Improving Identification of Latent User Goals through Search-Result Snippet Classification

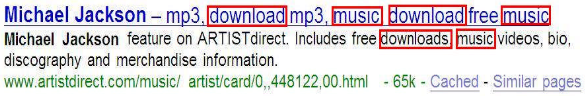
Kuan-Yu He, Yao-Sheng Chang, Wen-Hsiang Lu
Department of Computer Science and Information Engineering
National Cheng Kung University, Tainan, Taiwan
p7694170@mail.ncku.edu.tw, ys.chang1976@gmail.com, whlu@mail.ncku.edu.tw

Abstract

In this paper, we propose an enhanced approach to improving our previous method which employs syntactic structures (verb-object pairs) to identify latent user goals. Our new approach employs a supervised-learning method to learn hint verbs and considers URL information and title information to classify snippets into three coarse categories, which are resource-seeking, informational, and navigational. Also, we propose three different models to identify three different categories of specific latent user goals from the classified snippets.

1. Introduction

Search engines nowadays suffer from critical challenges in dealing with diverse short queries. However, it is apparently not easy for most of existing search engines to predict user goals behind ambiguous short queries or to associate user goals with queries given limited information in incomplete short queries. For example, when a user submits a query, “Michael Jackson”, the real user goal in his/her mind may be that “I want to download Michael Jackson’s music”. Figure 1 illustrates a search-result snippet retrieved with the query “Michael Jackson” from Google. However, the snippet of fitting a user goal is returned at about the position of the 100th rank since the search engine doesn’t know the above real user goal.



Michael Jackson – mp3, download mp3, music, download free music
Michael Jackson feature on ARTISTdirect. Includes free downloads, music videos, bio, discography and merchandise information.
www.artistdirect.com/music/ artist/card/0,,448122,00.html - 65k - Cached - Similar pages

Figure 1. A search-result snippet with the query “Michael Jackson”.

In the past few years, Broder [1] and a few researchers [2, 3] have pointed out the important issue that accurate understanding and modeling of user goals has a great benefit in applications to Web search ranking, click spam detection, Web search personalization, and other tasks. Recently, some researchers tried to identify user goals automatically. For example, Lee et al. [2] proposed to use past user-click behavior and anchor-link distribution as potential features for the prediction of informational goal and navigational goal. However, the above two works only consider the automatic classification of the two coarse categories of goals, namely navigational goal and informational goal. To our knowledge, there is no work focusing on dealing with the classification of the above three categories of user goals simultaneously. In fact the identification of resource-seeking goals is more difficult.

In the past years, some researchers utilized click-through data to effectively improve the performance of search results. Thus, in our previous work [4], we assume that certain frequently clicked snippets may contain certain latent terms related to real user goals. Based on the viewpoint of natural language processing (NLP), we assume that a user goal should be expressed with the form of a latent sentence in his/her mind. Generally, a typical sentence in Chinese/English consists of a subject, a verb, and an object (SVO syntactic structure). Also, we assume that the subject of the latent sentence in user mind is the user himself/herself and the composed verb-object pair can be used to represent a latent user goal. For the above query “Michael Jackson”, we predict that the latent sentence in the user mind is “I want to download Michael Jackson’s music,” and the latent user goal is the verb-object pair “download music”, which can be found around the query “Michael Jackson” in the snippet in Figure 1.

However, our previous method is not good at identification of informational and navigational goals. For example, most navigational queries have only a goal to visit a specific site, thus they don’t need other tedious verb-object pairs to be their goals. In fact, we need different strategies to deal with different categories of user goals. To improve our previous method, in this paper, we thus propose a two-stage approach which first classifies search-result snippets into three coarse categories and then identifies specific latent user goals from the classified snippets.

2. Improving User Goal Identification

2.1 Problems and Ideas

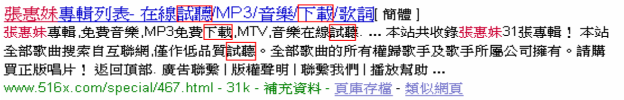
0-7695-3026-5/07 $25.00 © 2007 IEEE
DOI 10.1109/WI.2007.95

A probabilistic inference model which identifies latent user goals in the form of verb-object pairs by utilizing search-result snippets and syntactic structure (verb-object pairs) was proposed in our previous work. Please refer to [4] for more details.

Our previous method still suffers from two problems. First, a query may simultaneously include three different categories of user goals, which are navigational, information, and resource-seeking. Moreover, we found that the idea of using verb-object pairs as latent user goals is effective only for resource-seeking goal identification while it is not effective for the other two categories of user goal identification.

Thus, we propose a search-result snippet classification method and three specific latent user goal models to improve our previous method. First, before identifying specific latent user goals, it may be beneficial to classify search-result snippets in advance into the three coarse categories, which are navigational, informational and resource-seeking. Second, we need to develop other appropriate strategies to identify specific latent user goals in other two categories, which are informational and navigational.

2.2 Snippet Classification


張惠妹專輯列表- 在線試聽/MP3/音樂/下載/歌詞[簡體]

張惠妹專輯,免費音樂,MP3免費下載,MTV,音樂在線試聽. ... 本站共收錄張惠妹31張專輯！ 本站全部歌曲搜索自互聯網,僅作低品質試聽。全部歌曲的所有權歸歌手及歌手所屬公司擁有。請購買正版唱片！ 返回頂部. 廣告聯繫 | 版權聲明 | 聯繫我們 | 播放幫助 ...

www.516x.com/special/467.html - 31k - 補充資料 - 頁庫存檔 - 類似網頁

Figure 2. One resource-seeking snippet with the Chinese query "張惠妹" (A-Mei, a famous singer in Taiwan).

In our observation, search-result snippets in the three different categories have different properties.

First, for resource-seeking category, this category of snippets often intends to deliver users information about some resource-seeking oriented services. Figure 2 illustrates one retrieved resource-seeking snippets for the Chinese query "張惠妹" (A-Mei, a famous singer in Taiwan). We can obviously observe several resource-seeking relevant verbs in the snippets, e.g., "下載" (download) and "試聽" (listen). Thus, in this paper, we utilize some highly relevant hint verbs learned by using a supervised-learning method to classify resource-seeking snippets. In the process of supervised learning, 100 random queries from top 1000 high-frequency queries from 1999 Dreamer's query log was used. For each query, we retrieved 100 snippets returned by Google search engine and three participants were asked to judge and select highly relevant words within resource-seeking snippets. Thus, the trained hint verbs are obtained from these highly relevant words. Table 1 shows the statistics of the top 5 high-frequency hint verbs, which can clearly reveal many resource-seeking goals. Hence, we believe that it is feasible to classify resource-seeking snippets by the assistance of these specific hint verbs.

Table 1. Statistics of top 5 trained resource-seeking hint verbs from collected resource-seeking snippets

Trained Hint Verbs	Normalized Weight
下載 (download)	0.447
購物 (shop)	0.092
聊天 (chat)	0.047
拍賣 (sale)	0.036
查詢 (inquire)	0.036

Second, for navigational category, we observe that the URL of a navigational snippet is usually in the form of a host name of the website. For example, the URL of a snippet about the website of "Apple Computer" is http://www.apple.com/. In addition, the page in the website root or in the first hierarchy directory is also considered. Moreover, the query terms usually appear in the title of navigational snippets. In brief, information of URL and title is considered to determine whether one snippet is navigational.

Finally, for informational category, the snippets of this category are hard to classify due to the diverse contents in topics. Fortunately, the two classification methods for resource-seeking and navigational snippets perform well. Hence, we regard the remaining snippets, which can't be classified into resource-seeking and navigational categories, as informational snippets.

2.3 Latent User Goal Models

After the process of search-result snippet classification, we employ three different latent user goal models to identify specific latent user goals from the classified snippets. The reason is that the properties of snippets and latent user goals in different categories of snippets are different. Our new model for resource-seeking goals is similar to the previous one, and the other two different models are proposed in this paper.

- *Identification of Resource-Seeking Goal*

We observed that our previous probabilistic inference model performs well in identifying resource-seeking goals. Therefore, we employ a similar model to identify resource-seeking goals. The difference between the previous model and our new model is that the classified snippets are focused on resource-seeking snippets obtained by snippet classification, but not all snippets.

For resource-seeking goals, given a query q, we try to identify the resource-seeking goal (verb-object pairs) g_i from the classified resource-seeking snippets. We predict some probable resource-seeking goal g_i with respect to each classified resource-seeking snippet related to q. The probability of g_i is calculated as

$$P(g_i \mid q) = \sum_{s \in S} P(s \mid q) P(g_i \mid s, q) \quad , \qquad (1)$$

where s is one of the classified resource-seeking snippets S. Because the query q should be covered in the snippet s, Equation (1) can be approximated as

$$P(g_i \mid q) \approx \sum_{s \in S} P(s \mid q) P(g_i \mid s) \quad , \qquad (2)$$

Because the resource-seeking goal g_i is composed of a verb v_j and a noun n_k, and we assume v_j and n_k are independent in order to relax the constraint of v_j appearing prior to n_k in our previous model, Equation (2) can be calculated as

$$P(g_i \mid q) \\ = \sum_{s \in S} P(s \mid q) P(v_j, n_k \mid s) = \sum_{s \in S} P(s \mid q) P(v_j \mid s) P(n_k \mid s) \, , \qquad (3)$$

where $P(s|q)$ is regarded as the weight of each snippet s with respect to the given q and we assume that the relevance of each snippet is the same to q. That is, $P(s|q) = 1 / m$, if there are m snippets in the classified resource-seeking snippets. $P(v_j|s) = Nv_j / Nv_{all}$, where Nv_j and Nv_{all} are the number of verb v_j and all verbs in the snippet s, respectively. $P(n_k|s) = Nn_k / Nn_{all}$, where Nn_k and Nn_{all} are the number of noun n_k and all nouns in the snippet s, respectively. In addition, we only consider nouns with the tag "Na (common noun)" and verbs with the tags "Va (active intransitive verb)", "Vc (active transitive verb)" and "Ve (active transitive verb with sentential object)" according to our previous observation and experiments.

- *Identification of Informational Goal*

As for identifying informational goals, we employ the techniques of phrase extraction adopted in Zeng et al's search-result clustering method [5]. In this paper, we extract representative phrase, which could be unigram or bi-gram, as the query-related topics to describe informational goals. We think that nouns or noun phrases are most suitable to describe the topic of queries, and therefore these phrases are only composed of nouns. For each noun phrase denoted as n_p, there are five features, introduced in [5], *TFIDF* (Phrase frequency/Inverted Document frequency), *LEN* (Length), *ICS* (Intra-Cluster Similarity), *CE* (Cluster Entropy), and *IND* (Phrase Independence), used to determine the importance of n_p. Thus, the importance of n_p is calculated as

$$w(n_p) = \alpha \times TFIDF + \beta \times LEN + \gamma \times ICS + \kappa \times CE + \omega \times IND, \qquad (4)$$

where the value of the coefficients α, β, γ, κ, and ω are determined heuristically.

We regard the top n phrases with manually specified verb "了解" (know) as identified informational goals based on our assumption of the form verb-object pairs for user goals.

- *Identification of Navigational Goal*

As mentioned earlier, information of URL and title was used to calculate relevance score to each snippet. The model of navigational goals is as Equation (5).

$$Naviscore(s) = \tau \times uscore(s) + (1 - \tau) \times tscore(s) \, , \qquad (5)$$

where *Naviscore*(s), the navigational score function of a snippet s, is the combination of the URL score of s, *uscore*(s), and the title score of s, *tscore*(s). τ is assigned 0.7 in this paper by empirical test.

3. Experiments

3.1 Experimental Setup

Data Set: After manually removing pornographic queries, we randomly selected 100 and 300 queries from the top 1000 queries in Dreamer's query log for the experiments of snippet classification and user goal identification, respectively.

Evaluation Metric: Three participants are asked to provide relevance judgments for all experiments. We use precision as metric to evaluate the performance of snippet classification, and user goal identification.

3.2 Experimental Results

3.2.1 Precision of Snippet Classification

- *Effect of thresholds in resource-seeking snippet classification*

Figure 3 shows the precision of snippet classification with different thresholds in resource-seeking category. The threshold is used to determine whether a snippet is regarded as resource-seeking snippet according to the accumulated weights. Since we find only one specific website in navigational snippets to satisfy users' navigational goal, the precision of resource-seeking and informational categories is more critical to the overall performance of user goal identification than that of navigational category. Therefore, 0.017 is an empirically optimal threshold value for the overall performance since the average precision converges to a stable value of 0.66. We think there are two main reasons. First, the set of trained resource-seeking hint verbs is not large enough. Second, there are some resource-seeking snippets not containing hint verbs.

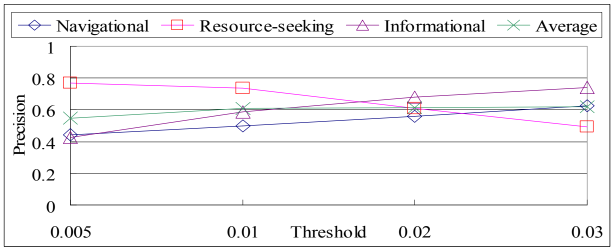


Figure 4. Precision of snippet classification with different threshold of resource-seeking-snippet classification

3.2.2 Performance of User Goal Identification

The comparison of the average precision of four kinds of user goal models is shown in Figure 4. Obviously, if we only compare the precision of resource-seeking goal identification with our previous model, our new model actually performs better. Furthermore, we also identify two other categories of user goals (informational and navigational) ignored by our previous methods. We can see that the identification of informational goals using our proposed new model also perform well. The precision of informational goal identification is better than that of resource-seeking goal identification since suitable informational goals only depend on the nouns or noun phrases but suitable resource-seeking goals depend on both the verbs and the nouns. Thus, for identification of resource-seeking goals, there are more semantic problems, which are still remained to be solved. Navigational goals are very different from other two categories of user goals. In general, users only want to browse a specific web site when they submit a navigational oriented query. Thus, the performance is effective at top-one precision of 0.46.

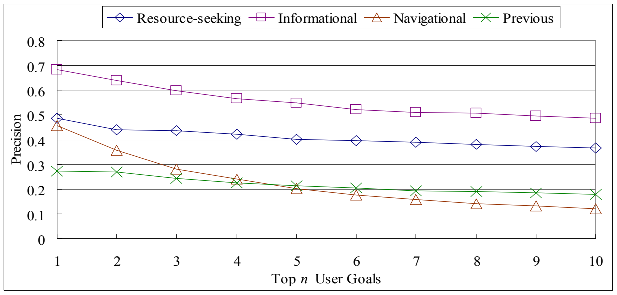


Figure 4. Comparison of precision of top n identified user goals.

- *Examples of Identified Resource-seeking Goals*

Table 2 demonstrates examples of identified resource-seeking goals for the English query "winamp", which is a music player software. The second and third identified user goals, "播放音樂" (play music) and "下載軟件"(download software) show that our methods should be feasible to the resource-seeking oriented queries. As for incorrect user goals such as "下載版本", it is identified due to high occurrence frequency of the verb-noun pair in search-result snippets. Obviously, the semantic correlation between the verb "下載" (download) and the noun "版本" (version) is incorrect.

Table 2. Some correct and incorrect examples of resource-seeking goals of the English query "winamp"

Correct User Goals	Incorrect User Goals
下載遊戲 (download games)	下載版本 (download version)
播放音樂 (play music)	支持音頻 (support channels)
下載軟件 (download software)	下載高速 (download speed)

4. Conclusion

We have presented a new method to improve our previous method which is effective to identify resource-seeking goals, but is not good at identification of navigational and informational goals. Our new method employed a supervised-learning method to learn hint verbs and consider URL information and title information to effectively classify snippets into three categories. In addition, we proposed three different models to effectively identify three different categories of specific latent user goals from classified snippets.

Although it is now unavailable to identify user goals in other languages using our proposed method, in fact, the techniques of identifying latent user goals we proposed in this paper might not be limited in Chinese. Of course, we will have a new challenge to adapt the verb-object pairs to certain languages. For example, object-verb pairs may be more suitable to Japanese.

5. References

[1] A. Broder. *A taxonomy of web search.* SIGIR Forum 36(2), 2002.

[2] U. Lee, Z. Liu and J.Cho. *Automatic Identification of User Goals in Web Search.* In Proceedings of the 14th International Conference on World Wide Web, 2005.

[3] D. E. Rose and D. Levinson. *Understanding User Goals in Web Search.* In Proceedings of the 13th International Conference on World Wide Web, 2004.

[4] Y.-S. Chang, K.-Y. He, S. Yu, & W.-H. Lu (2006). *Identifying User Goals from Web Search Results.* In Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence

[5] H. J. Zeng, Q.C. He, Z. Chen, W. Y. M and J. Ma, *Learning to cluster web search results.* In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004