SITAC: Discovering Semantically Identical Temporally Altering Concepts in Text Archives

Amal Kaluarachchi¹, Debjani Roychoudhury¹, Aparna S. Varde¹, Gerhard Weikum²
1. Department of Computer Science, Montclair State University, Montclair NJ, USA
2. Databases and Information Systems Group, Max Planck Institut für Informatik, Saarbrücken, Germany
(amilkal@hotmail.com, debjani23@gmail.com, vardea@montclair.edu, weikum@mpi-sb.mpg.de)

ABSTRACT
This paper demonstrates a system called SITAC based on our proposed approach to automate the discovery of concepts (called SITACs) in text sources that are identical semantically but alter their names over time. This system is developed to perform time-aware translation of queries over text corpora by incorporating terminology evolution, thus providing more accurate responses to users, e.g., query processing on Mumbai should automatically take into account its former name Bombay. The SITAC system constitutes a novel collaborative framework of natural language processing, association rule mining and contextual similarity.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications - data mining
General Terms: Design, Experimentation, Human Factors
Keywords: Association rules, information retrieval, query processing, ranking, temporal changes, text mining, web search

1. INTRODUCTION
Various time-stamped documents, e.g., weblogs, news articles and other pages are often found archived online. When these archives cover long time spans, the terminology in them could undergo significant changes. Thus, if users pose queries pertaining to historical information over such documents, the queries should be translated incorporating temporal changes, in order to provide accurate responses. We give some examples of such queries.
1. What has been the USA policy on the UK over 200 years?
2. How has the United States addressed Native American issues?
3. When did the USA have the minimum number of states?
4. Which Prime Minister of the UK has been most influential on American matters?
5. Who was the longest reigning Indian Prime Minister?

The answers to such queries can be found from various sources, e.g., the speeches of many presidents of the USA who address domestic and foreign policy matters. However, some of the former American presidents use other terms such as the Union when referring to the USA. Some of them also refer to Native Americans as Indians, obviously not the same as the term Indian when referring to the Prime Minister of India. Another example is the use of the terms Great Britain or the British Isles when referring to the UK. For instance, it is found that if Query 1 is executed on Google, the results contain documents with terms USA, UK or 200 years. Unless there is a document worded similar to USA foreign policy we do not get the exact answer even though that information is available in some article(s). The problem addressed in this paper thus goes beyond correlated terms in queries. We add significant changes. Thus, if users pose queries pertaining to American history, it is known that when some of the erstwhile presidents mentioned the Union, they meant the USA, as of today. We refer to such terms as SITACs, i.e., Semantically Identically Temporally Altering Concepts [6]. Examples of SITACs are: Person: Agnes Gonxha Bojaxhiu, Mother Teresa
Place: Ceylon, Sri Lanka; Kalikata, Calcutta, Kolkata

We present a solution to this problem of SITAC discovery, developing a system by the same name SITAC which is demonstrated in this paper. The SITAC system helps to answer queries pertaining to historical data with evolving terminology. The focus of SITAC to discover rules of the type \((C_1, T_1) \Rightarrow (C_2, T_2), \) i.e., concept \(C_1\) at time \(T_1\) implies concept \(C_2\) at time \(T_2\), to serve as the basis for time-aware query translation. This system is based on a novel solution approach that involves an integration of natural language processing, association rule mining and contextual similarity [5, 6]. It intelligently simulates human thinking for query processing, because humans intuitively tend to associate concepts that are semantically identical when answering questions about such temporally altering terms [6]. This SITAC system is briefly described below.

2. THE SITAC SYSTEM
The system architecture of SITAC is shown in Figure 1.

Figure 1: SITAC System Architecture

It has four modules: Concept Extraction, Rule Derivation, Ranking of Rules and Query Translation, as summarized next.
2.1 Module for Concept Extraction

In this part of the system, the text archives are preprocessed to extract information in the form of concepts in documents over time related by events. Thus, we mainly get the following:

- **Document**: Text source $D$ with time-stamp $T$
- **Concept**: Individual term $C$ (word or phrase)
- **Event**: The event $E$ relating concepts

This module of the system extracts concepts from time-stamped documents in the text archives. The concepts are primarily nouns and noun phrases referring to various entities such as person, place and organization. The events correspond to common verbs referred to by the concepts. This involves natural language processing exploiting semantic features in preparation of the data set for deriving rules [5]. More will be explained in the demo.

2.2 Module for Rule Derivation

This module performs association rule mining discovering rules of the type $(C_1,T_1)\Rightarrow(C_2,T_2)$ from corresponding time-stamped documents. The classical *Apriori algorithm* was found to be a well-suited for this purpose. Our transaction set for rule mining is built based on using linguistic properties such as subject, object, noun and verb. If an event is referred to by two distinct nouns, and such events occur multiple times, then we consider that the corresponding nouns (concepts) are related. This can be further explained using set theories. Consider the sets, $E$ for common event $E$ occurring (as identified by verbs). Based on this linguistic relationship of concepts we propose in this research, the following data sets that are generated from the text archives $\{EVENT, TIME1, TIME2, \ldots, TIME_n\}$ where $TIME_1 \ldots TIME_n$ have concepts that appear in the archives associated with events. These are then used to retrieve appropriate documents from a given corpus that contain the given and found words.

2.3 Module for Ranking of Rules

Once the SITACs are discovered, this module finds how strongly they are related. Thus, it serves to give a measurement to the temporal relationships captured by SITACs. Among many existing similarity measures, we select *Jaccard’s coefficient* as it is found to be the most useful in capturing contextual similarity, based on a literature survey e.g., [11]. In our problem, we employ this as follows. For two relationships, $R_1\{(Cx,Ts)\}$, $(Cy,Tt)\}$, $R_2\{(Cx,Ts)\}$, $(Cy,Tu)\}$, we count the other words (nouns, verbs, adjectives, etc) that are used with the concepts $Cx$, $Cy$, and $Cz$. As per Jaccard’s coefficient, we calculate the score for similarity $J$ as $J(Cx,Cy) = \frac{|Cx \cap Cy|}{|Cx \cup Cy|}$. We then proceed to rank the SITACs. Ranked SITACs are stored in a SITAC database.

2.4 Query Translation Module

This module works as follows. SITACs have been filtered by ranking and stored in database with some linguistic knowledge incorporating all parts of the speech with their time-stamps acquired during text parsing. The following piece of SQL code shows the example of the SITAC storage.

```sql
TBL_word : (time, word)
TBL_SITAC : (word, time, SITAC)
/* Find SITACs */
SELECT * FROM TBL_SITAC WHERE WORD = $W1 // $W1 is SITAC
/* Find Documents */
SELECT * FROM TBL_word
where word = $W1 AND
```

When a user enters a query, it goes through a parser and the system stores all words in an array after eliminating stop words (the, an) and common words (I, We). The query on $(SWI \ldots SWn)$ is translated to SQL as shown here. First the system checks for SITACs from TBL_SITAC. If any SITACs are found they are included into the list of words and then the TBL_Word file is searched which stores all words from the corpus with their time stamps. All words from user queries and associated SITACs of those words retrieved from the SITAC table are found from TBL_word. These are then used to retrieve appropriate documents from a given corpus that contain the given and found words.

3. SYSTEM DEMONSTRATION

We provide the demonstration of our SITAC system presenting various snapshots. A few of these are shown below while more will be available in a live demo.

![Figure 2: The SITAC Generator Interface](image-url)
object. The populated database is used to create data set which is then sent to the widely used WEKA data mining tool to extract SITACs using the Apriori algorithm for association rules. To automate all above processes we developed the SITAC generator using Java and MySQL shown in Figure 2. The user interaction with this interface will be explained in a live demo.

The text source for our demo shown here is the Gutenberg corpus [2] of the USA Presidents’ speeches from 1790 to 2006. It spans a time period starting from 1789 to 2006. This serves as the input to the Concept Extraction module.

Each speech is separated and subject to natural language processing with the help of Minipar to obtain parsed documents. Parsed documents are still in form of text. In order to apply data mining techniques, it is required to store them in dataset. Using the SITAC Generator those parsed documents are processed to the dataset shown in Figure 3 after exploring the semantic relationships.

In this example, Figure 3 corresponds to the complete data set and Figure 4 is the data set derived from the complete data set using “create data set” function in SITAC interface that embeds SQL queries for appropriate selection.

<table>
<thead>
<tr>
<th>concept</th>
<th>Word Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>United State</td>
<td>3 1 2 1 1 1</td>
</tr>
<tr>
<td>Union</td>
<td>1 1 1 1</td>
</tr>
<tr>
<td>Nation</td>
<td>1 1</td>
</tr>
</tbody>
</table>

**Figure 6: Example for Ranking of Rules**

Likewise, after obtaining the SITACs and their ranking, we store the results in a SITAC database to serve as the basis for answering queries on the text archives. Thus, a query on the USA uses the SITAC Union and is answered more appropriately.

SITACs stored in a database are then used to answer several user queries. When a concept is entered in a query, the system automatically adds its SITACs, if any, to the query and then feeds it to the search engine. The search engine developed in the SITAC system gives the relevant documents on the USA and the UK policies, including the documents with former names, e.g., Union.

On reviewing Gutenberg documents, it is found that not all the documents have the term USA though they are speeches from the USA Presidents. Thus, in regular search, if we use only the term USA in our query we are able retrieve only some of the documents. However, by adding the word Union to the query using the SITAC system, we are able to retrieve more relevant documents. We evaluated the results of the search before and after using SITAC in conjunction with a regular search engine (Google) and obtained results are given in Figure 7.

**Figure 7: Search Result Comparison Before and After SITAC**

Due to a greater number of relevant documents retrieved after SITAC than before, we obtain higher precision and recall using the SITAC approach than in the absence of the approach. This is noticed with various queries using different terms.

Several examples of such queries will be presented in a live demonstration of the system. We will also show more examples of parsing steps to discover various SITACs, allowing the users to interact with the SITAC system.
4. DISCUSSION ON THE SYSTEM

It is to be noted that the SITACs once discovered and stored in a SITAC database can be reused for further query processing. This serves as the basis for faster time-aware query translation, a recurrent operation to be performed each time a user query is posed. Although the discovery of the SITACs itself is a time-consuming process and is also corpus-specific, this involves a one-time operation, not a recurrent one. Thus, the SITAC system serves as an efficient means of time-aware query translation incorporating temporal terminology evolution.

In the absence of SITAC discovery, when the same queries are executed on existing search engines, we do not essentially get all the relevant documents pertaining to the user query that incorporate historical information. We have conducted evaluation using information retrieval measures precision and recall and found that the use of the SITAC system provides higher values, showing at least a 10% increase in precision and recall with respect to several queries [5]. Thus, the SITAC system also serves to enhance the accuracy of the information retrieval.

Our work includes the following technical contributions [5, 6] which will be illustrated in a live system demonstration.

1. Identifying the challenging problem of terminology evolution in text archives, motivating time-aware query translation
2. Introducing the terminology of SITACs to address the given problem and propose a solution accordingly.
3. Simulating human thinking by a methodology to discover the SITACs using the manner in which humans associate concepts.
4. Proposing a novel collaborative framework of natural language processing, association rules and contextual similarity as a learning technique to solve the given problem.
5. Solving various non-trivial subtasks such as defining adequate transactions required in association rule mining to capture the essence of the problem.
6. Developing the SITAC system as a software tool useful in information retrieval, enhancing precision and recall.
7. Evaluating this SITAC system with real online data and depicting its effectiveness.

This work would be of interest to database, information retrieval, data mining and machine learning professionals. It would particularly be appreciated by those conducting interdisciplinary work across multiple tracks.

5. RELATED WORK

Named entity recognition is addressed in works such as [3]. Research has been conducted on sequence classification and mining as in. [7,9]. Similarity measures over text and web documents have been studied in the literature, e.g., [4,11]. Norvag et al. [8] address the problem of mining association rules over collections of temporal documents defining inter-transaction associations. However, none of these works address time-aware query translation in text incorporating terminology evolution.

Berberich et al. [1] have addressed such temporal terminology evolution. They consider Hidden Markov Models and the frequency of co-occurrence terms between concepts. For example, the terms iPod and Walkman are generally used with words portable, music and earphones. This word overlapping is used to determine their semantic similarity. This requires a recurrent computation each time query processing is performed. In our system, we pre-compute and materialize by discovering SITACs in advance which is a one-time process. However, the SITAC system requires more storage, so we have a space-time trade-off.

Our earlier research [10] that led to this work involves the use of anticipated queries over the text corpora as a relatively fast method of discovering the most likely concepts from text archives that alter over time. This involves relatively less storage space. However, it may not always be possible to anticipate such user queries over text sources. Hence, this is even more corpus-specific and is suitable for use only when such anticipation seems feasible.

6. CONCLUSIONS

We have addressed the problem of discovering SITACs in text sources, i.e., Semantically Identical Temporally Altering Concepts, in order to perform time-aware translation of user queries. We have developed a system by the same name SITAC to solve this problem, based on a novel integrated framework of natural language processing, association rule mining and contextual similarity. It is found that the SITAC system helps to provide more accurate responses to queries. This paper presents a demonstration of the SITAC system. Ongoing work includes detailed comparative studies with the state-of-the-art.

7. ACKNOWLEDGMENTS

This work started when one of the authors, Aparna Varde, was a visiting researcher at Max Planck Institut für Informatik (MPI), Germany. We acknowledge the initial inputs of Srikantha Bedathur and Klaus Berberich from MPII. We also thank Anna Feldman from the Department of Linguistics and Jing Peng from the Department of Computer Science, both at Montclair State University, for their feedback on this work.

8. REFERENCES