Knowledge Extraction Using Rule Based Decision Tree Approach

P.Bhargavi B.Jyothi S.Jyothi K.Sekar
MotherTheresa Institute of Computer Applications,
Madanapalli, Andhra Pradesh, India

Summary

Text mining has been defined as "the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources" [6]. 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. This paper also concentrates to design text and data mining tool to extract the valuable information from curriculum vitae according to concerned requirements. The tool clusters the curriculum vitae into several segments which will help the public and private concerns for their recruitment. Rule based approach is used to develop the algorithm for mining and also it is implemented to extract the valuable information from the curriculum vitae on the web. Analysis of Curriculum vitae is until now, a costly and manual activity. It is subject to all typical variations and limitations in its quality, depending of who is doing it. Automating this analysis using algorithms might deliver much more consistency and preciseness to support the human experts. The experiments involve cooperation with many people having their CV online, as well as several recruiters etc. The algorithms must be developed and improved for processing of existing sets of semi-structured documents information retrieval under uncertainity about quality of the sources.

Keywords
Text mining, Text matching, Text Chunking, Text Mining Technologies and tools

1. Introduction

The research of text mining as a potential tool for TIES initiated mainly from interaction with technoloy policy and assessment Center(TPAC) at Georgia Institute of Technology [1], especially the research done by TPAC director [2]. In the process of performing the literature review, it was discovered that the office of the noval research(ONR) Science & Technology Division [3], itself has established a database of text mining related documents and useful links. Science & Technology(S&T) text mining is the application of text mining to highly detailed technical material(Kostoff [4],[5]). The three major components of S&T text mining are information retrieval, information processing and information integration.

2. What is text mining?

Text mining is used to extract useful information from text based files .It consists of the analysis of multiple text documents by extracting key phrases, concepts etc. and in the preparation of the text processed in that manner for further analysis with numeric data mining techniques(e.g., to determine co-occurences of concepts, key phrases, names, addresses, product names etc.). Figure 1 shows the sequence of flow in text mining.

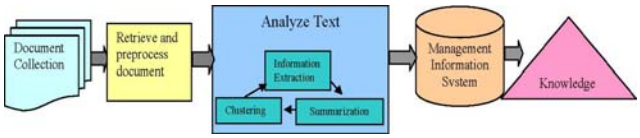

Fig1. An example of text mining

3. What is the purpose of text mining?

To discover and use knowledge that is contained in a document collection as a whole, extracting essential information from document collections and from a variety of different sources. Text mining lets executives ask questions of their text-based resources quickly extract information and final answers they never imagined.

4. The three steps to text mining

1. Preprocessing the text to distill the documents into a structured format.
2. Reducing the result into a more practical size
3. Mining the reduced data with traditional data mining techniques

Manuscript received July 5, 2008.
Manuscript revised July 20, 2008.

5. How Text Mining is different from Data mining ?

The difference between the two technology solutions is that while data mining extracts, analyzes and summarizes, numerical structured data, text mining handles large volumes of unstructured text based data. 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.

Currently there exist a range of tools to help organizations find and access data. The problem today is that there is too much information overload. Increasingly, companies are looking for a solution that will help them leverage their legacy data.

The main difference between text mining and data mining is that text mining considers all the information online determines key relationships and messages in that information, then helps users understand it. By contrast, data mining develops theories and / or models based on quantitative analysis of data.

6. Technology Foundation

Text mining makes use of 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. 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.

6.1 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.

6.2 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. 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.

6.3 Text 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. 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.

6.4 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,

6.5 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.

6.6 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.

7. Simple Application – Text Mining for short listing Curriculum Vitae

7.1 Objective of the system

The objective of the system is to analysis text under the resume documents to select best person from huge amount of resumes using rule based decision tree approach by converting text document and template text in to ASCII. Create rules using decision tree method to mine the text under documents. Displays the mined information under the table, now a table contains the selected information . Again the table is mined using data mining technique which contain only best top order list. The information provided by the system helps in making decision in order to select the best person quickly.

7.2 Methodology Applied – Rule based Decision Tree

7.2.1 Generating a decision tree

Derivation of the tree begins with the identification of the attribute .

7.2.2 The strengths of the decision tree methods

The decision trees are able to generate understandable rules. Decision trees perform classification without requiring much computation. Decision trees are able to handle both continuous and categorical variables. Decision trees provide a clear indication of which fields are most important for prediction or classification.

7.2.3 Decision tree applications

The decision tree method may be applied to any training problem and is thus suited for any data mining task. However, it has the unusual property of the producing as a by-product, an explicit and meaningful representation of the target function. In some cases, obtaining this by-product may actually be the main objective.

The tree provides a hierarchical representation of the distributional structure of the data. It also shows implicitly which variables are more significant with respect to classification decisions and it also suggests through the decision path groupings the way in which the variables are linked together. Decision trees thus have a range of potential uses further downstream in the data mining process.

The main steps in the complete algorithms are as follows:

(1) Initialize the decision tree as a root node and store all the examples there.
(2) Select the next node in the tree which contains non-uniform examples.
(3) Consider all possible decision branches that might be added to this node.
(4) Select the decision branch that divides the examples into maximally uniform groups.
(5) Add this decision branch to the node and allocate the examples to the new sub nodes according to their value of the tested attribute.
(6) Repeat from step 2 until all nodes contain uniformly classified examples.
(7) Finally, replace each leaf node with the appropriate classification decision, i.e., the class of the stored examples.

7.3 Output generation

Using the tree to classify a new example involves starting at the root and then descending through the tree, choosing branches at each stage according to the values contained in the example. When the tree descent process

reaches a tip node, the classification label associated with the node is produced as the classification of the example.

7.4 System Design and Development

7.4.1 Problem Analysis

In the first step, scan the documents, resumes and templates and convert them into ASCII files. In order to reduce the file size and more importantly to enable full text search, Intelligence information retrival(text mining) follow, in which the original word documents and templates are converted into ASCII files and based on knowledge approach (a set of decision rule) it will extracting needed information in a compact form. IIR also store the neeed text information in Data Base. In the last step from the selected list again extracting and producing top list in Data Base.
It is apparent that two kinds of information are available for analyzing curriculum vitae: selected list and the top list. Fig.2 shows the workflow in analysing curriculum vitae.

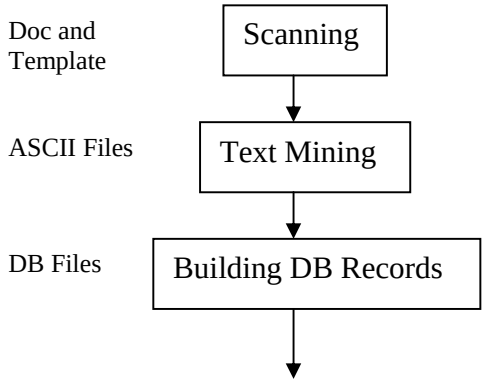

Fig 2: Workflow of the analyzing curriculum vitae

One straightforward approach is to extract the information from the applicant through the templates, will have the following drawbacks. Direct method can make errors in scutinizing the resume in raw list. The curriculum vitae are all correctly recognized but we will make errors in deciding area expertise. For some fancy curriculum vitae the information printed in special formats so that the system cannot recognize them at all.

7.4.2 Proposed Solution

7.4.2.1 Basic idea and problems

The intrinsic characteristic common for most curriculum vitae: The name of the person, age qualification and/or work experience. Based on the qualification matching, candidate can be called for an interview. Applying rule based decision tree intelligence information retrieval , the needed information is retrieved from the given resume. A huge amount of text string matches with the context, to accelerate the matches; a dynamic tree structured dictionary for each content page is build.

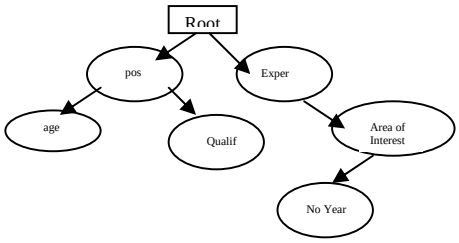

Fig 3: Tree Dictionary Structure Based on Resume

7.4.2.2 Graph representation of content pages

The resume is modeled as a directional graph with eachc node as a unique word. First lookup the vertex in the tree dictionary. Then find out the phrases starting from that word by traversing along all the edges from that vertex. Besides if both the start stare and end state are given, one phrase can be uniquely decided(range). This concept is used in text chunking.

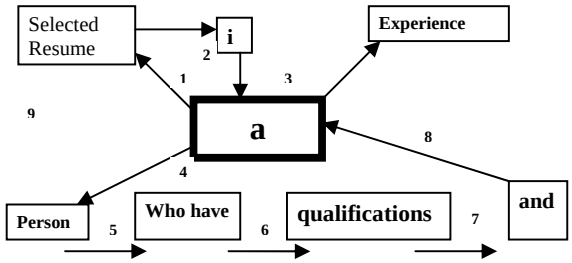

Fig. 4 Graph Description of Resume

7.4.2.3 Text match algorithm

In recent years, a lot of research is done in field of text mining, whose goal is to automatically spot important information from huge amount of text. One recucrring problem is to group documents by identifying similar phrases. Pattern detection ideasfrom data compression algorithm and suffix tree are used to locate common phrases among different documents. In pre analysis curriculum vitae, we cannot count on the multiple appearances of critical phrases. They may only show up

twice—one in the selected list and the other in the resume.

Step 1 : Build ASCII file, build tree dictionary and GR of resume.
Step 2 : Do text chunking of resume based on input values.
Step 3 : Score each body resume and make decision.

7.4.2.4 Text chunking

Locating key phrases in the resumes is referred as " Text Chunking", from natural language processing. In NLP text chunking involves dividing sentences into non overlapping segments. Based on the heuristics that a text segment in resume should have a corresponding maximal match in the body documents, design of the text chunking algorithm:

RangeList :={}
Foreach <word>, Bodydocuments
Find the longest match starting with word between Bodydocuments and resume. We put this match into a range(s,e).
If(s,e) overlaps with any existing range in rangelist, the two ranges are merged. Otherwise we inser it into the rangelist.
Endfor

The words in the same rectangle are put together after text chunking which cannot be grouped together due to OCR errors, is emphasized that no geometric layout information is used in the text chunking stage. The grouping result comes exclusively from text matches.

8. Major Vendors and Applications of Text Mining

Tables 1 and 2 list major vendors 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.

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

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

	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.

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

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 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" [13]. 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] Technology Policy and Assessment Center(TPAC) at Georgia Institute of Technology. http://www.tpac.gatech.edu

[2] Porter AL. Text Mining For Technology Foresight. http://www.tpac.gatech.edu/~darius/papers/foresight-outline.html , 2000.

[3]Office of Naval Research(ONR) Science & Technology http://www.onr.navy.mil/sci_tech/special/technowatch/default.htm

[4] Kostoff RN. Text Mining for Global Technology Watch

[5]Kostoff RN. Information Extraction from Scientific Literature with Text Mining,2001.

[6] A Roadmap to Text Mining & Web Mining http://www.cs.utexas.edu/users/perbronia

[7]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.

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

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

[10] 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.

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

[12] Hearst, M. What is text mining. *http://www.sims.berkeley.edu/~hearst/textmining.html*, (2004),

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

[14] 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.

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

[16]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.

[17] VantagePoint. http://www.thevantagepoint.com/pages/overview.html

[18] Van Germat J. Text Mining tools on the Internet. Intelligent Sensory Information Systems(ISIS) technical report series, September,2000.

[19] Watts RJ, and Porter AL. innovation Forecasting. http://www.tpac.gatech.edu/toa/inov.shml

[20] Porter AL and Detampel MJ. Technology Opportunities Analysis. Techcnology Forecasting & Social Change, Vol.49, 237-255,1995.

Dr. S.Jyothi , received the M.Sc(Maths) degree from Sri Venkateswara University. She received the Dr. degree from Sri Venkateswara University, India. She is working as an associate Professor in the department of computer science, school of mathematical and physical sciences, Sri Padmavathi Mahila Viswa Vidyalayam, Tirupati,India.

P.Bhargavi, received the M.Sc(Computers) and M.Tech degrees from Sri Krishnadevaraya .University and Vinayaka. missions, India in 1997 and 2005, respectively. She is doing her research in data mining. She is working as an associate professor at Madanapalli Institute of Technology & Science (from 2004) in the Dept. of Computer Science Engineering.

B.Jyothi, received the M.Sc(Maths) and M.Phil(Maths) degrees, from Alagappa university and Madurai Kamraj University, India in 2002 and 2004, respectively. She has been working as an Assistant Professor at Mother Theresa Institute of Master of Computer Applications, India since 2004.

K.Sekar, received the MCA and M.Tech degrees, from Madras University and Vinayaka Missions, India in 1998 and 2005, respectively. He is a research associate and is working as an associate professor at Madanapalli Institute of Technology & Science (from 2004) in the Dept. of Computer Science Engineering.