QUEST: Query Expansion Using Synonyms over Time*

Nattiya Kanhabua and Kjetil Nørvåg

Dept. of Computer Science
Norwegian University of Science and Technology
Trondheim, Norway

Abstract. A particular problem of searching news archives with named entities is that they are very dynamic in appearance compared to other vocabulary terms, and synonym relationships between terms change with time. In previous work, we proposed an approach to extracting time-based synonyms of named entities from the whole history of Wikipedia. In this paper, we present QUEST (Query Expansion using Synonyms over Time), a system that exploits time-based synonyms in searching news archives. The system takes as input a named entity query, and automatically determines time-based synonyms for a given query wrt. time criteria. Query expansion using the determined synonyms can be employed in order to improve the retrieval effectiveness.

1 Introduction

News archives are publicly available nowadays, e.g., Google News Archive and The Times Online. Nevertheless, searching for information in such resources is not straightforward because their contents are strongly time-dependent. To increase precision, a user can narrow down search results by extending query keywords with the creation or update date of documents (called temporal criteria). Two ways of obtaining temporal criteria relevant to a query are 1) having them provided by the user [1,6], or 2) determined by the system [5]. One way of increasing recall is to perform query expansion using synonyms. However, when queries are named entities (people, organizations, locations, etc.), a problem of expanding the queries is the effect of rapidly changing synonyms[1] over time, e.g., changes of roles or alterations of names. For example, "Cardinal Joseph Ratzinger" is a synonym of "Pope Benedict XVI" before 2005, and "United States Senator from New York" is a synonym of "Hillary R. Clinton" between 2001 and 2008. Instead of referring to a synonym alone, we have to always refer to an entity-synonym relationship because a term can be a synonym of one or more entities. In this paper, we present QUEST (Query Expansion using Synonyms over Time), a system that exploits changing synonyms over time in searching news archives. To the best of our knowledge, this has never been done before in the existing news archive search systems. Our system consists of two parts: 1) the offline module for extracting time-based synonyms as depicted in Fig. 1, and 2) the online module for searching news archive as

* This work has been supported by the LongRec project, partially funded by the Norwegian Research Council.

[1] In general, synonyms are different words with very similar meanings, but in our context synonyms are name variants (other names, titles, or roles) of a named entity.

J.L. Balcázar et al. (Eds.): ECML PKDD 2010, Part III, LNAI 6323, pp. 595–598, 2010.

illustrated in Fig. 2. With a web-based interface, the system can take as input a named entity query. It automatically determines time-based synonyms for a given named entity, and ranks the synonyms by their time-based scores. Then, a user can expand the named entity with the synonyms in order to improve the retrieval effectiveness.

Our news archive search system is mainly driven by entity-synonym relationships, which can be automatically created based on the whole history of Wikipedia. Evolving relationships are detected using the most current version of Wikipedia, while relationships for particular time in the past are discovered through the use of snapshots of previous Wikipedia versions. Using our approach, future relationships with new named entities can be also discovered simply by processing Wikipedia as new contents are added. Further, we employ the New York Times Annotated Corpus[2] in order to extend the covered time range as well as improve the accuracy of time of synonyms. The rest of the paper is organized as follows. In Sect. 2, we describe an approach to extracting synonyms from Wikipedia, and ranking synonyms based on their temporal characteristic. In Sect. 3, we outline the online search system and our proposed demo.

2 Extracting Time-Based Synonyms from Wikipedia

We extract entity-synonym relationships in an offline manner as depicted in Fig. 1. We downloaded the complete dump of English Wikipedia from the Internet Archive[3], which is composed of all pages and all revisions. Each revision of a page has the time period that it was in use before being replaced by the succeeding version. In other words, the associated time of a revision is the period when it was a current version.

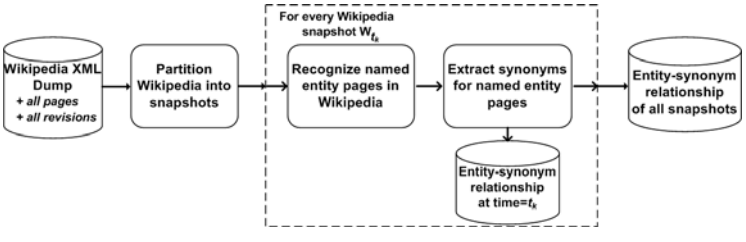

Fig. 1. Extracting time-based synonyms from the history of Wikipedia

First, we partition Wikipedia into snapshots $\{W_{t_1}, \ldots, W_{t_z}\}$ with *1-month* granularity. For each Wikipedia snapshot W_{t_k}, we identify all named entities in the snapshot W_{t_k} using the approach described by Bunescu and Paşca in [3]. After identifying an entity page p_e from a snapshot W_{t_k}, we will have a set of entity pages $\mathcal{P}_{e,t_k} = \{p_e | p_e \in W_{t_k}\}$. From this set, we will create a set of named entities E_{t_k} at time t_k by simply extracting a title from each named entity page $p_e \in \mathcal{P}_{e,t_k}$. For each named entity in E_{t_k}, we will find synonyms by extracting anchor texts from article links, as described by Bøhn and Nørvåg [2]. For a page $p_i \in W_{t_k}$, we extract all internal links in

[2] http://www.ldc.upenn.edu/Catalog/docs/LDC2008T19/new_york_times_annotated_corpus.pdf

[3] http://www.archive.org/details/enwiki-20080103

p_i, but only those links that point to an entity page $p_e \in \mathcal{P}_{e,t_k}$ are interesting. In other words, the system extracts as synonyms all anchor texts for the associated entity, and these synonyms are weighted by their frequencies of occurrence. We then obtain a set of entity-synonym relationships. By accumulating the set of entity-synonym relationships from every page $p_i \in W_{t_k}$, we will have a set of entity-synonym relationships at time t_k, i.e., a synonym snapshot $S_{t_k} = \{\xi_{1,1}, \ldots, \xi_{n,m}\}$. Named entity recognition and synonym extraction steps are processed for every snapshot W_{t_k}. Finally, we will have obtained the set of entity-synonym relationships from all snapshots $\mathbb{S} = \{S_{t_1}, \ldots, S_{t_z}\}$, and the set of synonyms for all entities $\mathcal{S} = \{s_1, \ldots, s_y\}$. Note that, the time periods of synonyms are timestamps of Wikipedia articles in which they appear, not the time extracted from the contents. To discover the more accurate time, we need to analyze a document corpus with the longer time period, i.e., the New York Time Annotated Corpus. Due to the size limitation of the paper, the reader can refer to [4] for more detail about improving time of entity-synonym relationships. Given a named entity e_i and temporal criteria $[t_a, t_b]$, we can retrieve a set of synonyms of e_i wrt. $[t_a, t_b]$ from $\mathbb{S}$. The synonyms can be ranked by time-based scores defined as a mixture model of a temporal feature and a frequency feature as follows.

$$\mathit{TB}(s_j, [t_a, t_b]) = \mu \cdot pf(s_j, [t_a, t_b]) + (1 - \mu) \cdot \overline{tf}(s_j, [t_a, t_b]) \quad (1)$$

where $pf(s_j, [t_a, t_b])$ is a time partition frequency or the number of time partitions (or time snapshots) in which a synonym s_j occurs within $[t_a, t_b]$. $\overline{tf}(s_j, [t_a, t_b])$ is an averaged term frequency of s_j in all time partitions within $[t_a, t_b]$, $\overline{tf}(s_j, [t_a, t_b]) = \frac{\sum_{t_i \in [t_a, t_b]} tf(s_j, p_{t_i})}{pf(s_j, [t_a, t_b])}$. μ underlines the importance of a temporal feature and a frequency feature, and μ=0.5 gave the best performance in our experiments.

3 Online Demo

The time-based synonyms extracted using our approach can be applied to any news archive collection. In this demo, we use the New York Times Annotated Corpus as an illustrative example of such a news archive. This collection contains over 1.8 million articles from January 1987 to June 2007. We use the enterprise search platform Solr from Apache Lucene. The system screenshots are shown in Fig. 2, and the online demo is publicly available at `http://research.idi.ntnu.no/wislab/quest/`. In this demo, we find over 2.5 million named entities and 3 million entity-synonym relationships. Given a query q and temporal criteria $[t_a, t_b]$, the system has to verify whether q is a named entity. We do this by searching Wikipedia with q, and the first page in the result list will be used as the associated named entity for q. Subsequently, the system determines synonyms for the associated named entity, and the user can select synonyms to expand the original q in order to improve the retrieval effectiveness. In addition, the user can choose whether to show time periods and scores associated to synonyms. In the following, we will give two search scenario as examples.

First scenario: A student studying the history of the Roman Catholic Church wants to know about the Pope Benedict XVI during the years before he became the Pope (i.e. before 2005). The student searches using the query "Pope Benedict XVI" and the publication dates "01/1987" and "04/2005". The system retrieves documents for the

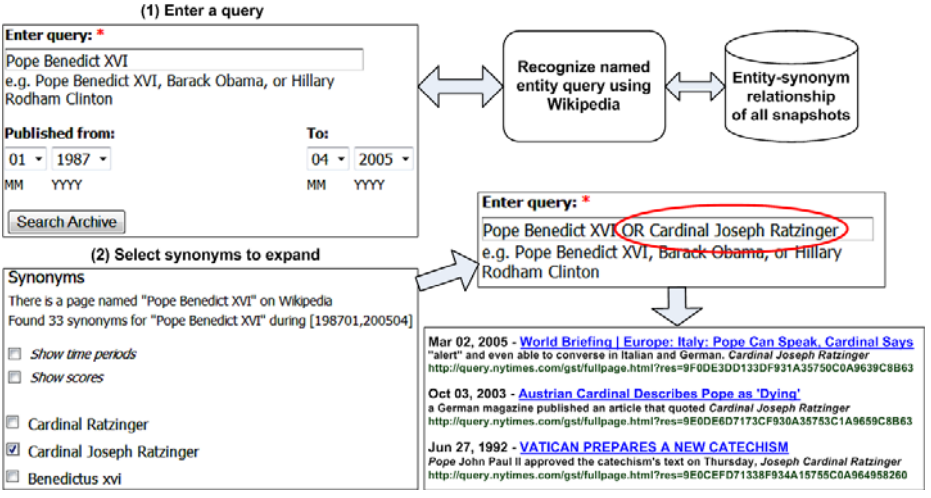

Fig. 2. QUEST online demo at http://research.idi.ntnu.no/wislab/quest/

query "Pope Benedict XVI", and also determines synonyms for the query wrt. time criteria. The student then selects the synonyms "Cardinal Joseph Ratzinger" to expand the query. The new query becomes "Pope Benedict XVI *OR* Cardinal Joseph Ratzinger". He performs search again, and the system retrieves documents which are relevant to both "Pope Benedict XVI" and "Cardinal Joseph Ratzinger".

Second scenario: A marketing journalist wants to search for past information about Kmart, or a chain of discount department stores in the United States. She enters the query "Kmart" and the publication dates "01/1987" and "01/2000". The system retrieves documents for the query "Kmart", and also determines synonyms for the query wrt. time criteria. She selects the synonyms "Kresge" to expand the query (Kmart was founded as the S. S. Kresge Company in 1899, and it was named to Kmart in 1962.). The new query becomes "Kmart *OR* Kresge". She performs search again, and the system retrieves documents which are relevant to both "Kmart *OR* Kresge".

References

1. Berberich, K., Bedathur, S.J., Neumann, T., Weikum, G.: A time machine for text search. In: Proceedings of SIGIR 2007 (2007)
2. Bøhn, C., Nørvåg, K.: Extracting named entities and synonyms from wikipedia. In: Proceedings of AINA 2010 (2010)
3. Bunescu, R.C., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of EACL 2006 (2006)
4. Kanhabua, N., Nørvåg, K.: Exploiting time-based synonyms in searching document archives. In: Proceedings of JCDL 2010 (2010)
5. Kanhabua, N., Nørvåg, K.: Determining Time of Queries for Re-ranking Search Results. In: Proceedings of ECDL 2010 (2010)
6. Nørvåg, K.: Supporting temporal text-containment queries in temporal document databases. Journal of Data & Knowledge Engineering 49(1), 105–125 (2004)