Discovering Trends in Text Databases

Brian Lent* and **Rakesh Agrawal** and **Ramakrishnan Srikant**
IBM Almaden Research Center
San Jose, California 95120, U.S.A.
lent@cs.stanford.edu, {ragrawal,srikant}@almaden.ibm.com

Introduction

We address the problem of discovering trends in text databases. Trends can be used, for example, to discover that a company is shifting interests from one domain to another. We are given a database $\mathcal{D}$ of documents. Each document consists of one or more text fields and a timestamp. The unit of text is a *word* and a *phrase* is a list of words. (We defer the discussion of more complex structures till the "Methodology" section.) Associated with each phrase is a *history* of the frequency of occurrence of the phrase, obtained by partitioning the documents based upon their timestamps. The frequency of occurrence in a particular time period is the number of documents that contain the phrase. (Other measures of frequency are possible, e.g. counting each occurrence of the phrase in a document.) A *trend* is a specific subsequence of the history of a phrase that satisfies the users' query over the histories. For example, the user may specify a "spike" query to finds those phrases whose frequency of occurrence increased and then decreased.

Approach

Our system uses several data mining techniques in novel ways and demonstrates a method in which to visualize the trends. We have two major mining components: phrase identification using sequential patterns mining (Srikant & Agrawal 1996) and trend identification using shape queries (Agrawal *et al.* 1995). We begin by cleansing and parsing the input data, and separating the documents based on their timestamps. We then assign a transaction ID to each word of every document treating the words as items in the data mining algorithms (the details of this assignment are discussed in the "Methodology" section). This transformed data is then mined for dominant words and phrases, and the results saved. The user's query is translated into a shape query and this query is then executed over the mined data yielding the desired trends. The final step in the process is to visualize the results. We give experiences from applying this system to the IBM Patent Server, a database of U.S. patents.

Related Work

An approach to discovering interesting patterns and trend analysis on text documents was presented in (Feldman & Dagan 1995). The text is first annotated with a set of concepts, organized as a hierarchy. Treating the concept hierarchy as a distribution of probabilities, they identify several model distributions distribution) to which a given concept hierarchy can be compared. Interesting concepts are those that differ from their model distribution. Analyzing trends involves the comparison of concept distributions using old data with distributions using new data.

In (Feldman & Hirsh 1996), the authors find associations between the keywords or concepts labeling the documents using background knowledge about relationships among the keywords. The purpose of the knowledge base is to supply unary or binary relations amongst the keywords labeling the documents.

Using words and phrases to describe themes and concepts in text documents has been studied by the information retrieval community. The work on Latent Semantic Indexing (LSI) (Deerwester *et al.* 1990) describes a mathematical model of relating word associations as weighted vectors that represent "concepts" found within the documents. Using LSI, a query can retrieve a document even when they share no words, but do share a similar concept. However, building the model takes $O(tk^4d)$ time, where t is the number of terms or words, k is the number the major concepts in the model (typically defined from 100 to 300), and d is the number of documents.

The use of phrases to build more advanced queries is discussed in (Croft, Turtle, & Lewis 1991). In this work, the authors identify phrases as concepts and as relationships between concepts. The usefulness of phrases is shown in (Lewis & Croft 1990) where the quality of text categorization is improved by us-

* Current address: Department of Computer Science, Stanford University. Continuing support has been provided by a graduate fellowship of the Department of Defense Office of Naval Research.

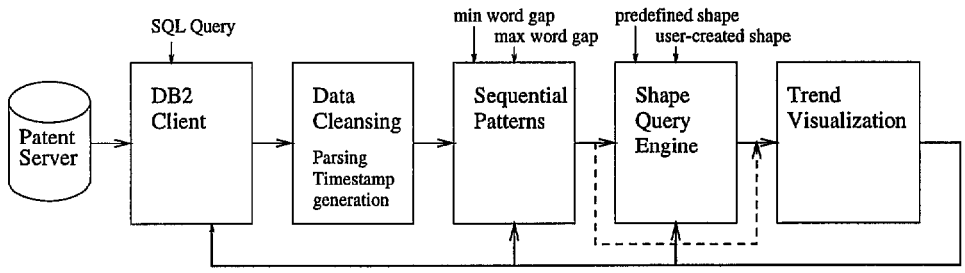

Figure 2: The PatentMiner system

Identifying trends

By maintaining a support history for each supported k-phrase we can query the set of histories to select those phrases that have some specific shape in their histories. We propose the use of a shape definition language called $\mathcal{SDL}$ (Agrawal *et al.* 1995) to define the users' queries and retrieve the associated objects. There are several benefits for using a shape query language such as $\mathcal{SDL}$ to identify trends: First, the language is small, yet powerful, allowing a rich combination of operators. Second, it is a fairly straightforward task to rewrite a shape the user may define graphically, as is done in our PatentMiner system described in the "Experience" section, into the $\mathcal{SDL}$ set of operators. Third, $\mathcal{SDL}$ allows a "blurry" match where the user may care about the overall shape but does not care about specific details of each interval of the shape. Finally, $\mathcal{SDL}$ allows itself to be implemented efficiently since most of the operators are designed to be greedy to reduce non-determinism which in turn reduces the amount of back-tracking that must be done across the histories.

Trends are simply those k-phrases selected by the shape query with the additional information of the time periods in which the trend is supported.

Experience: The PatentMiner System

Figure 2 shows a high-level view of our system to compute and visualize the word-phrase trends, which we now describe.

The PatentMiner prototype is a system we developed to discover trends among patents granted in different categories. The system is connected to an IBM DB2 database containing all granted U.S. Patents and patent data is retrieved using a dynamically generated SQL query based upon the selection criteria specified by the user. The system allows selection of patents in a specific classification or by keywords appearing in the title or abstract of the patents. Once retrieved, a histogram displaying the number of patents for each year is shown and the user may then specify a range of years upon which the system will focus.

Next, the user can choose the maximum and minimum gap desired between words in the phrases to be mined, as well as the minimum support all phrases must meet for each time period between the start and

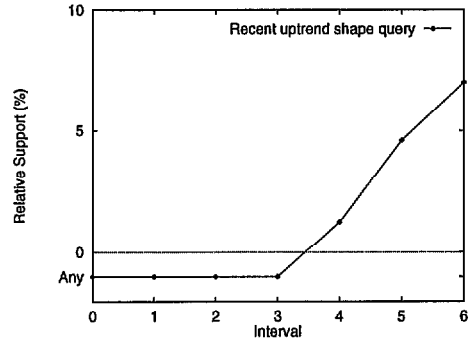

Figure 3: A recent Uptrend shape query

```
(shape strongUp ( ) (comp Bigup Bigup Bigup ))
(files "list_hist" )
(query (window 1) ((strongUp) (support 2 end)))
(quit)
```

Figure 4: Recent uptrend query

ending years. Finally, a shape matching the desired trend (such as "recent upwards trend", "recent spike in usage", "downwards trend", and "resurgence of usage") is selected and the mining process begins. Alternatively, users can define their own shape by using a visual shape editor. Once the phrases matching the shape query are found, they are presented in a visual display.

Once a shape query has been defined, either internally or using the graphical editor, a rewriting of the query into $\mathcal{SDL}$ (Agrawal *et al.* 1995) is performed. Given the shape query in Figure 3, the rewriting of this query into $\mathcal{SDL}$ is shown in Figure 4. The rewriting happens as follows. For every partitioned time period of documents there is a corresponding interval in the shape query graph that has associated with it beginning and ending relative levels of support. In the case where every interval has a specific beginning or ending value, the rewriting into SDL is straightforward in that the slope of each interval determines the basic shape query that is used for that interval. For example, intervals with a positive slope translate to an "up" shape of length one, while intervals with a negative slope trans-

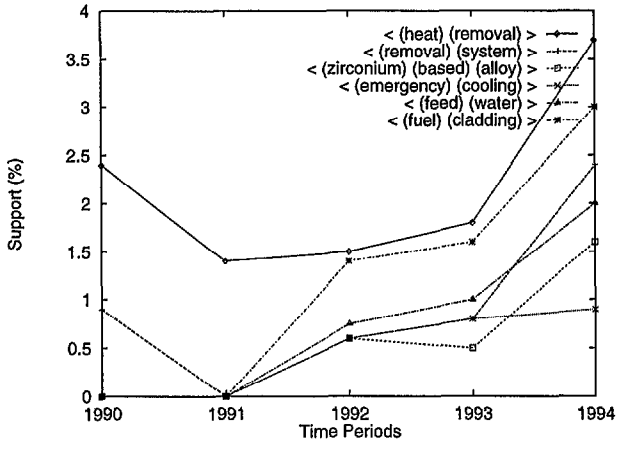

Figure 5: Some recent upwards trends

late to a "down" shape of length one. The concatenation of all of these base shapes then defines our SDL query. In the case where only some of the intervals of a shape query have been specified, as in intervals three to six in Figure 3, then the same concatenation occurs but the resulting SDL shape can have any support value match the unspecified intervals.

We present some of the trends our system found from U.S. Patents classified in the category "Induced Nuclear Reactions: Processes, Systems, and Elements" in Figure 5. These example phrases matched a shape query that represented an increasing trend of their usage in recent years. Without knowing a priori the kind of patents filed in this category, we are able to look at the trends and determine some of the popular topics of recently granted patents.

A potential problem with this system is that the number of phrases that match a query can be quite large. There are two types of pruning we use to reduce the number of phrases to a more reasonable number. The first form of pruning is to drop non-maximal phrases when their support is near that of a maximal phrase that is a superset. The second form of pruning involves the use of a syntactic hierarchical ordering of phrases. The intuition is that if phrase X is a syntactic sub-phrase of phrase Y, then the concept corresponding to X is usually a generalization of the concept corresponding to Y. Users initially see only the most general concepts, and can explore lower-level concepts by selecting some of the phrases.

Conclusion

We presented a system for identifying trends in text documents collected over a period of time. Our system uses several data mining techniques such as sequential patterns and shape queries in novel ways and demonstrates a trend visualization method. We described our experience in applying this system to the IBM Patent Server, a database of U.S. patents. Scaleup experiments show that our system, PatentMiner, scales approximately linearly with the number documents.

Acknowledgments We are grateful to the IBM Almaden Patent Server team, especially Laura Anderson, Steve Boyer and Tom Griffin for their ongoing contributions and suggestions.

References

Agrawal, R.; Psaila, G.; Wimmers, E.; and Zait, M. 1995. Querying shapes of histories. In *Proceedings of the 21st International Conference on Very Large Databases.*

Croft, W.; Turtle, H.; and Lewis, D. 1991. The use of phrases and structured queries in information retrieval. In *14th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 32–45.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41(6):391–407.

Feldman, R., and Dagan, I. 1995. Knowledge discovery in textual databases (KDT). In *Proceedings of the 1st International Conference on Knowledge Discovery in Databases and Data Mining.*

Feldman, R., and Hirsh, H. 1996. Mining associations in text in the presence of background knowledge. In *Proceedings of the 2nd International Conference on Knowledge Discovery in Databases and Data Mining.*

Gay, L., and Croft, W. 1990. Interpreting nominal compounds for information retrieval. *Information Processing and Management* 26(1):21–38.

Lewis, D., and Croft, W. 1990. Term clustering of syntactic phrases. In *13th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 385–404.

Renouf, A. 1993a. Making sense of text: automated approaches to meaning extraction. *17th International Online Information Meeting Proceedings* 77–86.

Renouf, A. 1993b. What the linguist has to say to the information scientist. *Journal of Document and Text Management* 1(2):173–190.

Salton, G.; Allan, J.; Buckley, C.; and Singhal, A. 1994. Automatic analysis, theme generation, and summarization of machine readable texts. *SCIENCE* 264(5164):1421–1426.

Salton, G.; Singhal, A.; Buckley, C.; and Mitra, M. 1996. Automatic text decomposition using text segments and text themes. In *Proceedings of Hypertext*, 53–65.

Srikant, R., and Agrawal, R. 1996. Mining sequential patterns: Generalizations and performance improvements. In *Proceedings of the Fifth International Conference on Extending Database Technology (EDBT).*