a method that utilizes click-through data for training, namely the query-log of the search engine in connection with the log of links the users clicked on in the presented ranking.As per reference paper [6] depending on the clicks relative preferences are obtained which are averagely accurate. The work in reference paper [7] presentsa method which is based on query clustering process. It facilitatesgrouping of semantic similar queries which could be identified. Referring to the work in reference paper no. [8],it proposes and evaluates a method for auto-segmentation of user's query streams into distinct units.Reference paper [9]states that query substitution is generation of a new query inorder to replace the user's old query.The work in reference paper [10] presentsincrease of precision retrieval, the new search engines input manually verified answers to frequently asked queries.

III. FRAMEWORK

The framework is divided into two parts. The upper part is to generate the feedback session and pseudo-document. The lower part is display re-ranked search results, whichare obtained from original search results.User search goals can be inferred from the upper part.

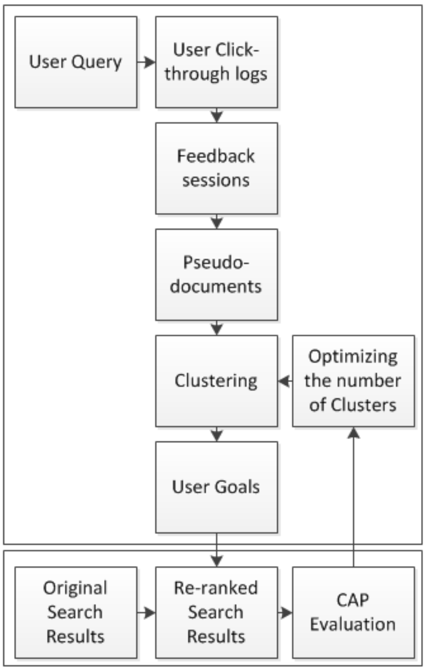
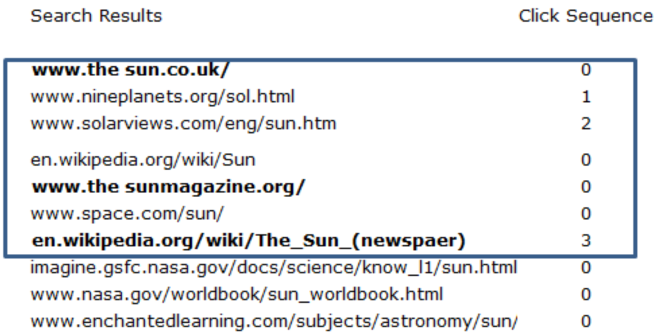


Fig1: Framework

IV. FEEDBACK SESSIONS AND PSEUDO DOCUMENTS

1. Feedback Session. A feedback session consists of both clickedand unclicked URLs and ends with the last URL that wasclicked in a single session.



Search Results	Click Sequence
www.the sun.co.uk/	0
www.nineplanets.org/sol.html	1
www.solarviews.com/eng/sun.htm	2
en.wikipedia.org/wiki/Sun	0
www.the sunmagazine.org/	0
www.space.com/sun/	0
en.wikipedia.org/wiki/The_Sun_(newspaer)	3
imagine.gsfc.nasa.gov/docs/science/know_l1/sun.html	0
www.nasa.gov/worldbook/sun_worldbook.html	0
www.enchantedlearning.com/subjects/astronomy/sun/	0

Fig.2: Feedback Session

Referring to Fig.2, '0' means 'Unclicked' URL. All 10 URLs are a part of single session. The rectangular represents a feedback session.

2. Pseudo Documents

Users generally tend to submit either a single word or may be few words to a search engine. Collectively, a set of these keywords can be called as 'goal texts' or 'pseudo-documents'.

3. Map Feedback Sessions to Pseudo-Documents

The feedback session consists of URLs, title and snippet. Using text processing techniques such as stemming, commonly used words such as nouns can be extracted as keywords and stored in pseudo-documents. Next time whenever user submits the same keyword which is present in pseudo-document, all the related URLs captured in feedback session will get displayed. But our aim is to further re-rankthese results into groups and to re-rank URLs in each group, for that clustering will be needed.

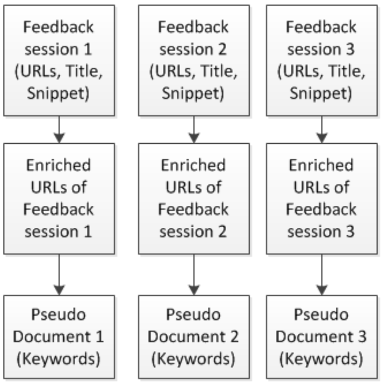

Feedback session 1 (URLs, Title, Snippet)
Feedback session 2 (URLs, Title, Snippet)
Feedback session 3 (URLs, Title, Snippet)
Enriched URLs of Feedback session 1
Enriched URLs of Feedback session 2
Enriched URLs of Feedback session 3
Pseudo Document 1 (Keywords)
Pseudo Document 2 (Keywords)
Pseudo Document 3 (Keywords)

Fig.2: Mapping Feedback session to Pseudo-document

Producing a pseudo-document includes two steps:
1) Using URLs in the feedback session. Each URL's title andsnippet are represented by a Term Frequency-InverseDocument Frequency (TF-IDF) vector [1], respectively, as in

$$T_{ui} = [t_{w1}, t_{w2}, …., t_{wn}]^T$$
$$S_{ui} = [s_{w1}, s_{w2}, …., s_{wn}]^T \qquad (1)$$

where T_{ui} and S_{ui} are the TF-IDF vectors of the URL's titleand snippet, respectively. u_i means the *i*th URL in thefeedback session. And$w_j (j=1,2,...,n)$ is the *j*th termappearing in the enriched URLs. t_{wj} and s_{wj}represent the TF-IDF value of the *j*thterm in the URL's title and snippet, respectively. Consideringthat URLs' titles and snippets have different significances,we represent the enriched URL by the weightedsum of $T_{ui}$and S_{ui}, namely

$$F_{ui} = w_t T_{ui} + w_s S_{ui} = [f_{w1}, f_{w2}, …, f_{wn}]^T \qquad (2)$$

where F_{ui} means the feature representation of the *i*th URLin the feedback session, and $w_t$and w_s are the weights of thetitles and the snippets, respectively. We set $w_s$to be 1 at first.Then, we stipulate that the titles should be more significantthan the snippets. Therefore, the weight of the titles shouldbe higher and we set w_t to be 2 in this paper. We also tried toset $w_t$to be 1.5, the results were similar. Based on (2), thefeature representation of the URLs in the feedback sessioncan be obtained. It is worth noting that although T_{ui} and $S_{ui}$are TF-IDF features, F_{ui} is not a TF-IDF feature. Each term of F_{ui} (i.e., f_{wj}) indicatesthe importance of a term in the *i*th URL.

*2) Forming pseudo-document based on URL representations.*Let F_{fs} be the feature representation of a feedback session,and $f_{fs(w)}$ be the value for the term *w*. Let $F_{u\acute{c}m}$(m=1, 2,…,M) and $F_{u\acute{c}l}$(*1*= 2,…,L) be the featurerepresentationsof the clicked and unclicked URLs in this feedbacksession, respectively. Let $f_{ucm}(w)$ and $f_{u\acute{c}l}(w)$ be the valuesfor the term *w* in the vectors. We want to obtain such aF_{fs} that the sum of the distances between F_{fs} andeach F_{ucm} is minimized and the sum of the distancesbetween F_{fs} and each $F_{u\acute{c}l}$ is maximized. Based on theassumption that the terms in the vectors are independent,we can perform optimization on each dimension independently,as shown in

$$F_{fs} = [f_{fs}(w_1), f_{fs}(w_2),…,f_{fs}(w_n)]^T$$

$$F_{fs}(w) = \arg\min_{ffs(w)} \{\sum_M [f_{fs}(w)-f_{ucm}(w)]^2 - \lambda \sum_L [f_{fs}(w)-f_{u\acute{c}l}(w)]^2\},\ f_{fs}(w) \epsilon I_c (3)$$

Let I_c be the interval $[\mu f_{uc}(w)-\sigma f_{uc}(w),\ \mu f_{uc}(w)+\sigma f_{uc}(w)]$ and $I_{\acute{c}}$be the interval $[\mu f_{u\acute{c}}(w)-\sigma f_{u\acute{c}}(w),\ \mu f_{u\acute{c}}(w)+\sigma f_{u\acute{c}}(w)]$, where $\mu_{fuc}(w)$ and $\sigma f_{uc}(w)$ represent the mean and mean square error of $f_{u\acute{c}}(w)$, respectively. If $I_c \underline{c} I_{\acute{c}}$ or $I_{\acute{c}} \underline{c} I_c$, we consider that the user does not care about the term *w*.In this situation, we set $f_{fs}(w)$ to be 0, as shown in

$$F_{fs}(w) = 0, I_c \underline{c} I_{ć} \text{ or } I_{ć} \underline{c} I_c \qquad (4)$$

λ is a parameter balancing the importance of clicked andunclicked URLs. When λin (3) is 0, unclicked URLs are nottaken into account. On the other hand, if λis too big,unclicked URLs will dominate the value of $f_{fs}(w)$. In thispaper, we set λ to be 0.5.

V. IMPROVED USER SEARCH RESULTSUSING PSEUDO-DOCUMENTS

Using pseudo-documents we can predict user search goals and thus obtain improved user search results.

As in (3) and (4), each feedback session is representedbya pseudo-document and the feature representation of thepseudo-document is F_{fs}. The similarity between twopseudo-documents is computed as the cosine score of $F_{fsi}$and F_{fsj} , as follows:

$$S_{imi,j} = \cos(F_{fsi}, F_{fsj}) = (F_{fsi} . F_{fsj}) / (|F_{fsi}|.|F_{fsj}|) \qquad (5)$$

And the distance between two feedback sessions is

$$D_{isi,j} = 1 - S_{imi,j} \qquad (6)$$

We cluster pseudo-documents by K-means clusteringwhich is simple and effective. Since we do not know theexact number of user search goals for each query, we set Kto be five different values (i.e., 1, 2,….,5) and performclustering based on these five values, respectively.

After clustering all the pseudo-documents, each clustercan be considered as one user search goal. The center pointof a cluster is computed as the average of the vectors of allthe pseudo-documents in the cluster, as shown in

$$F_{centeri}=\sum_{k=1}^{Ci} F_{fsk}/C_i \text{ , } (F_{fsk} c \; Clusteri)(7)$$

where $F_{centeri}$ is the *i*th cluster's center and C_i is the numberof the pseudo-documents in the ith cluster. $F_{centeri}$ isutilized to conclude the search goal of the *i*th cluster.

Keywords having highest values in the centre are the user search goals. They can also be used in future query recommendations.

VI. EVALUATION

We will use CAP to evaluate re-ranked results inorder to select the best cluster number.

*1. Re-ranking Web Search Results:*Web search results are re-ranked by categorizing URLs of original search results into groups (Business, Technology, Sports, etc…).

*2. Evaluation Criterion:*From user click-through logs, implicit relevance feedbacks can be obtained, namely 'Clicked URL' means relevant and 'Unclicked URL' means irrelevant.

*A. Average Precision.*A possible evaluation criterion is the average precision (AP) which evaluates according to user implicit feedbacks.

$$AP = (1/N^{+)} [\sum_{r=1}^{N} rel(r).(R_r/r)] \qquad (8)$$

where N^+ is the number of clicked documents in the retrieved ones, *r* is the rank, *N* is the total number of retrieved documents, *rel()* is a binary function on the relevance of a given rank, and R_r is the number of relevant retrieved documents of rank *r* or less. For example, considering the single feedback shown in Fig.2and we can compute AP as: {(1/4) * [(1/2)+(2/3)+(3/7)+(4/9)]} = 0.510. However, AP is not suitable for evaluating the re-structured or clustered searching results.

*B. Voted Average Precision (VAP).*Referring to fig.2, the URLs in a single feedback session are re-structured into two classes where the un-boldfaced ones are clustered into class 1 and boldfaced ones are clustered into class 2. Voted Average Precision (VAP) is the average precision of the class including more clicks namely votes. If the numbers of the clicks in two classes are the same, the bigger AP is selected as the VAP. Assume that one user has only one search goal, then ideally all the clicked URLs in a single session should belong to one class. And a good restructuring of search results should have higher VAP.

The URLs in the single session are restructured into two classes where the un-boldfaced ones are clustered into class1 and boldfaced ones are clustered into class 2.

For example, the VAP of the restructured search results is the AP of class 1, calculated by: VAP = (1/3)*[(1/1)+(2/2)+(3/6)] = 0.833.

However, VAP is still an unsatisfactory criterion. Considering an extreme case, if each URL in the click session is categorized into one class, VAP will always be the highest value namely 1 no matter whether users have so many search goals or not. Therefore, there should be a risk to avoid classifying search results into too many classes by error.To overcome the limitations of VAP, CAP is proposed.

*C. Classified Average Precision (CAP).*CAP selects the Average Precision of the class that user is interested in (i.e., with the most clicks/votes) and takes the risk of wrong classification into account. The influence of Risk on CAP is adjusted, which can be learned from training data. Finally, we utilize CAP to evaluate the performance of restructuring search results.

Considering another extreme case, if all the URLs in the search results are categorized into one class, Risk will always be the lowest namely 0; however, VAP could be very low. Generally, categorizing search results into less clusters will induce smaller Risk and bigger VAP, and more clusters will result in bigger Risk and smaller VAP. The proposed CAP depends on both of Risk and VAP.

$$CAP = VAP \times (1 - Risk)^{\gamma} \qquad (9)$$

γ is used to adjust the influence of *Risk* on CAP, which can be learned from training data.

$$Risk = \sum_{i,j=1(i<j)}^{m} dij / C_m^2 \qquad (10)$$

It calculates the normalized number of clicked URL pairs that are not in the same class, where *m* is the number of the clicked URLs. If the pair of the *i*th clicked URL and the *j*th clicked URL are not categorized into one class, d_{ij} will be 1; otherwise, it will be 0.

C^2_m = m(m-1)/2 is the total number of the clicked URL pairs. Then, the risk in can be calculated by: Risk = 3/6 = 1/2 = 0.5. The proposed CAP depends on both Risk and VAP.

VII. EXPERIMENTS

To implement the desktop based application, Google API was downloaded and added to the reference library so that application could fetch URLs from Google.

As a first step, user registered by clicking on 'Create New Account', filled-in all the details and thus create a user id and password.

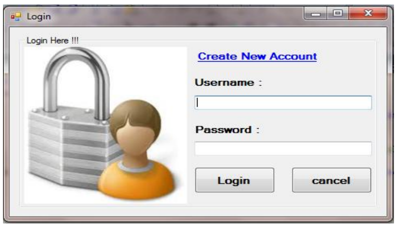
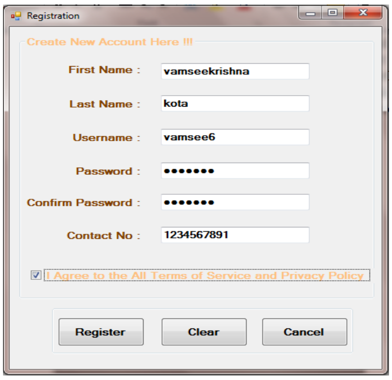


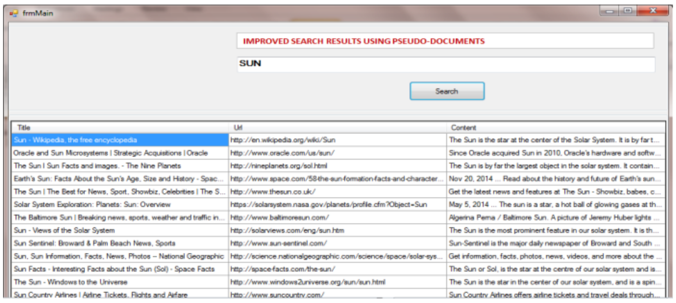
Fig.3: Login page



First Name : vamseekrishna
Last Name : kota
Username : vamsee6
Password : •••••••
Confirm Password : •••••••
Contact No : 1234567891

☑ I Agree to the All Terms of Service and Privacy Policy

Fig.4: Registration page



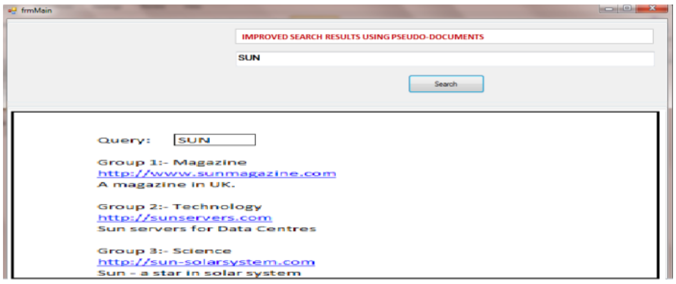
SUN

Title	Url	Content
Sun - Wikipedia, the free encyclopedia	http://en.wikipedia.org/wiki/Sun	The Sun is the star at the center of the Solar System. It is by far t...
Oracle and Sun Microsystems I Strategic Acquisitions I Oracle	http://www.oracle.com/us/sun/	Since Oracle acquired Sun in 2010, Oracle's hardware and softw...
The Sun I Sun Facts and images. - The Nine Planets	http://nineplanets.org/sol.html	The Sun is by far the largest object in the solar system. It contain...
Earth's Sun: Facts About the Sun's Age, Size and History - Spac...	http://www.space.com/58-the-sun-formation-facts-and-character...	Nov 20, 2014 ... Read about the history and future of Earth's sun...
The Sun I The Best for News, Sport, Showbiz, Celebrities I The S...	http://www.thesun.co.uk/	Get the latest news and features at The Sun - Showbiz, babes, c...
Solar System Exploration: Planets: Sun: Overview	https://solarsystem.nasa.gov/planets/profile.cfm?Object=Sun	May 5, 2014 ... The sun is a star, a hot ball of glowing gases at th...
The Baltimore Sun I Breaking news, sports, weather and traffic in...	http://www.baltimoresun.com/	Algerina Perna / Baltimore Sun. A picture of Jeremy Huber lights ...
Sun - Views of the Solar System	http://solarviews.com/eng/sun.htm	The Sun is the most prominent feature in our solar system. It is th...
Sun Sentinel: Broward & Palm Beach News, Sports	http://www.sun-sentinel.com/	Sun-Sentinel is the major daily newspaper of Broward and South ...
Sun, Sun Information, Facts, News, Photos -- National Geographic	http://science.nationalgeographic.com/science/space/solar-sys...	Get information, facts, photos, news, videos, and more about the ...
Sun Facts - Interesting Facts about the Sun (Sol) - Space Facts	http://space-facts.com/the-sun/	The Sun or Sol, is the star at the centre of our solar system and is...
The Sun - Windows to the Universe	http://www.windows2universe.org/sun/sun.html	The Sun is the star in the center of our solar system, and is a spin...
Sun Country Airlines I Airline Tickets, Flights and Airfare	http://www.suncountry.com/	Sun Country Airlines offers airline tickets and travel deals through...

Fig.5: Display of original search results

Table 1: Pseudo-document keywords for the user query 'Sun'

Query	Pseudo-document keywords
sun	magazine, server, solar system



SUN

Query: SUN

Group 1:- Magazine
http://www.sunmagazine.com
A magazine in UK.

Group 2:- Technology
http://sunservers.com
Sun servers for Data Centres

Group 3:- Science
http://sun-solarsystem.com
Sun - a star in solar system

Fig.6: Display of Improved Search Results using Pseudo-documents

As shown in Fig.5, for the first time when user made a search, using Google API, original results as displayed in Google were shown. These URLs were captured as usage logs, feedback session was generated, pseudo document was created and next time when the same user (using same user id and password) logged-in and made the same search, improved/re-ranked search results got displayed as shown in Fig.6.

Considering $\gamma = 0.7$ in (9), let us compare our method with the existing method for 10 random queries. The existing method is from reference paper[20].

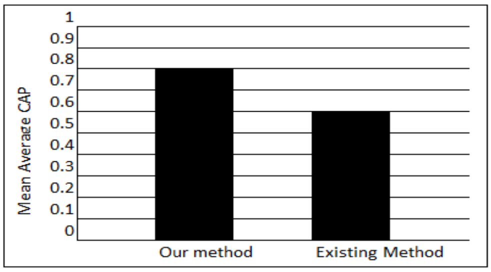


Fig.7: Our method Vs. Existing method

As shown in Fig.7, CAP of our method is better.

VIII. CONCLUSION

In reality, our approach can discover user search goals for some popular queries offline at first. Then, when users submit one of the queries, the search engine can return the results that are categorized into different groups according to user search goals online. Thus, users can find what they want conveniently.

REFERENCES

[1] Zheng Lu, Hongyuan Zha, Xiaokang Yang, Weiyao Lin and Zhaohui Zheng, "*A New Algorithm for Inferring User Search Goals with Feedback Sessions*", IEEE transactions on Knowledge and Data Engineering, Vol.25, no.3, pp. 502-513, 2013.

[2] H. Chen and S. Dumais, "*Bringing Order to the Web: Automatically Categorizing Search Results*," Proc. SIGCHI Conf. Human Factors in Computing Systems (SIGCHI '00), pp. 145-152, 2000.

[3] H.-J Zeng, Q.-C He, Z. Chen, W.-Y Ma, and J. Ma, "Learning to Cluster Web Search Results," Proc. 27th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '04), pp. 210-217, 2004.

[4] X. Wang and C.-X Zhai, "*Learn from Web Search Logs to Organize Search Results*," Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '07), pp. 87-94, 2007.

[5] T. Joachims, "*Optimizing Search Engines Using Clickthrough Data*," Proc. Eighth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (SIGKDD '02), pp. 133-142, 2002.

[6] T. Joachims, L. Granka, B. Pang, H. Hembrooke, and G. Gay, "*Accurately Interpreting Clickthrough Data as Implicit Feedback*," Proc. 28th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '05), pp. 154-161, 2005.

[7] R. Baeza-Yates, C. Hurtado, and M. Mendoza, "*Query Recommendation Using Query Logs in Search Engines*," Proc. Int'l Conf. Current Trends in Database Technology (EDBT '04), pp. 588-596, 2004.

[8] R. Jones and K.L. Klinkner, "*Beyond the Session Timeout: Automatic Hierarchical Segmentation of Search Topics in Query Logs*," Proc. 17th ACM Conf. Information and Knowledge Management (CIKM '08), pp. 699-708, 2008.

[9] R. Jones, B. Rey, O. Madani, and W. Greiner, "*Generating Query Substitutions*," Proc. 15th Int'l Conf. World Wide Web (WWW '06), pp. 387-396, 2006.

[10] J.-R Wen, J.-Y Nie, and H.-J Zhang, "*Clustering User Queries of a Search Engine*," Proc. Tenth Int'l Conf. World Wide Web (WWW '01), pp. 162-168, 2001.