Text Mining: Finding Nuggets in Mountains of Textual Data

Jochen Dörre, Peter Gerstl, Roland Seiffert
SWSD, IBM Germany
P.O. Box 1380, D-71003 Böblingen
{doerre,gerstl,seiffert}@de.ibm.com

ABSTRACT

Text mining applies the same analytical functions of data mining to the domain of textual information, relying on sophisticated text analysis techniques that distill information from free-text documents. IBM's Intelligent Miner for Text provides the necessary tools to unlock the business information that is "trapped" in email, insurance claims, news feeds, or other document repositories. It has been successfully applied in analyzing patent portfolios, customer complaint letters, and even competitors' Web pages. After defining our notion of "text mining", we focus on the differences between text and data mining and describe in some more detail the unique technologies that are key to successful text mining.

Keywords

Text mining, feature extraction, text categorization, clustering, customer relationship management

1. MINING TEXT

"There is gold hidden in your companies data" - and data mining promises to help you finding it. And in fact, many successful applications of data mining prove that this is true indeed. But data mining addresses only a very limited part of a company's total data assets: the structured information available in databases. Probably more than 90% of a companies data are never being looked at: letters from customers, email correspondence, recordings of phone calls with customers, contracts, technical documentation, patents, ... With ever dropping prices of mass storage, companies collect more and more of such data online. But what can we get from all this data? More often than not, the only way the data is made usable - outside of very specific applications for subsets of that data - is by making it accessible and searchable in a companies intranet. But today there is more you can do: text mining helps to dig out the hidden gold from textual information. Text mining leaps from old-fashioned information retrieval to information and knowledge discovery.

KDD-99 San Diego CA USA

The first challenge in text mining is that information in unstructured textual form is not readily accessible to be used by computers. It has been written for human readers and requires natural language (NL) interpretation. The information really is buried inside the text. Although a full-blown interpretation of even just factual knowledge stated in unrestricted natural language is still out of reach with current technology, there are tools using pattern recognition techniques and heuristics that are capable of extracting valuable bits of information from arbitrary free-text. Extracted information ranges here from identifying what companies or dates are mentioned in a document to summaries of a document.

But there is more to text mining than just extracting information pieces from single documents. Where the mining task in the sense of data mining comes in is when one has to deal with huge collections of documents. Tools performing that task generally support classification - either in a supervised or in an unsupervised fashion - on the documents seen as objects that are characterized by features extracted from their contents.

We use the "mining" metaphor for both knowledge discovery processes described above: the *extraction* of codified information (features) from single documents as well as the *analysis* of the feature distribution over whole collections to detect interesting phenomena, patterns, or trends. Any non-trivial application of "text mining" necessarily involves both of those mining phases.

2. HOW TEXT MINING DIFFERS FROM DATA MINING

Data mining generally involves the steps.

1. Identification of a collection
2. Preparation and feature selection
3. Distribution analysis.

As described in the last section, text mining essentially adds to the preparatory phase the very complex feature extraction function. Moreover, the cardinality of the feature set that can be extracted from a document collection usually is very high, easily running into several thousands. There are two consequences of this affecting the overall text mining process.

1. The feature selection task is quite different, since it is no longer feasible to have a human examine each feature to decide whether to use it or not. Different approaches to accomplish this task range from using simple functions, e.g., to filter out features that can be considered as noise, to complex analytical processes possibly involving human intervention.

2. The distribution analysis step must be able to handle highly dimensional, but sparsely populated feature vectors. This often requires special versions and implementations of the analytical algorithms used in data mining.

2.1 IBM Intelligent Miner for Text

In 1998, IBM for the first time introduced a product in the area of text mining: the Intelligent Miner for Text[1]. It is a software development toolkit - not a ready-to-run application - for building text mining applications. It addresses system integrators, solution providers, and application developers. The toolkit contains the necessary components for "real text mining": feature extraction, clustering, categorization, and more. But there are also more traditional components, e.g., the IBM Text Search Engine, the IBM Web Crawler, and drop-in Intranet search solutions. We will not describe the latter part of the product here, but just mention it. These components are essential to build applications that use the information generated in a mining process, e.g., in an Intranet portal for the company.

3. MINING WITHIN A DOCUMENT: FEATURE EXTRACTION

The task of feature extraction is to recognize and classify significant vocabulary items in unrestricted natural language texts. Examples of vocabulary found are shown in Figure 1. The process is fully automatic - the vocabulary is not predefined. Nevertheless, as the figure shows, the names and other multiword terms that are found are of high quality and in fact correspond closely to the characteristic vocabulary used in the domain of the documents being analyzed. In fact, what is found is to a large degree the vocabulary in which concepts occurring in the collection are expressed.

...	Dana
Certificate of deposit	debt security
Chronar	debtor country
CMOs	Detroit Edison
Commercial bank	Digital Equipment
Commercial paper	dollars of debt
Commercial Union Assurance	end-March
Commodity Futures Trading Commission	Enserch
Consul Restaurant	equity warrant
Convertible bond	Eurodollar
Credit facility	...
Credit line	...
Credit Lyonnais	
Credit Suisse	
Credit Suisse First Boston	

Figure 1 Some of the vocabulary found by feature extraction in a collection of financial news stories. The canonical forms are shown

In general, our implementation of feature extraction relies on linguistically motivated heuristics (cf. [2]) and pattern matching together with a limited amounts of lexical information, such as part-of-speech information. We neither use huge amounts of lexicalized information, nor do we perform in-depth syntactic and semantic analyses of texts. This decision allows us to achieve two major goals:

1. Very fast processing to be able to deal with mass data
2. Domain-independence for general applicability

The extracted information will be automatically classified into the following categories:

- Names of persons, organizations and places
  like Mrs. M. Albright, National Organization of Women Business Owners, or Dheli, India
- Multiword terms
  like joint venture, online document, or central processing unit
- Abbreviations
  like EEPROM for Electrical erasable programmable read-only memory
- Relations
  like Jack Smith-age-42, John Miller-own-Knowledge Corp., Janet Perna-General Manager-Database Management
- Other useful stuff: numerical or textual forms of numbers, percentages, dates, currency amounts, etc.

We always assign a so-called canonical form to each feature we find. Simple, but very useful examples include normalized forms of dates, numbers etc. This allows applications to use that kind of information very easily, even though, e.g., a number was written in words in a text. A canonical form also abstracts from different morphological variants of a single term, e.g., singular and plural forms of the same expression are mapped to the same canonical form.

More complex processing is done in the case of names. All the names that refer to the same entity, for example President Clinton, Mr. Clinton and Bill Clinton, are recognized as referring to the same person. Each such group of variant names is assigned a canonical name, (e.g., "Bill Clinton") to distinguish it from other groups referring to other entities ("Clinton, New Jersey"). The canonical name is the most explicit, least ambiguous name constructed from the different variants found in the document. Associating a particular occurrence of a variant with a canonical name reduces the ambiguity of variants. For example, in one document, "IRA" is associated with the Irish Republican Army, while in another it may be associated with an Individual Retirement Account.

Optionally, the tool computes statistical data on the distribution of features in documents and document collections. This allows, e.g., the clustering tool to judge the significance of a feature in a document with respect to the statistical background properties of the whole document collection.

4. MINING IN COLLECTIONS OF DOCUMENTS: CLUSTERING AND CATEGORIZATION

With the mapping of documents to feature vectors that describe them in place, we can perform document classification in either of two ways.

Clustering is a fully automatic process, which partitions a given collection into groups of documents similar in contents, i.e., in

[1] For more information also on other components, please refer to http://www.software.ibm.com/data/iminer/fortext

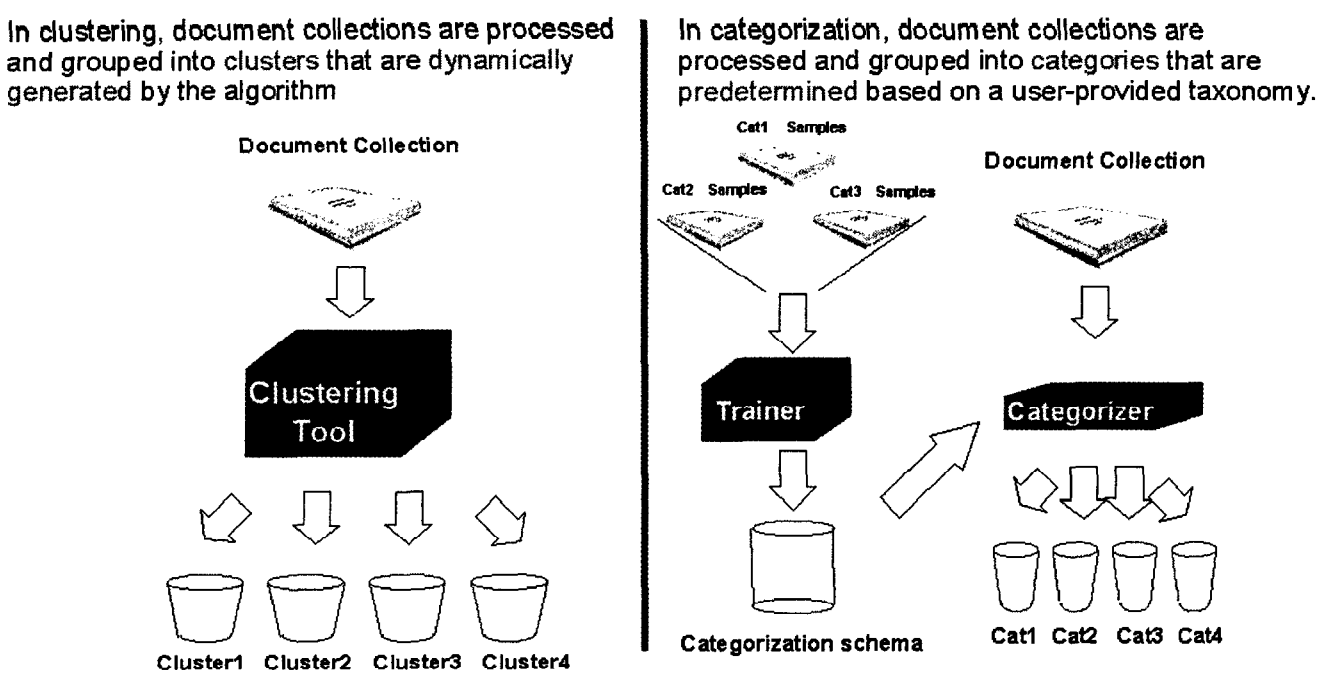

Figure 2 Clustering versus Categorization

their feature vectors. Intelligent Miner for Text includes two clustering engines employing algorithms that are useful in different kinds of applications. The Hierarchical Clustering tool orders the clusters into a tree reflecting various levels of similarity. The Binary Relational Clustering tool uses "Relational Analysis" (cf. [1]) to produce a flat clustering together with relationships of different strength between the clusters reflecting inter-cluster similarities. Both tools help to identify the topic of a group by listing terms or words that are common in the documents in the group. Thus, clustering is a great means to get an overview of the contents of a collection.

The second kind of classification is called (text) categorization. The Topic Categorization tool assigns documents to preexisting categories, sometimes called "topics" or "themes". The categories are chosen to match the intended use of the collection. In the Intelligent Miner for Text those categories are simply defined by providing a set of sample documents for each category. All the analysis of the categories, feature extraction and choice of features, i.e., key words and phrases, to characterize each category, is done automatically. This "training" phase produces a special index, called the categorization schema, which is subsequently used to categorize new documents. The categorization tool returns a list of category names and confidence levels for each document being categorized. Documents can be assigned to more than one category. If the confidence level is low, then typically the document would be put aside so that a human categorizer can make the final decision.

Tests have shown that, provided the set of defined categories does match the subject matter of incoming documents, the Topic Categorization tool agrees with human categorizers to the same degree as human categorizers agree with one another.

5. TEXT MINING APPLICATIONS

As is the case with data mining technology, one of the primary application areas of text mining is collecting and condensing facts as a basis for decision support. The main advantages of mining technology over a traditional 'information broker' business are:

- The ability to quickly process large amounts of textual data which could not be performed effectively by human readers.
- 'Objectivity' and customizability of the process - i.e. the results solely depend on the outcome of the linguistic processing algorithms and statistical calculations provided by the text mining technology
- Possibility to automate labor-intensive routine tasks and leave the more demanding tasks to human readers.

Taking advantage of these properties, text mining applications are typically used to:

- Extract relevant information from a document (summarization, feature extraction, ...)
- Gain insights about trends, relations between people/places/organizations, etc. by automatically aggregating and comparing information extracted from documents of a certain type (e.g. incoming mail, customer letters, news-wires, ...).
- Classify and organize documents according to their content; i.e. automatically pre-select groups of documents with a specific topic and assign them to the appropriate person.
- Organize repositories of document-related meta-information for search and retrieval

- Retrieve documents based on various sorts of information about the document content

This list of activities shows that the main application areas of text mining technology cover the two aspects (1) knowledge discovery (mining proper) and (2) information 'distillation' (mining on the basis of some pre-established structure). Due to the lack of space we will concentrate on a single application that uses IBM's Intelligent Miner for Text to support both aspects, discovery and distillation: customer relationship management.

5.1 Customer Relationship Management

Based on the Intelligent Miner for Text product IBM offers an application called CRI (Customer Relationship Intelligence) that is designed to specifically help companies better understand what their customers want and what they think about the company itself.

After selecting the appropriate set of input documents (e.g. customer complaint letters, phone call transcriptions, e-mail conversation) and converting them to a common standard format, the CRI application uses the feature extraction and clustering tools to derive a database of documents which are grouped according to the similarity of their content.

Depending on the purpose of the data analysis, the user might select different parameters for the preprocessing (e.g. concentrate on names and dates) and for the clustering step (e.g. use a more or less restrictive similarity measure).

When clustering customer feedback information, the result exposes groups of feedback that share important linguistic elements, e.g. descriptions of a difficulty customers have with a certain product or support organization. Information of this type can be used to identify problematic areas that need to be addressed. Sometimes the cluster itself may provide clues about how the problem could be solved, e.g. do the documents have something in common which is independent from the problem description such as the location, background, ... of the customers that raised the issue?

As a separate step after a set of useful clusters has been identified and, probably manually enhanced, the categorization tool can be used to assign new incoming customer feedback to the identified categories.

The CRI application incorporates both, the distillation and discovery aspects of text mining. A typical usage scenario starts with an unstructured collection of documents (transcripts, mails, and scanned letters), creating some interpretable structure by means of clustering which corresponds to the aspect of discovery. The refinement and extension of the clustering results by means of interpreting the results, tuning of the clustering process, and selecting meaningful clusters emphasizes the aspect of distillation.

6. CONCLUSION

In this paper we have described our notion of "text mining" and relevant core technology components. IBM's product for text mining applications, the Intelligent Miner for Text, enables customers to use these new technologies in practical text mining applications. We have shown this by describing a customer relationship management application that is based on IBM's text mining components. As this application shows, text mining to-date can be used as an effective business tool that supports the creation of knowledge by preparing and organizing unstructured textual data (discovery) and by supporting the extraction of relevant information from large amounts of unstructured textual data through automatic pre-selection based on user-defined criteria (distillation). Using automatic mining processes to organize and scan huge repositories of textual data can significantly enhance both the efficiency and quality of a routine task while still leaving the more challenging and critical part of it to the one who can do it best, the human reader.

7. REFERENCES

[1] F. Marcotorchino, "Block seriation problems: A unified approach", *Applied Stochastic Models and Data Analysis*, **3**, 73-91 (1987).

[2] N. Wacholder and Y. Ravin, *Disambiguation of Proper Names in Text*, Proc. of the 5th Conference on Applied Natural Language Processing, April 1997, Washington, D.C.