

Proposal

Text-mining and associative analysis: detecting topic trends in Technique Slivers

Jay Anderson, Madhumati Gundapuneni

Motivation and Objectives

We want to know what's going on in the minds of the Georgia Tech community: What are the students, professors, and alumni talking about and how have these thoughts changed over time? We intend to peer into the collective of Georgia Tech by analyzing anonymous comments for conversation threads and persistent topics.

The planned data sources are "Slivers", anonymous comments submitted to and published weekly in the Georgia Tech school newspaper, the Technique. Slivers are an excellent source of opinion data:

- They are anonymous, so users can offer candid criticism without fear of repercussion.
- They require negligible preparation for submission, comments can be spontaneous and barrier of entry is low.
- They are free and readily available.

On the other hand, while they may be edited for profanity and 'unprintable statements', Slivers are not spell-checked or edited for grammar which might make text comparison more difficult. We feel the benefits are more compelling.

By text-mining Slivers and identifying topics (frequent keyword combinations), we believe there is an unprecedented opportunity to discover the voice of 'Tech'.

Related work

To our knowledge, there has been no previous work analyzing the content of "Slivers." There are however, plenty of relevant projects in text-mining:

Text mining has often been used for text categorization, clustering of data, production of taxonomies, document summarization, analyzing new trends or patterns and drawing interesting relationships amongst entities.

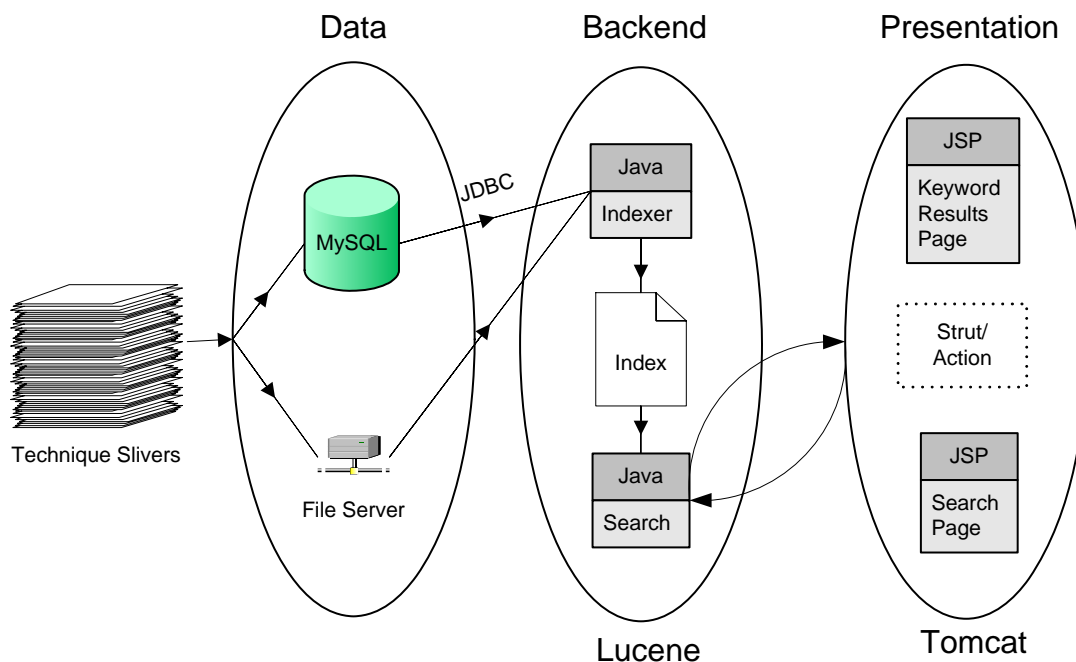
We have mainly concentrated on textual mining which involved extraction of keywords and determining the associations amongst various keywords

- The paper on "GIS: a biomedical text-mining system for gene information discovery" provides a GIS system which involves finding, counting and indexing (by using a domain specific lexicon) the keywords for biological functions, associated diseases and related genes in the abstracts. To draw inter-relation between genes, learning process is used on the training data containing information on related genes to infer sentence expression patterns which are fed into decision trees to yield gene functions that were previously unknown.

- Mark-A. Krogel and co. had predicted yeast gene regulation functions by extracting information from several abstracts per gene using a tool which found various search terms. Text classification was done by applying a stemming algorithm, and forming a TFIDF representation which was given as an input to SVM_{light} . Decision function values output by this function were then used for classification.
- Martin Rajman et. al. have provided data mining tasks for extracting information by associative extraction method where keywords have been extracted from indexed documents. Associative rules have been formulated for keyword sets with factors of support and confidence used to evaluate the association rules.
- Rainer Malik et. al. have used a combination of algorithms of text mining to extract keywords relevant for their study from various databases and also identified relationships between key terminologies using PreBIND and BIND system (Donaldson et al., 2003; Bader et al., 2003). Boosting classifier was used for performing supervised learning and used on the test data set.
- Usually in marketing, open- ended survey responses are analyzed which yield insights into customer's views or opinions that generally would not be detected by regular questionnaires. By studying words or terms associated with pros and cons of products, using text mining applications misconceptions or confusions related to the subject under study are often recognized.

Proposed work

The following is a high level image of the proposed architecture. Detailed summaries of each component immediately follow.



1. Text Gathering

As this is written, a consolidated database or repository for Slivers does not exist. However, the *Technique* is available in digital format (pdf). We will manually scan Sliver sections and gather them. The Sliver extracts will be stored as a repository of text documents in a file server and also processed and fed into a MySQL database.

2. Text Preprocessing

- Black box approach is used to extract deep meaning from documents with little human effort (to first read and understand those documents) to understand the context of Sliver extracts.
- An important pre-processing step before indexing would be stemming which implies reduction of the words to their root. Hence, different grammatical forms or verb forms are treated as one while indexing.
- Exclusion of numbers, certain special characters or words that are unimportant would be useful before indexing is done.
- Stop words such as “a”, “the”, “of”, “since,” etc. , i.e., words that are used very frequently but do not contribute to the information content would be removed.
- Using the thesaurus, synonyms of words can be extracted and words or phrases with the same semantic context can be categorized as single words.

Java Database Connectivity ([JDBC](#)) will be used to interface with database and Java is used as the back-end. We will use the tool [Lucene](#) for preprocessing, parsing the slivers and building an index for the words within the extracts.

3. Data Analysis

Once the input documents have been indexed and the initial word frequencies (by document) computed, a number of additional transformations can be performed to summarize and aggregate the extracted information. Log, binary and inverse document frequencies will be computed.

Lucene will be used to extract keywords along with their frequencies. Next we will use the Search JSP to interface with the Lucene backend for query combinations of top key words. This process will return a number of hits and a list of ranked documents. Different ranking techniques (Lucene default, tf/idf, bm25) will be used to identify combinations with high association scores. Our prediction is that combinations with the highest scores would be the most discussed topics in the slivers. We could also make interesting predictions or inferences by studying the inter-relationships or associations amongst different key words and phrases. We also plan to look at different time slices of the events that have happened in Gatech and observe how things have changed over a time period.

4. Visualization

We are going to create a web application using the Struts framework and Java Servlet Pages ([JSP](#)). The application will be hosted on a [Tomcat](#) server of our creation. One JSP will function as a search engine for the Slivers repository. Another JSP will present keyword combination query results, association scores, and relevant statistics.

5. Evaluation

After the platform has been completed, we will introduce, fabricated “controlled slivers” into the repository and observe how well the system’s reaction fits a model constructed from the training data.

We will calculate the recall, precise and F measure values to evaluate the relevance of the data items from Sliver. Also, use statistical test like regression using the program R to determine how well the association of the terms are relevant.

Plan of action

The following Gantt chart represents the project schedule. Each task has been assigned to Jay, Madhumati or both.

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
Madhumati	Sliver Collection	mySQL Database	mySQL Database	Tomcat Webserver	JSP Search + Topic Pages	Data Analysis	Final Project Report	Presentation
Jay	File Server	Lucene Indexer	Lucene Search					

Evaluation and Testing Method

Evaluation and testing is done to observe how efficiently the data has been mined by our tools. New extracts of “Slivers” can be used as a test data to check for the results which would be compared to the results obtained when the training data set was given as input.

We plan to use statistical measures of regression, leveraging the programming language R to observe if the association rules match predictions.

Also, the most important measures--precision, recall and the F measure--will be calculated.

Precision is defined as the percentage of retrieved documents that are relevant (i.e. the percentages of the association terms that are retrieved are relevant). Recall is defined as the percentage of relevant documents that are retrieved. (i.e. the percentages of the association terms which are relevant that are retrieved). The F-measure combines precision and recall into one measure, being the harmonic mean. It is computed as

$$\text{F-measure} = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

Bibliography

Rainer Malik, Lude Franke and Arno Siebes, “Combination of text-mining algorithms increases the Performance”, Bioinformatics, 2006

Youngja Park, Roy J. Byrd “Hybrid text mining for finding abbreviations and their definitions”, Genetic and Evolutionary Computation Conference, 2001

Mark-A. Krogel, Marcus Denecke, Marco Landwehr, and Tobias Scheffer Combining Data and Text Mining Techniques for Yeast Gene Regulation Prediction: A Case Study”, ACM SIGKDD, 2001

Brigitte Mathiak and Silke Eckstein “Five Steps to Text Mining in Biomedical Literature” , Proceedings of the Second European Workshop on Data Mining and Text Mining in Bioinformatics
Jung-Hsien Chiang, Hsu-Chun Yu and Huai-Jen Hsu, “GIS: a biomedical text-mining system for gene information discovery”, Bioinformatics, 2004

Aaron M. Cohen and William R. Hersh “A survey of current work in biomedical text mining” , 2004

Martin Krallinger and Alfonso “ValenciaText-mining and information-retrieval services for molecular biology”, *Genome Biology* 2005

Wikipedia entries

Lucene: <http://en.wikipedia.org/wiki/Lucene>

Apache Struts: http://en.wikipedia.org/wiki/Apache_Struts