ISSN: 2321-7782 (Online)
Volume 3, Issue 4, April 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Web Graph of Queries for Recommendation

Rupesh Patil[1]
Department Of Computer Engineering
Padmabhooshan Vasantdada Patil Institute of Technology
Pune, India

Gaurang Diwakar[2]
Department Of Computer Engineering
Padmabhooshan Vasantdada Patil Institute of Technology
Pune, India

Abstract: ***Now a days searching is the biggest thing in our Internet Life. Web Graph of Queries for Recommendation is the exponential extension of different contents over the internet or Web; Recommendation techniques have become absolutely necessary and increasing widely over the internet. Countless different kinds of suggestions or recommendations are made on the Web, for e.g. URL's, mails, maps, locations, query suggestions tags recommendations, movies, songs, wallpapers(pictures), magazines, videos, etc. So it does not matter what kind of data sources are used for the recommendations, so these data sources can be made in the form of different types of graphs generated in background. So in that paper our aim is to providing a general framework on Web graph of Queries for recommendation, 1. We first take the different queries from the user on that basis propose a novel diffusion method which generates similarities between different nodes and then generates recommendations. 2. After that we are going to show how to generalize different recommendation queries or problems into our graph diffusion framework. The advanced framework can be utilized in so many recommendation tasks on the Internet or World Wide Web, that includes URL's, videos, image recommendations, query suggestions, image annotations, expert finding, tag recommendations, etc. The experimental analysis on the small as well as large data sets shows the good future of our work.***

Keywords: Platform Java; Data Mining, Searching, Easy to access Information, Web Recommendations, Heat Diffusion, Query suggestions, Collaborative Filtering.

I. INTRODUCTION

The diverse and explosive growth of internet information for people, to organize as well as utilize the information effectively and efficiently becomes more and more difficult. So this is important for internet related applications since user-generated information is more freestyle and less structured, which increases the difficulty in mining useful information from these data sources. In order to satisfy the information needs of Web users and improve the user experience in many internet applications, Recommend System, have been well studied in all and widely spread in companies. So first one is the ambiguity which commonly exists in the language. The Queries containing ambiguous terms may confuse the algorithms which do not satisfy the information needs of user or employees. Then Another consideration is that, as reported is that users has to submit short queries consisting of only one or more terms under most circumstances, and short queries are more likely to be ambiguous. The analysis of a commercial search engine's query logs recorded over three months, we observe that 25 percent of Web queries are single term queries, and further 40.3 percent of Web queries contain only two to three terms.

Which is a technique that automatically predicts the interest of an active end user to collecting rating or frequencies information from other user? Underlying assumption of collaborative filtering that the user will prefer items which other similar users use in their day to day life. So based on this simple but effective intuition, the collaborative filtering has been mostly in some large, well-known commercial systems, including product recommendation at flip kart, wallpaper recommendation at Instagram. Collaborative filtering algorithm it requires a user item rating (frequency) matrix which shows user specific rating

preference to other user's characteristics. However, in most of the situations, rating data are always unavailable since information on the internet is less structured.

II. MOTIVATION

For the challenge of the searching method last challenge is that it is time consuming and inefficient to design different recommendation algorithm for different recommendation activities. Most of these recommendation problems have some common features, where a general framework is needed to unify the recommendation tasks on the Web. Moreover, most of existing methods are complicated and require tuning a large number of parameters. Aiming at solving the problems analyzed above, we propose a general framework for the recommendations on the Internet. This framework is built upon the heat diffusion on both undirected graphs and directed graphs, and has several advantages. It is a general method, which can be utilized to many recommendation tasks on the Web. It can provide latent semantically relevant results to the original information need. This model provides a natural treatment for personalized recommendations. The designed recommendation algorithm is scalable to very large data sets. The empirical analysis on several large scale data sets AOL through data and Instagram images tag data shows that our framework is effective and useful for generating high quality recommendation.

III. QUERY SUGGESTION AND COLLABORATIVE FILTERING

1) Query Suggestion

So this work is based on the observation that Internet queries and texts are highly similar with the word. Now these methods different kinds of data sources documents, texts, queries, URL's, different logs, for suggesting queries. Most of these methods are only designed for query suggestions; the extensibility of these methods is very limited. Then two query recommendation methods is based on click through data is proposed. The main disadvantage of these algorithms is that they ignore rich information embedded in the query bipartite graph, and consider only queries that appear in the query logs, potentially losing the opportunity to recommend highly semantically related queries to user or users. In local i.e. query dependent documents and global i.e. the whole corpus documents are employed in query expansion by applying the measure of global analysis to the selection of query terms in local feedback. Although experimental results show that this method is generally more effective than global analysis, it performs worse than the query expansion method proposed in based on user interactions recorded in user logs.

A goal of query suggestion is similar to that of query expansion query substitution, and query refinement, which all focus on understand the user search intentions and improving the queries submitted by users. The query suggestion is closely related to query expansion or query substitution, which extends the original query with new search terms to narrow down the scope of the search. But different from query expansion, query suggestion aims to suggest full queries that have been formulated by previous users so that query integrity and coherence are preserved in the suggested queries. Query refinement is another closely related notion; the objective of query refinement is interactively recommended the new queries related to a particular query.

2) Collaborative Filtering

Collaborative filter method in that model based approaches; training data sets are used to train a predefined model. E.g. of model based approach include the clustering model, the aspect models and the latent factor model. Kohrs and Merialdo presented an algorithm for collaborative filtering based on hierarchical clustering, which tried to balance robustness and accuracy of predictions, especially when few data were available. Hofmann proposed an algorithm based on a generalization of probabilistic latent semantic analysis to continuous-valued response variables. Recently, several matrix factorization methods have been proposed for collaborative filtering. These methods all focus on fitting the user-item rating matrix using low-rank approximations, and use it to make further predictions. The premise behind a low-dimensional factor model is that there are only a small number of factors influencing preferences, and that a user's preference vector is determined by how each factor applies to the different users. Neighborhood based approaches are the most popular prediction methods and are widely adopted

in commercial collaborative filtering systems. The most analyzed examples of neighborhood-based collaborative filtering include user-based approaches and item-based approaches. User-based approaches predict the ratings of active users based on the ratings of their similar users, and item-based approaches predict the ratings of active users based on the computed information of items similar to those chosen by the active user. User-based and item-based approaches often use the Pearson Correlation Coefficient algorithm and the Vector Space Similarity algorithm as the similarity computation methods.

IV. GRAPHS DIFFUSION

In that we are going to discuss all the part of the diffusion matrix on the graphs as well as heat diffusion methods. This model can be applied to both undirected graphs and directed graphs and at last we analyze the complexity of the models.

1. Random Jump

In the click through data, people of different cultures, genders, ages, and environments, may implicitly link queries together, but we do not know these latent relations. Another good example is the trust relations in a social network. On online social network sites, users always explicitly state the trust relations to other users. Actually, there are some other implicit hidden trust relations among these users that cannot be observed. Hence, to capture these relations, we propose to add a uniform random relation among different nodes. The heat can only propagate through the links that connect nodes in a given graph, but in fact, there are random relations among different nodes even if these nodes are not connected.

2. Directed and undirected Graphs Diffusion

In mathematics, and more specifically in graph theory, a graph is a representation of a set of objects where some pairs of objects are connected by links. The interconnected objects are represented by mathematical abstractions called vertices, and the links that connect some pairs of vertices are called edges. Typically, a graph is depicted in diagrammatic form as a set of dots for the vertices, joined by lines or curves for the edges. Graphs are one of the objects of study in discrete mathematics. The vertices belonging to an edge are called the ends, endpoints, or end vertices of the edge. A vertex may exist in a graph and not belong to an edge. V and E are usually taken to be finite, and many of the well-known results are not true or are rather different for infinite graphs because many of the arguments fail in the infinite case. The order of a graph is the number of vertices. A graph's size is the number of edges. The degree of a vertex is the number of edges that connect to it, where an edge that connects to the vertex at both ends a loop is counted twice.

$$f_i(t+\Delta t) - f_i(t) = \sum_{j:(j,i)\in E} \gamma(f_j(t) - f_i(t))\Delta t,$$

where E is the set of edges. To find a closed form solution to Eq. (2), we express it in a matrix form: $(\mathbf{f}(t+\Delta t) - \mathbf{f}(t))/\Delta t = \gamma \mathbf{H}\mathbf{f}(t)$, where $d(v)$ denotes the degree of the node v. In the limit $\Delta t \to 0$, it becomes $\frac{d}{dt}\mathbf{f}(t) = \gamma \mathbf{H}\mathbf{f}(t)$. Solving it, we obtain $\mathbf{f}(t) = e^{\gamma t \mathbf{H}}\mathbf{f}(0)$, especially we have

$$\mathbf{f}(1) = e^{\gamma \mathbf{H}}\mathbf{f}(0), H_{ij} = \begin{cases} -d(v_j), & j = i, \\ 1, & (v_j, v_i) \in E, \\ 0, & \text{otherwise}, \end{cases}$$

where $e^{\gamma H}$ is defined as $e^{\gamma \mathbf{H}} = \mathbf{I} + \gamma \mathbf{H} + \frac{\gamma^2}{2!}\mathbf{H}^2 + \frac{\gamma^3}{3!}\mathbf{H}^3 + \cdots$

$t+\Delta t$ and time t will be equal to the sum of the heat that it receives, deducted by what it diffuses. This is formulated as $f_i(t+\Delta t) - f_i(t) = -\gamma f_i(t)\Delta t + \sum_{j:(v_j,v_i)\in E} \gamma/d_j f_j(t)\Delta t$. Similarly, we obtain

$$\mathbf{f}(1) = e^{\gamma \mathbf{H}}\mathbf{f}(0), H_{ij} = \begin{cases} -1, & j = i, \\ 1/d_j, & (v_j, v_i) \in E, \\ 0, & \text{otherwise}. \end{cases}$$

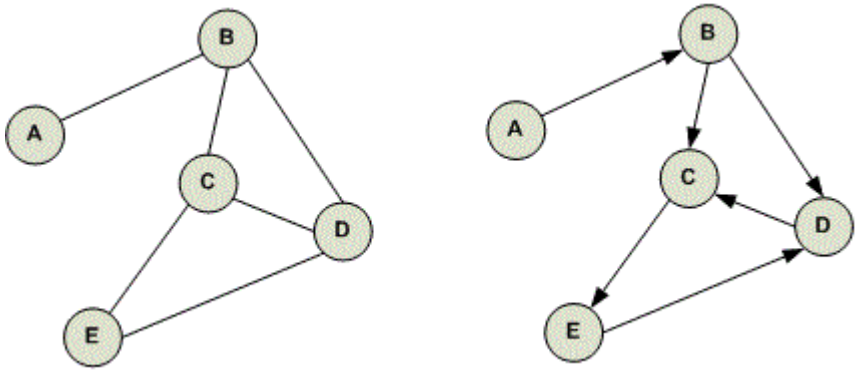

Fig. 1 Undirected graph and Directed Graph

3. Heat Diffusion

So in this paper, we use heat diffusion to model the similarity information propagation on Web graphs. In Physics, the heat diffusion is always performed on a geometric manifold with initial conditions. However, it is very difficult to represent the Web as a regular geometry with dimension. This motivates us to investigate the heat flow on a graph. The graph is considered as an approximation to the underlying manifold, and so the heat flow on the graph is considered as the heat flow on the manifold. Heat diffusion is a physical phenomenon. In a medium, heat always flows from a position with high temperature to a position with low temperature. Based approaches have been successfully applied in various domains such as classification and dimensionality reduction problems. Lafferty and Lebanon approximated the heat kernel for a multinomial family in a closed form, from which great improvements were obtained over the use of Gaussian or linear kernels. Kondor and Lafferty proposed the use of a discrete diffusion kernel for categorical data, and showed that the simple diffusion kernel on the hypercube can result in good performance for such data. Belkin and Niyogi employed a heat kernel to construct the weight of a neighborhood graph, and apply it to a nonlinear dimensionality reduction algorithm. Yang et al. proposed a ranking algorithm known as the Diffusion Rank using heat diffusion process; simulations showed that it is very robust to Internet threaten.

V. ACTUAL ANALYSIS

1. Process Summary

(a) Build the database from the historical click through dataset of the web which contains different query and the clicked links.

(b) Constructing the Query-URL Bipartite graph.

(c) Heat diffusion to model the similarity information propagation on Web graphs. Diffusion on directed Graph.

(d) Calculating Heat Diffusion for each query.

(e) Showing suggestions based on the heat values.

(f) Take dataset of social networking, in which queries are also related to other queries.

(g) Generate the Bipartite graph of social networking dataset. And follow steps 3,4,5 for recommendations.

2. Data Collection

This data set is the raw data recorded by the search engine, and contains a lot of noise which will potentially affect the effectiveness of our query suggestion algorithm. Hence, we conduct a similar method employed in [59] to clean up the raw data. We filter the data by only keeping those frequent, well formatted, English queries which only contain characters a to z. After cleaning and removing duplicates, we get totally 18, 30,277 unique queries and 7, 05,482 unique URLs in our data collection. After the construction of the query-URL bipartite graph using this data collection procedure, we observe that a total of

40,29,100 edges exist in the query-URL bipartite graph, which indicates that, on average, each query has 5.19 distinct clicks, and each URL is clicked by 6.60 distinct queries. So we are constructing our query suggestion graph based on the click through data of the AOL search engine. In total, this data set spans 3 months from 01 March, 2006 to 31 May, 2006. There are a total of 21, 57,492 lines of click through information, 4,802,520 unique queries, and 1,606,326 unique. The data record the activities of Web users, which reflect their interests and the latent semantic relationships between users and queries as well as queries and clicked Web documents. Each line of click through data contains the following information: a user ID, a query issued by the different users, a URL on which the user clicked, the rank of that URL, and the time at which the query was submitted for search. From a statistical point of view, the query word set corresponding to a number of Web pages contains human knowledge on how the pages are related to their issued queries. Thus, so in this paper, we utilize the relationships of queries and Web pages for the construction of the bipartite graph containing two types of vertices.

3. Construction of Graph

In order to compare our method with other approaches, we create a set of 200 queries as the testing queries, covering a wide range of topics, such as Computers, Arts, Business, and others. We cannot simply employ the bipartite graph extracted from the click through data into the diffusion processes since this bipartite graph is an undirected graph, and cannot accurately interpret the relationships between queries and URLs. Hence, we convert this bipartite graph into Matrix form. So in this converted graph, every undirected edge in the original bipartite graph is converted into two directed edges. The weight on a directed query-URL edge is normalized by the number of times that the query is issued, while the weight on a directed URL-query edge is normalized by the number of times that the URL is clicked.

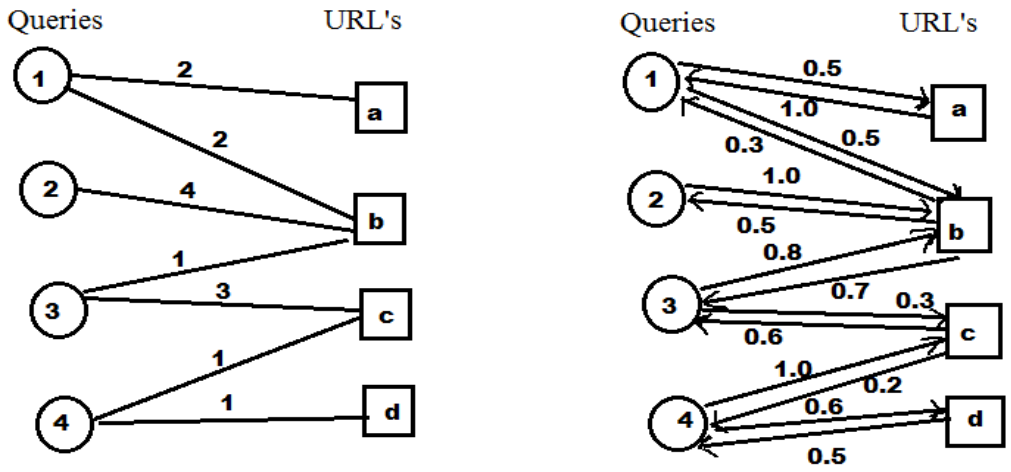

Fig. 2 Graph construction for query suggestion. (a) Query-URL bipartite graph. (b) Converted query-URL bipartite graph

So now we cannot simply employ the bipartite graph extract from the click through data into the diffusion process, this bipartite graph is an undirected graph, so it cannot accurately interpret the relationships between queries and URLs. From this we convert this bipartite graph into other. Converted graph, every undirected edge in the original bipartite graph is converted into two directed edges. The weight on a directed query-URL edge is normalized by the number of times that the query is issued, while the weight on a directed URL-query edge is normalized by the number of times that the URL is clicked by the users or different people on the Internet. The designed recommendation algorithm is scalable to very large datasets

4. Query Suggestion Algorithm

The algorithm is stated that query suggestion nothing but the taking the query from the user or the people all over the world who are working over the Internet, and then again passes that query to the machine for the further processing. A converted

bipartite graph G consists of query set V and URL set V. The two directed edges are weighted using the method introduced. Given a query q in V, a subgraph is constructed by using depth first search in G. The search stops when the number of queries is larger than a predefined number. Then third step is to analyzed above, set 1, and without loss of generality, set the initial heat value of query the choice of initial heat value will not affect the suggestion results. Start the diffusion process. Then the last part is output of the top n number of queries are with the largest values in the vector as the suggestion for the given queries.

(a) A converted bipartite graph

G= (V+U V*, E) set consist of query V+ and URL set V*.

(b) Given a query in V+, a sub-graph is constructed by using depth-first search. The search stops when the number of queries is larger than a predefined number.

(c) As analyzed above, =1, and without loss of generality, set the initial heat value of query q f (0) =1 (the choice of initial heat value will not affect the suggestion results.). Start the diffusion process using f (1) =e f (0).

(d) Output the Top-K queries with the largest value in vector f (1) as the suggestions.

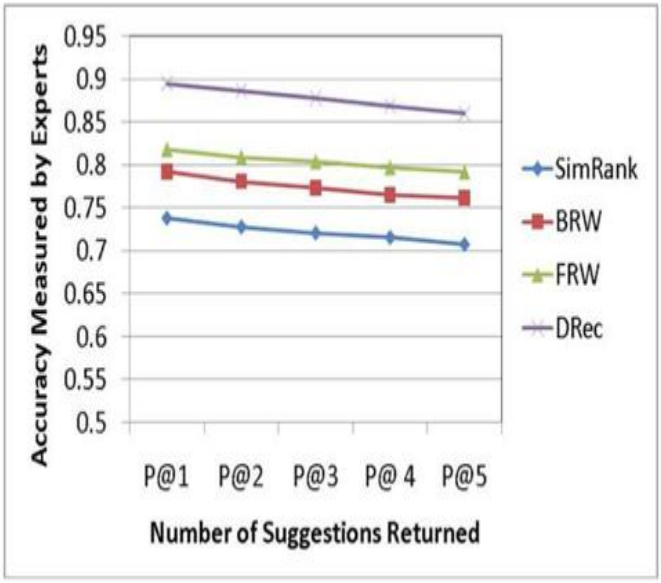

Fig. 3 Comparison is accurate by professionals

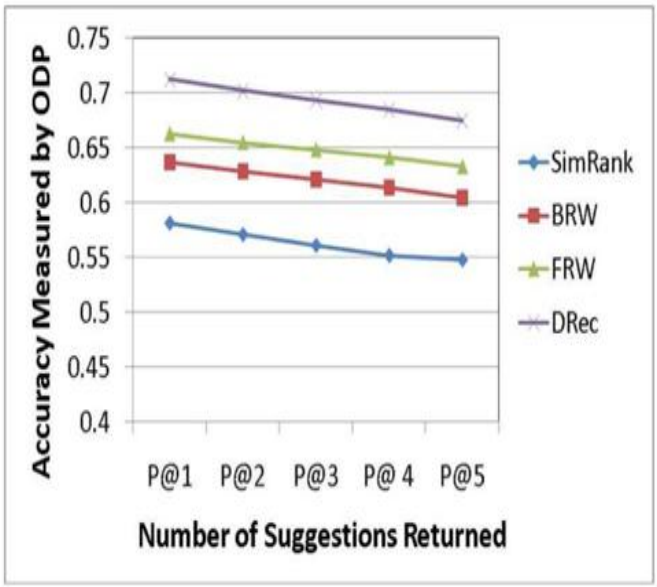

Fig. 4 Comparison is accurate by ODP

5. *Advantages*

(a) It is a general method, which can be utilized to many recommendation tasks on the Web.

(b) It can provide latent semantically relevant results to the original information need.

(c) This model provides a natural treatment for personalized recommendations.

(d) The designed recommendation algorithm is scalable to very large datasets.

(e) System works for social networking dataset also.

6. *State Chart Diagram*

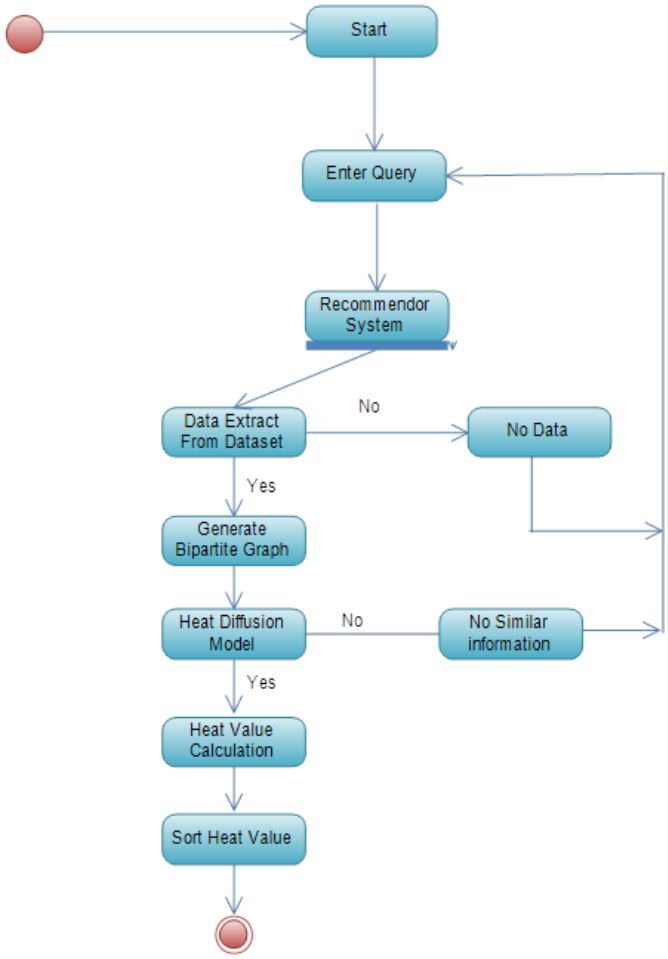

Fig. 5 Flow of Recommendation System

VI. CONCLUSION

So in this paper our work is to, we are presenting the novel framework for different recommendation on large scale Internet or Web graphs using heat diffusion method. So this is a general framework which can basically be adapted to the most of the Internet Search or Web graphs for the recommendation tasks all over the internet, such as query suggestions, URL's recommendations, personal recommendation for user or many more, etc. So the generated suggestions are basically related to the inputs provided by the users over the search engine on different locations. This experimental analysis on large scale of Web data sources shows the promising future of this kind of work.

ACKNOWLEDGMENT

The First and foremost, We would like to thank my Seminar guide, Prof. A.D. Bhosale V.A, for his guidance and support. I will forever remain grateful for the constant support and guidance extended by my guide, in makinh this work. Through our many discussions, he helped me to form and solidify ideas. The invaluable discussions, we had with him, the

penetrating questions he has put to us and the constant motivation, has all led to the development of the ideas presented in this work. With a deep sense of gratitude, we wish to express our sincere thanks to the, Prof. A.D. Bhosale for his immense help in planning and executing the works in time. Also, I would like to thank my parents for their continual encouragement and the positive support.

I would also like to thank my wonderful colleagues and friends for listening my ideas, asking questions and providing feedback and suggestions for improving my ideas distribution.

References

1. E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," SIGIR '07: Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 19-26, 2006.
2. E. Auchard, "Flickr to Map the World's Latest Photo Hotspots," Proc. Reuters, 2007.
3. R. TiberiBaeza-Yates and A. Tiberi, "Extracting Semantic Relations from Query Logs," KDD '07: Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 76-85, 2007.
4. R.A. Baeza-Yates, C.A. Hurtado, and M. Mendoza, "Query Recommendation Using Query Logs in Search Engines," Proc. Current Trends in Database Technology (EDBT) Workshops, pp. 588-596, 2004.
5. D. Beeferman and A. Berger, "Agglomerative Clustering of a Search Engine Query Log," KDD '00: Proc. Sixth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 407-416, 2000.
6. M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimension-ality Reduction and Data Representation," Neural Computation, vol. 15, no. 6, pp. 1373-1396, 2003.
7. J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence (UAI), 1998.
8. S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," Computer Networks and ISDN Systems, vol. 30, nos. 1-7, pp. 107-117, 1998.
9. J. Canny, "Collaborative Filtering with Privacy via Factor Analysis," SIGIR '07: Proc. 25th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 238-245, 2002.
10. H. Cao, D. Jiang, J. Pei, Q. He, Z. Liao, E. Chen, and H. Li, "Context-Aware Query Suggestion by Mining Click-Through and Session Data," KDD '08: Proc. 14th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, pp. 875-883, 2008.
11. P.A. Chirita, C.S. Firan, and W. Nejdl, "Personalized Query Expansion for the Web," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 7-14, 2007.
12. N. Craswell and M. Szummer, "Random Walks on the Click Graph," SIGIR '07: Proc. 30th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 239-246, 2007.

AUTHOR(S) PROFILE

Rupesh Sunil Patil, Pursuing Final Year Of Computer Engineering at Padmabhooshan Vasantdada Patil Institute Of Technology, Bavdhan ,Pune , Maharashtra , India affiliated by University of Pune. Completed Diploma in Computer Engineering at MAEER's MIT SSPP in 2008 to 2012.

Gaurang Krishna Diwakar, Pursuing Final Year Of Computer Engineering at Padmabhooshan Vasantdada Patil Institute Of Technology, Bavdhan ,Pune , Maharashtra , India affiliated by University of Pune. Completed Diploma in Computer Engineering at MAEER's MIT SSPP in 2008 to 2012.

