Information Mining with the IBM Intelligent Miner Family

Daniel S. Tkach

An IBM Software Solutions White Paper

February, 1998

First Edition (February 1998)

Abstract

Information mining refers to the process of extracting previously unknown, comprehensible, and actionable information from **any** source - including transactions, documents, e-mail, web pages, and other, and using it to make crucial business decisions.

The two most pervasive types of information are structured data and text, therefore information mining includes data mining and text mining.The IBM Intelligent Miner Family that comprises the IBM Intelligent Miner for Data and the IBM Intelligent Miner for Text, provide the most advanced and comprehensive set of solutions for information mining in the industry.

This document describes the information mining operations and techniques as they are implemented in the IBM Intelligent Miner Family, and highlights the applications that have proven to provide competitive advantages in many enterprises world-wide.

Notices

References in this publication to IBM products, programs, or services do not imply that IBM intends to make these available in all countries in which IBM operates. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any of the intellectual property rights of IBM may be used instead of the IBM product, program, or service. The evaluation and verification of operation in conjunction with other products, except those expressly designated by IBM, are the responsibility of the user.

IBM may have patents or pending patent applications covering subject matter in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:

IBM Corporation
IBM Director of Licensing
208 Harbor Drive
Stamford, Connecticut 06904
U.S.A.

Version Date: February 24, 1998

Trademarks

Intelligent Miner For Data, Intelligent Miner for Text,AIX, and Visual Warehouse are trademarks of International Business Machine Corporation.

Microsoft, Windows, and the Windows NT are registered trademarks of Microsoft Corporation.

UNIX is a registered trademark in the United States and other countries licensed exclusively through X/Open Company Limited.

Other company, product, and service names used in this publication may be trademarks or service marks of others.

Information Mining

The state of the art of information technology ensures that there is no shortage in the sources of information an organizations can tap to run their businesses. Operational systems record transactions as they occur, day and night, and store the transaction data in files and databases. Documents are produced and placed in shared files or in repositories provided by document management systems. The growth of the Internet, and its increased worldwide acceptance as a core channel both for communication among individuals and for business operations, has multiplied the sources of information and therefore the opportunities for obtaining competitive advantages.

Business Intelligence Solutions is the term that describes the processes that together are used to enable improved decision making. **Information mining,** comprising **data mining** and **text mining**, is one of those processes. It uses advanced technology for gleaning valuable insights from these sources that enable the business user making the right business decisions and thus obtaining the competitive advantages required to thrive in todays competitive environment.

We call **Information Mining** to the process of extracting previously unknown, comprehensible, and actionable information from **any** source - including transactions, documents, e-mail, web pages, and other, and using it to make crucial business decisions.

Data, Information, and Knowledge

Data is the raw material we get from the sources It can be a set of discrete facts about events, and in that case, it is most usefully described as *structured* records of transactions, and it is usually of numeric or literal type. But documents and Web pages are also a source of an *unstructured* data, delivered as a stream of bits which can be decodified as words and sentences of text in a certain language. Industry analysts estimate that unstructured data represent 80% of an enterprise information compared to 20% from structured data; it comprises data from different sources, such as text, image, video, and audio; text, is however, the most predominant variety of unstructured data.

Data has in itself very little meaning. When a customer goes to a gas station and buys gas, this transaction can partially be described by data: when the purchase was made, how many gallons were bought, how much was paid, and how it was paid (cash, credit, check, or ATM).

This data tells nothing about why the customer went to that service station and not another one, cannot predict how likely it is that he will come back, and whether the service station is well or badly run[1]. Data describes only a part of what happened; it provides no judgment or interpretation and no sustainable basis of action. But data is important to organizations because it is the essential raw material for the creation of information.

Information has been defined as “data endowed with relevance and purpose”[2]. Information is usually described in a communications schema involving messages between a source and a receiver. Information is most valuable when it makes a difference for the receiver to know about it. For instance, the information “you are 250 ft above the ground in the middle of the city of New York” has a very different relevance to the prospect tennant of a penthouse than to the pilot of an airplane. The relevance of information is therefore related to the context in which it is received. In a business environment, the context is provided by the business process in which the information will be used. Thus in a large bank, information about the trends of the stock market in Southeast Asia is far more relevant to a corporate loans risk assessment process than to home equity loans.

Textual data is usually semantically richer than numeric data because in text we rarely deal with individual words but with complete text documents that have meaning per se. For that reason, to access structured data we talk about “data access”, while to access a document we talk about “information retrieval”, because each text document provides information. This information may however not be relevant to the task at hand. If we want to retrieve from the Internet, documents referring to “Credit Suisse”, and we are not careful, we will end up with thousands of documents describing everything from credit cards to the making of cheese, Swiss style. And even if we know how to establish the condition of the search in order to look for those two words together, we will still end up with hundred on documents related to Credit Suisse operations, from which we have to sift those which are relevant to our task on hand. The enormous amount of information the user receives and has to digest gave recently birth to a new term “infoglut”. The business users are “drowning in information” while starving for the insights that will allow them to make better decisions.

[1] Davenport and Prusak, 1977
[2] Attributed to Peter Drucker

Insights are part of the universe of knowledge. Although the concept of knowledge is hard to pin down, and is the object of many controversies, the reality is that we all recognize knowledge when we see it, be it in the head of an expert that tells how to tackle a particular problem, or in a document that provides us with the required elements to solve that problem. A working definition provided by the META group is that knowledge is “a set of logical connections among pieces of information whose relationship is revealed through context/process familiarity " (10/97).

We talk of “tacit knowledge” - latent in the head of the expert - and “explicit knowledge”, when it has been transferred to an accessible media. Explicit knowledge can be declarative, that is, the knowledge is available but not organized to solve our specific problem or can be procedural, when the knowledge is stated as a sequence of steps that guaranty success if followed appropriately.

An example of declarative knowledge is a set of business rules, documented in the procedures manual of a department. The rules are clear and explicit, and their use relies on the capacity of the user to apply these rules correctly in each business situation. An example of procedural knowledge appears in work flow management: the sequence of steps to achieve the successful completion of a task is already defined and enforced by the computer.

The IBM Intelligent Miner Family is a set of offerings that enables the business professional and in general any knowledge worker to use the computer to extract meaningful information and useful insights from both structured data and text. Although the general problems to solve (e.g.. clustering, classification) are similar for the different data types, the technology used in each case is different, because it needs to be optimized to the media involved, the user needs, and to the best use of the computing resources. For that reason, the IBM Intelligent Family is comprised of two specialized products: the IBM Intelligent Miner for Data, and the IBM Intelligent Miner for Text. As it will be discussed later in this paper, both products can be used separately or concurrently to maximize the quality and value of the derived insights.

Information Mining Technology Overview

We have defined information mining as the process of extracting previously unknown, comprehensible, and actionable information from any source. This definition exposes the fundamental differences between information mining and the traditional approaches to data analysis such as query and reporting
and online analytical processing (OLAP) for structured data, and from full text search for textual data. In essence, information mining is distinguished by the fact that it is aimed at the discovery of information and knowledge, without a previously formulated hypothesis.

By definition, the information discovered through the mining process must have been previously unknown[3], that is, it is unlikely that the information could have been hypothesized in advance. For structured data,
the interchangeable terms “data mining” and “knowledge discovery in databases” describe a multidisciplinary field of research that include machine learning, statistics, database technology, rule based systems, neural networks, and visualization. Text mining technology is also based on different approaches of the same technologies.

Both data mining and text mining share key concepts of knowledge extraction, such as the discovery of which features are important for clustering, that is, finding groups of similar objects that differ significantly from other objects (also called segmentation in the case of structured data). They also share the concept of classification, which refers to. finding out to which class it belongs a certain database record, in the case of data mining, or to a document, in the case of text mining. The classification schema (also known as taxonomy) can be discovered automatically through clustering techniques (the machine finds the groups or clusters and assigns to each cluster a generalized title or cluster label that becomes the class name). In other cases the taxonomy can be provided by the user, and the process is called categorization.

The role of the machine learning technologies used in information mining is to enable the computer to recognize patterns that lead to useful, actionable, business insights. As an example, suppose that we want to detect credit cards frauds, that is, be able to identify transactions that could be made without the credit card owner’s consent. The computer algorithm analyzes the buying habits of the card owners and be able to recognize their patterns as well as deviations that could be attributed to fraud intents. The learning process includes a training stage, providing the mining program with positive and negative examples of valid transactions. After the training, the algorithm should be able to induce (learn) a definition of a valid transaction from the examples and to predict (classify) that a new transaction is or not a valid one, and therefore authorized it or not. The

[3] Cabena et al.

attributes that define the validity of a new transaction are discovered by the learning process: we will present to the computer all the attributes of the transaction, and the learning algorithm will decide which of them are meaningful for determining the validity of a transaction.

A distinctive feature of an information mining program is the quality of its algorithms, which determines the way the program can learn complete and consistent definitions of the concepts it deals with, such as "a valid transaction". The learned definition of a concept is *complete* if it recognizes correctly all the instances of the concept [4] ; in our case it means that it does not classify some legitime transactions as fraudulent. The definition of a concept is *consistent* when it does not classify any invalid transaction as a valid one.

The award winning IBM Intelligent Miner Family products feature a broad variety of high quality machine learning algorithms that is unique in the marketplace, both in their quality as well as in their scalability, that is, the ability of mining extensive amounts of data without a disproportional increase in the response time, and has enabled the deployment of successful real life applications of target marketing, market basket analysis, risk management, fraud management, customer relationship management, and competitive intelligence worldwide.

Data Mining with the IBM Intelligent Miner for Data

One of the key differences between of data and text mining approaches to knowledge discovery is due to the inherent differences in the way the sources need to be handled. As we discussed before, structured data has little semantic meaning, and when extracted from a normalized database there is no relationship among the attributes of a record that can be derived from the data itself. For instance, if the age of a person is recorded incorrectly, we cannot correct it based on the other attributes we are considering such as sex or type of payment chosen. Because the data mining process will reach incorrect conclusions based on wrong or missing data, the knowledge discovery process requires a previous stage in which the data is prepared for mining. This preprocessing of data does not occur in text mining: the correctness of a document is not an issue, only its content. Fuzzy logic and neural network techniques take care of misspellings, and the use of canonical expressions handle variations of the same concept, such as Mr. Clinton, President Clinton, and Bill Clinton, that may appear in a document or sets of documents

[4] Adriaans and Zantinge

In data mining the quality of the data to be mined is key to obtain meaningful results. Data preparation is an important step that may consume as much as 80% of the data mining efforts. Once the data source is selected, the data needs to be cleaned. This includes handling the duplication of records due to negligence, typing errors, or of changes in the environment, such as having records for the same customers with different addresses because the customer moved, misspelled the name, or gave a false address. Record fields may lack domain consistency, as it happens when people enter 11-11-11 for their birth date to accelerate the data entry process, or when 1991 is entered as 1901. Data may also be enriched with syndicated demographic data, and may need to undergo a number of transformations such as creating data aggregations, converting individual addresses to region numbers, changing dates to month numbers starting on a given date to perform time series analyses on the data[5], and other modifications.

The steps of the data mining process can be therefore summarized as follows:

- Determination of the business objectives: the business problem or challenge needs to be clearly defined. For instance, to segment a file of car insurance claims and analyze the profiles of the claimants in the segments.
- Data preparation: includes the determination of the data to be mined (the insurance claims file), and the preprocessing of the data, for instance, to handle missing values.
- Select the appropriate data mining technique, in our example, segmentation (clustering), based on the business objectives, the data characteristics, and the available computational equipment.
- Interpret and evaluate the results, using, for instance, visualization tools to analyze the results, and display and drill down on the customer segments. The objective could be to evaluate the significance of each variable (e.g. sex, age, number of dependents, and commute distance) for a given segment. To that end, the user could compare the variable distribution for a segment with the variable distribution for the overall population.

[5] Adriaans and Zantinge

Several iterations of one or more of these process steps may be required.

Data Mining Operations and Techniques

The major data mining operations are:

Predictive modeling: we use previous observations to build a model of a certain concept and we use later that model to predict if a new observation fits that concept. For instance, we could build a model of a loyal customer using our database, and later use this model to find out if certain customers are of the loyal type or if we are in danger to lose them.

Database segmentation or clustering: we partition a database into segments that contain similar records, that is, records that share a number of properties defined by the value of their attributes. Which attributes define the clustering is discovered by the mining process. For instance, "urban, single, wealthy males" may be an interesting segment for some marketing campaign, while "suburban professional mothers with two children or more" could be found to be an homogeneous sub population for another.

Link analysis: we want to find links between the individual records or sets of records in a database. These links are called *associations.* Variants of link analysis such as *associations discovery* are used to discover the relationship between products and services customers tend to purchase together or sequentially over time. Supermarkets use a variant of this operation called *market basket analysis* to discover which products to display together in an aisle.

Deviation detection: we want to understand why certain values, called outliers, exhibit deviations from some previously known expectation and norm.

Operations, Techniques and Algorithms

Some applications are implemented by one type of operation: for instance, target marketing is almost always implemented by means of database segmentation, while fraud detection could be implemented by any of the four operations depending on the nature of the problem and the input data[6].

These operations are implemented by techniques defined through algorithms, that is, predetermined sequences of calculations and logic prepared to obtain the desired result.

[6] Cabena et al.

When we do predictive modeling, classification is usually implemented using either *tree induction*, where the program builds a decision tree for classifying new cases, or *neural induction*, where a structure called neural network is trained to recognize a certain pattern. Neural networks have the advantages over decision trees that they are rather immune to noisy data (e.g. missing values). Their drawback is that they accept only numeric input, therefore categorical data, such as “male” or “female” must be recoded in numeric form (e.g. . male=0, female=1).

Value prediction is usually implemented through statistical techniques such as *linear regression* and *nonlinear regression.* A recent technique called *radial basis function,* implemented with neural networks, has demonstrated more robustness and flexibility than traditional regression approaches.

Database segmentation can be implemented using *demographic clustering* in which the records are compared with all the segments created by the data mining run. It uses techniques based on the *distance* between records which is based on the number of record fields that have similar values. A voting system called Condorset is used to assign a record to a cluster. Database segmentation can also be implemented using *neural clustering* a technique which employs a type of neural network called *Kohonen feature map* which clusters together similar customers, products, or behaviors, and defines the typical attributes of an item that falls in a given cluster or segment.

Link analysis uses simple counting techniques to uncover *association rules*, that is, rules that govern the affinities among the collection of items, and also for *sequential pattern discovery,* that is, the detection of patterns where the presence of some item in transactions is followed by the presence of other items in other transactions over a time period. (E.g. people who buy shirts will buy ties over a certain period of time). Graphical techniques are used in link analysis for *similar time sequence discovery,* that is, finding sequences similar to a certain sequence. For instance, when used in stock market analysis, this technique could help to relate stocks price behavior to sequences of values of market variables over time.

Deviation detection relies heavily on statistical analysis and visualization. Visualization techniques are among the most powerful devices for identifying hidden patterns in data. They are particularly useful for detecting phenomena that hold for a rather small subset of the data, and go undetected when statistical techniques are used. Therefore, visualization techniques are useful for detecting deviations, while statistics are used to measure their significance.

It is clear that a rich set of high quality algorithms is essential to obtain quality mining results. The award winning IBM Intelligent Miner for Data offers the most comprehensive set of high quality algorithms in the industry supporting all the mining operations required to obtain meaningful insights from the client's operational data. Moreover, these algorithms scale up, that is, they are capable of handling very high volumes of data with excellent response times.

Applications of Data Mining

Most current data mining applications fall under the categories of market management, risk management, or fraud management. Each application in a category uses defined sets of operations and techniques that have proven to clients and practitioners to yield the best results

Market management is an application area where data mining has been applied with singular success. **Database marketing** is the most widespread marketing management application. The objective is to drive effective targeted marketing and promotional campaigns though the mining of corporate databases that record customer product preferences and public information about customer demographics and lifestyles. The mining algorithms determine clusters of consumers sharing the characteristics that makes them attractive for marketing efforts. Database marketing includes cross selling, market basket analysis, and is a key component of customer relationship management applications.

Risk management applications help manage the risk associated with insurance because data mining is a useful technology to predict property or casualty losses for a given set of policy holders. In addition, understanding the total loss exposure can support improvement of the overall insured portfolio. More generally, risk management deals with the broader business risks arising from competitive threat, poor product quality, and customer attrition.

Attrition means loss of customers, especially to competitors. Attrition is an increasing problem in an increasingly competitive marketplace, and data mining is used in retail, finance, utilities, and telecommunication industries to predict likely customer losses by building models of vulnerable customers, that is, customers that exhibit characteristics typical of someone who is likely to leave for a competitive company.

Fraud management data mining applications have demonstrate their benefits in areas where many transactions are processed, making the respective company vulnerable to fraud. Health care, credit card services, and telecommunications, are at the forefront of using data mining to guard against fraud and potential fraud.

Business professionals that are in search of the insights that will provide their company with a competitive advantage understand clearly the business problem they are trying to solve, but are not necessarily aware of how the source data they need has to be handled, and the right techniques to obtain the results they are looking for.

Early approaches to data mining considered mandatory to support the business professional with specialists with skills related to mining technology and algorithms, generally professionals with degrees in Mathematics, Statistics, and related topics. Data mining projects were almost research projects, because the whole area was poorly understood, there were no major successful applications deployed, and the user interfaces were rather crude.

Today, data mining is far down the road, major successes have been achieved, and the experience acquired has led practitioners to define the following success factors for a data mining application:

The right application: the application should fulfill requirements derived from clearly understood business objectives and have the potential to make a significant impact on the business problem or market opportunity.

The right people: the basic team should include a business analyst and a data analyst. Data management specialists are also needed to facilitate access to physical data and metadata.

The right data: a clean supply of data from a limited set of data sources. In the ideal case, the data would come from the data warehouse.

The right tools: In addition to a comprehensive set of powerful algorithms the tool should provide extensive data preparation facilities, and should be able to use the hardware, operating system, and database resources efficiently. The scalability characteristics should allow the user to solve meaningful problems on small departmental servers while allowing for growth in order to solve complex problems involving very large amounts of data through optimized approaches and massive parallelism if required.

The productivity of the business professional and the impact on the enterprise business is clearly dependent on the speed of obtaining insights from the mining process. The technology issues should therefore not stand in the way of the business professional in order to obtain results.

The solution provided by IBM to facilitate the use of the Intelligent Miner for Data for the business professional is to provide excellent usability features, but most important, make available “ready to use” applications attacking the most pressing problems of key industries. The set of applications is called the **IBM Discovery Series** and it sits on top of the IBM Intelligent Miner for Data, solving the problem of the business professional that needs to determine which sequence of operations and techniques will provide the sought results.

Data Mining Summary

The IBM Intelligent Miner for Data is the most advanced data mining product in the marketplace. The applications provided under the IBM Discovery series enables the business expert to focus on the business results instead of worrying about the underlying mining techniques, allowing for substantial increases in the productivity of the mining process.

The IBM Global Business Intelligence Solutions (GBIS) division, as well as the IBM Global Services (IGS) organization are providing support to organizations using data mining helping them to take care of the success factors and to integrate the newly built data mining solutions into existing business applications, making data mining a repeatable, streamlined process. The IBM Consulting Group also helps building the necessary enterprise culture of data-driven business intelligence.

Text Mining with the IBM Intelligent Miner for Text

In 1995 analysts predicted that unstructured data, such as text, will become the predominant data stored online[7]. Today, just the Internet has over 500 million pages of information and is expected to reach 1500 million pages by the year 2000. The growth of the Internet, and the availability of very large amounts of documents stored online that contain information of great potential value, have created the need for tools to assist the users to extract the key concepts in the heap of information without having to read them all, and retrieve in a fast and effective way the information of their interest.

Another driver for developing tools for mining text resides in the fact that enterprises are becoming client-centric. Customer Relationship Management (CRM) is key in every industry; many interactions with the customer are text based or verbal, and mining these interactions provides

[7] Forrester

crucial business insights.This situation presents a huge opportunity to make more effective use of repositories of business communications, and other unstructured data, by using computer analysis.

The problem with textual information, however, is that it is not designed to be handled by computers. Unlike the tabular information typically stored in databases today, documents have limited internal structure, if any. Furthermore, the important information they contain is not explicit but is implicit, buried in the text. Hence the “mining” metaphor -- the computer rediscovers information that was encoded in the text by its author.

The IBM Intelligent Miner for Text has three major components: the **Advanced Search Engine** called TextMiner, the **Web Access Tools** which include an optimized web search engine called **NetQuestion** and a **Web Crawler,** and the **Text Analysis Tools**. All the tools in the Intelligent Miner for Text are designed to be used in building applications that deal with text. Many of these tools are information extractors which enrich documents with information about their contents, because the first step in text mining is usually to extract key features from texts to act as “handles” in further processing. Examples of features are the language of the text, company names, multiword concepts (such as “computer hardware”) or dates mentioned. After the feature extraction step, the next step may be to assign the documents to subjects from a cataloging scheme, then to index them for ad-hoc searching.

The general objective of a text mining system is to minimize the time a user spends in the steps leading to understanding the content of a document or a collection of documents. Text mining involves therefore two aspects: information retrieval, and text analysis. *Information retrieval* systems facilitate users finding the information they need. *Text analysis* tools help extracting key knowledge from text, organize documents by subject, and find predominant themes in a collection of document using supervised or unsupervised machine learning techniques. They also help the user to express their real knowledge needs, and provide navigational facilities. The IBM Intelligent Miner for Text provides tools to help the user efficiently in all the aspects related to text mining including full text search, text analysis, and web document query and retrieval.

Information Retrieval

Information retrieval is an operation with the goal to fulfill a given query (a request for information)[8]. As text queries may have to be performed against very large document sets with a response time acceptable to the user, we need an *index based* approach to achieve that speed level. A direct string comparison based search is unsuitable for the task.

[8] Novak

The index file for a collection of documents is built “off-line” (prior to the query. An index file is similar to a card catalog in a library. At query time, the system matches the query against representations of the documents, not against the documents themselves, much in the same way that we search for a book by consulting the catalog cards and not by browsing the books on the shelves. The system can retrieve the documents referenced by the indexes that satisfy the query if the user wants them. But it may not be necessary to retrieve the full document because in addition to a pointer to the document, index records may also contain all the important information of the documents they point to, allowing the user to extract information of interest without having to read the source document.

.

There are two basic retrieval models: a) *Boolean*, in which the document set is partitioned in two disjoint parts: one fulfilling the query and one not fulfilling it, and b) *relevance ranking based* in which all the documents are considered relevant to a certain degree. Boolean logic models use exact matching, while relevance ranking models use fuzzy logic, vector space techniques (all documents and the query are considered vectors in a multidimensional space, where the shorter the distance between a document vector and the query vector, the more relevant is the document), neural networks, and probabilistic schema. In a relevance ranking model, low ranked elements may even not contain the query terms. Advanced information retrieval systems, like the one provided by the IBM Intelligent Miner for Text, minimize the time the users requires to find the documents that are most relevant to their needs.

The Advanced Search Engine: Text Miner

TextMiner, the advanced fully-featured search engine delivered with Intelligent Miner for Text, allows for the construction of high-quality information retrieval systems. These systems may cover a wide range of possible applications for large collections of documents, which can be written in any of sixteen different languages and stored using multiple file formats. TextMiner does in-depth document analysis during indexing and allows for sophisticated query enhancement and result preparation to supply high-quality information retrieval. It is a fully-featured product providing full text search facilities in addition to the ability to index and search in many languages, use supporting thesauri, and process queries with a combination of natural-language and Boolean search conditions.

Search Engine Features

The basic full text search engine of TextMiner offers a number of advanced features such as result list clustering and relevance feedback which help a user to find the most relevant documents. making it one of the most advanced products of its kind on the market today. TextMiner is a client/server application and allows for a great number of concurrent clients performing searches and other tasks. A important feature is the online update capability, that is, TextMiner is able to perform indexing tasks without having to suspend searches.

Multiple search paradigms

TextMiner implements a number of different search paradigms inside the same search engine.The heart of the search engine is an index structure that supports **Boolean**, **free text**, and **hybrid** queries. **Phonetic** searches are possible on the same index structure. A special purpose index supports **fuzzy** searches and the double-byte character set-based languages Japanese, Chinese and Korean..

Boolean queries allow for conjunction, disjunction, and exclusion of search terms. Individual search terms may be single words or phrases. Boolean search allows for very exact specification of a user’s information need. Because of its complexity and the flexibility pure Boolean search is usually a typical “expert search", for example performed by a professional librarian.

Free text queries are based on the probabilistic retrieval model. Probabilistic retrieval covers a broad range of applications in information retrieval and has demonstrated very good performance in multiple large-scale comparative evaluations.. In general, pure free text queries are easy to use even by novice users of a retrieval system. .

Hybrid queries combine free text and Boolean queries in a unique, patented way. The goal is to overcome the problems of pure free text queries. Basically, a hybrid query is a free text query that restricts the result set to the documents that also match the Boolean part of the query. This allows for negative specifications in free text queries that are not supported by pure free text system. For example, it would be possible to search for documents about “market share of Japanese cars” (free text) but not “Toyota” (Boolean). Hybrid queries could be ideally used for “advanced search” functionality to allow an experienced user to perform more effective information retrieval tasks.

For any of these query types, additional expansions using synonyms, thesaurus information, or previously extracted features as described below may be applied to the search terms.

Through the construction of an additional **n-gram index**, TextMiner also supports **fuzzy searches**. Indexing and search is based on n-grams, that is, matching sequences of n characters. Searches may use exact or fuzzy matching. Fuzzy matching allows for a limited deviation of the search term from the index term in the document. For example, person names may be found without knowing the exact spelling in many cases. The input text is not processed linguistically for this index. This makes this kind of search language-independent.

Phonetic searches are supported in a way such that all occurrences of similar-sounding words are also retrieved. This is particularly useful whenever the exact spelling of a term to be searched is not known.

Linguistic processing for queries

Linguistic processing of search terms is controlled by the user through qualifiers in the query language. Query processing aims at making search terms broader so that more of the relevant documents in a collection are retrieved. This can be achieved by expanding a query to include synonyms or related terms from a user-supplied thesaurus.

Web Search

NetQuestion is the search engine included in the web tools provided with the IBM Intelligent Miner for Text. It is designed to build global WWW search services or centralized intranet search services. It is built on the same technology that TextMiner uses, but it is optimized to handle the large amounts of information that are typically stored on Web sites, which provides for faster indexing and response to large volumes of queries.

Use of Advanced Search in IBM Products

TextMiner is the integrated full text search facility imbedded in IBM DB2 Universal Database V5 and in IBM Digital Library V2. NetQuestion is the search engine for IBM Lotus Go, IBM's corporate home page ***www.ibm.com*** and IBM's 1500-server intranet, Network Computing Framework.

Text Analysis

Text analysis is performed by the set of operations on text that are more akin to data mining. It includes finding the key concepts in the document grouping a set of documents based on similarities of the key concepts they contain, and defining a user developed schema (taxonomy) and have the system classify the documents according to this schema.

To analyze documents they have to be accessible to the tools. To that end, the documents may be retrieved through search, or may simply be gathered in a repository. The Text Analysis Tools that are part of the IBM Intelligent

Miner for Text do not require starting a search phase as a first step to analysis, however, a combination of text search and text analysis provides a very powerful approach to knowledge discovery in text.

Functions in this grouping analyze text to select features for further processing. They can be used by application builders.

Language Identification

The language identification tool in the IBM Intelligent Miner for Text can automatically discover the language(s) in which a document is written. It uses clues in the document's contents to identify the languages, and if the document is written in two languages, say French and Dutch, it determines the approximate proportion of each one. The determination is based on a set of training documents in the languages. Its accuracy as shipped in the tool suite is usually close to 100% even for short text. The tool can be extended to cover additional languages or it can be trained for a completely new set of languages. Its accuracy in this case can be easily higher than 90%, even with limited training data. Applications of the language identification tool include: automating the process of organizing collections of indexable data by language; restricting search results by language; or routing documents to language translators.

Feature Extraction

The feature extraction component of the Intelligent Miner for Text recognizes significant vocabulary items in text. The process is fully automatic -- the vocabulary does not need to be predefined. Nevertheless, 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. This makes Feature Extraction a powerful Text Mining technique.

Among the features automatically recognized are

- Names, of people, organizations and places
- Multiword terms
- Abbreviations
- Other vocabulary, such as dates and currency amounts

For instance, analyzing a group of financial news stories, the extractor recognized **credit facility, credit line, Credit Lyonnais, and Credit Suisse** as four separate concepts. On the other hand, and by use of a canonical form for each concept, "President Clinton", "Mr. Clinton", and “Bill Clinton”, were recognized as the same entity, and different from Clinton, NJ.

Clustering

Clustering is a fully automatic process which divides a collection of documents into groups. The documents in each group are similar to each other in some way. When the content of documents is used as the basis of the clustering (as in the Intelligent Miner for Text), the different groups correspond to different topics or themes that are discussed in the collection. Thus, clustering is a way to find out what the collection contains. To help to identify the topic of a group, the clustering tool identifies a list of terms or words which are common in the documents in the group.

Clustering can also be done with respect to combinations of the properties of documents, such as their length, cost, date, etc. The clustering tools in the Intelligent Miner for Data are applicable to this kind of problem.

An example of clustering in Text Mining is to analyze e-mail from customers to discover if there are some important themes that have been overlooked. The effect of clustering is to segment a document collection into subsets (the clusters) within which the documents are similar in that they tend to have some common features. Clustering can be used to

- Provide an overview of the contents of a large document collection
- Identify hidden similarities
- Ease the process of browsing to find similar or related information.

Categorization

Categorization tools assign documents to preexisting categories, sometimes called “topics” or “themes”. The categories are chosen to match the intended use of the collection. By assigning documents to categories, the Intelligent Miner for Text can help to organize them. While categorization cannot take the place of the kind of cataloging that a librarian can do, it provides a much less expensive alternative.

For example, documents on an intranet might be divided into categories like “Personnel policy”, “Lab News” or “Commuter information” By using automatic categorization, documents can be assigned to an organization scheme which makes it easier to find them by browsing, or by restricting the scope of a text search. Classification of e-mail, for instance as “urgent”, “can wait”, and “junk”, allows the users to prioritize effectively their work without having to read all the messages.

.

Enhancing Search using the Text Analysis Tools

The linguistic processing of documents and queries is further enhanced for English by applying a number of text analysis techniques. When building a *feature index*, TextMiner discovers and extracts names of persons, places, or organization, domain-specific multi-word terms, and abbreviations. The extracted information can be used to expand queries later on and therefore significantly enhance the recall quality of the retrieval system.

Result list clustering groups the results of a query into sets of related documents. This eases the user's comprehension of search results. The result list clustering uses the same technology as the clustering tool provided in the tools suite. For performance reasons it has been limited to a maximum of 200 documents.

Relevance feedback allows the user to mark documents on the result list of a query as relevant or irrelevant. The information provided is used to reformulate the query. The result of that query will then be more focused on documents related to the user's information interest, that is, the precision of the retrieval system will be improved.

The Web Tools: NetQuestion, the Web Crawler and the Web Crawler Toolkit

The Web tools provide technologies to build intelligent Internet/intranet Web sites. They allow companies to leverage the use of Internet and intranets to gain access to relevant information. Information analysts can use these tools to crawl and collect information available on the Internet or intranet servers. The information collected that way is examined and analyzed to obtain meaningful insights for business decision making, such as competitive or marketplace trends. The tools support push and pull mode of information access.

Net Question

NetQuestion is a powerful, full-text search engine that can be used to build a global WWW search service or a centralized intranet search service.
It is designed and optimized to handle the large amounts of information that are typically stored on Web sites. Therefore, the document analysis and
query processing are more streamlined, compared to TextMiner, to provide for faster indexing and response to large volumes of queries. NetQuestion, however, features the same online update mechanism for indexing as used by TextMiner and other important components, such as client/server handling and queue mechanisms.
.

Documents to be indexed by NetQuestion can be provided in either plain text or text with HTML markup. Sample CGI scripts and HTML forms to develop a search interface are provided with the system. Administration can be performed through command line functions.

.

Since NetQuestion does not use dictionaries, it can be used for all single-byte character set languages, Boolean queries allowing for phrase and proximity searches as well as for front-, middle-, and end-masking using wild cards. Precise term searches are optimized for Web applications in both Internet and intranet environments.

NetQuestion features high speed indexing and retrieval where one precise index is built. An optimized and reduced index spans about 35% to 40% of the document size. An ultra compact index of about 10% of the document size can be built for users not needing full context information (no need for proximity and free-text searches.)

Net Question also provides sophisticated lexical affinities-based ranking for free-text and hybrid queries, advanced relevance ranking, and detection of misspellings in documents, expanding the search request accordingly. NetQuestion can, in some circumstances, even pick up misspellings of words that are not in a dictionary, such as brand names or new technology terms. For instance, during indexing NetQuestion notices that one of the occurrences of "Toyota" is misspelled as "Toyotta." If someone later tries to search for "Toyota", NetQuestion automatically adds "Toyotta" to the query. .In addition, linguistically fuzzy searches are available for English documents

The Web Crawler.

The Web Crawler is a robot that starts at one or more Web sites and follows selected HTML links. It retrieves objects of any content type and language, such as HTML, text, images, audio, or video and stores them to the local file system for further processing. For example, an indexer can use HTML and other text documents to build an index of documents. Types and number of levels of HTML links can be selected through a customization step. The Web Crawler can monitor Web-page activities and changes to optimize retrievals.

The Web Crawler Tool kit

.

In addition to the Web Crawler which is a ready-to-run implementation, the IBM Intelligent Miner for Text includes a **Web Crawler Tool kit** allowing the users to develop Web crawlers according to their needs. The tools use

DB2 to store metadata information, therefore a restricted use version of DB2 Universal Database is included with the product.

The Web crawlers can run on a single machine and can also generate a user-specified number of crawler copies that run in parallel or on multiple machines. These copies can be configured to independently crawl disjoint subsets of very large Webs. The individual crawl results consisting of data objects and their metadata can be shared for subsequent processing across machines either through shared file systems only, or through shared file systems and the parallel query capability of DB2.

Text Mining Applications

We can find applications for text mining in multiple areas of an enterprise. Many of them have been characterized as “front office” applications; including customer relationship management, e-mail processing, help desk operations, contract management, and sales force automation. Most knowledge users in an enterprise benefit from “knowledge delivery applications”, that is, applications that focus on knowledge discovery, retrieval, and delivery to the user, either in pull or in push modality. A last category of applications are the “business intelligence delivery” applications, such as competitive analysis, trends watch, patent analysis, and corporate image analysis.

For instance, customers call an enterprise to place a order, complain about services or products, request some information, or to provide suggestions. Call analysis allows the enterprise to increase their knowledge about customers, understand better their complains, and to learn the perception of these customers of the enterprise and of its competitors.
In other areas of the company, such as R&D and marketing, it is very interesting to know which are the promising new technologies, which applications result from a new technology, which companies are most active in their area of business, and how the enterprise compares to its competitors.

Information Mining Across the Boundaries: CRM

Customer Relationship Management applications are a good example of the advantages of combining data and text mining in a single application.

It is productive, for instance, to provide such functions as automatic customer satisfaction indicators, automatic calls or complaints routing, build a typology of complain letters, and all the other instruments whose business value is to maintain customer loyalty, and get a better customer understanding.

Using data and text mining allows the user to analyze at the same time the customers signaletic data (such as age, sex, number of children) together with the full text information provided by their complaints letter, opinion survey entry fields, and business transaction related text and data. The business insights obtained from the combined mining of text and data are far more powerful that the ones obtained from one data type alone.

Text Mining Summary

Text mining is key to addressing the information overload problem, and it is very effective for providing new and improved understanding of customer requirements and satisfaction. Leading edge technologies of the IBM Research division included in the IBM Intelligent Miner for Text make this text mining product the most advanced in the market place providing the power of advanced mining algorithms while ensuring an excellent performance and scalability of the applications. The IBM Global Business Intelligence Solutions (GBIS) division has extensive experience in the design and development of text mining applications for the utilities, telecomunications, government, finance, and farmaceutical industries, and is helping customers world-wide to deploy these applications and obtain rapidly competitive advantages from the knowledge discovered in text.

References

1. Davenport and Prusak, *Working Knowledge*, Harvard Business School Press, 1998
2. Peter Drucker, cit. 1.
3. Cabena et al., *Discovering Data Mining*, PTR-ITSO, 1998
4. Adriaans and Zantinge, *Data Mining*, Addison-Wesley, 1997
5. Adriaans and Zantige, op. cit.
6. Cabena et al., op. cit.
7. The Forrester Report, *Coping with Complex Data*, *April 1995*
8. Hans Novak, *Text Search Architecture*, IBM Technical Report, 1997

Additional Information

This white paper is one of a suite of Business Intelligence papers available from IBM.

To interactively view IBM Business Intelligence white papers or to download Adobe Acrobat PDF files for viewing and printing, access this web site: **http://www.software.ibm.com/data/**

There you will find information about the IBM Business Intelligence products and solutions. Selecting the ***"Publications"*** link on this page will guide you to the white paper section.