IMPACT: International Journal of Research in
Engineering & Technology (IMPACT: IJRET)
ISSN(E): 2321-8843; ISSN(P): 2347-4599
Vol. 2, Issue 5, May 2014, 301-310
© Impact Journals

AUTOMATIC DERIVATION AND ANNOTATION OF USER SEARCH GOALS WITH FEEDBACK SESSIONS

K. SOUNDARARAJAN[1] & SUMA R[2]

[1]Professor, Department of Computer Science, Mahendra Engineering College, Salem, Tamil Nadu, India

[2]Assistant Professor, Department of Computer Science, St. Joseph's College of Engineering and Technology, Palai, Kerala, India

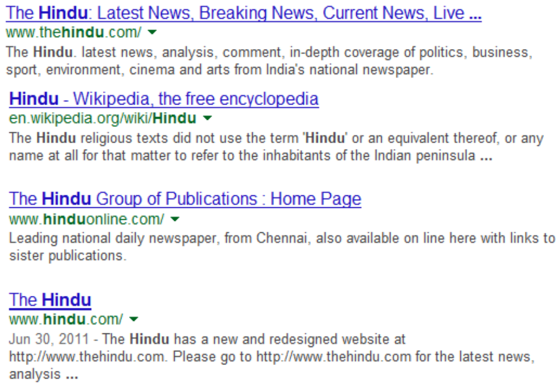
ABSTRACT

The search goals of different users may be different depending up on their need. The analysis and inference of user search goal helps to improve the search engine relevance. In this paper we introduce a method to predict the user search goals of the query that a user give by clustering the existing feedback details. We use Hybrid Bee Algorithm and cluster validity indexes for the effective clustering of the feedback. We also use an automatic annotation approach to align the data units on a result page into different groups such that the data in the same group have same meaning.

KEYWORDS: Hybrid Bee Algorithm, Cluster Validity Indexes, Annotation, User Search Goals

INTRODUCTION

In Web search engines queries are submitted to obtain the needed web pages. But the meaning of a query may be wider or different user may give the same query for different information. For example, a person is giving the query 'Hindu', it may be meant for "The Hindu" news paper or "Banaras Hindu University' or the religion 'Hindu'. An example for the query given in the search engine 'Google' is given below.



The **Hindu**: Latest News, Breaking News, Current News, Live ...
www.the**hindu**.com/

The **Hindu**. latest news, analysis, comment, in-depth coverage of politics, business, sport, environment, cinema and arts from India's national newspaper.

Hindu - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/**Hindu**

The **Hindu** religious texts did not use the term '**Hindu**' or an equivalent thereof, or any name at all for that matter to refer to the inhabitants of the Indian peninsula ...

The **Hindu** Group of Publications : Home Page
www.**hindu**online.com/

Leading national daily newspaper, from Chennai, also available on line here with links to sister publications.

The **Hindu**
www.**hindu**.com/

Jun 30, 2011 - The **Hindu** has a new and redesigned website at http://www.thehindu.com. Please go to http://www.thehindu.com for the latest news, analysis ...

Figure 1: Example for the User Search of Query 'Hindu'

If the necessary page is not available in the staring pages, the user has to search in other pages or by giving some other query words. The user has to spend a great amount of valuable time by giving different form of queries. Hence it is necessary to capture the different user goals to retrieve the exact information which he or she needs. User search goals are the different information that the user wish to obtain for a particular query. Finding out and analysis of the user search goals have a lot of advantages in improving the search engine relevance. The analysis can also be used for re-ranking of the result obtained in a user search in such that the most wanted pages can be displayed initially and then the remaining.

Different researches are going on in the field of query goal analysis and query result re-ranking. In our method we propose an algorithm which does the following steps.

Feedback session only include the URLs, it consist of the clicked URL and Unclicked URL links. Usually language because users will scan the URLs single by single from top to down, we can believe that in addition the three clicked URLs, the four unclicked ones in the rectangular box have also been browsed and evaluated by the user and they should reasonably be a part of the user feedback. Inside the feedback session, the clicked URLs tell what users require and the unclicked URLs reflect what users do not care about. It should be noted that the unclicked URLs after the last clicked URL should not be included into the feedback sessions since it is not certain whether they were scanned or not. Each feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. Therefore, for inferring user search goals, it is more efficient to analyze the feedback sessions than to analyze the search results or clicked URLs directly.

Our system has the following steps:

First we proposed a method to infer user search goals for a query by clustering the similar documents from the web search.

To map feedback sessions from the pseudo documents we proposed a novel optimization method to collect the similar pages of links to satisfy user goals and retrieve the user information.

The K means clustering algorithm can be used to cluster the similar web pages or URL's.

The result of clustering will be more efficient by measuring the semantic similarity between the retrieved results which makes result better than the normal keywords information. Measuring the semantic similarity between the pseudo terms we proposed a correlation based similarity measure between the normal pseudo documents terms. In k means clustering results are difficult and select the cluster centroid values becomes difficult in clustering of the document. Finally Measure the clustering results we use classified average precision (CAP) to evaluate the performance of the restructured web search results. It demonstrates that the evaluation criterion can help us to optimize the parameter in the clustering method when inferring user search goals.

FRAMEWORK OF THE WORK

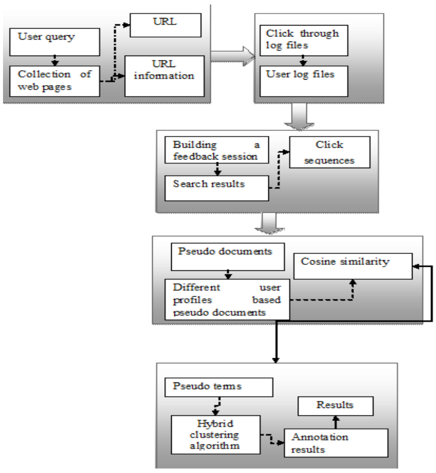


Figure 2: Framework of the Approach

On giving the query by the user, the clicks through logs of previous clicks are analysed. A feedback session is created based on the previous user clicks as shown in diagram.

Feedback sessions are analysed to create the pseudo documents and the pseudo documents are clustered together to find the cosine similarity. The similar documents are clustered to form different groups such that similar documents comes under the same category. Each group is given a common annotation name.

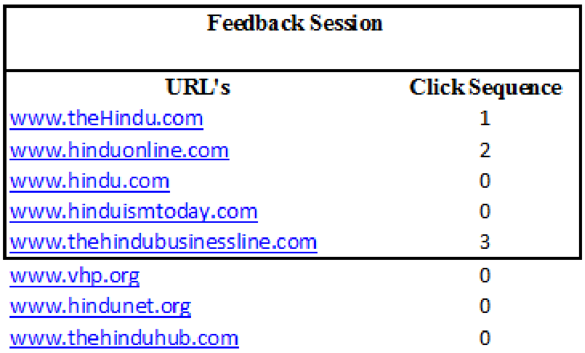
The details of the algorithms and steps are explained in the following sessions.

COLLECTION OF WEB PAGES WITH QUERY

The first phase is the collection of the web pages with similar key terms of the given query. For example, when the user gives a query 'hindu', the search engine collects all the web pages based on query, with link pages clicked by user. All the links and the contents from the link that contains information about the link pages are copied in to the database. This creates a click through log. It gives a feedback about the different user clicks. From the click through log we can analyse which are the links that are clicked by the user for a particular query and which are the links that the user doesn't care about for a particular query. This click through log is the main source of user feedback creation.

FEEDBACK SESSION REPRESENTATIONS

The feedback session represents the sequence of consecutive queries to satisfy a single information requirement. The investigation of the clicked URL's of the click through data, results in inferring user search goals for a particular query. The feedback session is a comprehensive study about the entire session. Consequently the single session which contains simply one query distinguishes from the conservative session. The feedback session contains information about both clicked and unclicked links for a particular query. It shows the order in which the links are clicked for a particular query. If a particular link is not clicked its click sequence is marked as 0, else the corresponding order will be entered. This list will be prepared until the last clicked URL. After the last clicked URL we are not sure whether the users have watched the URL or not. Hence we treat them as not seen by the user.



Feedback Session

URL's	Click Sequence
www.theHindu.com	1
www.hinduonline.com	2
www.hindu.com	0
www.hinduismtoday.com	0
www.thehindubusinessline.com	3
www.vhp.org	0
www.hindunet.org	0
www.thehinduhub.com	0

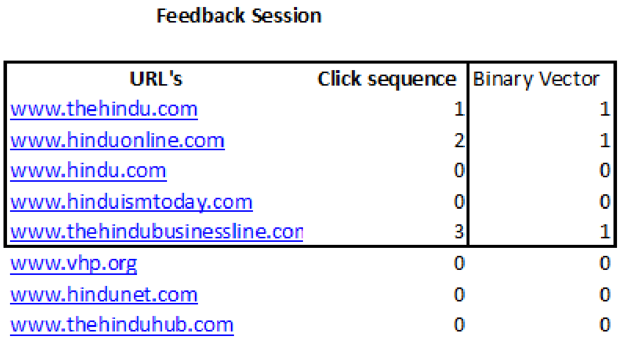
Figure 3

Figure 3 The feedback session. The 0 in the click sequence represent the unclicked URL's and numbers represent their click order. The URL's in the rectangular window shows the feedback session and URL's outside the rectangular window shows the URL's that are not certain whether the user has seen or not.

Each feedback session can tell what a user requires and what he/she does not care about. Moreover, there are plenty of diverse feedback sessions in user click-through logs. We assume that the entire URL's before the last user click are scanned and evaluated by the user. Only the required URL is clicked by the user and the unclicked URL’s are one that is not required by the user. Hence while creating the feedback session all the clicked and unclicked URL’s before the last click are considered. In figure, the list of items which are outside the window are links which are not scanned or clicked by the user. The click sequence of the unclicked URL is given as zero and for the clicked one, the sequence represent then order in which the URL is selected by the user. The feedback session gives a detailed knowledge about what the user needs and what he/she doesn’t care about.

MAPPING FEEDBACK SESSIONS TO PSEUDO DOCUMENT

There is one feedback session for a single query session. Similarly there are a number of feedback session for different query sessions and different click through logs. We have to consider all the feedback session to make the click through logs effectively for predicting the user goals. Some demonstration method is needed to explain feedback sessions in an additional efficient and logical way. There can be a lot of kind of feature representation of feedback sessions. Binary vector technique to characterize a feedback session search consequences are the URLs return by the search engine when the question “the sun” is submit, and “0” represent “unclicked” in the click sequence. The binary vector [0110001] can be second-hand to symbolize the feedback session, where “1” represent “clicked” and “0” represents “unclicked.



Feedback Session

URL's	Click sequence	Binary Vector
www.thehindu.com	1	1
www.hinduonline.com	2	1
www.hindu.com	0	0
www.hinduismtoday.com	0	0
www.thehindubusinessline.con	3	1
www.vhp.org	0	0
www.hindunet.com	0	0
www.thehinduhub.com	0	0

Figure 4

Figure 4: The representation of the binary vector. Binary vector value is ‘1’ for the clicked URL and ‘0’ for the unclicked URL.

BUILDING PSEUDO DOCUMENTS

In the primary step, we augment the URLs with extra textual by extracting the titles and snippets of the returned URLs as appear in the feedback session. Each URL in a feedback session is represented by a little text paragraph that includes its title and snippet. This paragraph is modified by making some alterations like removing the stop words, conversion of all the characters to a particular case etc.

Then, a number of textual processes are implemented to text paragraphs, such as transforming all the letters to lowercases, stemming and removing stop words. Finally, every URL’s title and snippet are generated by a Term Frequency-Inverse Document Frequency vector (TF-IDF, short for **term frequency–inverse document frequency**, is a numerical statistic that reflects how important a word is to a document in a collection or corpus. It is often used as a weighting factor in information retrieval and text mining vector, correspondingly).

$$T_{u_i} = \{T_{W_1}, T_{W_2}, \dots \dots, T_{W_n}\}^T$$

$$S_{u_i} = \{S_{W_1}, S_{W_2}, \dots \dots, S_{W_n}\}^T$$

Where

T_{u_i} - TF-IDF vectors of the URL’s title

S_{u_i} - TF-IDF vectors of the URL’s snippet.

u_i- ith URL in the feedback session.

$W_{j\,=}${1; 2; . . . ; n} –jth term appear in the enriched URLs. Each term in the URL is defined as a word or a numeral in the vocabulary of document collections. T_{wj} and S_{wj} characterize the TF-IDF significance of the jth term in the URL’s title and snippet, correspondingly. Taking into consideration that URLs’ titles and snippets have dissimilar significances, we symbolize the enriched URL by the weighted sum of T_{ui} and S_{ui}, namely,

$$F_{u_i} = T_{u_i}\omega_t + S_{u_i}\omega_s = \{f_{W_1}, f_{W_2}, \dots \dots, f_{W_n}\}^T$$

Where, F_{ui} means the feature representation of the ith URL in the feedback session, and weights ω_t of the titles and ω_s snippets respectively.

FORMING PSEUDO-DOCUMENT BASED ON URL REPRESENTATIONS

In order to obtain the feature demonstration of a feedback session, we suggest an optimization method to merge both clicked and unclicked URLs in the feedback session. Attain such a F_{fs} with the purpose of the calculation of the distance between F_{fs} and each F_{ucm} is minimize and the sum of the distance between F_{fs} and each F_{ucl} is maximize. Based on the supposition that the terms in the vectors are self-governing, we perform optimization on each dimension separately,

$$F_{f_s} = [f_{f_s}(\omega_1), \dots \dots f_{f_s}(\omega_n))]^T$$

Then the similarity between the pseudo-documents are evaluated as the cosine similarity score of

$$Sim_{i,j} = \cos\left(f_{fs_i}, f_{fs_j}\right) = \frac{f_{fs_i} f_{fs_j}}{|f_{fs_i}||f_{fs_j}|}$$

$$Dis_{i,j} = 1 - Sim_{i,j}$$

K MEANS CLUSTERING

In this research we cluster pseudo-documents by K-means clustering which is straightforward and efficient. Because we do not recognizable with the precise figure of user search goal for every query, we position K to be five different values.

$$F_{center_i} = \frac{\sum_{k=1}^{C_i} F_{fs_k}}{C_i}, (F_{fs_k} \subset Cluster\ i)$$

Where $F_{centeri}$ –is the ith cluster’s center and C_i is the numeral of the pseudo-documents in the ith cluster. $F_{centeri}$ is utilized to finish the investigation goal of the ith cluster. Finally, the conditions with the highest values in the $F_{centeri}$ are second hand as the keywords to represent user search goals, it is a keyword based explanation is that the extracted keywords be able to in addition be utilized to form a more significant query in query suggestion and thus can represent user information needs most effectively.

HYBRID K-MEANS AND BEES ALGORITHM

K-mean algorithm requires total number of cluster, *k* beforehand in order the algorithm operates correctly. This pre-requisite value is needed to ensure the algorithm works on the tested data. In this paper, a *test-and-generate* approach is applied to estimate total number present in a data. A hybrid Bees Algorithm and cluster validity index are used for this purpose. The modified Bees algorithm is used to find near optimal cluster centres (centroids) whereas cluster validity index is employed to examine 'goodness' of the generated clusters.

This is done by evaluating one by one possible solution from lower bound until the final boundary. True centroids of each clusters is vital for this proposed approach. The K-Bees algorithm is applied to find near-optimal centroids.

- **Finding Near-Optimal Centroids Using Hybrid K Means and Bees Algorithm**

True centroid is important in this approach. Four different synthetic data sets are used to evaluate K-Bees Algorithms in finding near-optimal centroids. A test is undertaken to evaluate the proposed hybrid in locating a near true centroid. For this purpose an adapted hybrid technique has been applied in this work.

- **Cluster Validity Index**

Validity index generally is targeted to minimise distances of intra-cluster of every object in of their cluster to their nearest centroid. Nonetheless, validity index attempt to maximise of inter-cluster distances between each centroid. Inter-cluster index is calculated using Equ as below:

$$IterCI = \frac{\sum_{i=1}^{k} \left\| c_i - c_{totl} \right\| \cdot n_i}{n \cdot k}$$

where *i c* is centroids of cluster *i*, *totl c* is mean total centroids of all centroids in the data, *i n* is total number of object in cluster *i*, *n* is the total number of object in the data, *k* total number of clusters.

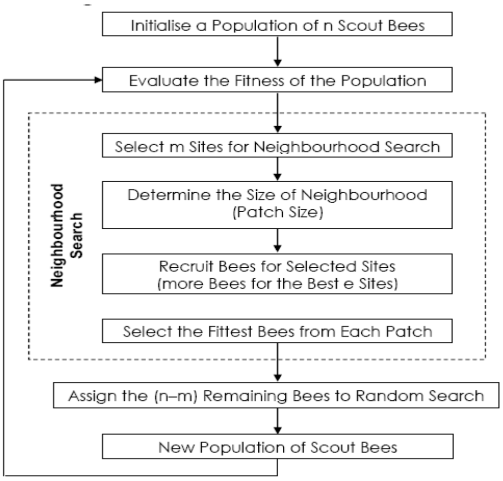

Initialise a Population of n Scout Bees
Evaluate the Fitness of the Population
Neighbourhood Search
Select m Sites for Neighbourhood Search
Determine the Size of Neighbourhood (Patch Size)
Recruit Bees for Selected Sites (more Bees for the Best e Sites)
Select the Fittest Bees from Each Patch
Assign the (n–m) Remaining Bees to Random Search
New Population of Scout Bees

Figure 5: The Working of Bees Algorithm

ANNOTATION OF RESULTS

Data alignment is to put the data units of the same concept into one group so that they can be annotated holistically. Whether two data units belong to the same concept is determined by how similar they are based on the features.

Presentation Style Similarity (SimP)

It is the average of the style feature scores (FS) over all six presentation style features (F) between d_1 and d_2

$$SimP(d_1, d_2) = \sum_{i=1}^{6} FS_i / 6$$

Data Type Similarity (SimD)

It is determined by the common sequence of the component data types between two data units. The longest common sequence (LCS) cannot be longer than the number of component data types in these two data units. Thus, let t_1 and t_2 be the sequences of the data types of d_1 and d_2, respectively, and TLen(t) represent the number of component types of data type t, the data type similarity between data units d_1 and d_2 is

$$SimD(d_1, d_2) = \frac{LCS(t_1, t_2)}{Max(Tlen(t_1), Tlen(t_2))}$$

Adjacency Similarity (SimA)

The adjacency similarity between two data units d_1 and d_2 is the average of the similarity between d_p^1 and d_p^2 and the similarity between d_s^1 and d_s^2 that is

$$SimA(d_1, d_2) = (Sim'(d_1^p, d_2^p) + Sim'(d_1^s, d_2^s))/2$$

PERFORMANCE EVOLUTION

The evaluation of user search goal inference is a major problem, since user search goals are not predetermined and there is no ground truth. It is necessary to develop a metric to evaluate the performance of user search goal inference objectively. In this module finally measure the performance of the hybrid kmeans + annotation and existing pseudo-documents based clustering Measure the performance of the system with parameters like Classified Average Precision (CAP), Voted AP (VAP) which is the AP of the class including more clicks namely, risk to avoid classifying search results and average precision (AP).

Average Precision (AP)

In order to be appropriate the assessment method to large-scale data, the solitary sessions in user click-through logs are second-hand to reduce physical work. Since beginning user click-through logs, we can get implied significance feedbacks, specifically “clicked” means applicable and “unclicked” means inappropriate. A probable evaluation principle is the average precision (AP) which evaluate according to user implicit feedbacks. AP is the average of precisions compute at the position of each applicable document in the ranked sequence

$$AP = \frac{1}{N^+}\sum_{r=1}^{N} rel(r)\frac{R_r}{r}$$

where N^+ is the numeral of applicable (or clicked) documents in the retrieved ones, r is the rank, N is the total numeral 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.

Classified Average Precision (CAP)

Extend VAP by introducing the above Risk and propose a new criterion Classified AP (CAP)

$CAP = VAP * (1\text{-}risk)^{\gamma}$

Where γ is used to adjust the influence of Risk on CAP.

Risk

VAP is still an unsatisfactory criterion. Taking into consideration an extreme case, if every URL in the click session is categorized into one class, VAP will forever be the highest value that is 1 no matter whether user contain so many investigate goals or not. Consequently present be supposed to be a risk to avoid classify exploration results into too many classes by error. They propose the risk as follows:

$$Risk = \frac{\sum_{i,j=1(i<j)}^{m} d_{ij}}{C_m^2}$$

CONCLUSIONS

The Novel approach has been proposed to infer user search goals for a query by clustering its feedback sessions represented by pseudo-documents. First, we introduce feedback sessions to be analyzed to infer user search goals rather than search results or clicked URLs. Both the clicked URLs and the unclicked ones before the last click are considered as user implicit feedbacks and taken into account to construct feedback sessions. Therefore, feedback sessions can reflect user information needs more efficiently. Second, we map feedback sessions to pseudo- documents to approximate goal texts in user minds. The pseudo-documents can enrich the URLs with additional textual contents including the titles and snippets. In this work the pseudo documents are clustered based on the hybrid k means clustering method. Based on these pseudo-documents, user search goals can then be discovered and depicted with some keywords. Finally, a new criterion CAP is formulated to evaluate the performance of user search goal inference. At last we are annotating the clustering search result. From the experiment the proposed system is improves the performance of the system.

ACKNOWLEDGEMENTS

The heading of the Acknowledgment section and the References section must not be numbered. Causal Productions wishes to acknowledge Michael Shell and other contributors for developing and maintaining the IEEE LaTeX style files which have been used in the preparation of this template. To see the list of contributors, please refer to the top of file IEEETran.cls in the IEEE LaTeX distribution.

REFERENCES

1. 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.
2. 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.
3. 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.
4. 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.
5. U. Lee, Z. Liu, and J. Cho, “Automatic Identification of User Goals in Web Search,” Proc. 14th Int’l Conf. World Wide Web (WWW ’05), pp. 391-400, 2005.
6. B. Poblete and B.-Y Ricardo, “Query-Sets: Using Implicit Feedback and Query Patterns to Organize Web Documents,” Proc. 17th Int’l Conf. World Wide Web (WWW ’08), pp. 41-50, 2008.
7. 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.
8. 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.
9. D. Shen, J. Sun, Q. Yang, and Z. Chen, “Building Bridges for Web Query Classification,” Proc. 29th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’06), pp. 131-138, 2006.
10. S. Beitzel, E. Jensen, A. Chowdhury, and O. Frieder, “Varying Approaches to Topical Web Query Classification,” Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development (SIGIR ’07), pp. 783-784, 2007.
11. C.-K Huang, L.-F Chien, and Y.-J Oyang, “Relevant Term Suggestion in Interactive Web Search Based on Contextual Information in Query Session Logs,” J. Am. Soc. for Information Science and Technology, vol. 54, no. 7, pp. 638-649, 2003.
12. T. Joachims, “Evaluating Retrieval Performance Using Clickthrough Data,” Text Mining, J. Franke, G. Nakhaeizadeh, and I. Renz, eds., pp. 79-96, Physica/Springer Verlag, 2003.
13. 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.
14. 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.
15. M. Pasca and B.-V Durme, “What You Seek Is what You Get: Extraction of Class Attributes from Query Logs,” Proc. 20th Int’l Joint Conf. Artificial Intelligence (IJCAI ’07), pp. 2832-2837, 2007.

16. R. Baeza-Yates and B. Ribeiro-Neto, Modern Information Retrieval. ACM Press, 1999.

17. X. Li, Y.-Y Wang, and A. Acero, “Learning Query Intent from Regularized Click Graphs,” Proc. 31st Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’08), pp. 339-346, 2008.

18. D. Beeferman and A. Berger, “Agglomerative Clustering of a Search Engine Query Log,” Proc. Sixth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’00), pp. 407-416, 2000.

19. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, “Context-Aware Query Suggestion by Mining Click-Through,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (SIGKDD ’08), pp. 875-883, 2008.

20. 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.Std. 802.11, 1997.