Incorporating Terminology Evolution for Query Translation in Text Retrieval with Association Rules

Amal Kalurachchi[1], Aparna S. Varde[1], Srikanta Bedathur[2], Gerhard Weikum[2], Jing Peng[1], Anna Feldman[1]
1. Department of Computer Science, Montclair State University, Montclair, NJ, USA.
2. Databases and Information Systems Group, Max Planck Institut für Informatik, Saarbrücken, Germany.
amalkal@hotmail.com, vardea@montclair.edu, bedathur@mpi-sb.mpg.de, weikum@mpi-sb.mpg.de, pengj@montclair.edu, feldmana@montclair.edu

ABSTRACT
Time-stamped documents such as newswire articles, blog posts and other web-pages are often archived online. When these archives cover long spans of time, the terminology within them could undergo significant changes. Hence when users pose queries pertaining to historical information over such documents, the queries need to be translated taking into account these temporal changes in order to provide accurate responses to users. For example, a query on Sri Lanka should automatically retrieve documents with its former name Ceylon. We call such concepts SITACs, i.e., Semantically Identical Temporally Altering Concepts. In order to discover SITACs, we propose an approach based on a novel framework constituting an integration of natural language processing, association rule mining and contextual similarity as a learning technique. The proposed approach has been experimented with real data and has been found to yield good results with respect to efficiency and accuracy.

Categories and Subject Descriptors: H.2.8 [Database Management]: Database Applications – data mining, H.3.1. [Content Analysis and Indexing] – linguistic processing, H.3.3. [Information Search and Retrieval] – search process

General Terms: Algorithms, Design, Experimentation, Human Factors, Performance

Keywords: Association Rules, Contextual Similarity, Natural Language Processing, Ranking, Search, Web IR

1. INTRODUCTION
Large Internet Archives, e.g. archive.org, have billions of articles stored online. For instance, the London Times archive contains documents since 18th century. As a consequence of these archives spanning long time periods, the terminology within them could evolve significantly. Hence, when users enter queries over such document collections, the queries need to be translated, taking into account this temporal terminology evolution in order to give more accurate responses to the user queries. We present a few motivating examples of such queries.

1. How many states were in the USA during independence?
2. When was the constitution of the USA established?
3. What has been the USA policy on the UK over 200 years?

These queries would be entered on a search engine using appropriate keywords or sentences. In response to these queries, multiple text documents need to be retrieved. An example of a document with answers to queries 1 and 2 is Lincoln's presidential address [2]. However, he refers to the *USA* as the *Union*. Query 3 can be answered from speeches of many presidents [2] who use terms such as *British Isles* and *Great Britain* in referring to the *UK*. If this query is executed on a search engine such as Google, the results consist of documents containing the terms *USA, UK* or *200* years. Unless there is a document worded similarly to USA foreign policy we do not get the exact answer even though that information is available in some article. It is also to be noted that this is not just an issue of synonymy, e.g., the terms *USA* and *Union* would not be detected in the literature as obvious synonyms. Moreover, users tend to formulate queries with current terminologies. Reviewing Google trends, we can see the name *Sri Lanka* has been mainly used when searching documents related to *Sri Lanka*. But there are many relevant documents available with its former name *Ceylon*. Any article related to Sri Lanka, generated before 1948 had no information about the term *Sri Lanka*. Though query expansion and refinement techniques are currently being employed in information retrieval systems, they do not always inject semantic and temporal concepts in user queries. Many existing techniques primarily refer to the correlations between words. The specific issue we address is incorporating temporal terminology evolution in query translation for better information retrieval. This research addresses two main problems in information retrieval over text.

1. Finding concepts that have changed over time.
2. Incorporating this knowledge in responding to queries.

In our work, we refer to the concepts changing over time as SITACs. The definition of SITAC is as follows: *The term SITAC is an acronym for a Semantically Identical Temporally Altering Concept. It refers to a concept whose names change over time, although they, in principle, refer to the same entity. SITACs could be under different categories such as persons, places, organization and item names.* Examples of SITACs include (Hillary Clinton, Hillary Rodham) and (Sri Lanka; Ceylon).

This problem is further analyzed as follows. While users request for accurate search results based on the way humans think, search engines generally provide results based on the term frequencies of the words in input query. By entering queries "When did *Sri Lanka* get its independence?", "When did *Ceylon* get its independence" and "When did *Sri Lanka Ceylon* gets independence" on a variety of search engines, we can see how varied the results are. Obviously, users can enter first and second queries, but third query is different. By experimenting with such kind of queries, we find that the third query produces better results than first and second. This leads us to formulating a solution based on how humans tend to associate concepts.

2. FRAMEWORK AND ALGORITHMS
In order to discover SITACs in text archives for time aware query translation, we propose an approach by the same name, SITAC. The SITAC approach constitutes a novel collaborative framework of natural language processing, association rule mining and contextual similarity as a learning technique. To explain the SITAC approach, we consider two example sentences.

1. Sri Lanka hangs like a jewel off India's tip, surrounded by the Indian Ocean.

CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.
ACM 978-1-4503-0099-5/10/10.

2. Ceylon is called the pearl of the Indian Ocean by explorers who came to Asia in 15th and 16th centuries.

Upon reading these two sentences, a human can identify that Sri Lanka and Ceylon have something in common. That judgment is made by understanding common words associated with Sri Lanka and Ceylon. In this research, we simulate that human judgment of identifying such concepts by using association rules. Other machine learning techniques such as clustering and classification have also been experimented, but given the type of knowledge we are trying to discover, association rules have been found the most suitable.. This is justified by the fact that such rule mining is in line with the logic of humans associating concepts. Furthermore, we propose the following a heuristic: *If one or more concepts (nouns) are referred to by similar events (verbs) those concepts are semantically related.* To enhance the logic here, we also consider objects, adjectives and bigrams along with verbs.

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 solution framework proposed in this work involves integrating natural language processing, association rule mining and contextual similarity. It is summarized in Figure 1 which constitutes the architecture and explained in the parts below, forming steps of the approach.

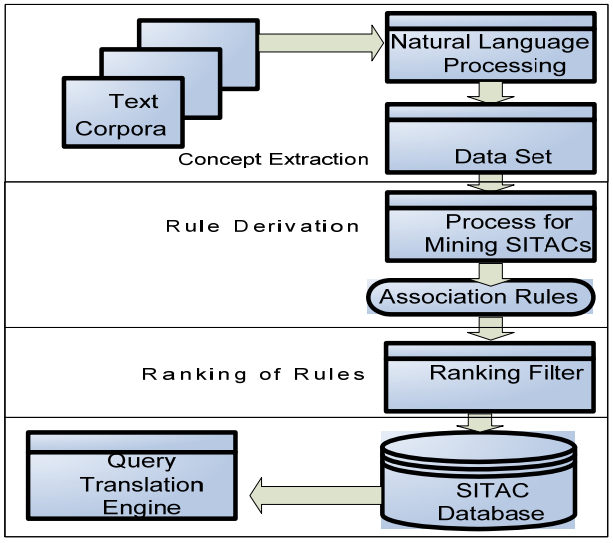

Figure 1 : SITAC System Architecture

Concept Extraction: In this step, text archives are processed to extract information as concepts in documents over time, related by events. So, we have:
Document: Text source *D* with time-stamp *T*
Concept: Individual term *C* (word or phrase)
Event : The event *E* relating concepts
Other neighboring words: Objects, adjectives, bigrams etc.

The concepts are primarily nouns and noun phrases referring to entities such as persons and places. Events correspond to verbs referred by concepts. Other words refer to works in the neighborhood that are linguistically related. This step involves exploiting semantic features from a linguistic perspective.

Rule Derivation: This step discovers rules of the type *(C1,T1)=>(C2,T2)* from corresponding time-stamped documents. We use the classical *Apriori algorithm* to mine association rules, for which we define transactions with respect to the text archives. Our transaction set 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 occurring multiple times, then we consider that those two nouns (concepts) are related. This can be further explained using set theories. Consider the sets Events *{E1,E2,E3..,En}* and Concepts with time stamps *{(C1,T1), (C2,T2)…,(Cp,Tp).* If *(Ci,Ti)* and *(Cx,Tx)* are referred to by *Er,* a distance value *r* is assigned to that relationship such that initially every pair of *C,T* has a very high value of *r* and each appearance of *(Ci,Tj)* and *(Cx,Ty)* together in a relationship *r.* Pairs *(Cx,Cy)* with smallest *r* values are considered SITACs. Thus we get rules of the type *(Cx,Tx) => (Cy,Ty).*

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. *{EVENT, TIME1,TIME2….,,TIMEn} where TIME1…. TIMEn* will have concepts that appeared in the text archive associated with the events listed under the *EVENT* attribute

Ranking of Rules: Once the SITACs are discovered, this step 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 used in information retrieval, we select *Jaccard's coefficient* because it is found to be the most useful in capturing contextual similarity as per a study of the literature e.g., [10]. In our problem, we employ this as follows. For two relationships, *R1{(Cx,Ts), (Cy,Tt)}, R2{(Cx,Tt) , (Cz,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)=(Cx∩Cy)/(CxUCy), J(Cx,Cz)=(Cx∩Cz)/(CxUCz)* and so forth, such that *Cx ∩ Cy)* is the count of other words used with both the concepts Cx and Cy, while *Cx U Cy* is the count of other words used with either concept Cx or concept Cy or both. Now, we argue that *J(Cx,Cz) > J(Cx,Cy)* means *Cx* is more related to *Cz* than *Cy* based on adapting the definition of Jaccards' coefficient. This logic is used to rank the SITACs.

Query Translation: The SITACs discovered are then used for query translation 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 storage. When a user enters a query, it goes through a parser and stores all words in an array after eliminating stop words (the, an) and common words (I, We). The user query (W1, W2.., Wn) is translated to SQL as follows:

```
/* Find SITACs * /
SELECT   *  FROM Tbl_SITAC
        WHERE  WORD  = <W1>
/*Find Documents */
SELECT  * FROM Tbl_word
  where word =<W1> AND
   time in (SELECT time from Tbl_word where  word=<W2> AND time
in (SELECT time from Tbl_word where word=<Wn>)))
```

Combining results of the queries, documents can be appropriately retrieved from a given corpus. The two algorithms below summarize all the steps involved in finding SITACs and utilizing them effectively in responding to user queries.

Algorithm1: Discovering and Ranking SITACs from Text
Input: Text Corpus with time-stamped documents
Processing: *D[i]= {D1,D2,..Dn}* // Documents by time
For *i = 1* to *n* {
Run parser on *S[i]* -> *F[i]* // Text files after parsing }
For *i = 1* to *n* { // Generate each instance from parsed files
E=" " ; C=" " // *E*=Event *C*=Concept
While (not End_of_File *F[i])* {
If (*F[i]*.Readline() has token *"V:")* //*V*=Verb, *N*=Noun
{*E*=word before *"V:"* token, *C*=word next to *"N:"* token
P=P+ {i,E,C} // *P*=Processed dataset: rows, columns }}}
Transpose *P* with instance *i* as columns and *E,C* as rows
Run Apriori association rule mining algorithm
For-each transaction {
Use relationship distance *r* to get SITAC pairs}
For-each SITAC pair { // *J* = Jaccard's coefficient score
J(Cx,Cy) = (Cx ∩ Cy) / (Cx U Cy)
J(Cx,Cz) = (Cx ∩ Cz) / (Cx U Cz)
J(Cx,Cz)>J(Cx,Cy)// Cx is more related to *Cz* than *Cy }*
Rank SITACs using *J* values
Output: Ranked SITACs stored in SITAC Database

Algorithm2: Using SITACs for Translating User Queries
Input: User query
Processing:
Q={W1,W2,W3,W4…Wn} //Q is the user query
Parse Q and remove stop words
For each Q{Wn}{
Search word = w1 in SITAC database
Q = Q + SITAC.SITAC // SITAC in the user query
time = SITAC. time // Store year of SITAC }
For each Q{Wn}{
SELECT time form Tbl_word where exists word ….. // SELECT statement to find words }
Output: List of documents contains words from user query

3. EXPERIMENTAL EVALUATION

The SITAC approach was subjected to experimental evaluation using real text archives in order to assess its effectiveness. The text source for experiments shown here was the Gutenberg corpus [2] of the USA Presidents' speeches from 1790 to 2006. These were separated in order to feed the next step of parsing. Those separated documents were subject to natural language processing with the help of Minipar [11] to obtain parsed documents as illustrated in Figure 2. Parsed data was processed to the dataset shown in Figure 3 after exploring the semantic relationships. This formed the output of the concept extraction step.

```
> fin   C:i:V   congratulate
congratulate    V:s:N   I
congratulate    V:subj:N    I
congratulate    V:obj:N people
people  N:det:Det   the
people  N:mod:Prep  of
of  Prep:pcomp-n:N  United States
United States   N:det:Det   the
United States   N:lex-mod:U United
```

Figure 2: Partial Dump of Minipar NLP output

verb	1790	1791	1792	1793
respect	right		sanction	
expect	right	peace	it	
require	safety		occasion	prompt
add	sanction		information	
feel	satisfaction			
derive	satisfaction	satisfaction	consolation	
have	secretary		tribe	United States
direct	Secretary of War	operation	fund	
call for	session		occasion	

Figure 3: Partial Snapshot of Concept Extraction

1. 1795=Union => 1958=United States
2. 1872= Union => 1995=United States
3. 1958= Nation => 1999= United States
4. 1952=war => 1999=terrorist
5. 1952=war 1952= weapon => 1999=terrorist

Figure 4: Arbitrary Sample of Rule Derivation

We used this transaction set as an input to the rule derivation step which deployed WEKA[12] to derive association rules. We got several rules of which we list an arbitrary sample in Figure 4.
The SITACs discovered through association rule mining were then subject to a contextual similarity analysis using Jaccard's Coefficient during the step involving ranking of rules.

Consider any 2 rules: Eg. *Union => USA* and *Union => EU ;* USA, Union and EU have following set of associated words;
W(USA)={abandon, abet, ability, ..}, W(Union)={abandon , abhors, ability, ...}, W(EU)={absentee, abet, ability,..}

As per our experiments, similarity based on Jaccard's coefficients, J is calculated as follows.
Consider. J(USA, Union).
The union of the terms (USA,Union) = 6351
The intersection of the terms (USA,Union)=1622
Hence, J(USA,Union)= 0.255393
Likewise, consider J(EU,Union)
The union of the terms(EU,Union) = 2320
The intersection of the terms (EU,Union)=356
Hence, J(EU,Union)= 0.153448
Thus, after obtaining the SITACs and their ranking, we stored the results in a SITAC database to serve as the basis for the query translation step. Thus, a query on the *USA* would use the SITAC *Union*, thus being answered more accurately; SITACs stored in a database were actually used to answer several such user queries. When a concept was entered in a query, the system automatically included its SITACs, if any, to translate the query and then fed it to the search engine to convey the response.

Likewise, after obtaining the SITACs and their ranking, we stored the results in a SITAC database to serve as the basis for the query translation step. Thus, a query on the *USA* would use the SITAC *Union*, thus being answered more accurately. SITACs stored in a database were actually used to answer several such user queries. When a concept was entered in a query, the system automatically included its SITACs, if any, to translate the query and then fed it to the search engine to convey the response.

We used precision and recall in our research in order to assess the accuracy of our approach. As widely known in information retrieval, precision refers to the exactness of the document retrieval process while recall refers to the completeness of the document retrieval process. Getting higher numbers for both increases the accuracy of the information retrieval. Consider the Gutenberg corpus shown here that contains documents from speeches of American presidents. On reviewing the documents, it was found that not all the documents have the term *USA*. Thus, in regular search, if we used only *USA* in our query we were able retrieve only some of the documents. However, by adding the word *Union* to the query, we were able to retrieve more relevant documents. Likewise, the SITAC approach also considered *UK* and *British Isles* which were used interchangeably in different speeches. Therefore, due to a greater number of relevant documents retrieved, we obtained higher precision and recall after using the SITAC approach than in the absence of the approach. Figure 6 plots a chart showing precision and recall with and without using the SITAC approach considering sample queries.

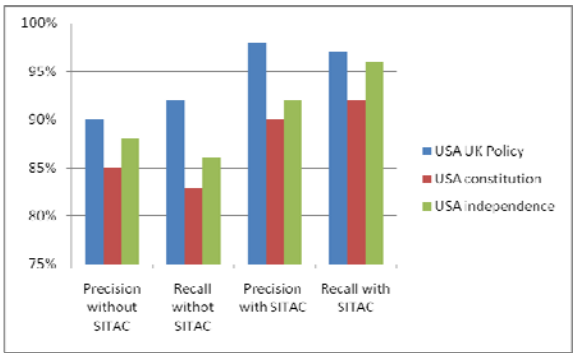

Figure 6: Precision and Recall with and without SITAC

From this chart we can see that higher precision and recall values are obtained with the SITAC approach than without. We also found that response time with the SITAC approach almost as fast as without it, since SITACs are pre-computed and directly used for query translation, thus providing efficiency. Thus, we claim that information retrieval is enhanced by incorporating temporal terminology evolution using the SITAC approach. This can be used in conjunction with existing search engines.

4. RELATED WORK

Some of the literature in text mining addresses word sense disambiguation and document classification, e.g., Michael Lesk's algorithm [4] which states that when 2 words are used in close proximity in a sentence, they must be addressing a related topic, also if each word addresses the same topic, then their dictionary definitions must use common words. Lin et al. [5] introduce an unsupervised method for word sense disambiguation which is able to tag test data with appropriate senses from machine readable dictionaries using syntactic dependency and semantic similarity as disambiguation information. The basic principle of this method is the observation that two occurrences of the same word have identical meaning if their local context is the same. In the work of Lu et al. [6], it is stated that users frequently use short queries in Internet searches, which make it difficult to retrieve relevant documents. Those user queries can be expanded by adding semantically related words to the same as in the work of Mihalcea's et al. [7]. Named entity recognition is addressed in works such as [3]. However, none of the above research focuses on the specific issue of terminology evolution in text.

Berberich et al. [1] have addressed temporal terminology evolution using Hidden Markov Model (HMM). They consider frequency of co-occurrence terms between concepts. For example, the terms *iPod* and *Walkman* are mostly 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 approach, we pre-compute and materialize by discovering SITACs in advance which is a one-time process. The discovered SITACs can directly be used for query translation. However, our approach involves a trade-off in terms of consuming additional storage space and being very corpus-specific.

The mining of sequential patterns has been studied in the literature. Parthasarthy et al. [9] perform the mining of subsequences that are frequent using minimum support levels and extend this paradigm to sorting. Their goal is to reduce input-output and computation needs in handling incremental updates to the data, while mining, since data sources could undergo changes. The data in this work consists of examples, each represented as sequences of events, each event having a set of predicates. Norvag et al. [8] define temporal association rules for document collections. They have 5 types of rules: episode rules, sequence rules, trend dependencies, calendar rules and inter-transaction rules that capture different kinds of temporal relationships within documents Though we can draw some analogy with the works here, their goals are quite different from ours. They do not consider time-aware query translation in particular.

5. CONCLUSIONS

In this research, we have addressed the problem of terminology evolution in text archives spanning long time periods. This motivates the need for time-aware query translation in order to provide accurate responses to user queries over such text sources. We have proposed a solution that involves discovering Semantically Identical Temporally Altering Concepts or SITACs in text archives, with the goal of performing the required query translation. Our proposed solution approach by the same name SITAC constitutes an integration of natural language processing, association rule mining and contextual similarity, for translating and answering user queries appropriately. We have evaluated the SITAC approach using real online text archives. It was found to enhance precision and recall. We claim that our approach would be useful in web-based information retrieval over text archives that contain historical data and require intelligent time-aware query translation to enhance user responses.

6. ACKNOWLEDGMENTS

This work was initiated when one of the authors, Aparna Varde, was a visiting researcher at Max Planck Institut für Informatik. We thank our colleagues there and also in the Department of Computer Science at Montclair State University.

7. REFERENCES

[1] Berberich, K. et al. "Bridging the Terminology Gap in Web Archive Search!", SIGMOD's WebDB 2009.
[2] Gutenberg Corpus, U.S. "Presidential Inaugural Addresses" www.gutenberg.net
[3] Hasegawa, T. et al. "Discovering Relations among Named Entities from Large Corpora", ACL 2004.
[4] Lesk, M. E. "Automatic sense disambiguation using machine readable dictionaries", International Conference on Systems Documentation, 1986.
[5] Lin, D. "Using syntactic dependency as a local context to resolve word sense ambiguity", ACL 1997.
[6] Lu, X. A. et al. "Query expansion/reduction and its impact on information retrieval effectiveness", TREC-3, 1994.
[7] Mihalcea, et al. "Page rank on semantic networks, application to word sense disambiguation", COLING 2004.
[8] Norvag, K. et al."Mining Association Rules in Temporal Document Collections", Dept. of Computer and Information, Systems, NTNU, Norway, 2006.
[9] Parthasarathy, S., et al. "Incremental and Inter-active Sequence Mining", CIKM 1999/
[10] Strehl A. et al. "Impact of Similarity Measures on Web-page Clustering", AAAI 2000.
[11] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm
[12] University of Waikato, New Zealand, WEKA.