

Quantify Query Ambiguity using ODP Metadata

Guang Qiu, Kangmiao Liu, Jiajun Bu*, Chun Chen, Zhiming Kang
 College of Computer Science
 Zhejiang University
 Hangzhou 310027, China

*Corresponding Author, +86 571 87952148
 {qiuguang,lkm,bjj,chenc,kzm}@zju.edu.cn

ABSTRACT

Query ambiguity prevents existing retrieval systems from returning reasonable results for every query. As there is already lots of work done on resolving ambiguity, vague queries could be handled using corresponding approaches separately if they can be identified in advance. Quantification of the degree of (lack of) ambiguity lays the groundwork for the identification. In this poster, we propose such a measure using query topics based on the topic structure selected from the Open Directory Project (ODP) taxonomy. We introduce clarity score to quantify the lack of ambiguity with respect to data sets constructed from the TREC collections and the rank correlation test results demonstrate a strong positive association between the clarity scores and retrieval precisions for queries.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Query formulation

General Terms

Algorithms, Experimentation

Keywords

ambiguity, quantification, ODP, rank correlation test

1. INTRODUCTION

Query ambiguity has long been of interest in information retrieval. In retrieval systems, words are the only evidence systems have about what user means. However, because of the problem of synonyms and homonyms, one query may contain various topics when no other prior knowledge is available.

Ambiguity of queries prevents retrieval systems returning reasonable results for every query. Failures cause the user to mistrust the system and discontinue use ([5]). Therefore, it is necessary for a retrieval system to maintain consistent performance. As there are already lots of researches done on resolving the ambiguity ([2], [3]), these proposed methods can be employed to handle the ambiguous queries separately if they can be identified in advance. Quantification of query ambiguity lays the groundwork for the identification. Our

present work in this poster provides such an approach to quantifying based on query topics. It differs from approaches proposed in [1] in computing complexity and is much closer to the definition of ambiguity.

The topic structure defined in our poster is extracted from the top 16 categories in the ODP taxonomy (<http://dmoz.org>). ODP has also provided a search service which returns topics for issued queries. We make use of this service to assign topics for each term of a query automatically¹. Clarity score is calculated based on the returned topics to measure the degree of lack of ambiguity. A high clarity score indicates few topics contained in the query while a low score implies multiple topics.

In our experiments, we evaluate our measure for query ambiguity through a rank correlation test with retrieval precisions. Evaluations of query precisions are conducted separately when different models are employed to do the retrieval. The experimental results show that clarity scores of queries have strong positive association with their retrieval performance.

2. QUANTIFY QUERY AMBIGUITY

2.1 Data Sets

We have constructed two data sets in our work: a query set Q consisting of 100 queries which are extracted from the titles of the 100 *topics*² used in the TREC 2003 and 2004 Novelty tracks ([4]); a document collection C composed of the documents provided for each *topic* in the two tracks.

2.2 Obtain Query Topics

Given a query q composed of n terms $\{t_1, t_2, \dots, t_n\}$, each associated with a topic set S_i fetched from ODP search, we suggest that topic of q could be obtained from three different sources: topics that occur in at least two term topic sets in the intersecting set, topics in some term topic set S_k and novel topics which are contained in none of the term topic sets.

A classification on queries in set Q to examine the occurrence distribution of queries of these three types is conducted for validation. We first assign each query the correct topic from the 16 ones. A simple mapping strategy is adopted to do the assignment automatically. Each query (i.e. title of

¹We failed to assign topics for the entire query in most cases (Only 19 out of 100 queries succeeded in getting results using this service).

²"topic" in italic refers particularly to meanings in TREC context.

a *topic* in TREC context) provided by TREC has already been categorized as *Event* or *Opinion*. In our mapping strategy, we categorize a query marked with *Event* or *Opinion* as *News* or *Society* in our topic structure respectively. Results show that in most cases (70%), query topics can be selected from the intersecting term topics (22% and 8% for the other two types respectively). Therefore in our poster, we propose our measure for query ambiguity only for queries whose topics can be identified through that way.

2.3 Quantification Measure

The idea of our measure is quite straightforward that large size of intersecting set implies numbers of potential topics for the query which thus indicates ambiguous. The number of topics in intersecting set is taken as the dominant factor in quantifying query ambiguity.

Given a query q with two terms t_1 and t_2 , we first construct two topic sets for these two terms through the ODP search, named as $\{topics\}_1$ and $\{topics\}_2$ respectively. Intersection of these two sets makes $\{topics\}_{intersect}$:

$$\{topics\}_{intersect} = \{topics\}_1 \cap \{topics\}_2 \quad (1)$$

For queries consisting of more than two terms, topics in the intersecting set (i.e. $\{topics\}_{intersect}$) are those being contained in at least two term sets. Consequently, we define the lack of ambiguity of the query q according to the number of topics in $\{topics\}_{intersect}$. Specifically, we use clarity score to quantify the lack of ambiguity i.e. the larger the clarity score of a query holds, the less ambiguous the query is. Clarity score is calculated as follows:

$$clarity\ score = F(|\{topics\}_{intersect}|) \quad (2)$$

It follows the relation that clarity score decreases as the size of the set $\{topics\}_{intersect}$ increases. Different functions $F(x)$ s can be adopted to describe this kind of relationship, such as the naive one $F(x) = 1/(x + 1)$ as adopted in our experiment. In fact, as we do evaluations using the Spearman rank correlation test which is irrelevant of exact scores, a function is appropriate as long as it is capable of giving the ranks.

3. EXPERIMENTS

In our experiment, we employ the same strategy as in [1] to estimate the association between clarity scores and retrieval performance in retrieval scenario. We measure the rank correlation between clarity score ranking and precision ranking for queries from two query sets (the entire query set Q of 100 queries and a subset Q' of 70 queries whose topics come from the intersecting set of term topics) respectively. The retrieval is done with three different retrieval models implemented in the Lemur toolkit (<http://www.lemurproject.org>) with respect to documents in the collection C . Result relevance is judged using the judgment files provided by TREC.

3.1 Results and Analysis

Table 1 shows correlation values on the two query sets with precision rankings derived from different retrieval models. The results overall demonstrate a strong positive association between clarity scores and precisions of queries. For example, an average correlation of 0.805 between the clarity scores and precisions is achieved for query set Q' , which shows perfect agreement in these two rankings. The lowest ranking correlation value 0.594 is obtained for queries in Q

Table 1: Correlation values between clarity scores and precisions in query sets Q and Q' .

Collections	TFIDF	Okapi	KL.dir	Average
Query Set Q	0.594	0.596	0.597	0.596
Query Set Q'	0.797	0.809	0.808	0.805

when the retrieval task is done using TFIDF retrieval model. However, it also shows positive association between the two rankings according to the declaration of the Spearman rank correlation test.

Another observation is that correlation values for set Q' are higher than those for set Q . The reason is that topics of some queries in Q can not be derived from the intersecting set of terms and thus our quantification measure for the ambiguity is no longer suitable for those queries. However, correlation value of all queries (i.e. the set Q) remains encouraging (0.596) which indicates our measure for ambiguity is reasonable for general queries.

4. CONCLUSIONS

In this poster, we have proposed a measure to quantify the query ambiguity based on the topic structure selected from the ODP taxonomy. We suggest and verify an assumption that topics of a query can be derived from three different sources. Our quantification measure is proposed for queries whose topics are from the intersecting topic set of terms. Although the measure is suggested for specific queries, experiments demonstrate encouraging results for general queries. The rank correlation test of clarity scores and precisions shows a strong positive association between these two rankings which indicates that our measure for ambiguity quantification is reasonable.

5. REFERENCES

- [1] S. Cronen-Townsend and W. B. Croft. Quantifying query ambiguity. In *Proceedings of Human Language Technology*, pages 94–98, 2002.
- [2] M. Sanderson and K. van Rijsbergen. The impact on retrieval effectiveness of skewed frequency distributions. *ACM Transactions on Information Systems*, 17(4):440–465, 1999.
- [3] H. Schutze and J. Pederson. Information retrieval based on word senses. In *Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175, 1995.
- [4] I. Soboroff. Overview of the trec 2004 novelty track. In *Proceedings of the Thirteenth Text REtrieval Conference (TREC 2004)*, NIST Special Publication 500-261, 2004.
- [5] E. Voorhees. Overview of the trec 2003 robust retrieval track. In *Proceedings of the Twelfth Text REtrieval Conference Proceedings (TREC 2003)*, NIST Special Publication 500-255, 2003.