

# ***Text Mining: A Powerful Tool for Knowledge Management***

by Dr. Antonis Spinakis, Managing Director of QUANTOS SARL.

## **1. Definition**

Text Mining (TM) is the process of extracting novel, undetected and unstructured knowledge “hidden” in a large collection of unstructured text documents, using advanced technology. It enables the knowledge worker to uncover relationships in a text collection and to explore them in order to discover new knowledge. TM is particularly relevant today because of the enormous amount of knowledge that resides in text documents whether on the Internet, within the enterprise, elsewhere, or any combination of these sources.

It is similar to Data Mining in that both deal with large amounts of data and aim at knowledge discovery within that data. Data Mining, however, focuses on discovery within already structured collections — in databases, data warehouses, and other corporate and external repositories. TM, by contrast, concentrates on the ever-increasing flow of text data of all kinds that come to the attention of the knowledge worker. In just dealing with the Web, for example, the average knowledge worker needs tools that help him or her cut through the irrelevant or already known information and focuses their attention on the truly important or novel. Additional sources such as news feed, email, work product, letters and publications add to the need to intelligently mine textual data.

To sum up with, TM is of critical value to any organization that needs to process text data and to make that information available across its organizational structure.

## **2. Knowledge Management and TM**

Knowledge Management (KM) could be broadly defined as the process of capturing a company’s collective expertise wherever it resides- in databases, on paper, or in people’s head-and distributing it to wherever it can help produce the biggest payoff. One of the main components of KM is Knowledge Discovery, described as the process of discovering information when you are not exactly sure what information you have. Examples of Knowledge Discovery include:

- What new markets are there for my existing products?
- What is really out there on my Internet/Intranet?
- What are my competitors doing?
- What do my customers think about my products and services?
- What are the new developments in my market?

- Who is doing research that might be related to my project?

TM is the necessary tool of the Knowledge Discovery process since it enables us to extract knowledge “hidden” in text documents.

### **3. Functionalities / Tools / Activities**

A company possesses and monitors electronic information in both structured and unstructured forms. TM provides the capability to exploit the vast amount of unstructured business information, both documents and Web pages, in data repositories on Intranets and the Internet. It is estimated that unstructured information represents 80% of the total business information available to a company. To address this "data digestion" problem, sophisticated solutions are required that turn unstructured data into information that knowledge workers can use to make informed decisions and solve business problems.

To become more specific, TM process encompasses a variety of functionalities that enable the user to explore information hidden in textual data. The most fundamental of these functionalities are reported in the sequel.

- *Information and Event Extraction.*

This facility can, automatically, find and index key phrases in texts, such as company names, personal names, product names, dates, and monetary expressions, while it can also detect duplicate documents in an archive.

- *Search and retrieval of information hidden on unstructured text documents*

This eases the process of browsing to find similar or related information

- *Semantic analysis of documents.*

To identify hidden structures between groups of text documents

- *Clustering and classification (of documents).*

To provide a summary of the contents of a large document collection. For instance, a summary of the contents of a customer feedback collection could indicate where a company's products or services need improvement.

Generally, we could say that TM can be used anywhere there is a large amount of text which needs to be analysed.

### **4. Benefits**

The benefits that one can derive from using TM solution include:

- *Increased productivity of knowledge workers:* TM makes it easier for knowledge workers to navigate within a huge collection of texts by following the key concepts that are important to them and to discover the knowledge they need.

- *Increased value of corporate information:* It allows companies to more effectively gather and use the corporate knowledge “buried” in large collections of unstructured data.
- *Facilitates better and faster decision- making.*
- *Lower costs than other text processing technologies:* TM process text automatically eliminating costly setup and configuration.

## **5. Real-Life Uses (Applications)**

Some real-life uses that TM solutions offer, could be described as:

- *E-mail Monitoring and Management*

The explosive growth of electronic communication (e-mails) has produced many new challenges for companies. It widened and accelerated their communication with the rest of the world but on the other hand it, also, make them vulnerable to a mass attack of information.

TM enables companies to deal with the huge number of e-mails they receive by routing the message to the appropriate person that should deal with it.

Many companies make concerted efforts to make their work environment a healthy one for their employees. One area that is often overlooked in the company's email system through which a variety of inappropriate and offensive material may pass. Samples of this material include profanity, pornography, sexual harassment, racist remarks, chain letters, viruses, and miscellaneous non-business conversations. Capturing and removing such traffic from the system protects both employees and employers.

In addition to general firm-preservation issues, special rules may apply to the contents of email traffic within particular industries. For example, many financial services firms are required by law that their communications with customers are in full compliance with securities laws and regulations. They are obliged to go through tedious and expensive manual review procedures to identify improper or illegal communications. For both general purpose and industry specific rules, TM offers them a better solution. It provides them the ability to automatically monitor its incoming and outgoing e-mail to make sure that its employees are following the law in their customer communications. It can identify what kind of potential violation is involved and also automatically quarantine the messages until a human can review it. It can, also, read and classify incoming, outgoing, and internal email messages based on the appropriateness of the conversation, and alert compliance officers to potential problem messages before they can be delivered.

- *Document Management (Knowledge Discovery from Text)*

Often, companies have accumulated a large text archive related to their area of interest. This data may include internal documents such as research reports, sales information, and product documentation, as well as external documents like competitor backgrounds and news releases. Being able to efficiently and accurately

catalogue, index, and retrieve key information from these documents can be extremely useful to a company.

First, just being able to track mentions of key company names, person names, locations, and domain specific concepts can be used to discover "hot" areas in your industry. Relationships such as "Mr. Smith became CEO of XYZ Corp." and "XYZ Corp is opening a new branch office in SomeCity, USA" can be extracted from text documents and provide insight into what key players are doing.

This technology may be particularly useful in areas such as litigation support and law enforcement. For litigation support, lawyers, para-legals, or office support personnel could sift through volumes of on-line data for references to particular legal citations, statute references, law firms, judges, companies, dates, dollar amounts, or court decisions. This information can make preparation for a particular case, or deciding on whether or not to take a case a simpler, more efficient process.

Similarly, law enforcement officials can piece together a coherent scheme of associations and links among individuals and organizations. These entities and links between them can be viewed in a highly intuitive visualization tool that allows the user to drill down and view the actual text, which produced the relationship. These displays can immediately focus the attention of a law enforcement official on entities related to others that they are investigating.

- *Market Research*

A marketer can use online TM to gather statistics on the occurrence of words, phrases, or themes that will be useful for estimating market demographics and demand curves.

- *Automated Help Desk*

TM can automatically identify messages coming from customers via email or WWW forms, categorize them, and route them for an appropriate action. This will not only reduce the cost of customer care, but also increase the quality of customer care by enabling timely and individualized care.

For instance, messages that are determined to be complaints can be sent to the customer service department for timely handling. Technical questions about a product can be forwarded to the appropriate tech-support organization or automatically responded to with a pre-defined answer. Sales related inquiries could be answered automatically with appropriate product information or sent to the sales manager if it is an important strategic sales lead.

- *Business Intelligence*

Today, there are an ever-increasing number of information sources available for a research analyst to stay abreast of the latest technology, latest news, and their established as well as emerging competition. The ability to sift relevant, timely information from this "info-glut" can save a company both time and money and help business intelligence analysts perform their duties more effectively. On the other hand, it enables companies to gather information about their markets and their competitors.

## **6. Questions to Answer for TM**

### *1. How do you define TM?*

TM is the process of extracting novel, undetected and unstructured knowledge “hidden” in huge collection of text documents, which are stored in text databases or available in web pages.

### *2. Which are the most important service providers in TM?*

After a search in the Internet for TM we detected some companies that have developed software tools for TM. Companies such as, Semio Corporation, IBM, Megaputer Inc., Autonomy Corporation, Data Junction, have developed their own software tools.

### *3. What kind of services do they offer?*

The mentioned companies offer TM software tools that operate over text documents stored in text bases or Internet.

### *4. Which are the Software tools used, if any?*

TextAnalyst by Megaputer Inc., Agentware by Autonomy Corp., Intelligent Miner for Text by IBM, SemioMap by Semio Corp. are some mentioned software tools that are used for TM.

### *5. Which are the methods contained in the software tools?*

The mentioned software tools use Natural Language Processing, Textual Information Retrieval and Statistical Textual Analysis techniques.

### *6. Which are the customers?*

Possible customers are corporations that need to acquire information hidden in text documents or to handle large collections of text documents.