Query Auto-Completion Based on Word2vec Semantic Similarity

To cite this article: Taihua Shao et al 2018 J. Phys.: Conf. Ser. 1004 012018

View the article online for updates and enhancements.
Query Auto-Completion Based on Word2vec Semantic Similarity

Taihua Shao†*, Honghui Chen† and Wanyu Chen†

†Key Laboratory of Information Systems Engineering, National University of Defense Technology, Changsha, Hunan, China

*E-mail: hunsth@outlook.com

Abstract. Query auto-completion (QAC) is the first step of information retrieval, which helps users formulate the entire query after inputting only a few prefixes. Regarding the models of QAC, the traditional method ignores the contribution from the semantic relevance between queries. However, similar queries always express extremely similar search intention. In this paper, we propose a hybrid model FS-QAC based on query semantic similarity as well as the query frequency. We choose word2vec method to measure the semantic similarity between intended queries and pre-submitted queries. By combining both features, our experiments show that FS-QAC model improves the performance when predicting the user’s query intention and helping formulate the right query. Our experimental results show that the optimal hybrid model contributes to a 7.54% improvement in terms of MRR against a state-of-the-art baseline using the public AOL query logs.

1. Introduction
Query Auto Completion (QAC) has its task to help the user formulate her query while she is typing only the prefix, e.g. several characters [1]. Its main purpose is to predict the user’s search intention by completing the query with fewer characters. The emergence of Google Instant is a major trigger for query auto-completion research. After that, how to build a better query completion becomes very important, which determines the time users wait for to get intended information [2].

With a QAC system, queries that match the query prefix are pre-generated and stored in an efficient data structure. Figure 1 provides a QAC instance. When one user inputs several characters, the search engine will provide a set of query completions matching user's input with a drop-down menu in real time [3].

A basic method in most QAC systems is to extract the query and its prefix from the search records for a period of time, and to count up the number of times that the query is submitted in the previous history, i.e. the frequency [4]. Then, queries can be sorted according to their frequencies. This method assumes that the future frequency of query is constant with the current frequency. The performance of a QAC system with this method is in good satisfaction but far less than the best. Because it fails to consider other important factors like time and semantic similarity relationship between queries, etc [5] [6]. However, these information always affect the query auto-completion ranking performance [7]. To smooth the deficiencies of QAC research, we study on how to consider these information in a QAC system.
We firstly propose a query auto-completion model S-QAC based on semantic similarity feature between queries, which is calculated by word2vec method. Next, by combining the frequency feature from traditional MPC model [2] and semantic similarity feature from our S-QAC model, a hybrid query auto-completion model FS-QAC is proposed.

We obtain the optimal hybrid model and analyze its performance by adjusting the weight of each feature. We also discuss the performance of our query auto-completion model under different prefix length. The results show that combining the frequency and semantic similarity does help to improve the performance of query auto-completion.

Our contributions can be summarized as follows:

- A new method considering semantic similarity feature in QAC system with the help of word2vec method, which indicates that semantic similarity is efficient for improving the QAC performance in terms of Mean Reciprocal Rank (MRR);
- A hybrid model combining both frequency and semantic similarity features for QAC, whose experiment results show that our optimal hybrid model significantly outperforms the state-of-the-art MPC method.

2. Approach
In this section, we describe our semantic-similarity-based model S-QAC, as well as a hybrid model FS-QAC, which considers both frequency and semantic similarity features. We introduce the related concepts of our models, and give concrete steps of modeling.

2.1. S-QAC model
Many methods have been proposed to measure the semantic similarity between two words. After comparing various semantic similarity measuring methods, we build the S-QAC model, which considers semantic similarity between the candidate queries and their previous queries submitted in the same session, on the basis of word2vec method.

Word2vec method [8] would firstly train the input corpus, namely, the sentences consisted by word collection: term₁, term₂,... termₙ. Its purpose is to predict the terms around through the term vector representation outputed by the model [9]. This goal can be achieved by maximizing the average log probability, i.e.:

\[
L = \max \left( \sum_{\text{term}_i \in \text{C}} \log P(\text{term}_i | \text{context(\text{term}_i)}) \right)
\]  

Where C is the size of the context used for training. It can be seen that the word2vec method highly depends on the training text dataset [10]. That's to say, the queries to be represented must appear in the training corpus before.

The pre-trained Google News model provided by Google includes hundreds of millions of text statements. However, in real world information retrieval system, queries submitted by users could sometimes be rarely used in the pre-trained model. In order to solve this problem, we first train our experimental AOL dataset [11]. After generating AOL dataset based model, we combine it with the pre-trained Google News model to ensure each query in test process could be represented as vector.

With the two word2vec models above, we could measure the semantic similarity between the candidate queries and the previous queries submitted in the same session. We could obtain two
semantic similarity scores of a certain candidate query: Score\(_{A}(q_c, q_x)\) and Score\(_{G}(q_c, q_x)\). Among them, Score\(_{A}(q_c, q_x)\) represents the score between the candidate query \(q_c\) and the query \(q_x\) in current session calculated by AOL word2vec model while Score\(_{G}(q_c, q_x)\) is the semantic similarity score calculated by Google News word2vec model. They were synthesized as follows:

\[
\text{ScoSim}(q_c, q_x) = \omega \cdot \text{Score}_{A}(q_c, q_x) + (1 - \omega) \cdot \text{Score}_{G}(q_c, q_x)
\]  

(2)

Where \(\omega\) is the weight parameter controlling the contributions between the similarity calculated by the two word2vec models.

As the prefix \(p\) is also a role reflecting user's query intention. We synthesize it into our model to improve the accuracy. The formula is as follows:

\[
\text{Score}_S(q_c) = \sum_{x=1}^{N} \gamma_x \cdot \text{ScoSim}(q_c, q_x) + \delta \cdot \text{ScoSim}(q_c, q_x)
\]  

(3)

Where \(N\) is the number of queries that had already been submitted in current session. \(\gamma\) and \(\delta\) are weight parameters.

2.2. **FS-QAC model**

Besides the semantic similarity score of each candidate query \(q_c\) obtained in § 2.1, it is reasonable to assume that the query with higher frequency in the search history is more likely to be submitted again. Thus, in this study, we propose to combine the semantic similarity feature and the frequency feature by using the linear fitting formula in our hybrid model FS-QAC as follows:

\[
\text{Score}_{FS}(q_c) = \alpha \cdot \text{Score}_S(q_c) + (1 - \alpha) \cdot \text{Score}_F(q_c)
\]  

(4)

Where \(\alpha\) denotes the tradeoff between query semantic similarity and query frequency. By adjusting \(\alpha\), we can linearly integrate two features. After groups of experiments, we could discover the optimal tradeoff for our FS-QAS model. Algorithm 1 details the major steps of our hybrid model FS-QAC.

**Algorithm 1** Frequency and Semantic Similarity Based Query Auto-Completion (FS-QAC).

**Input:** AOL dataset: \(Q\); Completion List: \(L\); Current Session Queries List: \(C\);
**Output:** Frequency Score List: \(F\); Semantic Similarity Score List: \(S\); Re-ranked List: \(R\);

1: for each \(a \in L\) do
2: \(\text{Score}_F(a) = \text{count}(a \text{ in } Q)\); \(F\).append(Score\(_F\) (a))
3: end for
4: AOL.model = models.save.Word2Vec(Q);
5: model.load(AOL.model + Google News.model)
6: for each \(a \in L\) do
7: \(\text{for each } b \in C\) do
8: \(\text{Score}_S(a) = \text{model.similarity}(b, a)\)
9: end for
10: \(S\).append(Score\(_S\) (a))
11: end for
12: \(\text{Score}_{FS}(a) = \alpha \cdot \text{Score}_S(a) + (1 - \alpha) \cdot \text{Score}_F(a)\)
13: \(\text{dic} = \{ a \in L : \text{Score}_{FS}(a)\}\)
14: \(R = L\).sorted(dic; \(\text{key} = \text{dic.getitem}; \text{reserve} = \text{True}\)
15: return Re-ranked List: \(R\);

3. **Experiments**

We begin with the model summary and detail the research questions that we aim to answer. We then describe the experiment setup for our study.
3.1. Model summary
The baseline to be compared is the Most Popular Completion (MPC) model, which is simple and direct, but widely used. This model predicts users' submission probability according to the frequency of queries in long-term search logs based on the maximum likelihood estimation method.

The models to be discussed in our paper are:
- **MPC**: a traditional query auto-completion model based on the frequency of queries, which is the baseline to be compared in our study;
- **S-QAC**: a query auto-completion model based on semantic similarity feature between candidate queries and the previous queries submitted in the same session;
- **FS-QAC**: a hybrid query auto-completion model proposed by us which combining both frequency and semantic similarity features;

The research questions guiding our experiments are:
- **RQ1**: Could semantic similarity feature improve the performance of QAC system? (See § 4.1)
- **RQ2**: How could semantic similarity feature improve the performance of QAC system together with frequency feature? (See § 4.2)
- **RQ3**: How does prefix length influence the performance of models? (See § 4.3)

3.2. Dataset and parameters
Our experiments was carried out on the AOL dataset [11]. AOL is one of few datasets that contain real query data. The number of query records could fully guarantee statistical significance. Queries from AOL were obtained in real search engine from March 1, 2006 to May 31st, 2006. We process queries submitted within 30 minutes as a session. For the sake of temporal coherence, we divide the AOL dataset into two parts. 75% of the data is used as the train set, and the remaining 25% as the test set. According to the prefix length, we divide test dataset into 5 different data blocks, their prefix lengths are 1, 2, 3, 4 and 5 respectively. Table 1 details the statistics of the dataset used.

| Variables          | Train     | Test     |
|--------------------|-----------|----------|
| Queries            | 2,908,666 | 1,121,126|
| Unique queries     | 452,980   | 452,980  |
| All Prefixes       | 2,430,110 | 867,344  |
| Prefix-1           | 216,925   | 75,455   |
| Prefix-2           | 384,909   | 132,786  |
| Prefix-3           | 574,980   | 203,830  |
| Prefix-4           | 644,897   | 232,613  |
| Prefix-5           | 608,399   | 222,660  |

To train our word2vec model, we set the vector space dimension size as 200, the least number of occurrences min_count as 1, the number of parallel training process workers as 4. To simplify our models, we set coefficients $\gamma$ and $\delta$ as $1/(N+1)$ to make $\text{ScoSim}(q, Q_i)$ the arithmetic mean of all relevant semantic similarity scores of $q$, and the linear integration parameter $\alpha$ as 0.5. $\alpha$ is also a linear integration parameter which control the proportion of semantic similarity in FS-QAC model. In this study, we adjust the $\alpha$ from 0 to 1 with the step increase 0.1 to conduct 11 groups of experiments.

In order to evaluate the effectiveness of query auto-completion method, we choose Mean Reciprocal Rank (MRR) to evaluate the performance of query auto-completion. MRR performed well in many other studies that allow returning multiple results. It is often used to measure the effectiveness of query auto-completion experiments.
In the drop list of query auto-completion system, if query is in the position rank, then it corresponds to the score RR=1/rank. Finally, the Mean Reciprocal Rank score is the mean of all RRs for all prefixes. The formula is as follows:

$$\text{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\text{rank}_i}$$  \hspace{1cm} (5)

Where Q is the sample dataset, |Q| represents the number of samples in Q, and rank, represents the position of the query in the query completion list.

4. Results and Discussion

4.1. Performance of S-QAC Model

To answer RQ1, we examine the query auto-completion performance (MRR) of the baseline as well as our S-QAC model under various prefix lengths. See Figure 2 for the results.

As it illustrates in Figure 2, when prefix length is 1, 2 or 5, the MRR of S-QAC model is higher than MPC model. Especially when prefix length is short, S-QAC model performs significantly better than MPC model. As for the whole MRR of test dataset, we have calculated that the performance of S-QAC model is 3.58% higher than that of MPC model, where MRR_{S-QAC}=0.6712 and MRR_{MPC}=0.6480. The result shows that our S-QAC model based on semantic similarity is helpful to improve the performance of QAC system especially when there is not enough input prefix for the QAS task.

![Figure 2. MRR comparison between S-QAC and MPC.](image1)

![Figure 3. MRR of FS-QAC under different tradeoff.](image2)

4.2. Performance of FS-QAC Model

In order to verify the performance improvement of our proposed hybrid model FS-QAC in RQ2, we conduct 11 groups of experiments and calculate the MRR of FS-QAC model under different tradeoff parameter $\alpha$. The results is showed in Figure 3.

As shown in Figure 3, our proposed hybrid model FS-QAC improves the performance of QAC system, and outperforms the plain model MPC ($\alpha=0$) and S-QAC ($\alpha=1$). When the tradeoff varies from 0 to 0.6 with the step increase 0.1, growth rate of MRR increases with $\alpha$. When the tradeoff
varies from 0.6 to 1, growth rate of MRR decreases with $\alpha$. With different tradeoff, the performance improvement rate is different, among which the best performance is when $\alpha=0.6$, where $\text{MRR}_{FS-QAC}=0.6969$, the growth rate against baseline is 7.54%. The result shows that our proposed hybrid model FS-QAC could significantly improve the performance of QAC system by well synthesizing frequency feature and semantic similarity feature.

4.3. Discussion on Different Prefix Length
To answer RQ3, we examine the MRR of our FS-QAC model with different tradeoff parameter under various prefix lengths. See Figure 4 for the results.

As it illustrates in Figure 4, when the prefix length is short ($\leq 3$), the MRR values of the models under different tradeoff are higher than the results when the prefix length is long ($>3$). With the growth of prefix length, the performance tends to be stable. It indicates that when prefix length is short, the proportion of semantic similarity feature could obviously increase the MRR value of QAC system, as the performance of our FS-QAC models ($\alpha=0.5$) outperforms those models with $\alpha<0.5$ especially when the prefix length is short ($\leq 3$). However, when the prefix length is long ($>3$), there is no too much differences among those models with various values of parameter $\alpha$ in terms of MRR.

5. Conclusions and Future Work
We propose a model FS-QAC based on hybrid ranking scores by combining two features, i.e. query semantic similarity and the query frequency. We calculate the semantic similarity score by integrating Google News model and AOL dataset model. Our experiments show that the semantic similarity together with query frequency can provide a satisfactory idea to predict the user's query intention and help formulate the right query. We verify the effectiveness of our best performer FS-QAC, showing significant improvement over the baseline MPC model in terms of MRR. The results show that the optimal hybrid model contributes to a 7.54% improvement in MRR against well-known baseline.

As future work, we plan to evaluate our models on other dataset so as to verify the effectiveness of our proposal. In addition, we also want to investigate the sensitivity of involved parameters, e.g. the personalized indicators of query [12].

References
[1] Cai F and Rijke M D 2016 A survey of query auto completion in information retrieval. *Foun. Tre. in Inf. Retr.* **10**(4) 273-363
[2] Bar-Yossef Z and Kraus N 2011 *Proc. Int. Conf. on WWW* (Hyderabad: ACM) pp 107-116
[3] Shokouhi M and Radinsky K 2012 *Proc. of Int. Acm Sigir Conf. on Res. Dev. in Inf. Retr.* (Portland: ACM) pp 601-610.
[4] Santo G D, McCreadie R, Macdonald C and Ounis I 2015 *Proc. of Int. Acm Sigir Conf. on Res. Dev. in Inf. Retr.* (Santiago: ACM) pp 775-778
[5] Cai F and Rijke M D 2016 *Proc. of Int. Acm Sigir Conf. on Res. Dev. in Inf. Retr.* (Pisa: ACM) pp 993-996
[6] Cai F, Liang S and Rijke M D 2014 *Proc. of 23rd ACM Int. Conf. on Inf. Kno. Mana.* (Shanghai: ACM) pp 1599-1608
[7] Jiang J Y, Ke Y Y, Chien P Y and Cheng P J 2014 *Proc. of 37th Int. Acm Sigir Conf. on Res. Dev. in Inf. Retr.* (Gold Coast: ACM) pp 445-454
[8] Mikolov T, Sutskever I, Chen K, Corrado G and Dean J 2013 *Proc. of 26th Int. Conf. on Neur. Inf. Proc. Sys. (vol 2)* (Lake Tahoe: Curran Associates Inc) pp 3111-3119
[9] Mikolov T, Chen K, Corrado G and Dean J 2013 Efficient estimation of word representations in vector space. *DBLP: J. CoRR abs/1301.3781*
[10] Cai F and Rijke M D 2016 *J. Inf. Proc. and Mana.: an Int. J. arch.* **52**(4) 628-643
[11] Pass G, Chowdhury A and Torgeson C 2006 *Proc. of 1st Int. Conf. on Sca. Inf. Sys.* (Hong Kong: ACM) p 1
[12] Shokouhi M 2013 *Proc. of 36th Int. Acm Sigir Conf. on Res. Dev. in Inf. Retr.* (Dublin: ACM) pp 103-112