Querying by examples

Arlind Kopliku — Mohand Boughanem — Karen Pinel-Sauvagnat

Institut de Recherche en Informatique de Toulouse, UMR 5505 CNRS, SIG-RFI, 118 route de Narbonne F-31062 Toulouse Cedex 9, France
Arlind.Kopliku@irit.fr, Karen.Sauvagnat@irit.fr, Mohand.Boughanem@irit.fr

1. Introduction

It is common for humans to identify some content by listing examples of similar content. Some movies of a Tarantino, a movie producer can be used to identify more movies of the same producer. Querying by examples is an alternative way of querying which allows to identify more content as well as to expand knowledge. We experiment this approach over a noisy collection of extracted lists from the Web.
Querying by examples demands set expansion. This differs from previous work as we do not focus on the expansion process, rather than on the query by examples approach. We use a collection of candidate expansion sets, while we focus on the number of queries that can be answered and the quality of results.
Query expansion is a known domain(Wang *et al.,* 2008). Some consider named entities as content can be grouped into well defined categories(Pantel *et al.,* 2009, Sekine *et al.,* 2008). Google Sets and Google Squared are examples that use query expansion. Given a number of seed examples $s_1, s_2, ...s_k$, the goal is an extended complete set $o_1, o_2, ...o_l$, where $s_1, s_2, ...s_k \in o_1, o_2, ...o_l$. If the candidate sets are known apriori, the task can be simplified to a simple matching process which consists in testing whether the query items belong to the candidate sets.
We consider as quality candidate sets, sets with items of the same type, which we also refer to as quality sets. To extract this sets, we use HTML lists. Although HTML lists contain many quality sets, the We show that there are many of them in HTML lists. Still, HTML lists represent a noisy collection.
The more candidate sets, the more example based queries can be satisfied. Still, noisy sets without elements of the same type are not good matches. It is important to measure how sensible is the query by examples to noise.

2. Dataset

Our dataset consists of a filtered subset of HTML lists extracted from about 100,000 french pages of Exalead search engine index.
HTML lists represent a huge collection over the Web. We estimate to have about 3.4 lists per page. Initially we filtered out potentially noisy lists such as lists with one element and lists with empty elements. A subset of 2000 lists was assessed by human judges. 8.22% of the lists was judged as qualitative, where a quality list is intended to be a list of named entities of the same type.

3. Observations and results

A quality set S of n items has $C_{n,l}$ subsets of l items. This means that there are totally $C_{n,1} + C_{n,2} + ... + C_{n,n-1} = 2^n - 2$ subsets. Each of the subsets is a good query for S. For a list of 40 items there are about $2^{40} - 2 \approx 10^{12}$ good queries.
The more candidate sets there are, the more queries can be satisfied. But, introducing sets which do not contain items of the same type might reduce chances to find a good match. We try to evaluate if a choice of good examples avoids wrong matches.
We take subsets from quality sets and we treat them as queries. These queries match at least with their source set. Even if they are not issues from real users, they are legitime queries for set expansion. We can call them good queries.
For good queries with two items, there were no matches in non qualitative sets. It is clear that for longer queries there will be no matches, too. On the other hand, good queries had sometimes multiple matches within quality lists. This means that they match at least one more quality set, except the source set.
We conclude that a good query is more probable to match with quality sets, rather than non quality sets. Thus, querying by examples works well even in the presence of many non qualitative lists. The approach itself filters out bad matches.
For future work, we condider repeating the same experiment with a bigger collection of quality sets. At the same time, it might be interesting to use queries issued by human users instead of artificially chosen example queries.

4. References

Pantel P., Crestan E., Borkovsky A., Popescu A.-M., Vyas V., « Web-Scale Distributional Similarity and Entity Set Expansion », *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Singapore, p. 938-947, 2009.

Sekine S., Sudo K., Nobata C., « Extended Named Entity Hierarchy », *Proceedings of LREC 2002*, 2008.

Wang R. C., Cohen W. W., « Iterative Set Expansion of Named Entities Using the Web », *ICDM '08: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, IEEE Computer Society, Washington, DC, USA, p. 1091-1096, 2008.