A New Model for Measuring Similarity of Web Queries and Its Application in Query Expansion[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*

llmeng@deit.ecnu.edu.cn, runqinghuang@gmail.com, jzgu@ica.stc.sh.cn

Abstract

The similarity of web queries plays an important role in capturing frequently asked questions, most popular topics of search engine or automatic query expansion. Accurate measurement of similarity between queries is crucial. The paper presents a new model for similarity metric of web queries using user logs and applied it into information retrieval for query expansion. Different from previous works, in the new model not only word form, but also semantic information has been taken into account. Experiments show that using the new model in query expansion actually improved recall of 8.1 percent and precision of 9.2 percent, which indicates the good performance.

Keywords: *similarity of web queries, word form, semantic information, query expansion*

1. Introduction

The similarity of web queries is a relative new research field. Recently years it attracts great concern and has been successfully applied in user query expansion [1, 2, 3], query recommendations [4, 5], document clustering [6] and question answer system [7]. It shows its talents and makes these applications more effective. Therefore it is necessary to design accurate methods for improving the performance of the bulk of applications relying on it. This paper proposes a new model for similarity metric of web queries using user logs. Both word form of two queries and their semantic information have been taken into account. Experiments show that using the new model in query expansion actually improved recall of 8.1 percent and precision of 9.2 percent, which indicates the good performance.

The rest of this paper is as follows: in section 2 previous works are discussed. A new web queries similarity model is presented in Section 3. In section 4 queries clustering are discussed to find the same or similar topics. Section 5 shows the evaluation of the new model, including experiments, data analyzing, and the achievements. Conclusion and future Work is described in Section 6.

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

2. Related Work

It is a relatively new research field that measuring the similarity between two queries. Some algorithms have been proposed. All the algorithms discussed in the paper are all based on user query logs. In user query logs, all users' query history will be recorded in detail. Despite the formats of logs in different search engines are slightly different, the basic information will be recorded, including the access time, session ID or IP address, keywords, clicked URL and so on. A classical format of query logs is shown as Table 1.

Table 1. The format of user query logs

Item	Content
Id	Session ID. For example, the ID number that system automatically assigns to user or IP address.
Keywords	The keywords that user input in the search engine.
Order	The order that user clicked on URL.
URL	The sequence of URLs that user clicked.
Rank	The order of clicked URL in the research result.
Datetime	The datetime that user clicked on the URL.
Submitter information	The information of browser.

In the search result list each record represents a web page or 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's query log denotes an implicit relevance feedback, base on which some algorithms have been proposed to determine the similar query. Bruno M. Fonseca uses association rule to measure the similarity of queries [8]. In his research it takes query as item and session as transaction of association rule mining. In Ji-Min Wang's research, a new method for discovering related Web queries was presented [9]. 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. Ji-Rong assumed that [7]:

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

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

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

3. A New Algorithm for Similar Queries Metric

All the measures above are simple and effective. However, they are all focus on the level of syntax, and ignored semantic information. In this section, a new algorithm will be presented. Definitions of related concept in the following algorithms are as follows:

(1)Term: a basic logical unit that user input in the search engine.

(2)Query: the keywords that user input in the search engine, including one or more terms.

$$Query = \{Term_1, Term_2, \ldots\ldots, Term_n\}$$

It is noticed that for any two queries $Query_p$, $Query_q$,

$Query_p = \{Term_{p1}, Term_{p2}, \ldots\ldots, Term_{pn}\}$

$Query_q = \{Term_{q1}, Term_{q2}, \ldots\ldots, Term_{qn}\}$

There are four cases between the relationships of $Query_p$ *and* $Query_q$:

Case 1： $Query_p = Query_q$

Case 2： $Query_p \subset Query_q$ or $Query_q \subset Query_p$

Case 3： $Query_p <> Query_q$, and $Query_p \cap Query_q <> \Phi$

Case 4： $Query_p \cap Query_q = \Phi$

For any $Term_{pi}$ in $query_p$, and any $Term_{qk}$ in $query_q$, there are two cases:

Case A： $Term_{pi} = Term_{qk}$

Case B： $Term_{pi} <> Term_{qk}$, but there are semantic associations between $Term_{pi}$ *and* $Term_{qk.}$

Based on the above observations, the new algorithm is based on the following assumptions.

(1) If two queries contain the same terms, they convey the same or similar information needs. The more terms in common, the more similar they are. For example,

$Query_p = \{computer, mouse\}$

$Query_q = \{computer, keyboard\}$

$Query_p \cap Query_q = \{computer\}$

Thus, $Query_p$ and $Query_q$ are similar.

(2) If two queries are semantic associated, and the degree of similarity is greater than a certain threshold, they convey the same or similar information needs. For example,

$Query_p = \{prolog\}$

$Query_q = \{computer\ language\}$

$Query_p \cap Query_q = \Phi$

However, prolog is a kind of computer language. Thus, $Query_p$ and $Query_q$ are similar.

Based on the assumption, the new algorithm is defined as follows:

$$\begin{cases} sim\ (Query_p, Query_q) \\ = \dfrac{N}{\max(N_1, N_2))} + (1 - \dfrac{N)}{\max(N_1, N_2)}) * \dfrac{sim_{semantic}(Queryp', Queryq')}{N_1 * N_2 - N^2} & if \quad Queryp <> Queryq \\ sim\ (Query_p, Query_q) = 1 & if \quad Queryp = Queryq \end{cases} \tag{1}$$

where N is the number of the same Terms of $Query_p$, $Query_q$; N_1, N_2 is the number of terms in $Query_p$, $Query_q$ respectively; $sim_{semantic}(Query_p', Query_q')$ is the semantic similarity of queries $Query_p'$,$Query_q'$, which denotes in the collection of $Quey_p$ (or $Quey_q$), after removal the same Terms, the semantic similarity of the remains and the terms of $Quey_q$ (or $Quey_q$). It is defined as:

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

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 in the new model not only word form, but also semantic information has 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 [10]. It focuses on the word meanings instead of word forms. In WordNet Nouns, verbs, adverbs and adjectives are organized by a variety of semantic relations into synonym sets (synsets), which represent one concept. These relations will be associated with words and words to form a hierarchy structure, which makes it a useful tool for computational linguistics and natural language processing. It is commonly argued that language semantics are mostly captured by nouns or noun phrases so that the study only focus on noun in semantic similarity calculating. 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 taxonomy in WordNet. In the taxonomy the deeper concept is more specific and the upper concept is more abstract.

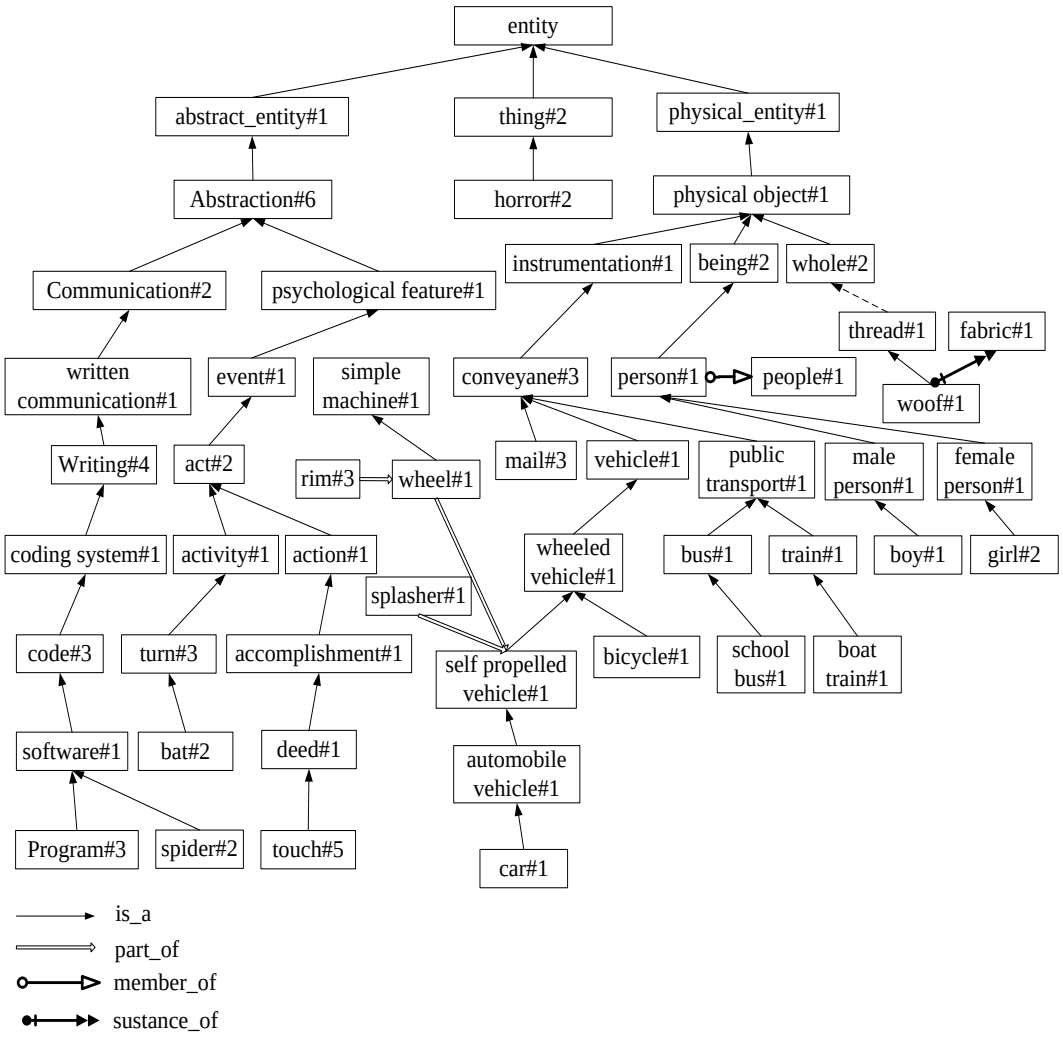

Figure 1. A fragment of taxonomy in WordNet

Some measures have been proposed for calculating the semantic similarity of two concepts [11, 12, 13, 14]. In this study the measure proposed by Meng has been used [14]. It is defined as:

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

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 need to compute the semantic similarity matrix as Table 2.

Table 2. semantic similarity matrix

Word Pairs		*Word*$_2$			
		c_{21}	c_{22}	……	c_{2j}
Word$_1$	c_{11}	$sim_{11\text{-}21}$	$sim_{11\text{-}22}$	……	$sim_{11\text{-}2j}$
	c_{12}	$Sim_{12\text{-}21}$	$sim_{12\text{-}22}$	……	$sim_{12\text{-}2j}$
	……	……	……	……	……
	c_{1i}	$sim_{1i\text{-}21}$	$sim_{1i\text{-}22}$	……	$sim_{1i\text{-}2j}$

Where C_{1i} is the sense of word1, and C_{2j} is the sense of word2. In the result, we took the most similarity pair of sense:

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

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

4. Queries clustering

As mentioned above, similar queries denote the same or similar topic. Our study obtains the topics by query clustering. Because there are a wide variety of queries, 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 number of clusters manually.

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

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

$$sim(Cluster_p, Cluster_q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} sim(Query_i, Query_j)}{m \times n} \tag{7}$$

Where $Query_i$, $Query_j$ are 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 $query_j$.

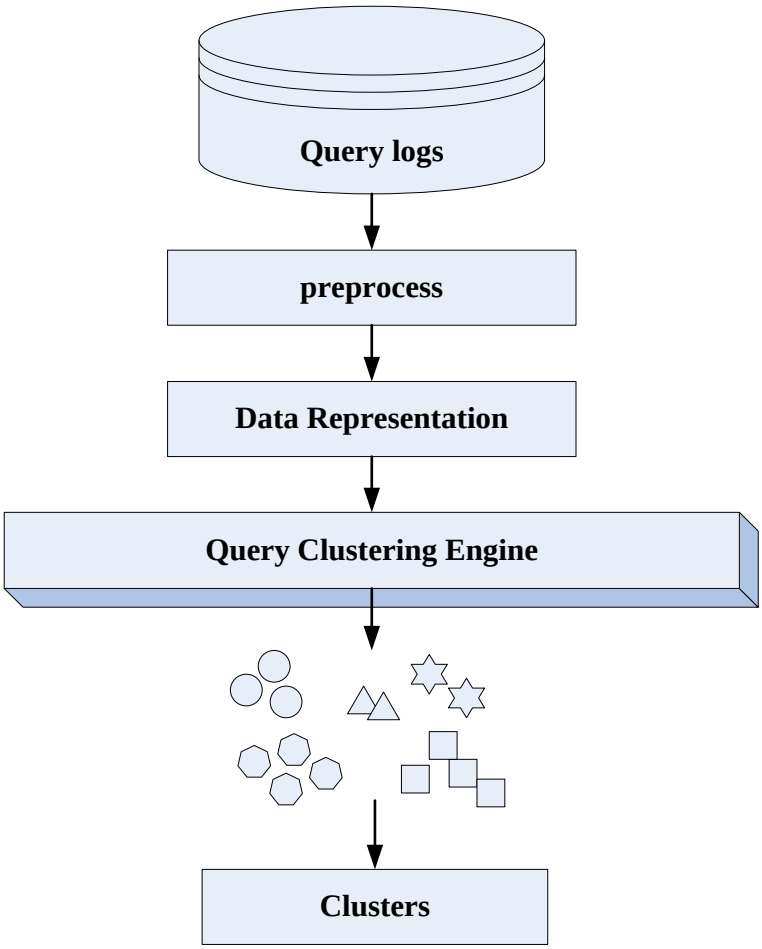

Figure 2. Flow chart of the clustering process

5. Evaluation

In this section, we evaluate our new model by experiments.

5.1. Data set

For evaluating the performance of our new model, 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 can be obtained. The latest version is AltaVista_2003. Unfortunately most pages in 2003 are not existed. Therefore, we build a search engine with Nutch, and ask 20 users to use the search engine randomly. And collect query logs from December 6, 2012 to December 26, 2012. After preprocessing the data, removing incomplete ones, uncivilized ones, a total of 8144 URLs are left.

5.2. Results analysis

Fourteen pairs of web queries similarity results are selected randomly, which is shown in Table 2.

Table 2. The similarity of Web queries

$Query_p$	$Query_q$	New_{term}
software	virus	0.74673300
bicycle	car	0.68007700
communist	Engels	0.84922200
hard disk	input device	0.29986400
java	software system	0.15769300
eye	neck	0.60595300
rocket	rocket, satellite	0.62134600
Atomic bomb	Nuclear weapon	0.97368000
Atomic bomb	Nuclear weapon, country	0.51927700
beijing	hangzhou	0.42877800
cell phone, inventor	mobile phone, inventor	0.73751300
rheumatoid arthritis	rheumatoid arthritis, cause	0.68047300
apple	potato	0.45753600
apple, fruit	orange, fruit	0.85157900

From table 2, it is noticed that only four Query pairs have the same term, they are {rocket; rocket, satellite},{cell phone, inventor; mobile phone, inventor }, { rheumatoid arthritis; rheumatoid arthritis, cause },{ apple fruit; orange fruit }。Other ten pairs do not contain the same term, they are {software; virus}, {bicycle; car}, {communist, Engels}, {hard disk; input device}, {java, software system}, {eye, neck}, {Atomic bomb, Nuclear weapon}, {Atomic bomb; Nuclear weapon, country}, {Beijing, hangzhou}, {apple, potato}. According to the new model in section 3, there are six pairs in the ten ones are highly semantic related. And the similarity is superior to 0.5.

In the clustering result, part of clustering figure is shown in Figure 3. Here the threshold is 0.6.

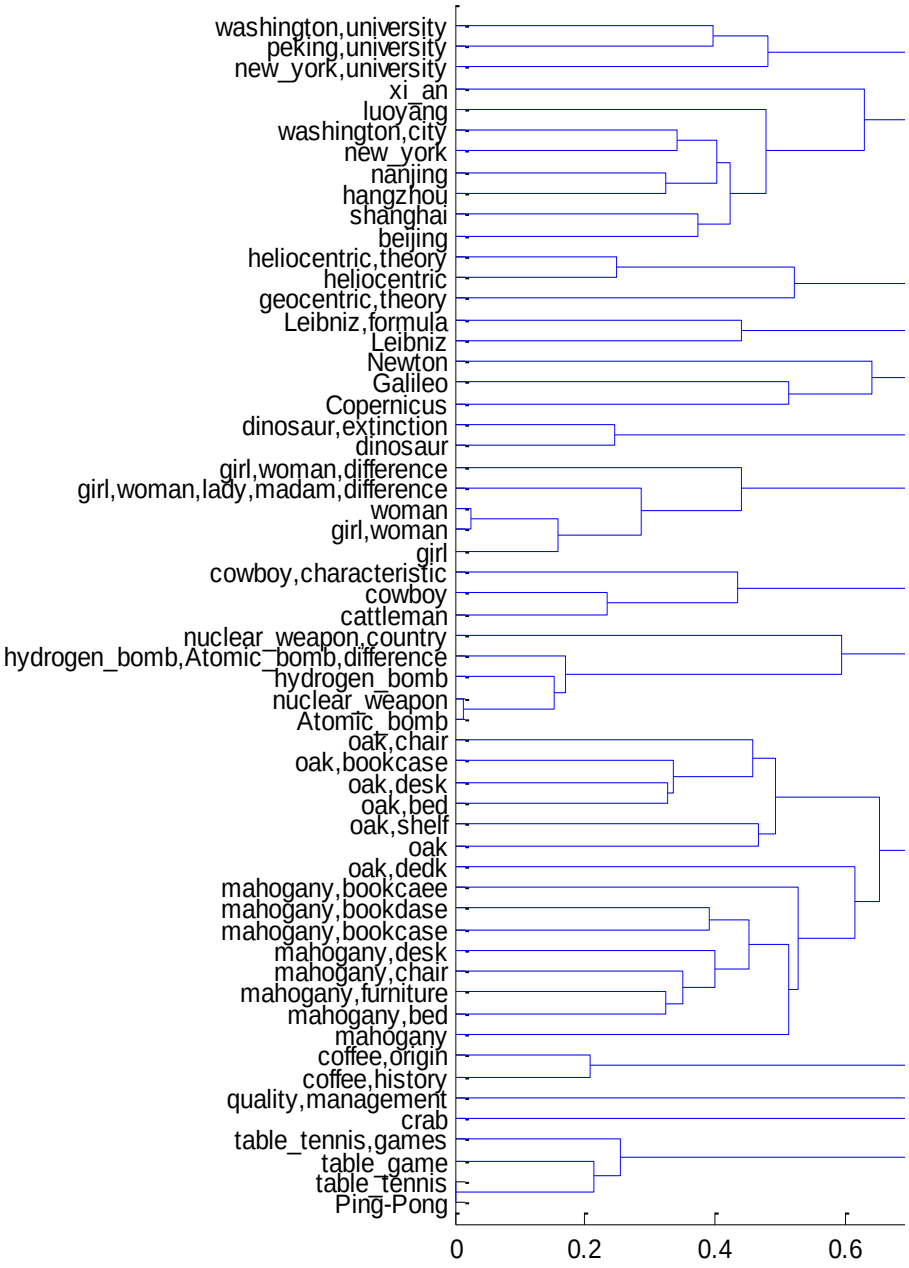

Figure 3. Part of queries clustering result

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

The proportion of clustered web queries with different threshold is shown in Figure 4.

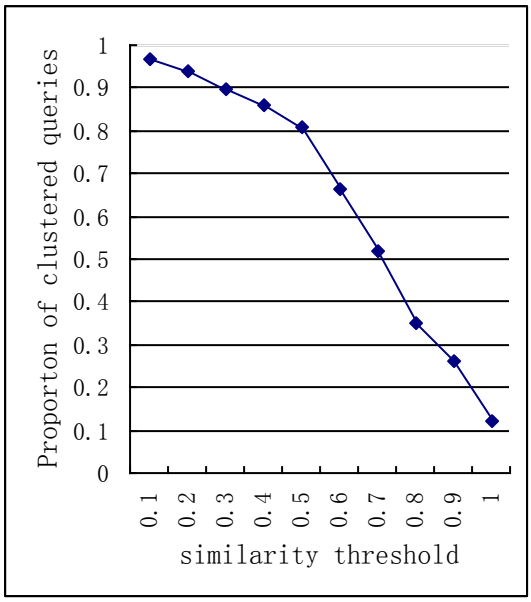

Figure 4. Proportion of clustered web queries

From Figure 4, we can see that with the increase of threshold, the proportion of clustered queries decrease.

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

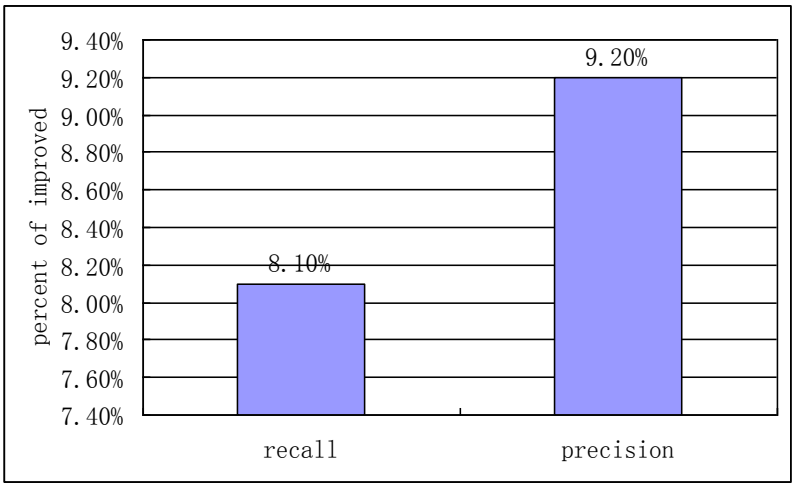

Figure 5. The improvement of recall and precision

From Figure 5, we can see that both average recall and average precision have been increased. The recall has increased 8.1%, and the precision has increased 9.2%, which indicates the good performance of our new model.

6. Conclusion and Future Work

This paper presents a new model for measuring similarity of web queries using user logs and applies it in query expansion. Different from previous works, in the new model both word form of two queries and their semantic information have been taken into account. A search engine with Nutch is built to evaluate the new model. First we cluster similar queries. The queries in the same cluster reflect similar topics. Then using Query clustering result expand initial query. Experiments show that using new model in query expansion significantly improved recall of 8.1%, and precision of 9.2%, which indicates the good performance of our new model. In future work, we will attempt to use this model 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, **(1997)** July 27-31, Philadelphia, PA, USA.
[2] E. De Lima and J. Pedersen, "Phrases recognition and expansion for short, precision-biased queries based on a query log", In Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, **(1999)** August 15-19, Berkeley, CA, USA.
[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, **(2005)** March 21-23, Santiago de Compostela, Spain.
[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, **(2000)** August 20 - 23, Boston, MA, USA.
[7] 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, **(2001)** September 9-13, New Orleans, Louisiana, USA.
[8] 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, **(2003)** November 10-12, Santiago.
[9] 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.
[10] C. Fellbaum, ed., "WordNet: An Electronic Lexical Database", MIT Press, Cambridge, USA, **(1998)**.
[11] P. Resnik, "Using information content to evaluate semantic similarity", Proceedings of the 14th International Joint Conference on Artificial Intelligence, **(1995)** August 20-25; Montréal Québec, Canada.
[12] D. Lin, "An information-theoretic definition of similarity", Proceedings of the 15th International Conference on Machine Learning, **(1998)** July 24-27; Madison, Wisconsin, USA.
[13] 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, **(1997)** August 22-24; Taipei, TaiWan.
[14] 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

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, query recommendation and knowledge engineering.

Runqing Huang

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

Prof. 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.