Learning to Measure Quality of Queries for Automatic Query Suggestion

Xian Chen
Dept. of Internet & Multimedia Eng.
Konkuk University, Seoul Korea
+82-2-450-4071
chenxian@konkuk.ac.kr

Hyoseop Shin
Dept. of Internet & Multimedia Eng.
Konkuk University, Seoul Korea
+82-2-2049-6117
hsshin@konkuk.ac.kr

ABSTRACT

Users tend to use their own terms to search items in structured search systems such as restaurant searches (e.g. Yelp), but due to users' lack of understanding on internal vocabulary and structures, they often fail to adequately search, which leads to unsatisfying search results. In this case, search systems should assist users to use different terms for better search results. To address this issue, we develop a scheme to generate suggested queries, given a user query. At first in doing this, in this paper we propose a scheme to evaluate queries (i.e. user queries and/or suggested queries) based on two measures: 1) if the query will return a sufficient number of search results, 2) if the query will return search results of high quality. Furthermore, we present a learning model to choose among alternative candidate queries against a user query. Our experiments show the proposed method is feasible and scalable.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Algorithms.

Keywords

Quality of queries, query suggestion, learning to measure

1. INTRODUCTION

The generic Internet search does not confront empty search results, such as web search, which always return millions of web pages (e.g. Google). However, due to users' lack of understanding on internal vocabulary and structures, if they use their own terms to search items in controlled vocabulary structure-based search systems, such as restaurant search on Yelp, they may obtain empty, too many or unsatisfying results. In this case, the system should suggest better query terms to users. Although generic search engines (e.g. Google) may also have query suggestion schemes based on most users' co-click frequency, they may not fully utilize controlled and structured domain vocabulary in specific areas (e.g. restaurant search), in which users will expect search results of a higher level than generic search systems provide. To assist users use better search terms, we develop a scheme to suggest alternative queries, given a user's query. For doing this, we evaluate the quality of user's query and/or suggested queries. Hence, we propose two measures: 1) whether the query will return a sufficient number of search results, 2) and/or whether the query will return search results of high quality. Based on the two measures, we build query suggestion and measurement models in one of restaurant search systems in Korea. Furthermore, we present a learning model to select alternative candidate queries against a user query.

2. Query Suggestion

For restaurant search, the most common types of search term are location and food. Figure 1 shows our query suggestion model how to deal with both. According to different inputs, the model suggests different options. First, the model checks the types of search terms: Original query is: 1) single food; 2) single location; 3) food with location. For case 1) and 2), the model has four choices: relax to bigger scope terms (R), change to similar terms (C), tighten (T) to smaller scope ones and do nothing (N); for case 3), since there are four types of actions: N, R, C and T and two types of terms, thus the model have 4^2 options to suggest, for instance, RC means relaxing food with changing location. This scheme is scalable for more types of terms, e.g., n types, 4^n options to suggest.

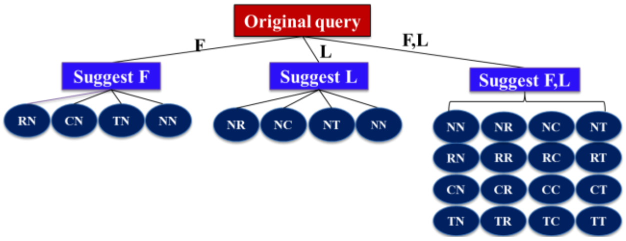

Figure 1. Query Suggestion Processing

In previous work, knowledge-based recommender systems [1] proposed to relax query by removing users' constraints; Query recommender systems [2] provided to build hierarchy structures for each attribute, in order to relax users' conflictive requirements and avoid empty result. However, we build hierarchical structures of location and food, not only for relaxing, but also for tightening and changing. For location, we build hierarchy structures in both vertical and horizontal level. Vertical level hierarchy is used to relax and tighten while horizontal level is used to change with neighbor areas. Meanwhile, we get food hierarchy structure by applying hierarchical clustering for relaxing and tightening. To support changing similar food, we calculate food similarity by building restaurant-food matrix, applying collaborative filtering, non-negative matrix factorization and latent semantic analysis, and we empirically select the best results from three of them.

3. Query Measures

In controlled vocabulary structure-based searches, there are few researches about query measurement. Here, We evaluate queries based on two measures:

- Quantity: The query should return a sufficient number of results. For controlled vocabulary structure-based search, the

RecSys 2014 Poster Proceedings, October 6–10, 2014, Foster City, Silicon Valley, USA.

number of returned results can be neither empty nor too much. Since the number of restaurants diversely distribute in the whole Korea, we can set neither a fixed number nor an interval to measure the quantity of results.

- Quality: The query should return search results with high quality. To define the high quality restaurant (HQ-rest), we collected blogs from NAVER blogosphere in Korea (http://blog.naver.com/) and ranked restaurants according to power bloggers' contribution [3] through social network. However, we cannot use the same criterion to measure HQ-rest in different locations, since both the restaurant distribution and the HQ-rest distribution are unbalanced in the country. (E.g. the high ranked restaurants in capital city (i.e., Seoul) are totally different with those in a small town.) Hence, we define HQ-rest based on different level of areas, such as province-level, city-level, town-level, district-level, valley-level and even street-level. The formula is as follows:

$$\sum_{i=1}^{K} score_i \geq \rho \sum_{r=1}^{R} score_r \qquad (1)$$

We sort R restaurants in one area by ranked scores, we keep on adding from the highest score until the accumulation score is more than some ratio of total sum score, such as ρ, a constant, ranges between (0,1). These top K is selected as this level HQ-rest for the area. In our experiment, we set ρ = 0.5.

4. Learning to Measure

Neither quantity nor quality can be measured by simple threshold or an interval. Therefore, we build a learning model to classify query into good or bad, with six criterions in terms of quantity and quality. The criterions are as follows:

Quantity:

- The number of returned restaurants;
- The number of matched restaurants among top 10 in returned list: matched restaurant is defined as the restaurant, which satisfies user's requirements. (E.g., user inputs "chicken" and try to find fried chicken restaurant, but the returned results include other types of chicken, e.g. chicken soup. In this case, we count the number of fried chicken restaurants as the number of matched restaurants.)

Quality:

- The ranked position of first matched restaurant;
- The maximum score of returned restaurants;
- The number of HQ-rest against area level among returned restaurants;

Others:

- The location level.

We make training data by evaluating 302 queries with human effects. 76 are evaluated as good, the other are as bad. We build learning model for several algorithms. The accuracies of 10-folds cross validation, precision, recall and F-score are shown in Table 1. The average accuracy is more than 77%. We draw the distribution of first matched position with number of HQ-rest for each location level in Figure 2. It indicates: #HQ-rest in high-level location is more than low-level location; the first matched position ranges [1,3] for most of good (red) queries, but we cannot use only this criterion to evaluate query, since between [1,3], first matched positions of bad (blue) queries distribute everywhere. That is why we design several criterions. Learning algorithms with more criterions help us deal with the complicated cases. This ambiguous distribution also encourages us to analyze criterions more detail in the future.

Table 1. Accuracies for machine learning models

Algorithm	**Precision**	**Recall**	**F-score**
Naïve Bayes	58.5%	97.6%	73.1%
SVM	71.4%	47.1%	56.7%
Decision Tree	82.5%	94.1%	87.9%
Neural Network	79.3%	76.5%	77.8%
Random Tree	86.9%	91.3%	89%

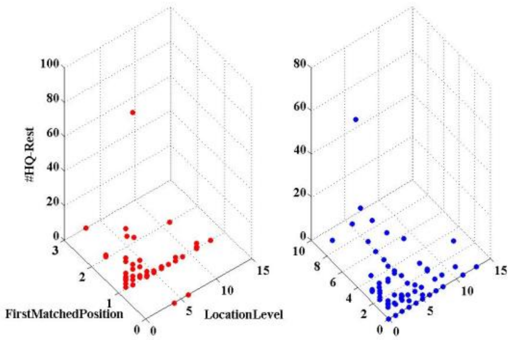

Figure 2. The relationship among three criterions

5. Conclusion

In this paper, to assist users to search better results in restaurant searches, we propose a scheme to suggest alternative queries and present a learning model to measure queries in terms of quantity and quality. Query suggestion can be extended for adding more types of search terms. The query measurement experiments showed that our method is feasible and scalable. Our future work includes: applying probabilistic model to rest-food matrix (e.g. PLSA, LDA), accumulating more training set and applying heuristic, ranking methods to query measurement.

6. ACKNOWLEDGMENTS

This research is supported by Mid-career Researcher Program through NRF grant funded by the MEST (2011- 0029729).

7. REFERENCES

[1] Burke, R., Knowledge-based Recommender Systems. *Encyclopedia of Library and Information Systems* (A. Kent, ed.). 69, 32 (2000).

[2] Mirzadeh, N., Ricci, F. and Bansal, M, support user query relaxation in a recommender system. *Lecture Notes in Computer Science*. 3182 (2004), 31-40.

[3] Shin, H. and Lee, J., Impact and Degree of User Sociability in Social Media. *Information Sciences*. 196, 1 (Aug, 2012), 23-46.