International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.6, No.5 (2013), pp.401-410
http://dx.doi.org/10.14257/ijsip.2013.6.5.35

A New Algorithm of Web Queries Clustering Using User Feedback[1]

Lingling Meng[1], Runqing Huang[2] and Junzhong Gu[3]

[1]Computer Science and Technology Department, Department of Educational Information Technology, East China Normal University, Shanghai, 200062, China
[2]Shanghai Municipal People's Government, Shanghai, 200003, China
[3]Computer Science and Technology Department, East China Normal University, Shanghai, 200062, China
[1]llmeng@deit.ecnu.edu.cn, [2]runqinghuang@gmail.com, [3]jzgu@ica.stc.sh.cn

Abstract

Web queries clustering attract great concern recently. It is crucial for capturing frequently asked questions in question-answering system, most popular topics in search engine or automatic query expansion in information retrieval. The paper presents a new algorithm of web queries clustering using user click information in the query logs and applies it into query expansion. Different from previous work, in the new algorithm both word form and semantic information have been taken into considered. Experiments show that the new algorithm works effectively.

Keywords: *web queries clustering, user feedback, word form, semantic information*

1. Introduction

Nowadays with the development of science and technology, more and more information are available to us in the Web. At the same time it is growing that people use search engine to get information. In the search result list each record represents an entrance, providing the title, URL, summary and so on. Users can judge whether the record contains the contents they are interested. Therefore, the user log denotes an implicit relevance feedback. Recently years it attracts great concern. Many researchers analyses the similarity of queries and apply it in user query expansion [1-3], query recommendations [4, 5], document clustering [6], document classification [7], question answer system [8]. It shows its talents and makes these applications more effective. This paper proposes a new query clustering algorithm using user feedback of query logs to discover similar topics and applies it in query recommendation. In the new algorithm both word form and semantic information have been taken into considered. Experiments show that the new algorithm works effectively.

The rest of this paper is as follows: in Section 2 related works are presented. A new algorithm of web queries clustering is proposed in Section 3. Section 4 shows the evaluation of the new algorithm, including experiments, data analyzing, and the achievements. Conclusion and future Work is described in Section 5.

[1] The work in the paper was supported by Shanghai Scientific Development Foundation (Grant No.11530700300)

ISSN: 2005-4254 IJSIP

2. Related Works

Web queries clustering are relatively new research field. The queries in the same cluster indicate the same or similar topics. It is the key issue that how to measure the similarity of queries. Some algorithms have been proposed.

Ricardo Baeza-Yates first built a term-weight vector for each query [5]. Each term was weighted according to the number of occurrences and the number of clicks of the documents in which the term appears. He measured the similarity of two queries as the similarity of their trace vectors using the cosine function.

Befferman and Berger also proposed a query clustering technique based on common clicked URLs [6]. It viewed the user query logs as a bipartite graph, with the vertices on one side corresponding to queries and on the other side to URLs. One can apply an agglomerative clustering algorithm to the graph's vertices to identify related queries and URLs.

Ji-Rong and Jian-Yun presented a similar queries clustering algorithm to recommend URLs to frequently asked queries of a search engine [8]. It assumed that:

(1)If two queries contained the same or similar terms, they denoted the same or similar information needs.

(2)Two queries were similar if they leaded to the selection of the same or similar document.

The function of similar queries was defined by combining both assumptions linearly.

Bruno M. Fonseca used association rule to measure the similarity of queries [9].In his research it took query as item and session as a set of transactions of association rule mining. Each transaction represented a session in which a single user submits a sequence of related queries in a time interval.

In Ji-Min Wang's research, a new method for discovering related Web queries was presented [10]. First, some statistical characteristics of a candidate query for a given query were extracted from the log files, such as the numbers of different users submitted, the numbers of the candidate query submitted as well as the returned result clicked, the numbers of common terms and common URLs clicked between the candidate query and the given query. Then these candidate queries were ranked with a linear regression model learned from human labeled training data.

Zaiane and Strilets [11] proposed a method to recommend queries based on seven different notions of query similarity. The method was intended for a meta-search engine. It not only took the keywords, phrases of the query or common clicked URLs into considered, but also took the content and title of the URL's in the result of a query into considered. However none of their similarity measures considered user preferences in form of clicks stored in query logs.

3. A New Algorithm of Web Queries Clustering

3.1. Similar Queries Metric

All the measures above are simple and effective. However, they all focus on the level of syntax, and ignored semantic information. In this section, a new query clustering algorithm will be presented. Related definitions are as follows:

Firstly, according to TF-IDF, feature terms of user clicked documents are extracted. It is commonly argued that language semantics are mostly captured by nouns or noun phrases so that the study only focus on noun.

Then the feature terms of user clicked documents with different queries are compared. Next the new algorithm is based on the following assumptions.

(1) If two clicked documents contain the same feature terms, they convey the same or similar information needs and two documents are similar. The more terms in common, the more similar they are.

(2) If two clicked documents don't contain the same feature terms, however the feature terms of the two documents are semantic associated, and the similarity value is greater than a certain threshold, the two documents are similar.

Based on the assumption, the similarity of two documents is defined as follows:

$$\begin{cases} sim\ (d_i, d_j) \\ = \dfrac{N}{\max(N_1, N_2)} + (1 - \dfrac{N}{\max(N_1, N_2)}) * \dfrac{sim_{semantic}(d_i', d_j')}{N_1 * N_2 - N * N} & if \quad d_i <> d_j \\ \\ sim(d_i, d_j) = 1 & if \quad d_i = d_j \end{cases} \tag{1}$$

Where N is the number of the same terms in document d_i, d_j; N_1, N_2 is the number of feature terms in document d_i, d_j respectively, $sim_{semantic}(d_i', d_j')$ is the semantic similarity of d_i', d_j', which denotes in the feature terms collection of d_i or (d_j), after removal the same terms, the semantic similarity of the remains and the terms of d_j or (d_i).

$$sim_{semantic}(d_i', d_j') = \sum_{k=1}^{N_1 * N_2 - N^2} x_k \tag{2}$$

Where x_k is the semantic similarity value in the collection *Sims* and it is satisfied with $x_k >= x_{k+1}$.

$$Sims = \{x_k \mid x_k = sim_{semantic}(Term_{pi}, Term_{qj}), k = 1,2,....., N_1 * N_2\} \tag{3}$$

$$Rank = \{x_k \mid x_k > x_{k+1}\} \tag{4}$$

It is noticed that the new algorithm not only word form, but also semantic information have been taken into account. Its values are range from 0 to 1.

However, how to obtain the semantic similarity of two terms is another problem. In the paper, we get semantic similarity with the help of WordNet. WordNet is the product of a research project at Princeton University which has attempted to model the lexical knowledge of a native speaker of English [12]. It focuses on the word meanings instead of word forms. In WordNet nouns, verbs, adjectives and adverbs are grouped into sets of synsets, which are interconnected via a variety of relations. The semantic relations for nouns include Hyponym/Hypernym (is-a), Part Meronym/Part Holonym (part-of), Member Meronym/Member Holonym (member-of), Substance Meronym/Substance Holonym (substance-of) and so on. Figure 1 illustrates a fragment of WordNet. In the taxonomy the deeper concept is more specific and the upper concept is more abstract.

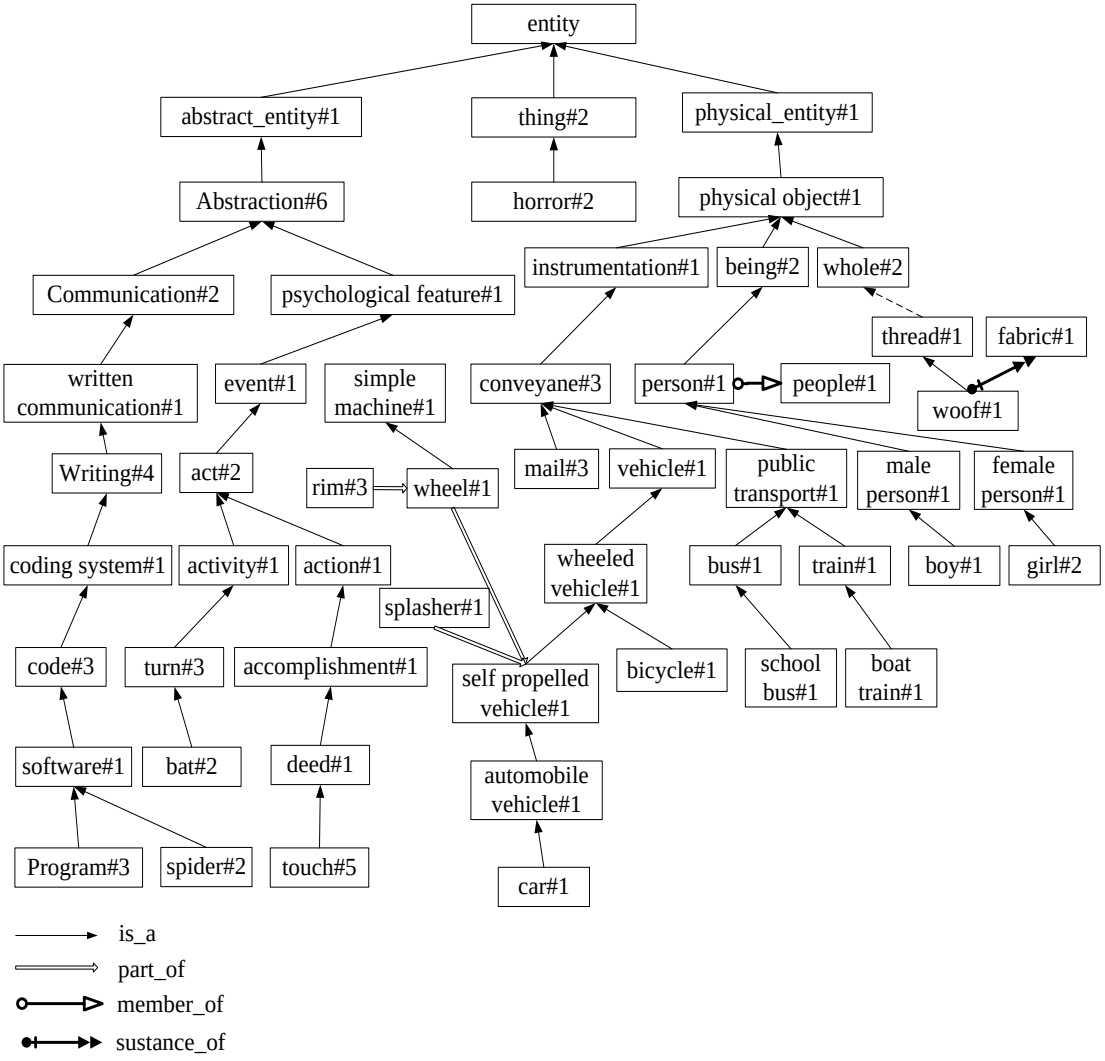

Figure 1. A fragment of Taxonomy in WordNet

Some measures have been proposed for calculating the semantic similarity of two concepts [13-16]. In this study the measure proposed by Meng has been used [16]. It is defined as:

$$sim\ (c_i, c_j) = e^{(\frac{2*IC(lso(c_i, c_j))}{IC\ (c_i)+IC\ (c_j)})} - 1 \tag{5}$$

Where sim (c_i,c_j) is semantic similarity between concept c_i and concept c_j; lso(c_i,c_j) is the most specific common subsumer of c_i and c_j; IC(c_i), IC(c_j) is the information content that contained in concept c_i, c_j respectively.

Because either or both of the words have more than one sense in WordNet, we take the most similarity pair of sense:

$$sim(word_1, word_2) = \max_{(i,j)}[sim(c_{1i}, c_{2j})] \tag{6}$$

Where c_{1i} is the sense of $word_1$, and c_{2j} is the sense of $word_2$.

Let $U_p(.)$ and $U_q(.)$ be the set of documents the system presents to the user as search results for the queries $Query_p$ and $Query_q$ respectively. The document set that users clicked on for the queries $Query_p$ and $Query_q$ may be seen as follows:

$U_p(.)=\{d_{p1}, d_{p2}, \ldots\ldots, d_{pm}\}$

$U_q(.)=\{d_{q1}, d_{q2}, \ldots\ldots, d_{qn}\}$

Then we need to compute the document similarity matrix:

$$A(\text{Query}_p, \text{Query}_q) = \begin{array}{c} \\ d_{p1} \\ d_{p2} \\ \ldots\ldots \\ d_{pm} \end{array} \begin{array}{c} \begin{array}{cccc} d_{q1} & d_{q2} & \ldots\ldots & d_{qn}. \end{array} \\ \begin{bmatrix} sim(d_{p1}, d_{q1}) & sim(d_{p1}, d_{q2}) & \ldots\ldots & sim(d_{p1}, d_{qn}) \\ sim(d_{p2}, d_{q1}) & sim(d_{p2}, d_{q2}) & \ldots\ldots & sim(d_{p2}, d_{qn}) \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots\ldots \\ sim(d_{pm}, d_{q1}) & sim(d_{pm}, d_{q2}) & \ldots\ldots & sim(d_{pm}, d_{qn}) \end{bmatrix} \end{array} \tag{7}$$

Then the similarity of $Query_p$ and $Query_q$ is defined as:

$$\begin{aligned} &sim(Query_p, Query_q) \\ &= \frac{1}{2} * [\frac{\sum_{i=1}^{m} \max(sim(d_{p_i}, doc_{q1}), sim(d_{pi}, doc_{q2}), \ldots\ldots, sim(d_{pi}, doc_{qn}))}{m} \\ &+ \frac{\sum_{j=1}^{n} \max(sim(doc_{p1}, doc_{qj}), sim(doc_{p2}, doc_{qj}), \ldots\ldots sim(doc_{pm}, doc_{qj}))}{n}] \end{aligned} \tag{8}$$

3.2. Queries Clustering

As mentioned above, similar queries denote the same or similar topic. Our study obtains the topics by query clustering. There are many clustering algorithms available to us. Because query logs usually are very large, the system does not know how many topics there will be exist. Therefore it is required the clustering algorithm does not need users to set the resulting form of clusters manually, such as the number or the maximal size of clusters.

After comprehensive comparison, the bottom-up hierarchical clustering method is adopted. The clustering process is shown in Figure 2.

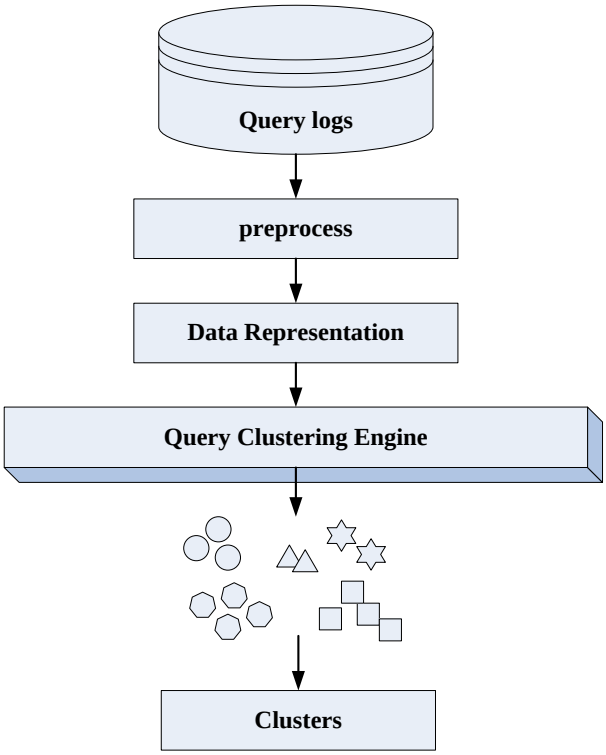

Figure 2. Flow Chart of the Clustering Process

For any two clusters $Cluser_p$, $Cluster_q$, cluster function is defined as follows:

$$sim(Cluster_p, Cluster_q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} sim(Query_i, Query_j)}{m \times n} \quad (9)$$

Where $Query_i$, $Query_j$ are any two queries; m is the number of queries in $Cluster_p$; n is the number of queries in $Cluter_q$; $Sim(Query_i,Query_j)$ is the similarity of $query_i$ and $query_j$.

4. Evaluation

In this section, the new algorithm is evaluated by experiments.

4.1. Data Set

For evaluating the performance of our new algorithm, a dataset is necessary. Because of commercial factors, most of search engine would not like to share their query logs. Only the logs of three companies that are Excite, AlltheWeb, AltaVista are available. The latest version is AltaVista_2003. Unfortunately most pages in 2003 are not existed. Therefore, we build a search engine with Nutch, and ask 200 uses in different major to use the search engine randomly. And collect query logs from December 6, 2012 to January 26, 2013. After preprocessing the data, removing incomplete ones, uncivilized ones, a total of 40728 URLs are left.

4.2. Results Analysis

The proportion of clustered web queries are shown in Figure 3.

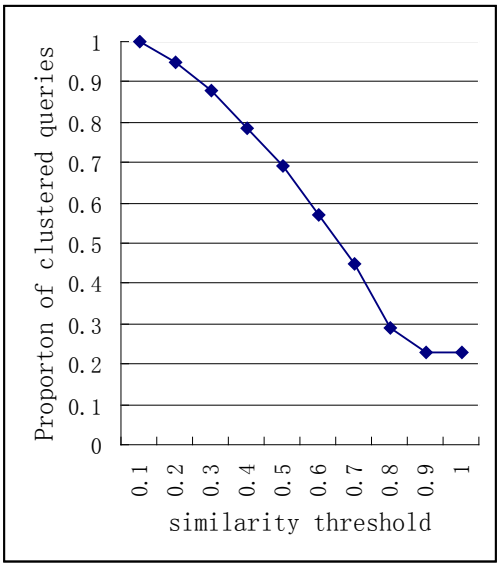

Figure 3. Proportion of Clustered Web Queries

From Figure 3, we can see that with the increase of threshold, the proportion of clustered queries decrease.

In the clustering result, part of clustering figure is shown in Figure 4, where the similarity threshold is 0.55.

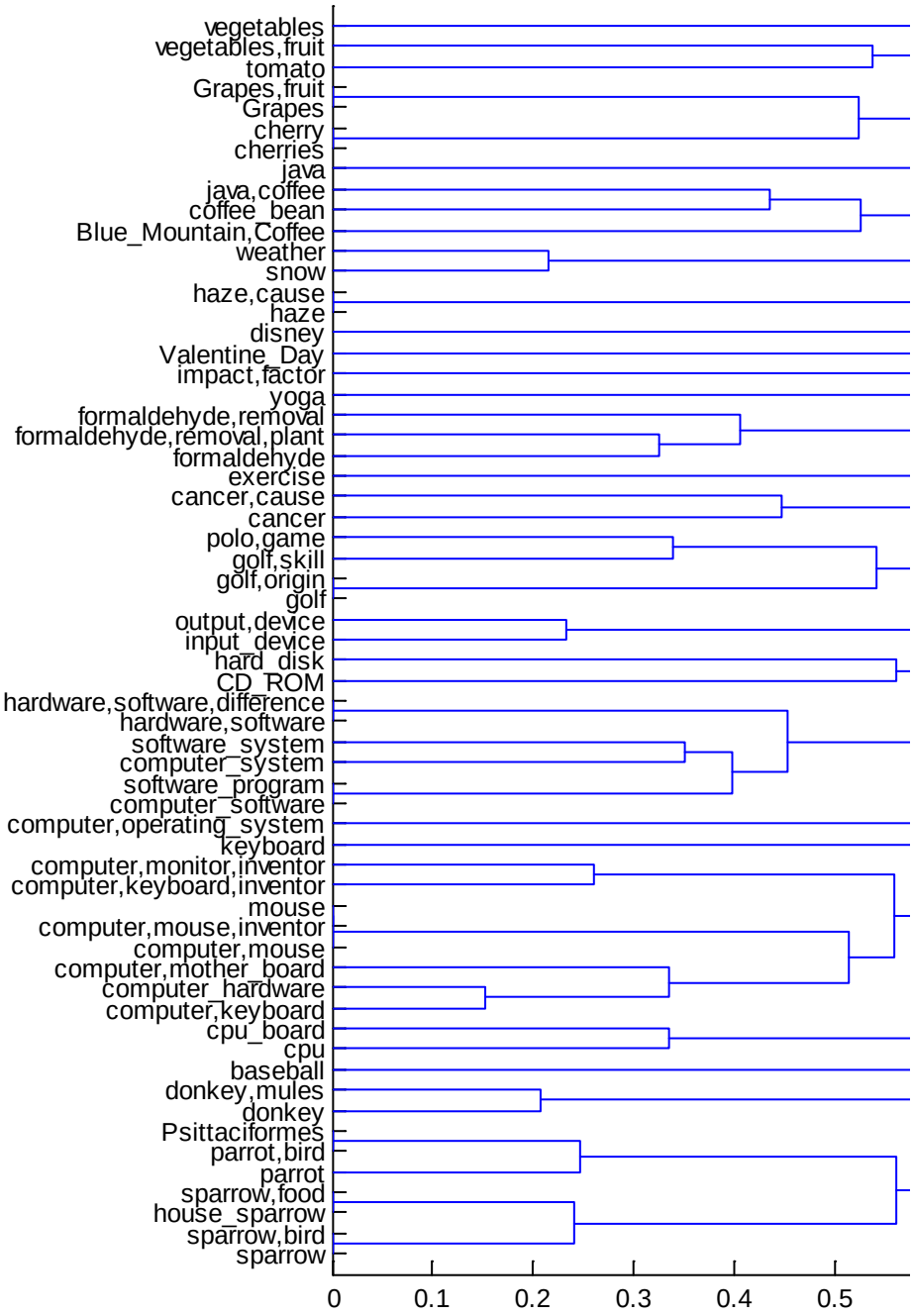

Figure 4. Part of Queries Clustering Result

Notes: X-axis is the semantic distance of the clusters
Y-axis is the queries.

In each cluster, we select two the most similarity queries to expand initial query. The result is shown is Figure 5.

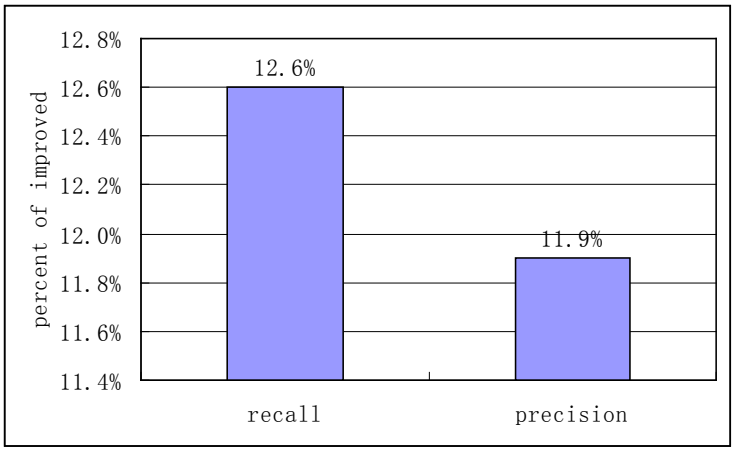

Figure 5. The Improvement of Recall and Precision

From Figure 5, we can see that both average recall and average precision have been increased. The recall has increased 12.6%, and the precision has increased 11.9%, which indicates the good performance of our new model.

5. Conclusion and Future Work

This paper proposes a new query clustering algorithm using user feedback of query logs to discover similar topics for query expansion. Different from previous works, in the new algorithm not only word form of two queries, but also their semantic information have been taken into account. A search engine with Nutch is built to evaluate the new algorithm. Firstly we cluster similar queries. The queries in the same cluster reflect the same or similar topics. Furthermore query clustering result is used to expand initial query. Experiments show that using new algorithm in query expansion significantly improved recall of 12.6%, and precision of 11.9%, which indicates the good performance of our new algorithm. In future work, we will attempt to use this algorithm in document clustering, question-answer system and so on.

References

[1] L. Fitzpatrick and M. Dent, "Automatic feedback using past queries: social searching", Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Philadelphia, PA, USA, **(1997)** July 27-31.
[2] E. De Lima and J. Pedersen, "Phrases recognition and expansion for short, precision-biased queries based on a query log", Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA, **(1999)** August 15-19.
[3] Z. Liu, S. Natarajan and Y. Chen, "Query Expansion Based on Clustered Results", Proceedings of the VLDB Endowment, vol. 4, no. 6, **(2011)**.
[4] M. Diligenti, M. Gori and M. Maggini, "A unified representation of web logs for mining applications", Information Retrieval, vol. 14, no. 3, **(2011)**.
[5] R. Baeza-Yates, "Applications of Web Query Mining", Proceedings of the 27th European conference on Advances in Information Retrieval Research, Santiago de Compostela, Spain, **(2005)** March 21-23.
[6] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, Boston, MA, USA, **(2000)** August 20-23.

[7] S. M. Beitzel and E. C. Jensen, "Automatic Classification of Web Queries Using Very Large Unlabeled Query Logs", ACM Transactions on Information Systems, vol. 25, no. 2, **(2007)**.
[8] J.-R. Wen, J.-Y. Nie and H.-J. Zhang, "Query Clustering Using Content Words and User Feedback", Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, USA, **(2001)** September 9-13.
[9] B. M. Fonseca, P. B. Golgher, E. S. de Moura and N. Ziviani, "Using association rules to discovery search engines related queries", Proceedings of the 1st Conference on Latin American Web Congress, Santiago, **(2003)** November 10-12.
[10] J. Wang, B. Peng and T. Meng, "Discovering Related Web Gueries Based on Search Engine's User Log", JournaI of Beijing University of Posts and TeIecommunications, vol. 28, no. S2, pp. 44-48.
[11] O. R. Zaiane and A. Strilets, "Finding similar queries to satisfy searches based on query traces", Proceedings of the Workshops on Advances in Object-Oriented Information Systems, Montpellier, France, **(2002)** September 2-5.
[12] C. Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press, Cambridge, USA, **(1998)**.
[13] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, Montréal Québec, Canada, **(1995)** August 20-25.
[14] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, Madison, Wisconsin, USA, **(1998)** July 24-27.
[15] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", Proceedings of International Conference on Research in Computational Linguistics, Taipei, TaiWan, **(1997)** August 22-24.
[16] L. Meng and J. Gu, "A New Method for Calculating Word Sense Similarity in WordNet", International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 5, no. 3, **(2012)**.

Authors

Lingling Meng is a PhD Candidate of Computer Science and Technology Department, and a teacher of Department of Educational Information Technology in East China Normal University. Her research interests include intelligent information retrieval, ontology construction and knowledge engineering.

Runqing Huang has a PhD from Shanghai Jiao Tong University. He works in Shanghai Municipal People's Government, P. R. China. His present research interests include modeling strategic decisions, economic analysis, electronic government and Logistics.

Junzhong Gu is Supervisor of PhD Candidates, full professor of East China Normal University, head of Institute of Computer Applications and director of Lab, Director of Multimedia Information Technology (MMIT). His research interests include information retrieval, knowledge engineering, context aware computing, and data mining.

