Tapping into the Power of Text Mining

Weiguo Fan[1]
Department of Accounting and Information Systems
Virginia Polytechnic Institute and State University

Linda Wallace
Department of Accounting and Information Systems
Virginia Polytechnic Institute and State University

Stephanie Rich
Department of Computer Science
Virginia Polytechnic Institute and State University

Zhongju Zhang
School of Business
University of Connecticut

February 16, 2005

Note: This article is accepted for publication at *the Communications of ACM*.

[1] Corresponding author. Address: 3007 Pamplin Hall, Blacksburg, VA 24061; Telephone: (540) 231-6588; Fax: (540) 231-2511; E-mail: wfan@vt.edu

Tapping Into the Power of Text Mining

1. Introduction

In 2001, Dow Chemicals merged with Union Carbide Corporation (UCC), requiring a massive integration of over 35,000 of UCC's reports into Dow's document management system. Dow chose ClearForest, a leading developer of text-driven business solutions, to help integrate the document collection. Using technology they had developed, ClearForest indexed the documents and identified chemical substances, products, companies, and people. This allowed Dow to add more than 80 years' worth of UCC's research to their information management system and approximately 100,000 new chemical substances to their registry. When the project was complete, it was estimated that Dow spent almost $3 million less than what they would have if they had used their own existing methods for indexing documents. Dow also reduced the time spent sorting documents by 50% and reduced data errors by 10-15% [2].

The Dow-ClearForest scenario is just one example of how the world is changing when it comes to the efficient and effective management of electronic information. In the future, books and magazines will become a part of history as electronic documents become the primary means of written communication. And, as research in all areas of life continues, many fields will become so overwhelmed with information that it will become physically impossible for any individual to process all the information on a particular topic. Massive amounts of data will be in cyberspace, creating a huge demand for the recently born field of text mining.

Text mining has been defined as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" [6]. The situation described in the opening paragraph is just one example of how text mining technology can be applied in a practical business situation. Many other industries and areas can also benefit from the text mining tools that are being developed by a number of companies. This paper provides an overview of the text mining tools and technologies that are being developed and is intended to be a guide for organizations who are looking for the most appropriate text mining techniques for their situation.

Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases or XML files, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents, HTML files, etc. As a result, text mining is a much better solution for companies, such as Dow, where large volumes of diverse types of information must be merged and managed. To date, however, most research and development efforts have centered on data mining efforts using structured data.

The problem introduced by text mining is obvious: natural language was developed for *humans* to communicate with one another and to record information, and computers are a long way from comprehending natural language. Humans have the ability to distinguish and apply linguistic patterns to text and humans can easily overcome obstacles that computers cannot easily handle such as slang, spelling variations and contextual meaning. However, although our language capabilities allow us to comprehend unstructured data, we lack the computer's ability to process text in large volumes or at high speeds. Herein lays the key to text mining: creating technology that combines a human's linguistic capabilities with the speed and accuracy of a computer.

Figure 1 depicts a generic process model for a text mining application. Starting with a collection of documents, a text mining tool would retrieve a particular document and preprocess it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. Three text analysis techniques are shown in the example, but many other combinations of techniques could be used depending on the goals of the organization. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

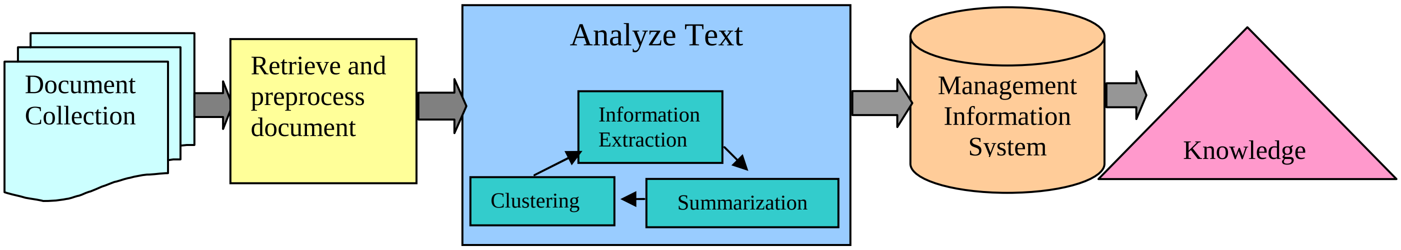


Figure 1. An example of text mining

2. Technology Foundations

Although the differences in human and computer languages are expansive, there have been technological advances which have begun to close the gap. The field of natural language processing has produced technologies that teach computers natural language so that they may analyze, understand, and even generate text. Some of the technologies that have been developed and can be used in the text mining process are information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering. In the following sections we will discuss each of these technologies and the role that they play in text mining. We will also illustrate the type of situations where each technology may be useful in order to help readers identify tools of interest to themselves or their organizations.

Information Extraction

A starting point for computers to analyze unstructured text is to use information extraction. Information extraction software identifies key phrases and relationships within text. It does this by looking for predefined sequences in text, a process called pattern matching. For example, given the sentence “Area relatives of a man being held hostage in Iraq waited for word about him Saturday as militants threatened to decapitate him, another American and a Brit unless demands were met within 48 hours”, information extraction software should identify two American hostages and a British hostage, militants, and the relatives of one of the hostages as people; Iraq as the place; and Saturday as the time. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information.   This technology can be very useful when dealing with large volumes of text. Almost all text mining software uses information extraction since it is the basis for many of the other technologies discussed below.

Topic Tracking

A topic tracking system works by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user. Yahoo offers a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available. Topic tracking technology does have limitations, however. For example, if a user sets up an alert for “text mining”, s/he will receive several news stories on mining for minerals, and very few that are actually on text mining. Some of the better text mining tools let users select particular categories of interest or the software automatically can even infer the user’s interests based on his/her reading history and click-through information.

There are many areas where topic tracking can be applied in industry. It can be used to alert companies anytime a competitor is in the news. This allows them to keep up with competitive products or changes in the market. Similarly, businesses might want to track news on their own company and products. It could also be used in the medical industry by doctors and other people looking for new treatments for illnesses and who wish to keep up on the latest advancements. Individuals in the field of education could also use topic tracking to be sure they have the latest references for research in their area of interest.

Summarization

Text summarization is immensely helpful for trying to figure out whether or not a lengthy document meets the user’s needs and is worth reading for further information. With large texts, text summarization software processes and summarizes the document in the time it would take the user to read the first paragraph. The key to summarization is to reduce the length and detail of a document while retaining its main points and overall meaning. The challenge is that, although computers are able to identify people, places, and time, it is still difficult to teach software to analyze semantics and to interpret meaning. Generally, when humans summarize text, we read the entire selection to develop a full understanding, and then write a summary highlighting its main points. Since computers do not yet have the language capabilities of humans, alternative methods must be considered.

One of the strategies most widely used by text summarization tools, sentence extraction, extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position information are also used for summarization. For example, summarization tools may extract the sentences which follow the key phrase “in conclusion”, after which typically lie the main points of the document. Summarization tools may also search for headings and other markers of subtopics in order to identify the key points of a document. Microsoft Word’s AutoSummarize function is a simple example of text summarization. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary.

Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic. If organizations, medical personnel, or other researchers were given hundreds of documents that addressed their topic of interest, then summarization tools could be used to reduce the time spent sorting through the material. Individuals would be able to more quickly assess the relevance of the information to the topic they are interested in.

Categorization

Categorization involves identifying the main themes of a document [10] by placing the document into a pre-defined set of topics. When categorizing a document, a computer program will often treat the document as a “bag of words.” It does not attempt to process the actual information as information extraction does. Rather, categorization only counts words that appear and, from the counts, identifies the main topics that the document covers. Categorization often relies on a thesaurus for which topics are predefined, and relationships are identified by looking for broad terms, narrower terms, synonyms, and related terms. Categorization tools normally have a method for ranking the documents in order of which documents have the most content on a particular topic.

As with summarization, categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first. Categorization can be used in a number of application domains. Many businesses and industries provide customer support or have to answer questions on a variety of topics from their customers. If they can use categorization schemes to classify the documents by topic, then customers or end-users will be able to access the information they seek much more readily.

Clustering

Clustering is a technique used to group similar documents, but it differs from categorization in that documents are clustered on the fly instead of through the use of predefined topics. Another benefit of clustering is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. If someone goes to www.clusty.com, powered by Vivisimo, and type in “Saturn” in the search field, the returned topics include planet, photo, car and performance. This clustering tool allows the user to quickly narrow down the documents by identifying which topics are relevant to the search and which are not. Clustering technology can be useful in the organization of management information systems, which may contain thousands of documents, such as the Dow and ClearForest example described previously.

Concept Linkage

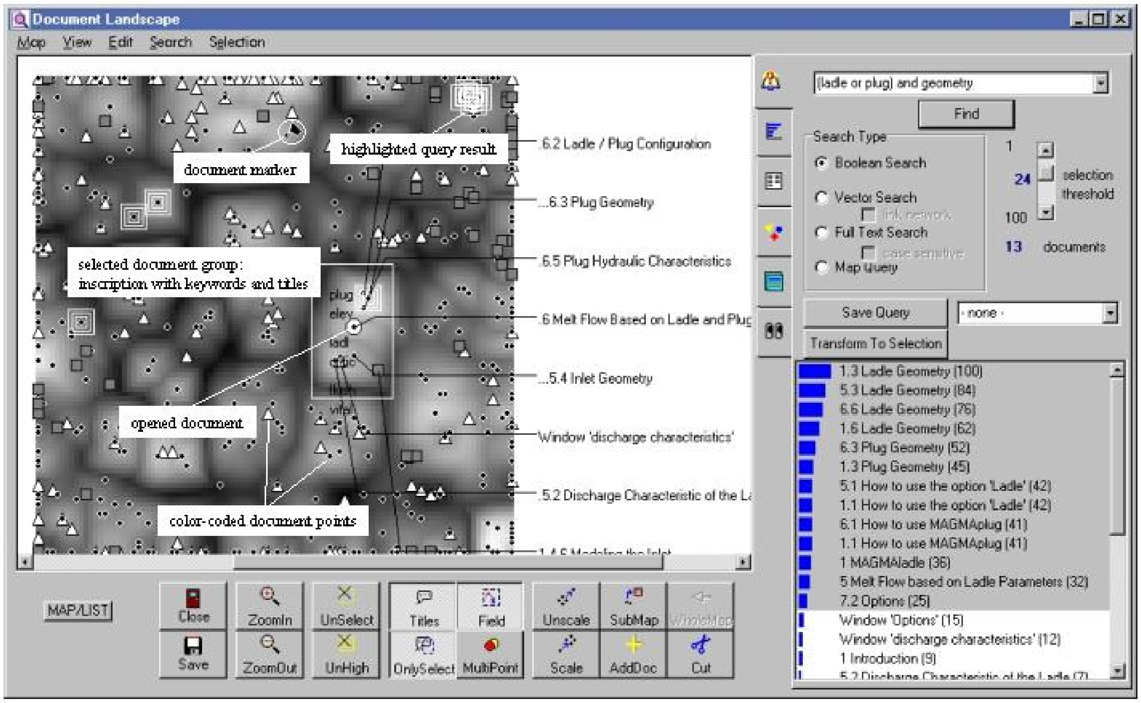
Concept linkage tools connect related documents by identifying their commonly-shared concepts and help users find information that they perhaps wouldn’t have found using traditional searching methods. It promotes browsing for information rather than searching for it. Concept linkage is a valuable concept in text mining, especially in the biomedical fields where so much research has been done that it is impossible for researchers to read all the material and make associations to other research. Ideally, concept linking software can identify links between diseases and treatments when humans can not. For example, a text mining software solution may easily identify a link between topics X and Y, and Y and Z, which are well-known relations. But the text mining tool could also detect a potential link between X and Z, something that a human researcher has not come across yet because of the large volume of information s/he would have to sort through to make the connection.

A well known non-technological example of this is Dan Swanson's research in the 1980's that identified magnesium deficiency as a contributor to migraine headaches [9]. Swanson looked at articles with titles containing the keyword "migraine", then from those identified keywords that appeared often within the documents. One such term was "spreading depression." He then looked for titles containing "spreading depression" and repeated the process with those documents. Then, he identified "magnesium deficiency" as a key term, and hypothesized that magnesium deficiency was a factor contributing to migraine headaches. There were no direct links between the two, and no previous research had been done suggested the two were related. The hypothesis was only made from linking related documents from migraines, to spreading depression, to magnesium deficiency. The direct link between magnesium deficiency and migraine headaches was later proved accurate by actual scientific experiments, showing that Swanson's linkage methods could be a valuable process in future medical research.

The work Swanson did by hand mimics the concept linkage technology that text mining products provide today and shows how valuable these products can be in medical fields. Experiments similar to Swanson's have been replicated through the use of automated tools that can be applied to text mining [4]. In the near future we expect that text mining tools with concept linkage capabilities will help researchers discover new treatments by associating treatments that have been used in related fields.

Information Visualization

Visual text mining, or information visualization, puts large textual sources in a visual hierarchy or map and provides browsing capabilities, in addition to simple searching. Informatik V's DocMiner [7] is a tool that shows mappings of large amounts of text, allowing the user to visually analyze the content. The user can interact with the document map by zooming, scaling, and creating sub-maps. Information visualization is useful when a user needs to narrow down a broad range of documents and explore related topics.



(ladle or plug) and geometry

Search Type
● Boolean Search
○ Vector Search
□ link network
○ Full Text Search
□ case sensitive
○ Map Query

1
24 selection threshold
100
13 documents

- none -

1.3 Ladle Geometry (100)
5.3 Ladle Geometry (84)
6.6 Ladle Geometry (76)
1.6 Ladle Geometry (62)
6.3 Plug Geometry (52)
1.3 Plug Geometry (45)
5.1 How to use the option 'Ladle' (42)
1.1 How to use the option 'Ladle' (42)
6.1 How to use MAGMAplug (41)
1.1 How to use MAGMAplug (41)
1 MAGMAladle (36)
5 Melt Flow based on Ladle Parameters (32)
7.2 Options (25)
Window 'Options' (15)
Window 'discharge characteristics' (12)
1 Introduction (9)
5.2 Discharge Characteristic of the Ladle (7)

Figure 2. Doc Miner's interface

The government can use information visualization to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them with a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own. Text mining has been shown to be useful in academic areas [1], where it can allow an author to easily identify and explore papers in which s/he is referenced.

Question Answering

Another application area of natural language processing is natural language queries, or question answering (Q&A), which deals with how to find the best answer to a given question [8]. Many websites that are equipped with question answering technology, allow end users to “ask” the computer a question and be given an answer. MIT has been accredited with implementing the first web-based natural query answering system called “START” (available at http://www.ai.mit.edu/projects/infolab/).

Q&A can utilize multiple text mining techniques. For example, it can use information extraction to extract entities such as people, places, events; or question categorization to assign questions into known types (who, where, when, how, etc.). In addition to web applications, companies can use Q&A techniques internally for employees who are searching for answers to common questions. The education and medical areas may also find uses for Q&A in areas where there are frequently asked questions that people wish to search.

3. Major Vendors and Applications

Tables 1 and 2 list major vendors[2] who have developed text mining technologies along with the features implemented in their tools. Some companies, such as ClearForest, focus exclusively on text mining tools, whereas in larger companies, such as IBM and SPSS, text mining tools are only a small portion of the software they market.

[2] It should be noted that the DocMiner example shown in Figure 2 is an academic tool, and is not offered for commercial re-sale so it is not included in the Tables.

	Company							
Feature	Inxight	Autonomy	Clearforest	SAS	Convera	Megaputer	SPSS	IBM
information extraction	x	x	x	x	x	x	x	x
topic tracking		x						
summarization	x	x			x	x		x
categorization	x	x	x	x	x	x	x	x
concept linkage		x	x	x				
clustering		x			x	x		x
information visualization	x						x	
question answering		x				x		

Table 1. List of text mining technologies offered by commercial vendors.

Company	Website	Product Names
Inxight	www.inxight.com	SmartDiscovery, VizServer
Autonomy	www.autonomy.com	IDOL Server, Retina
Clearforest	www.clearforest.com	ClearForest Text Analysis Suite
SAS	www.sas.com	SAS Text Miner
Convera	www.convera.com	Retreival Ware
Megaputer	www.megaputer.com	TextAnalyst
SPSS	www.spss.com	LexiQuest, Clementine
IBM	www.ibm.com	Intelligent Miner for Text, TAKMI

Table 2. List of vendor websites and the names of the text mining products that they offer.

The key to selecting a good text mining tool is finding a company that markets the technologies that meet your needs. The previous sections have provided several examples of how some industries may choose to apply text mining technology. Table 3 shows some additional applications. Table 3 does not contain all examples of industries or applications of text mining techniques, but rather represents some of the most likely applications in the areas of medical, business, government and education. More “x’s” can be added to the table and more industries can be added as the advantages and applications of text mining continue to increase. Data mining has been shown to be useful in the areas of telecommunications, geospatial data sets, biomedical engineering, and climate data [5], so there is definite potential for extending text mining to these areas and many others in the future.

	information extraction	topic tracking	summarization	categorization	clustering	concept linkage	information visualization	question answering
Medical:								
FAQ's	x			x		x		x
Drug design	x				x	x		
New treatment		x				x		
Business:								
Competitive Analysis		x	x					
Media impact / analysis		x						
Current Awareness		x						
Intellectual property infringement	x	x			x			
Customer support for FAQ's	x			x	x			x
Social network detection							x	
Content personalization		x			x			
Government:								
Homeland security: detecting terrorist networks	x	x			x	x	x	
Law enforcement: crime detection / prevention	x	x			x	x	x	
Education:								
Research on a topic		x	x	x				
Citation analysis	x				x		x	
FAQ's	x			x	x			x

Table 3. Some examples of where text mining tools can be applied to the fields of medicine, business, government, and education.

4. Conclusion

As the amount of unstructured data in our world continues to increase, text mining tools that allow us to sift through this information with ease will become more and more valuable. Text mining tools are beginning to be readily applied in the biomedical field, where the volume of information on a particular topic makes it impossible for a researcher to cover all the material, much less explore related texts. Text mining methods can also be used by the government's intelligence and security agencies to try to piece together terrorist warnings and other security threats before they occur. Another area that is already benefiting from text mining tools is education. Students and educators can find more information relating to their topics at faster speeds than they can using traditional ad hoc searching.

And perhaps the largest target for text mining developers right now is the business world. There are many businesses today with overwhelming amounts of information that they don’t use because they have no reasonable way of analyzing it. Text mining tools can help these businesses analyze their competition, customer base and marketing strategies, thereby allowing them to financially profit from the text mining software purchase. In order to successfully deploy a new text mining project, companies need to be sure to

- clearly define their goal and expectation from the text mining project. The goal should be in line with the company strategic goal and vision. The descriptions of the techniques contained in this article provides information about the possibilities of what this software can do in order to help a company set reasonable goals for a project.
- perform return on investment (ROI) analysis to justify both the tangible and intangible benefits to the company. Dow Chemicals was able to quantify many of the benefits of the software that they used in order to justify their investment in Clearforest. A clear cost justification may be necessary in order to receive the necessary top management support for a text mining project.
- talk to different vendors and their clients about their deployment experience and product support. We’ve provided a list of some of the more well-known vendors that can serve as a starting point for this investigation.
- integrate the text mining project with existing information technology (IT) infrastructure. For example, some companies may be able to integrate text mining software with their existing data warehousing infrastructure to provide more powerful business intelligence support.
- hire and train the right IT professionals. Text mining is an evolving field. New text mining techniques are under development and text mining products are being added to the market regularly. Companies must ensure that their IT personnel are being educated with the essential knowledge to make full use of the text mining software. This paper provides an overview of some of the latest techniques being

used at this time. But there is no question that other techniques and software will be added in the future.

In fact, one of the future trends for text mining applications appears to involve the integration of data mining and text mining into a single system. The combination of data and text mining is referred to as “duo-mining” [3]. SAS and SPSS have begun recommending duo-mining to their customers as a way of giving them the edge on consolidated information for better decision making. This process combination has proven to be especially useful to banking and credit card companies. Instead of only being able to analyze the structured data they collect from transactions, they can add call logs from customer services and further analyze customers and spending patterns from the text mining side. These new developments in text mining technology that go beyond simple searching methods are the key to information discovery and have a promising outlook for application in all areas of work. Companies with document collections that are collecting dust should invest in text mining applications that will help them better analyze their documents and provide pay-back with the useful information they can provide.

References

1. Bollacker, K.; Lawrence, S.; and Giles, C.L. A system for automatic personalized tracking of scientific literature on the web, *Proceedings of the ACM JCDL*, 1999.

2. Clearforest Dow chemicals case study. *http://www.clearforest.com/Customers/Dow.asp*, (2004),

3. Creese, G. Duo-Mining: combining data and text mining, *DM Review*, No. September, (2004), http://www.dmreview.com/article_sub.cfm?articleId=1010449.

4. Gordon, M.D.; Lindsay, R.; and Fan, W. Literature-based discovery on the WWW. *ACM Transactions on Internet Technology (TOIT)*, 2, 4, (2002), 262-275.

5. Han, J.; Altman, R.B.; Kumar, V.; Mannila, H.; and Pregibon, D. Emerging Scientific Applications in Data Mining. *CACM*, 45, 8, (2002), 54-58.

6. Hearst, M. What is text mining. *http://www.sims.berkeley.edu/~hearst/text-mining.html*, (2004),

7. Informatik http://www-i5.informatik.rwth-aachen.de/lehrstuhl/projects/DocMINER/DocMINER.html, 2004,

8. Radev, D.R.; Libner, K.; and Fan, W. Getting answers to natural language queries on the Web. *Journal of the American Society for Information Science and Technology (JASIST)*, 53, 5, (2002), 359-364.

9. Swanson, D.R. Two medical literatures that are logically but not bibliographically connected. *JASIS*, 38, 4, (1987), 228-233.

10. Yang, Y., and Pedersen, J.O. A comparative study on feature selection in text categorization, *the Proceedings of the 14th International Conference on Machine Learning*, Morgan Kaufmann, 1997, 412-420.