# A NEW FRAMEWORK TO EXTRACT KNOWLEDGE BY TEXT MINING TOOLS

Domenico CONSOLI[*]

**Abstract:** Nowadays, enterprises are invaded from a large amount of unstructured information in textual documents, web-pages, e-mails, chats, forums, blogs. In recent years the number of documents available in electronic form, has grown almost exponentially. Therefore it's important use a technological platform to extract and manage useful knowledge for business goals. The goal of the knowledge management is to provide, in all corporate levels, in the right format, the right information at the right time. The aim of this paper is the presentation of a new framework to manage unstructured knowledge by text mining technology. With text mining tools, we can obtain high performances and discover interesting hidden relationships among business data. Text mining technology uses semantic engine and artificial intelligence algorithms to mine, extract and classify the knowledge. The knowledge extracted is useful for Business Intelligence tools used from top manager in the strategic planning.

**Keywords:** Text mining, knowledge management

## 1. Introduction

According to Merrill Lynch (Blumberg and Atre, 2003), the 85% of business data is hidden in unstructured textual documents, such as e-mails, web pages, notes, news, chats, reports, letters, surveys, papers and only 15% of data, inside the organization, is structured. The major input for any kind of decision, strategic, tactical or operational, is the information and with 15% of structured information it's impossible to have high quality decisions. Therefore for enterprise it's necessary to have a framework that processes unstructured information to extract useful knowledge for strategic business planning.

From a survey of IDC (Rizzotto, 2006), the analysis and research of successful information is about 50%. The weight of failure research is about 9% which can cause the rewriting of existing documents (8%) and the conversion of contents in new formats (10%).

In past years, enterprises have improved the infrastructure (data warehouse) and invested in tools of Business Intelligence and Business Analytics to support the management of structured information. Nowadays companies could invest in Text Mining (TM) technology to discover hidden knowledge from unstructured information. Extracting hidden knowledge from texts it's very important for decision making.

This paper presents the following structure: in the next section is described the information overload phenomenon while in the third section we show the enterprise knowledge management. In the fourth section we present the original framework to extract, inside organizations, useful knowledge for business decisions. In the fifth section we give a brief description of application of text mining technologies. The sixth and seventh sections are useful to describe information retrieval and business intelligence. In the last section, performances of classification and research algorithms are shown. Finally some conclusion is drawn.

## 2. Information overload

In recent years the number of documents, available in electronic format, has grown almost exponentially. The world produces between 1 and 2 exabytes (1018 bytes) of information per

[*] Dr. Domenico Consoli, Department of Business and Law, University of Urbino "Carlo Bo", Italy, domenico.consoli@uniurb.it

year (Lyman et al., 2003). The digital universe is growing by 60% every year. In the 2011 only half of the information will be kept, the rest will not find space on storage devices (Gantz and Reinsel, 2009). The active mailboxes are rose from 253 million in the 1998 to 1.6 billion in 2006 and in the 2006, 161 exabytes (10246 bytes) of information, that will become 998 Exabyte's in 2010 (Assinform, 2007), have been created and copied.

The number of Internet users who visited a social network or a blog increased by 24 percent, from April 2009 to April 2010, while the average person spent 66 percent of time on those sites (Nielsen Company, 2010).

About 75 percent of all Internet users visited a social network or blog in April 2010, and they spent 22 percent of their online time – more than 110 billion minutes – on those sites; that is more than twice the number of minutes that Nielsen recorded last year. The average person spent 5 hours, 51 minutes on these sites in April, compared to 3 hours, 31 minutes per person during April of the previous year.

Most electronic documents, written in different formats and languages, which contain 80% of the relevant information for a company, are available on web sites and departmental networks.

Therefore it's evident that we are invaded from an information overload. Information overload is a term used by Alvin Toffler (1990) to describe the difficulty of a person, in understanding an issue and making decisions, caused from the presence of too much information. Toffler writes about sensory overload, in the information age, that causes disorientation and lack of responsiveness. Too much information makes us stressed, confused and disoriented.

An other term used to describe the information overload and any undesirable side effect, caused by information technology and its applications, is information pollution or info pollution (Orman, 1984) (De Rosnay, 2002).

The term has taken a particular importance when Jakob Nielsen (2003), an expert leader in web usability, published a series of articles on this topic.

Info pollution is the contamination of redundant and irrelevant information. The spread of unnecessary and unwanted information could have a negative effect on human activities. This is one of the negative effects of the information revolution.

Some author argues that we have information overload of global proportions to same level of threats polluting the environment. The info pollution is a big problem and is growing rapidly, especially with web 2.0 tools (chat, blog, forum, wiki), e-mail, instant messaging (IM) and RSS feeds. Although the technology has clearly exacerbated the problem, it isn't the only cause of pollution. There is also info pollution when the quality of the information is reduced. Information is sometimes partial, incomplete and asymmetric and can continuously change. For example laws, regulations are submitted on rapid changes and revisions and they are often outdated.

The increasing time to process information easily causes the loss of productivity and income. Imperfect decision-making will also increase the risk of many mistakes (Orman, 1984).

The digital inclusion reduces the gap between rich and poor countries but causes the growth of information and it makes more difficult to separate valuable from worthless documents.

Sometimes, the level of pollution depends on which environment tools are used. For example, e-mail is more polluting in a business environment than in a private environment. Some technology, such as instant messaging, is seen as particularly intrusive or pollutant.

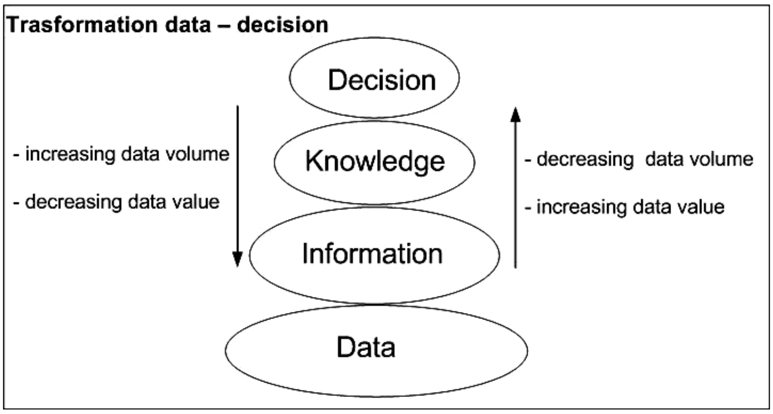
To face infopollution we could use different solutions (Nielsen, 2003):

- using a control panel to determine an information priority
- using technologies without frequent breaks
- controlling public e-mail usage and developing a strategy to guarantee the integrity of the information
- writing clearly and concisely to minimize overload information
- minimizing interruptions and spending time in productive and creative activities. Interrupt-driven works reduce the productivity.

## 3. The enterprise knowledge management

The knowledge is one of the most valuable assets that creates added value. The transformation from data to knowledge and decision is represented in Figure 1.

**Figure 1**. Transformation data-decision.



In this figure is shown the path from a large amount of low-value data to a small volume of high-value data (decision).

Without knowledge it’s impossible to create innovation, change and renewal. The knowledge is an important resource for the company, but its use is often ignored and rarely is organized and managed by an efficient system. The enterprise knowledge, crossing all business processes, should not be centralized only in certain areas but disseminated and available wherever. Knowledge Management (KM) is a business strategy that provides the right information at the right time and in a right format to all corporate level (strategic, tactical, operational). In searching information, we observe only the emerging of an iceberg. Often, our approach is a kind of serendipity because we discover information that we didn’t search or know. To acquire knowledge it’s necessary to enable networks of relationships among people that will enrich and integrate it.

The KM is an important methodology to exploit opportunities (before competitors) and to better fight threats of external environment.

A KM system supports different stages to manage the knowledge: acquisition, creation, transformation, classification, storage, distribution and utilization.

KM is the business discipline that enables organizations in improving the capacity to spread information to all functional area and to transform it in business asset. The challenge is to successfully integrate processes and information flows.

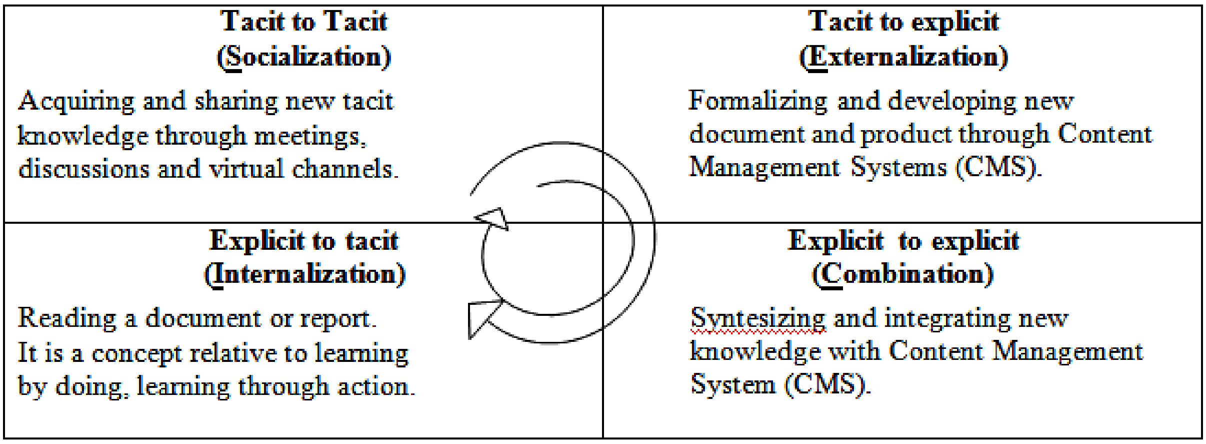
In KM we can consider two types of approach:

- strategic: define the processes of knowledge management in relation to strategic goal
- tactical: identify the type of knowledge necessary for solving a specific business problem.

Inside organizations there is a lot of tacit knowledge. It's a task of KM system to transform it in explicit and useful knowledge. The Socialization Externalization Combination Internalization (SECI) model of Nonaka-Takeuci (1995) (Figure 2) is useful to explicit and save, in specific repositories, this tacit knowledge.

The process for explicating knowledge (Consoli, 2010) is a spiral path where the phases of Socialization, Externalization, Combination and Internalization are continually repeated to obtain an explicit knowledge.

**Figure 2.** SECI model

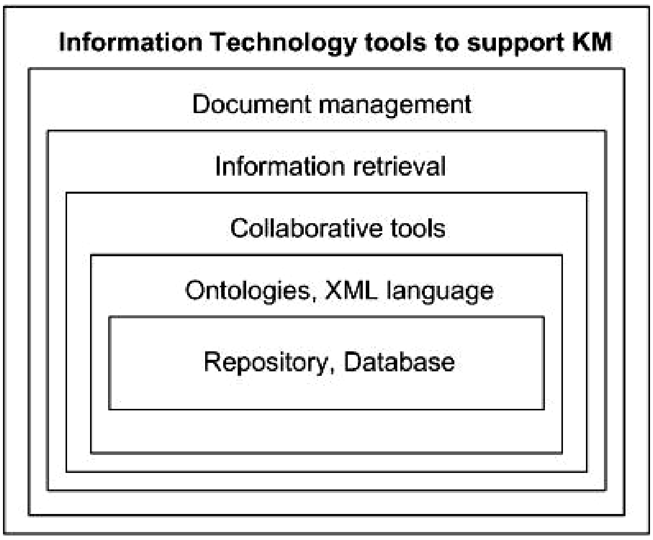


| | |
|---|---|
| **Tacit to Tacit**<br>**(Socialization)**<br>Acquiring and sharing new tacit knowledge through meetings, discussions and virtual channels. | **Tacit to explicit**<br>**(Externalization)**<br>Formalizing and developing new document and product through Content Management Systems (CMS). |
| **Explicit to tacit**<br>**(Internalization)**<br>Reading a document or report. It is a concept relative to learning by doing, learning through action. | **Explicit to explicit**<br>**(Combination)**<br>Syntesizing and integrating new knowledge with Content Management System (CMS). |

**Source**: Nonaka and Tacheuchi , 1995 (adapted by author)

Information Technology (IT) provides many tools for increasing the capacity to map, codify and transfer the knowledge (Figure 3).

**Figure 3**. Information Technology to support KM



Information Technology tools to support KM

Document management

Information retrieval

Collaborative tools

Ontologies, XML language

Repository, Database

The main tools are:

- repositories and databases to expand the memory capacity and efficient storage
- ontologies and languages, such as XML, to increase the ability to encode and describe the knowledge
- collaborative tools to maximize the ability to create and share knowledge inside virtual communities of practice
- tools for information retrieval and text mining to facilitate the recovery and discovery of useful and hidden knowledge. These technologies can use automatic crawlers to gather information and algorithms to classify information
- document management technology to publish and facilitate the dissemination of knowledge

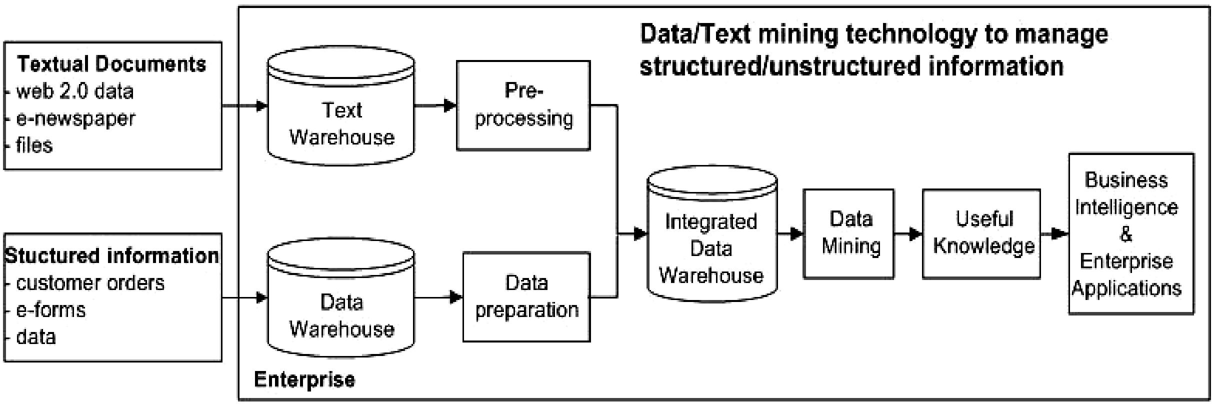
## 4. Enterprise framework to extract useful knowledge

Enterprise Information System (EIS) currently processes only structured information that corresponds to 15% of the information that circulate in the company; the remaining 85% is unstructured information. From the annual survey Teradata (2006), on a sample of about 1200 managers from 23 countries (56% from the Americas, 23% from Asia-Pacific and 21% from Europe, Middle East and Africa –EMEA-), 70% of interviewees who work in America and Asia-Pacific and 36% of interviewees of EMEA said that the use of unstructured data, can be a high competitive advantage to support business decision.

Therefore, for enterprise, it is necessary to use a framework that processes this unstructured information for business strategic plans.

To process enterprise knowledge, structured and unstructured, we have designed the following framework (Figure 4 ):

**Figure 4**. Framework to process enterprise knowledge.



Textual and structured information are saved in respective databases. After pre-processing or data preparation enterprise information is registered in an integrated data warehouse. By algorithms of Data Mining (DM) (Kurgan and Musilek, 2006), hidden relationship and useful knowledge are extracted from data contained in data warehouse. This knowledge is useful for Business Intelligence and Enterprise Applications.

In the figure only DM module is represented because unstructured information is converted, by pre-processing module, in structured data and therefore in an adapt format (dataset) to apply techiniques of Data Mining or Knowledge Discovery in Databases (KDD) (Canuto et al., 2007). Data Mining is based on statistical, neural networks and artificial intelligent algorithms able to identify new logical relationships among contents of the dataset.

The process of KDD is characterized by a structure accepted by the scientific community, composed of four steps: document acquisition, document preprocessing, mining, analysis and results interpretation.

Documents are acquired, by techniques and algorithms of crawling, in various formats of different sources (web, intranet, textual databases). Documents acquired, usually converted in a standard format, are stored in a repository.

To classify documents it's possible to use classification and clustering algorithms. Clustering (Hruschka et al., 2009) is a technique used to group similar documents depending on dominant characteristics. In the hierarchical approach, the top of the hierarchy is a set of all documents belonging to a collection, the leaves of the tree represent individual documents, while intermediate levels contain sets of documents less populated.

By refinements we can obtain groups of documents with similar contents or topics, associations between documents and trends on contents.

The system provides the correlation between specific events such as the participation of a company in conventions or conferences, the presence of patents, etc.. and can identify, for example, which relationships exist between companies and specific production methods.

Text Mining (TM) (Berry and Castellanos, 2007) is a set of techniques obtained as a generalization and contextualization of those used in KDD but with information sources in textual format.

In Table 1 the comparison, between DM and TM technology, is shown.

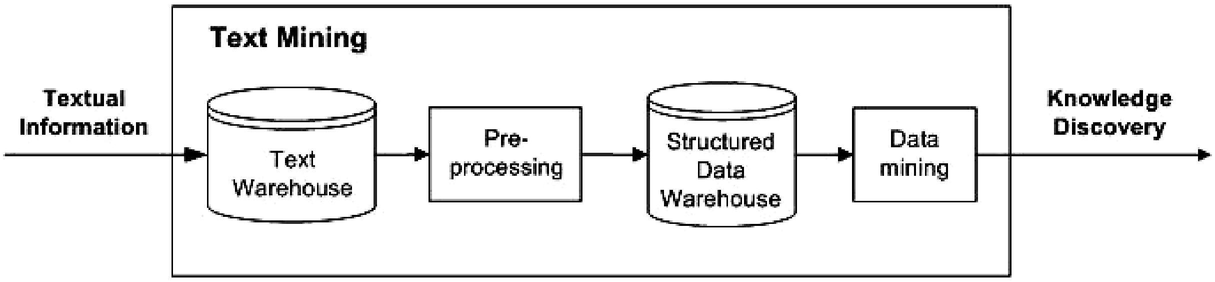
**Table 1**. Comparison between Data and Text Mining

| | **Data Mining** | **Text Mining** |
|---|---|---|
| **Object of analysis** | structured data | text |
| **type of structure** | relational database | textual file |
| **goal** | classification and forecasting | Information retrieval |
| **period** | since 1994 | Since 2000 |

The main difference is on data type to analyse: structured or unstructured.

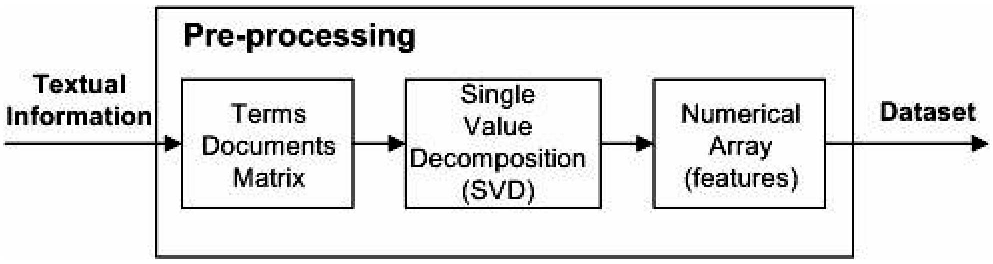
The Text Mining technology is represented in Figure 5.

**Figure 5.** Text Mining



After the pre-processing phase, where textual data is transformed in structured data that is saved in a data warehouse, DM techniques can be applied. The pre-processing module is represented in the following Figure 6.

**Figure 6.** Pre-processing of textual information



In the figure we can distinguish the following stages:

- Terms-Documents Matrix. Each element of this matrix represents how many times single words occur in different documents. Each numerical value identifies the significance or the attitude of the word (lexical unit) to be representative of specific document (context unit).
- Single Value Decomposition (Li and Zhang, 2007): compression of original matrix using main factors. In SVD algorithm, the indicator Term Frequency-Inverse Document Frequency (TF-IDF) is used. The TF-IDF (Qu et al., 2008) is a statistical measure used to evaluate, by weights, how a word is important in a specific document of corpus. The importance increases proportionally to the number of times that a word appears in the document but is inversely proportional to number of times that the word appears in the entire corpus. In the pre-processing module the frequency of words (Li et al., 2008) and weights of single words are two parameters very important.
- Numerical data, contained in the matrix, represents the feature of document. It's the dataset useful to apply Data Mining techniques. For describing a document it's possible to use an array containing a list of the main features or keywords with a numerical weight indicating the importance of each term. The documents are described with a small set chosen from many terms.

During the pre-processing stage, each document is analysed to extract its features. Extractors of features are distinguished on accuracy and technologies used. For example, it's possible to use stemmer, lemmatizer, Part of Speech (POS) (Tanawongsuwan, 2010), taggers to extract lemma of words like nouns, verbs, adverbs, adjectives, etc…

## 5. Applications of Text Mining technology

Text Mining technology (Bolasco et al., 2005) allows to monitor unstructured information sources and identify automatically useful knowledge for enterprise goals. Web pages, patents, scientific publications, e-mails, phone calls, can be an inexhaustible source that could provide new interpretative keys to business. Text Mining is the answer to information overload in the automatic management of information. Text mining is important to identify patterns that helps to better understand events and to provide better services to enterprises.

TM is used in several areas: risk analysis, fraud prevention, insurance charges, money laundering, environmental monitoring, definition of macroeconomic models, analysis of the impact of promotions on sales, quality control, forecasting demand, curriculum vitae management, analysis of costs and charges of services. By TM we can automatically classify articles of newspaper for topic area, regional, national and community laws for subject and create a database to customer support.

The TM can be used for the analysis of competition in different areas of the market. In this case, TM tools allow to check, in function of certain parameters, business activities in a specific area.

Text Mining techonology allows also these operations:

- thematic indexing
- text summarizing
- forecasting

The thematic indexing (Gelernter and Lesk, 2009) uses the knowledge of the words meaning of a text to identify topics in a document: for example, documents about aspirin or ibuprofen can be classified under analgesic category. Indexing is often implemented using multidimensional taxonomies. A taxonomy is a representation of hierarchical knowledge often called ontology. Thematic taxonomies allow to search a document by topics rather than keywords.

Text summarization (Abulaish et al. 2009) is the reduced description of contents of a document. The main topics of many documents can be described with only 20% of the original text, without losing its meaning. Like clustering, there are different approaches and algorithms to summarize: some algorithm uses morphological analysis of terms to identify the words most frequently used and eliminating those insignificants; other algorithms weigh terms more used in the opening or closing sentences, while other approaches search keys that identify important phrases such as “in conclusion” or “basic concept”, etc…

The forecasting by analytic tools (Sush et al., 2008) is essential for the success in the market because it allows to simulate future scenarios and make future plans. It represents a resource necessary for companies to capitalize data collection and data warehousing.

Analytics could identify hidden correlations between data, analyse time series, determine trends and behaviors, simulate business scenarios, segment and classify customers by specific scores and ranks.

## 6. Information retrieval on unstructured texts

Information Retrieval techniques (Zhai, 2008)(Kurland and Lee, 2009) have evolved over time. From the first full text searches on keywords we have used advanced tools that combined search with keywords and statistical algorithms for a more focused recovery. In recent years, for research and analysis, semantic engines, a new class of linguistic products, are used. Semantic technology adds to basic linguistic analysis (morphological and grammatical) a substrate of logic for the disambiguation of polysemous words with different meanings. Semantic engines allow to capture key elements of each document, removing any ambiguity, flexion and alteration, classifying each word, emphasizing the morphological and semantic attributes (gender for names, mode/time for verbs,…).

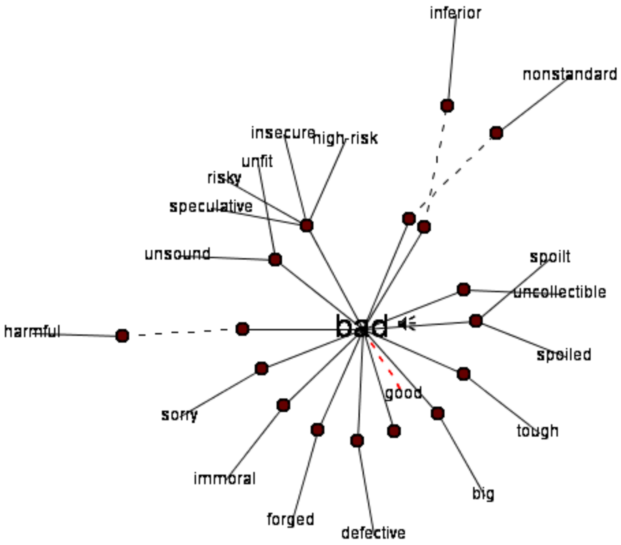
The structured data has a single meaning; a text or a phrase has a precise meaning only within a linguistic context or speech. Semantic engines search words in that linguistic context. Models that work on computational linguistics have found large growth in recent years, with applications in new areas of investigation. These models use multidimensional statistical techniques, Natural Language Processing (NLP) (Wilks and Brewster, 2009) and neural networks to detect association patterns of words that have a high probability of semantic meaning.

For example, these tools applied in biomedical publications, have highlighted a possible link between the metabolism of magnesium and chronic headaches. This data was confirmed to a clinical level but the primary information has emerged from an intelligent textual analysis on generic clinical data.

An other way of evolutes information retrieval is the use of semantic networks (Zhuge and Sun, 2010). The research systems of information within a text, are based on the assumption that, in our brain, meanings of words are arranged in a grid, a road map where single terms are represented by a node (e.g. first, dish, language) and their connections by paths (i.e. "first" is linked to "dish" and "language" but these two terms are not linked together).

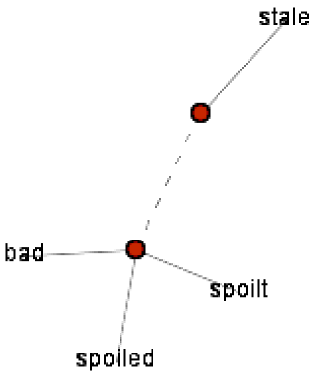
Semantic networks or road maps are used to simulate the similarities between linguistic units. For example, in the Figure 7, we show the semantic path, obtained by the web tool visual thesaurus, relative to bad term. The term bad, as we can see, returns several synonyms. In relation to food, if we limit the synonym research clicking on "spoiled", the number of synonyms becomes smaller (Figure 8). The distance between two terms represents the proximity/similarity index.

**Figure 7.** Synonyms of term bad



**Source:** VisualThesaurus, 2008

**Figure 8**. Synonyms of term spoiled



**Source:** VisualThesaurus, 2008

## 7. Business Intelligence

Business Intelligence (BI), term coined by Howard Dresner, Gartner analyst, in the early '90s, is a set of models, methods and tools to extract useful knowledge for decision making.

Nowadays, the context where enterprise works is the following:

- too much information, mainly due to exponential growth of the Internet, with complex problems of information retrieval and content quality
- multi-dimensionality of information
- globalization and its complexity due to overcoming of market barriers and to development of alliance networks
- organizational rigidity with a top-down and hierarchical structure that contrasts a quickly and flexibly use of information.

In the past, the Forrester Research company (Butt et al., 2001) presented the following results of his survey on time spent from employees: the 40% for information searching, 80% to prepare data, 20% for data analysis, 85% to manually manage unstructured information and the 70% of contents recreated rather than reused.

From a study commissioned by British Aerospace: the 80% of employees lost, on average, 40 minutes a day, in looking for information, 60% spend more than an hour in duplicating work already done and the 8% is the loss of productivity.

Therefore it's very important an intelligent enterprise solution to reduce the time spent in all these operations, resolve information overload and improve productivity and business results.

The traditional BI systems are based on data warehouse that provide factual data and answer to question on what has happened but usually fail to understand why. The BI advanced, that uses data/text mining technologies (Baars and Kemper 2008), it's able to discover causes of new events. For example when the margins on product sale decrease, over time, new tools of BI are able to analyze possible causes: competitors that have launched an alternative product with a lower price or higher quality, a policy of cross-selling that has obtained a reduction of margins, new competitors, etc….

BI covers different application areas: supply, call center, customer care, prediction and strategy optimization. BI solutions support enterprises to monitor review, loyalty and behavior trend of the customer and improve Customer Relationship Management (CRM), Customer and Market Intelligence and the knowledge sharing among all enterpise stakeholders. Nowadays we live in a collaborative era where co-design, co-creativity, co-operation are very strategic and important assets for the development of business strategies.

Semantic engines used by BI tools, enable the analysis of so-called "Internet-invisible" promoting the discovery of useful information that may include competitors, markets, financial information, etc…

BI uses also analytic tools that are essential for the success in the market because they allow to forecast and plan timely interventions.

Business Intelligence is important also for monitoring and displaying Key Performance Indicators (KPIs) (Tseng, 2008). A KPI compares the main business factor with its expected value to understand and graphically show trend and current status.

To align processes to business and optimize the processes management, BI uses a new generation of tools named Business Performance Management (BPM) (Chowdhary, 2006).

## 8. Performances of classification and retrieval algorithms.

In the development of classification or predictive model it's important to test the accuracy of algorithms. In the classification algorithms of Data Mining (Mastrogiannis et al., 2009) and Text Mining (Malik et al., 2006) we must consider the phases of training and testing. By training set a mathematical model is built and then we test the model with a part of the dataset named test set. Then the mathematical model is applied on new data. At the end, we create the matrix of confusion, from which we calculate the accuracy (a) observing the percentage of tests on total that have been classified correctly.

One of the most common sources of error is the overtraining: the model predicts very well data of test set but incorrectly new data. Not all algorithms used to implement a transaction are equal. Different algorithms can vary the range of accepted input, transparency of the mining output and the ability to manage large volumes of data.

Regarding information retrieval it is important to measure the precision (p) and recall (r). The precision measures whether documents returned meets the requirements. The recall measures the percentage of correct documents found on total of relevant ones.

In the research field there are quantitative problems relative to the growth of volume of documents and qualitative problem relative to the difficulty to retrieve and manage information in automatic modality.

Traditional research tools showed, over time, its limitations:

- lack of precision or inability to identify only relevant documents on the basis of selection criteria because there are many different meanings associated to a single word.
- limited recall or inability to extract all relevant documents. It's impossible to extract all documents relative to a specific concept unless to specify, as keywords, all related terms.

Tools based on linguistic analysis reduced these limitations while tools based on semantic analysis have better performance because they are able to understand more clearly the meaning of a word in a specific context in which each term is used. Semantic engines generally have an interpretation module and a lexicon to find relationships between terms.

Semantic tools limit the negative impact of the explosion of information providing to users most useful knowledge. By these tools, with a better disambiguation of terms, it's possible to extract relevant content of quality.

Even the activity of automatic mining benefits from this greater text understanding, allowing the identification of specific entities but also of relationships between information and concepts in different parts of the same document or in different documents.

Recently retrieval and classification algorithms (Mastrogiannis et al. 2009), with specific refinements, achieve an accuracy and precision around 90%.

## Conclusions

Enterprises are invaded from a large amount of unstructured information mainly with the exponential growth of web 2.0 tools (chat, forum, blog, wiki). This textual information is important for decision making and strategic plannings. So it's necessary to use a technological platform for managing this high-value knowledge for business goals.

Solutions based on data/text mining tools, optimizing unstructured information management, play an important role in facilitating the extraction of useful knowledge from contents disseminated in different business areas. This knowledge can be exploited by tools of Knowledge Management and Business Intelligence. By tools of Business Intelligence it's possible to analyse and synthesize relevant information for business strategies. Managers daily need to process unstructured information, coming from e-mails, reviews, news, reports, etc…, of customers, markets and competitors. The development of good algorithms of classification, prediction and information retrieval, with high values of accuracy and precision, improves the management of knowledge inside organizations.

**References**

Abulaish M., Jahiruddin S. and Dey L. (2009) "A Relation Mining and Visualization Framework for Automated Text Summarization". In Proceedings of the 3rd international Conference on Pattern Recognition and Machine intelligence (New Delhi, India, December 16 - 20, 2009). S. Chaudhury, S. Mitra, C. A. Murthy, P. S. Sastry, and S. K. Pal, Eds. Lecture Notes In Computer Science, vol. 5909. Springer-Verlag, Berlin, Heidelberg, pp. 249-254

AITech-Assinform (2007) "Assinform report, ICT and multimedial contents", Milano, Italy

Baars H. and Kemper H. (2008) "Management Support with Structured and Unstructured Data-An Integrated Business Intelligence Framework" Inf. Sys. Manag. 25, 2 (Mar. 2008), pp. 132-148

Berry M. W. and Castellanos M., editors (2007) "Survey of Text Mining II: Clustering, Classification, and Retrieval", Springer

Blumberg R. and Atre S. (2003) "The problem with unstructured data", DM Rev. February 2003

Butt J., Rutstein C., Gilett F. and Khawaja S. (2001) "Turning Data Into Dollars", Forrester Research, May 2001

Canuto A. M., Campos A. M., Bezerra V. M. and Abreu M. C. (2007) "Investigating the use of a multi-agent system for knowledge discovery in databases", Int. J. Hybrid Intell. Syst. 4, 1 (Jan. 2007), pp. 27-38

Chowdhary P., Mihaila G. and Lei H. (2006) "Model Driven Data Warehousing for Business Performance Management", In Proceedings of the IEEE international Conference on E-Business Engineering (October 24 - 26, 2006). ICEBE. IEEE Computer Society, Washington, DC, pp. 483-487

Consoli D. (2010) "The multidimensional model of knowledge management in the competitive enterprise". In Proceeding of 12th IS MM&T 2010, Sunny Beach, Bourgas, Bulgaria, Journal of International Scientific Pubblication: Materials, Method & Technologies, ,Vol. 4, p. 2, 2010, pp. 5-29.

De Rosnay, J. (2002) "Les risques de l'infopollution", Transversales, Science Culture, Nouvelle série n°1, Mai, 2002

Gantz J. and Reinsel D. (2009) "As the economy contracts, the digital universe expands", IDC Multimedia white paper, ECM, may 2009

Gelernter J. and Lesk M. (2009) "Text mining for indexing", In Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries (Austin, TX, USA, June 15 - 19, 2009). JCDL '09. ACM, New York, NY, pp. 467-468

Hruschka E. R., Campello R. J., Freitas A. A. and De Carvalho A. C. (2009) "A survey of evolutionary algorithms for clustering". Trans. Sys. Man Cyber Part C 39, 2 (Mar. 2009), pp. 133-155

Kurgan, L. A. and Musilek, P. (2006) "A survey of Knowledge Discovery and Data Mining process models". Knowl. Eng. Rev. 21, 1 (Mar. 2006), pp. 1-24

Kurland, O. and Lee L. (2009) "Clusters, language models, and ad hoc information retrieval", *ACM Trans. Inf. Syst.* 27, 3 (May. 2009), pp. 1-39

Li Y. and Zhang H. (2007) "Two properties of SVD and its application in data hiding", In Proceedings of the intelligent Computing 3rd international Conference on Advanced intelligent Computing theories and

Applications (Qingdao, China, August 21 - 24, 2007). D. Huang, L. Heutte, and M. Loog, Eds. Lecture Notes In Computer Science. Springer-Verlag, Berlin, Heidelberg, pp. 679-689.

Li Y., Chung S. M., and Holt J. D. (2008) " Text document clustering based on frequent word meaning sequences.", Data Knowl. Eng. 64, 1 (Jan. 2008), pp. 381-404

Lyman P., Varian H.R., Charles P., Good N., Jordan L.L. and Pal J. (2003) "How much information?", http://www2.sims.berkeley.edu/research/projects/how-much-info-2003

Malik R., Franke L. and Siebes A. (2006) "Combination of text-mining algorithms increases the performance", Bioinformatics 22, 17 (Aug. 2006), pp. 2151-2157

Mastrogiannis N., Boutsinas B. and Giannikos I. (2009) "A method for improving the accuracy of data mining classification algorithms", Comput. Oper. Res. 36, 10 (Oct. 2009), pp. 2829-2839

Nielsen Company (2010) "Understanding the Value of a Social Media Impression: A Nielsen and Facebook Joint Study", New York, US, 2010

Nielsen J. (2003) "IM, Not IP (Information Pollution)", Queue 1, 8 (Nov. 2003), pp. 76-75

Nonaka I. and Takeuci H., (1995) "The Knowledge Creating Company: How Japanese Companies Create the Dynamics of Innovation", Oxford University Press, New York, 1995

Orman L. (1984) "Fighting Information Pollution with Decision Support Systems", Journal of Management Information Systems, 1(2), pp. 64-71

Rizzotto F. (2006) "White paper: Qualità e valore nella gestione dell'informazione non strutturata: gli strumenti basati sull'analisi semantica", IDC company, 2006

S. Bolasco, A. Canzonetti, F. M. Capo, F. della Ratta-Rinaldi and B. K. Singh (2005) "Understanding text mining: A pragmatic approach" in Knowledge Mining, ser. Studies in Fuzziness and Soft Computing, S. Sirmakessis ed., Springer Verlag, 2005, vol. 185, pp. 31–50.

S. Qu, S. Wang, and Y. Zou (2008) "Improvement of text feature selection method based on tfidf ", in FITME '08: Proceedings of the 2008 International Seminar on Future Information Technology and Management Engineering. Washington, DC, USA: IEEE Computer Society, 2008, pp. 79-81.

Suh J. H., Park C. H. and Jeon S. H. (2010) "Applying text and data mining techniques to forecasting the trend of petitions filed to e-People", Expert Syst. Appl. 37, 10 (Oct. 2010), pp. 7255-7268

Tanawongsuwan P. (2010) "Part-of-Speech Approach to Evaluation of Textbook Reviews", in Proceedings of the 2010 Second international Conference on Computer and Network Technology (April 23 - 25, 2010). ICCNT. IEEE Computer Society, Washington, DC, pp. 352-356

Teradata (2006) "Insights from the Fifth Annual Teradata Survey Validate a Global Phenomenon", Enterprise Decision-Making survey, 2006 Report, Teradata

Toffler A. (1990) "Powershift: Knowledge, Wealth and Violence at the Edge of the 21st Century", Bantam Books, 1990

Tseng S. (2008) "Knowledge management system performance measure index", Expert Syst. Appl. 34, 1 (Jan. 2008), pp. 734-745

Wilks Y. and Brewster C. (2009) "Natural Language Processing as a Foundation of the Semantic Web", Found. Trends Web Sci. 1, 3$#8211;4 (Mar. 2009), pp. 199-327

Zhai C. (2008) "Statistical Language Models for Information Retrieval A Critical Review", Found. Trends Inf. Retr. 2, 3 (Mar. 2008), pp. 137-213

Zhuge H. and Sun Y. (2010) "The schema theory for semantic link network", Future Gener. Comput. Syst. 26, 3 (Mar. 2010), pp. 408-420