Intelligent Time-Aware Query Translation for Text Sources

Amal Kaluarachchi[1], Aparna S. Varde[1], Jing Peng[1], Anna Feldman[2,1]

1. Department of Computer Science, Montclair State University, Montclair. NJ, USA
2. Department of Linguistics, Montclair State University, Montclair. NJ, USA

(kaluarachca1@mail.montclair.edu, vardea@montclair.edu, pengj@montclair.edu, feldmana@montclair.edu)

Abstract

Time-stamped documents such as newswire articles, blog posts and other web-pages are often archived online. Since these archives cover long spans of time, the terminology in them could undergo significant evolution. In answering user queries over such text, it is desirable that the system be intelligent enough to incorporate historical information. For example, a query on Sri Lanka should automatically retrieve documents with its former name Ceylon. Hence, temporal terminology evolution needs to be taken into account to translate these queries. This has become vital today because users expect that computer systems have the intelligence to find all related information pertaining to their queries. In this research we attempt to discover such concepts that evolve over time and use those discovered concepts to provide time-aware responses to user queries. Our solution and evaluation are summarized in the paper.

Problem Definition

We call the evolving concepts *SITACs*, i.e., Semantically Identical Temporally Altering Concepts. Examples of SITACs include:

Person: Hillary Clinton ->Hillary Rodham@1974, First Lady@1993-2001
Place: Sri Lanka -> Ceylon @ 1972
United States ->Union during civil war

The goal of this work is that queries involving SITACs should be translated and answered appropriately, e.g., "When did Sri Lanka get its independence?" (Sri Lanka received independence in 1948 when it was called Ceylon, and the system must have this knowledge in providing the answer.) Clearly, this is not just an issue of synonymy.

Proposed Solution

In our research, the focus in on finding a method to discover SITACs from a given text corpus using linguistic properties of concepts (i.e verb, noun, subject, object etc.), putting them in data sets with different matrices followed by extensive data mining tasks. Concepts involved in this context have two categories: 1) concepts in anticipated queries (supervised approach) and 2) concepts in any generic query (unsupervised approach).

Association rule mining is the primary focus of the proposed solution because we try to simulate the manner in which humans mentally associate concepts that evolve over time. Once the rules are identified, similarity and correlation measurements such as Jaccard's coefficient are used to rank the rules. The proposed solution in this research thus includes the following.

1. Discovering knowledge in the form of SITACs in text archives with the following subtasks:
   a. Preparing datasets from text archives, a challenging task involving linguistics, especially the appropriate harnessing of natural language processing techniques
   b. Defining adequate transactions to capture the essence of the problem and deriving association rules on the generated data sets which is also a non-trivial task
2. Answering user queries in an intelligent manner using temporal knowledge of concepts obtained on discovering SITACs, thereby developing a query translation engine.
3. Ranking the responses to user queries appropriately incorporating factors that human users would consider.

Prior work in this area has been done by our colleagues [4,6] assuming anticipated queries found suitable at the initial stage of the research. Our further research covers the anticipated queries as well as general queries. We also work on implementation of the SITAC approach to develop a query translation engine for resolving time-aware queries. Our aim is to discover rules of the type (C1, t1) => (C2, t2) where C1 and C2 are concepts and t1 and t2 are corresponding time stamps for C1 and C2. The basic logic proposed in this work is explained as follows.

A transaction defined for the purpose of association rule mining in this problem consists of two or more concepts (as identified by nouns) {C1, C2 … Cn} that are referred to by any common event E occurring (as identified by verbs). Based on this linguistic relationship of concepts that we propose in this research, the following data sets will be generated from the text archive.

a. {EVENT, YEAR1,YEAR2….,,YEARn}
YEAR1.. YEARn will have concepts that appeared in the text archive associated with the events listed under the EVENT attribute

b. {CONCEPT, YEAR1,YEAR2….. YEARn}
YEAR1.. YEARn will have events that appeared in the text archive associated with the concepts listed under the CONCEPT attribute

c. {CONCEPT, YEAR1,YEAR2….. YEARn}
YEAR1.. YEARn will have bi-grams that appeared in the text archive associated with the concepts listed under the CONCEPT attribute

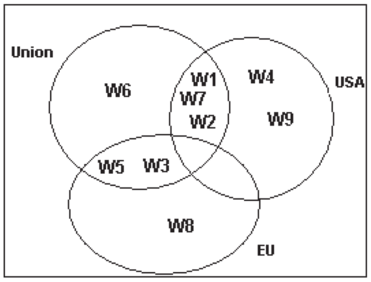
Copyright © 2009, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this research we are interested in exploiting the relationships in different part of speech. By doing that we find events (verb) used by concepts (nouns).

In order to improve the quality of the results obtained from mining the rules, we propose to rank the rules based on the Jaccard's coefficient for similarity. On studying the literature, e.g., [6], this was found to be very effective in capturing contextual similarity. In our problem, we deploy this as follows. Consider that we have rules such as:

Union => USA and Union => EU

To apply Jaccard's coefficient, we do the following, with reference to Figure 1.



W1="accession", W6="government" W7="blessed" and there is long list of words. (W1..Wn).

Figure 1: Using Jaccard's coefficient for similarity

Based on this, Jaccard's coefficient similarity score JS can be calculated as follows.

JS(Union,USA) = (Union $\cup$ USA) / (Union $\cap$ USA)

Thus, JS (Union, USA) = 3/8 = 0.38,

and JS (Union, EU) = 2/8 = 0.33

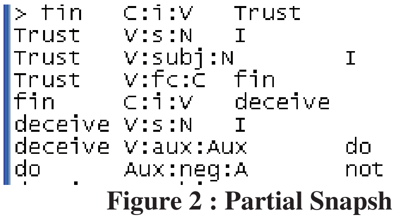
We find that JS (Union, USA) > JS (Union, EU)

Thus, (Union, USA) has a higher rank than (Union, EU)

This final similarity scores of all SITACs derived from association mining rules will be stored in a knowledge base to resolve future queries.

Experimental Evaluation

We summarize our experimental evaluation using the Gutenberg corpus with the USA presidential speeches. We preprocessed this data using a Java program developed by us and sub documents were generated based on the time they were produced. Those set of documents were syntactically parsed with Minipar (References) to obtain the transactions as we defined. Figure 2 shows a sample output from the Minipar parser used in our work.



```
> fin    C:i:V    Trust
Trust    V:s:N    I
Trust    V:subj:N          I
Trust    V:fc:C   fin
fin      C:i:V    deceive
deceive  V:s:N    I
deceive  V:aux:Aux         do
do       Aux:neg:A         not
```

Figure 2 : Partial Snapshot of Parsing

Then we programmatically analyzed the output from Minipar to generate datasets which were used as the input to the widely-used data mining tool WEKA, i.e., the Waikato Environment for Knowledge Analysis, which includes the code for implementing the Apriori algorithm for association rule mining. The rules obtained were subject to ranking using the method we described.

After ranking, we got a set of rules as summarized in the sample shown in Figure 3. These rules were used to populate a knowledge base consisting of SITACs to serve as the basis for intelligent time-aware query translation.

1. 1795=Union ==> 1958=United States
2. 1872= Union ==> 1995=United States
3. 1958= Nation ==> 1999= United States
4. 1995=work ==> 1999=teacher
5. 1895=Administration 1999=teacher ==> 1958=work
6. 1952=war ==> 1999=terrorist
7. 1952=war 1952= weapon ==> 1999=terrorist

Figure 3: Sample of Rules for Knowledge Base

Related Work

WordNet [7] provides correlated concepts but without the temporal factor. We can use inputs from this for our supervised approach with anticipated queries [4,6]. There are contextual similarity measures such as SimRank [2], named entity recognition methods, e.g., [1] and inter-transaction associations on temporal document collections as in [3]. We can draw an analogy here, though we make an additional contribution in terms of intelligent query processing by the discovery of semantically identical temporally altering concepts for historical time-aware query translation.

.

Conclusions

The artificial intelligence contributions of this research are:

1. Identifying the problem of time-aware query translation in text archives and defining SITACs with the goal of intelligent query processing.
2. Proposing a methodology to discover the SITACs by simulating human thinking based on an integration of natural language processing and association rule mining addressing its non-trivial subtasks.
3. Ranking the SITACs exploiting linguistic properties and Jaccard's coefficient and developing a knowledge base of ranked SITACs to serve as the basis for answering user queries in an intelligent manner.

Acknowledgements

We thank Srikanta Bedathur and Gerhard Weikum from the Max Planck Institute for Informatics, Germany and. Debjani Roy from Computer Science at Montclair State University for their inputs.

References

[1] Hasegawa, T., Sekine S.,Grishman R., "Discovering Relations among Named Entities from Large Corpora",ACL (Aug 2004), pp. 415-422.

[2] Jeh., G., Widom., J.: "SimRank: A Measure of Structural-Context Similarity". KDD (Jul 2002), pp. 538–543.

[3] Norvag, K., Eriksen, T.O. , Skogstad, K.I : "Mining Association Rules in Temporal Document Collections", Dept. of Computer and Information, Systems (2006), NTNU, Norway.

[4] Roychoudhury D.., Varde A., "Terminology Evolution in Web and Text Mining Using Association Rules", Dept. of Computer Science (May 2009), Montclair State University, NJ.

[5] Strehl A. Ghosh, J. and Mooney R.: "Impact of Similarity Measures on Web-page Clustering", AAAI, (Jul 2000), pp. 58-64.

[6] Varde A., Bedathur S., Berberich K, Weikum G., "Time Aware Query Translation over Text Archives", Max Planck Institute for Informatics (Jul 2008), Saarbrucken, Germany.

[7] WordNet http://wordnetweb.princeton.edu/perl/webwn