Understanding Factors that Affect Text Mining Results

Cherie Courseault Trumbach
University of New Orleans
2000 Lakeshore Dr
New Orleans, LA 70148
(504) 240 - 1456
ctrumbac@uno.edu

Alan L Porter
Search Technology
4960 Peachtree Industrial Blvd
Norcross, GA 30071-
(770) 441- 1457
aporter@searchtech.com

ABSTRACT
Text mining draws on various information resources analyzing both pre-determined fields and free text. This paper identifies the factors that impact the nature of end results when choosing a text mining approach. While it is not intended as a survey paper, this paper provides insight into the organization of the overall text mining domain. It draws upon current research examples for: Retrieval, Processing, Data Cleansing, Mining, and Visualization.

Keywords
Text mining, Information Retrieval, Data Cleansing, Extraction, Data Visualization

1. Introduction

Text mining is showing explosive growth. The large number of returns from searches frustrates user attempts to identify relevant information in a sea of hits. As a result, researchers are attempting to find new ways of processing the search results. The primary domains in which text mining approaches are being applied are the web; research publication databases, such as *Medline*; patent databases; and news source databases. This paper sorts text mining activities into major categories, then addresses the factors within each category that affect the end result. Individual algorithms are provided as an example of a particular approach. This paper has implications for tool developers and implementers who have tended to focus on choosing from among algorithms within one approach.

The overall text mining process can be roughly partitioned into five major technique elements: Retrieval, Processing, Cleansing, Mining, and Visualization (Figure 1). However, like most technologies, research domains begin to overlap over time. We offer Figure 1 as a conceptual benchmark for our discussion, recognizing that it is not an adequate categorization of evolving text mining practices.

2. Retrieval

"Document Retrieval" is the term often used to describe the act retrieving documents from a document collection. Much

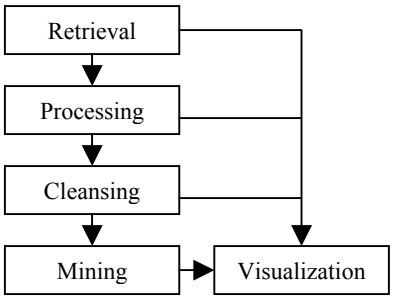

Figure 1. The Text mining Process

of the research in document retrieval is being done in support of Web search engines. In 2004, 37% of the *EI Compendex* documents on information retrieval contained either the word "Internet," "web," or "w w w". Initially, retrieval was based solely on Boolean search methods. However, the recent trend in document retrieval is to incorporate methods that were initially considered post-retrieval techniques into the retrieval process. For example, clustering may be used both to assist the user narrowing the results to focus on their area of interest or for query expansion by returning documents in a cluster with the search term [1,2,3]. Query expansion may also take place using thesaurus functions [4]. Kaji et. al. present a method for generating a thesaurus using term clustering as a means to traverse a domain-specific corpus [3]. Other researchers are looking to improve the document set by returning extracted entities into the search [5]. The results have had mixed results, depending on the entity that is returned. Overall, the trend in retrieval is combining methods and including user interaction.

3. Processing

Processing entails parsing individual terms from the text and tagging them into an appropriate category. Primarily, parsing entails first utilizing a Parts-of-Speech tagger to distinguish nouns, verbs, etc. Nouns provide the most interest for text data miners, because it is nouns that capture domain specific concepts[3]. However, identifying an actual noun phrase is challenging. Parsers may use Natural Language Processing techniques including NP Chunking to identify noun phrases. Search Technology's

VantagePoint uses Natural Language Processing to extract phrases from abstracts. It identifies and tags term by their parts of speech but does not include word sense disambiguation, distinguishing when the same word can be used as different parts of speech.

Many users require knowing the difference between an organization and a person. For these purposes, recall is much more important than precision because, in general, it would not be challenging to remove names from the "Proper Name" group, but missing proper names requires that an individual sort through the entire word list to identify proper names missed by the tool. Some of the tools that perform entity extraction are SRA's NetOwl, Inxight's Thingfinder, and IBM's Intelligent Miner. A free extraction tool is available for research from Sheffield Natural Language Processing Group named ANNIE. These extraction tools boast wonderful results in the 90% area for recall and precision in MUC and TREC data, which includes primarily newspaper sources. However, for less predictable formats, such as publication abstracts or web sites, the effectiveness drops considerably. In these areas, the drop in proper noun recall may cause problems. Another issue with many of the software products currently on the market, however, is that identifying identical entities only takes place within a document. For example, NetOwl uses anaphora resolution to link a last name listed in a document with a full name in the same document. The same is true for company acronyms and company full names. However, if the acronym or last name is in a different document, then the association is missed.

4. Cleansing

Once a list of noun phrases is developed, the list must be cleansed. Data Cleansing consists of the algorithms and methods that determine the final information used in analysis. Data Cleansing determines the quality of the information that is fed into the actual mining algorithms and ultimately the structure of the end result.

Cleansing is based on two principles: selection and compression. Selection is identifying which terms to include in analysis and at times determining the significance of that term in the document. For example, VantagePoint only uses words that meet a minimum frequency for clustering. Ahonen-Myka et al. uses only maximal frequent sequences of words based a frequency threshold[6]. In order to bolster the frequency of terms in abstracts or full text documents, compression is used. Compression is grouping together synonymous words/phrases. The most basic type of compression involves the variations of the same phrases such as "management of technology" and "technology management" or "computer" and "computers." VantagePoint's List Cleanup function uses a stemming algorithm and shared words in reverse order to improve the compression. At a more sophisticated level is the compression attempts to bring to together terms such as "Internet commerce" and "web commerce." Ahonen-Myka et al. described using the concept of equivalence class, which they defined as sets of phrases that occur together in the same documents frequently enough. Phrases belonging to some equivalence class are replaced by the name of the class. Another approach offered by Courseault combine words that share words in common such as "engineering science" and "general engineering science" [7]. In this approach, conflicts when a phrase shares words with different groups of words are resolved using a similarity measure, analyzing the context of the phrases.

5. Mining

In this paper, we will discuss the most common approaches in text mining: link analysis and cluster analysis. Link Analysis is the linking of information within documents. The most basic type of link analysis shows networks of word relationships, usually involving co-occurrence of some sort. It is a knowledge-poor statistical approach. Depending on the number of links, these networks can get very large and complex. The more powerful Link Analysis tools involve linking particular types of verbs with the doer and the object(s) of that action. SRA international incorporates this type of link analysis in order to identify links between entities in text and to identify key events in text. Hearst [8] is also doing some work in this area. This powerful approach, based on computational linguistics, requires a significant amount of training for individual domains. It can be used to develop a disease hypothesis or uncover a social impact that is not contained in any one document. Kostoff [9] is pursuing similar objectives, but his approach is a more statistically based effort.

Clustering is based on the co-occurrence of words, and it would, therefore, seem that the actual clustering algorithm chosen will not bring about large differences in the actual clusters developed [10,11]. This statement does not necessarily mean that the results are not somewhat different. The details of the chosen clustering algorithm are important to the end result and must be determined by the end goal. The difference is primarily based on factors such as whether the clusters are term clusters or document clusters or whether the location of certain words or documents have a distinct location in the space or can have multiple locations.

The cluster research contains a plethora of clustering techniques and additions to well known methods designed to improve the ability to find either documents or bits of information, as well as to provide a general landscape of

the documents. Hierarchical methods group items in a treelike structure. In contrast, non-hierarchical methods simply break the corpus into subsets [12]. Partitioning clustering divides the data into disjoint sets. Statistical clustering methods, such as factor analysis or Kohonen Self Organizing Maps, use similarity measures to partition documents[13,14]. In general, each of these methods is based on term frequency of co-occurrence. Shah offers one unique method. In this method, the semantic relationships between words in the document are captured [15].

Most clustering methods use document clustering as a way to maneuver through documents, especially as clustering is being promoted as a visualization method for document retrieval [16]. The increased number of internet sites have sparked a greater interest in this area [17]. In addition to finding the structure of document relationships, clustering can be utilized in a number of different ways to mine text. Watts, Courseault, and Kapplin present an algorithm based on combining various clustering techniques is used to find emerging technologies in a corpus containing over 10,000 publication records from a functional search [18]. This method is based on the term clustering approach taken in the VantagePoint software package. Term clustering is focused on displaying the relationship between concepts within documents rather than just the documents themselves. As stated in Section 2, clustering can also be reapplied to the original document set in order to improve information retrieval.

6. Visualization

Visualization can be a challenge that hinders the effectiveness of text mining techniques. An effective interface should allow the user to review, manipulate, search, explore, filter, and understand large volumes of data [19]. The visualization literature covers two main areas: general visualization principles and task completion

Robertson, Card, and Mackinlay address the issue of increasing the speed of information access to complete work processes[20]. They suggested various types of visualizations depending on the type of data. They suggested using a cone tree to represent hierarchical structures, a perspective wall for linear structures, a data sculpture for continuous data, and an office floor plan for spatial data. Text analysis could result in any of these data types. Hierarchical data can also be represented using geographical maps using metaphors for states, counties, and cities[21,22].

The number of dimensions is an important topic in text mining visualization. 3D is becoming a popular element in more complex visualization tools. However, the value of 3D or the optimal use of 3D is yet to be determined. Sebrechts et al compared 3D, 2D, and text versions of the visualization tool NIRVE on a corpus of documents that had been clustered. The text presentation had the fastest average completion time. However, as the participants gained experience, the 3D representation showed significant improvement. Color is a way to add a dimension. The most valuable use of color appears to be when it is used to represent concepts[23,24]. The effectiveness of color was found to decrease once there were more than five concepts shown [25]. Animation adds a time dimension. A work by Baker and Bushell which uses animation for cloud formation could be adapted for the formation of clusters [24].

In text mining clustering is an important concept that requires visualization, especially for large cluster maps. One of the challenges is providing the ability to maneuver through a cluster map, focusing on the details of the cluster, while maintaining perspective in relation to the entire cluster. Kosara et al presents a concept called Semantic Depth of Field in which the surrounding context to a key sentence is blurred[26]. The idea of using blur and cues can also be applied to finding documents in a tree structure.

7. Conclusion

There are many factors that affect text mining results and decisions must be made at each step of the process in order to ensure that the method matches the objectives of the user. At the Retrieval stage, it must be determined whether the objective is to capture all of the records, then use a post retrieval method for review to capture concepts or if the objective is to simply obtain a few "best" records out of the set. If the user is simply looking for records, at this point they are done. However, if they want the system to help in the analysis, more decisions must be made. First, the type of extraction must be determined. In extraction, three primary decisions need to be made. First, how is a "word" going to be identified. The complexity ranges from natural language processing or simple windows containing a certain number of words. Second, does the user need to simply identify the part of speech or do they need to identify specific information about a noun entity? Third, the system must determine which words to be included in further analysis. After word selection is made, the actual analysis can take place. Here again, more decisions, the goal of the analysis must drive the method and tool use. Is the goal to gather a landscape of the documents or to of the underlying document concept, to provide overview information, patterns or hidden relationships, or the needle in the haystack, or to find links between bits of information within documents. At the visualization stage, decisions must be made about the number of dimensions, whether certain information should be highlighted in its context. The Visualization should driven by the type of analysis.

The purpose of text mining is to provide insight into a large amount of text. Developers should look to combine methods and provide flexibility in the tools in order to enhance the capabilities of their tools and enable the user to apply the methods that fit their goals.

8. REFERENCES

[1]D. Roussinov and J. L. Zhao, "Automatic discovery of similarity relationships through Web mining"*Decision Support Systems*, pp. 149-166,vol.35,2003.

[2]T. T. Quan, S. C. Hui, and T. H. Cao, "A fuzzy FCA-based approach for citation-based document retrieval"*2004 IEEE Conference on Cybernetics and Intelligent Systems, Dec 1-3 2004*, pp. 577-582,2004.

[3]H. Kaji, Y. Morimoto, T. Aizono, and N. Yamasaki, "Navigation in an Association Thesaurus Automatically Generated from a Corpus"*Proceedings of the 16th Joint Conference on Artificial Intelligence*,1999.

[4]L. F. Soualmia, C. Barry, and S. J. Darmoni, "Knowledge-based query expansion over a medical terminology oriented ontology on the web" pp. 209-213,vol.2780,2003.

[5]C. McCabe, "Advancing Information Retrieval through Databases, Fusion, and Information Extraction"7/2000.

[6]H. Ahonen-Myka, "Discovery of Frequent Word Sequences in Text"*The ESF Exploratory Workshop on Pattern Detection and Discovery in Data Mining*, Retrieved from http://www.cs.helsinki.fi/u/hahonen/publications.html on 9/2002.

[7]C. R. Courseault, "A Text Mining Framework Linking Technical Intelligence from Publication Databases to Strategic Technology Decisions"*PhD Dissertation*,5/2004.

[8]M. Hearst, "Untangling Data Mining"*Proceedings of ACL '99: the 37th Annual Meeting of the Association for Computational Linguistics*,1999.

[9]R. N. Kostoff, J. A. Block, J. A. Stump, and K. M. Pfeil, "Information content in Medline record fields"*International Journal of Medical Informatics*, pp. 515-527,vol.73,2004.

[10]C. R. Courseault, "TPAC internal report" Retrieved from www.tpac.gatech.edu on 2004.

[11]C. H. Q. Ding, "Document Retrieval and Clustering from Principal Componenet Analysis to Self-Aggregation."*Proceedings from the 9th International Workshop on Artificial Intelligence and Statistics*,2003.

[12]A. V. Leouski and W. B. Croft, "An Evaluation of Techniques for Clustering Search Results"1996.

[13]M. Halkidi and M. Vazirgiannis, "Clustering Validity Assessment: Finding the Optimal Partitioning of a Data Set"2001.

[14]R. T. Freeman and H. Yin, "Adaptive topological tree structure for document organisation and visualisation"*Neural Networks*, pp. 1255-1271,vol.17,2004.

[15]C. Shah, "Automatic Organization of Text Documents in Categories Using Self-Organizing Map (SOM)"*IEEE's Regional Student Paper Contest*,2002.

[16]B. G. T. Lowden and J. Robinson, "An Analysis of File Space Properties using Clustering"*Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics*,vol.5,2002.

[17]O. Zamir and O. Etzioni, "Web document clustering: a feasibility demonstration"*Proceeding of the 21st annual internation ACM SIGIR conference on Research and Development in information retrieval*, pp. 46-54,1998.

[18]R. J. Watts, C. Courseault, and S. Kapplin, "Identifying Unique Information Using Principal Component Decomposition. Management of Technology: the Key to Prosperity in the 3rd Millennium"2000.

[19]N. Gershon and S. G. Eick, "Information Visualization: The Next Frontier"*Journal of Intelligent Information Systems*, pp. 29-31,vol.11,8/1998.

[20]G. G. Robertson, S. K. Card, and J. D. Mackinlay, "Information Visualization Using 3D Interactive Animation"*Communications of the ACM*, pp. 57-71,vol.36,4/1993.

[21]S. I. Fabrikant, "Evaluating the Usability of the Scale Metaphor for Querying Semantic Spaces, Spatial Information Theory: Foundations of Geographic Information Science"*Conference on Spatial Information Theory*, pp. 156-171,2001.

[22]S. I. Fabrikant, "Visualizing Region and Scale in Information Spaces"*Proceedings of the 20th International Cartographic Conference ICC 2001*, pp. 2522-2529,2001.

[23]J. Stasko, R. Catrambone, M. Guzdial, and K. McDonald, "An Evaluation of Space-Filling Information Visualizations for Depicting Hierarchical Structures" pp. 1-24,2/2000.

[24]M. P. Baker and C. Bushell, "After the storm: considerations for information visualization. Computer Graphics and Applications"*IEEE*, pp. 12-15,vol.15,5/1995.

[25]C. Perez and A. De Antonio, "3D visualization of text collections: An experimental study to assess the usefulness of 3D"*Proceedings - Eighth International Conference on Information Visualisation, IV 2004, Jul 14-16 2004*, pp. 317-323,vol.8,2004.

[26]R. Kosara, S. Miksch, and H. Hauser, "Focus + Content Taken Literally"*IEEE Computer Graphics and Applications*, pp. 22-39,1/2002.