Chapter 21

Text mining tools

A. Zanasi
TEMIS SA, France.
Modena & Reggio Emilia University, Italy.

Introduction

Several tools are already available in the text mining (or more generally *unstructured data*) quickly growing market. We recognize the text mining market as composed by *pure players* (companies which develop text mining software, *e.g.* Clearforest, Inxight, Temis), *indirect players* (which integrate text mining into their offering, *e.g.* IBM, SAS, SPSS), *partial players* (companies which use text mining to improve their core business, *e.g.* Fast, Verity). A section is dedicated to each player which provided us with their company description, and a cumulative section listing the most known players in the text mining arena. Hundreds of other companies are already working successfully in the worldwide market. For a list of them, go to www.kdnuggets.com.

1 Megaputer intelligence

1.2 Company description

Megaputer Intelligence (www.megaputer.com) was founded in 1993. The company provides a family of tools for the analysis of structured data and text.

1.3 Products

Megaputer Intelligence offers two products: *PolyAnalyst for Text*™, aimed primarily at the analysis of incident reports, survey responses and customer communications data and *TextAnalyst*™ .

PolyAnalyst for Text™ performs semantic text analysis, record coding, identification and visualization of patterns and clusters of information, automated or manual taxonomy creation and editing, taxonomy-based forced or self-learning

WIT Transactions on State of the Art in Science and Engineering, Vol 17, © 2005 WIT Press
www.witpress.com, ISSN 1755-8336 (on-line)
doi:10.2495/978-1-85312-995-7/21

categorization of documents, text OLAP, link analysis, and interactive visual processing of various combinations of structured and unstructured data.

***TextAnalyst*™** automatically distills semantics of text, creates ontologies, summarizes, clusters, and categorizes documents. Visual representation of results with drill-down capabilities facilitates quick navigation of the knowledgebase. The performed analysis is domain- and language-independent, while customizable dictionaries help optimize the system efficiency in any specific area.

1.4 Incorporation of domain knowledge

PolyAnalyst for Text incorporates domain knowledge in the analysis through:

- The Dictionary module (helps define domain-specific synonyms, typical misspells, inseparable phrases and stop-words).
- The Semantic Editor module (specifies hierarchical semantic relationships between terms. For example, the user can define that 'pike' represents a type of road in the considered context, and not a weapon).
- A language of regular expressions (allows the user to define structural patterns that represent objects of similar semantic nature in the text).

In *TextAnalyst*, the VocEdit module helps define user-preferred terms and domain-specific synonyms.

1.5 Exporting discovered knowledge

PolyAnalyst for Text can export the discovered knowledge to:

- HTML-based business reports capturing the discovered graphs, patterns and conclusions
- Graphical objects readily rendered by Microsoft Office tools
- CSV tables
- External databases through OLE DB connection

TextAnalyst can export the discovered knowledge to:

- Excel and CSV format (for generated taxonomies)
- HTML format (for knowledge bases)
- Microsoft Word (for summaries)

1.6 Supported languages

PolyAnalyst for Text currently works only with English. Support for other European languages is expected in 2005.
TextAnalyst supports English, German, Spanish, French, Russian, Italian, Portuguese, Dutch, Swedish and Greek.

1.7 IT requirements

PolyAnalyst for Text requires Intel PC platform, with at least 256 MB RAM, and Microsoft Windows. It is offersed in both standalone and client/server configurations. Integrators can utilize all functions of *PolyAnalyst* provided in the corres-ponding API library of COM-based components.

PolyAnalyst for Text is designed for the analysis of relatively short text records.

TextAnalyst requires Intel PC platform, with at least 128 MB, and Microsoft Windows. It is a standalone application.

Standalone desktop *TextAnalyst* can process up to 50 MB of text. To process enterprise-level volumes of text, *TextAnalyst SDK* v3 should be utilized.

1.8 Customer base

Over 800 customers including eleven Fortune100 companies. The most known of them are: Department of Homeland Security, IBM, Siemens, McKinsey, Boeing, Northrop Grumman, US Navy, Chase Manhattan, Bank of Russia, Pfizer, Ask Jeeves, Centre for Disease Control.

1.9 Partners

MedStat, Hewlett-Packard, Microsoft, MicroSystems, Learning Tech.

1.10 Supported applications

PolyAnalyst for Text supports all major commercial RDBMS, spreadsheet and statistical systems. A customized version of *TextAnalyst* supports Lotus Notes.

2 SAS

2.1 Company description

Headquartered in Cary, North Carolina, SAS Institute is the largest privately held software company in the world. SAS serves more than 39,000 business, government and university sites in 118 countries (www.sas.com).
Contact Name:

- Text Mining Product Managers:
  Manya Mayes (Manya.Mayes@sas.com) for the Americas
  Bernd Drewes (Bernd.Drewes@eur.sas.com) for Europe, Middle East, Africa
- Enterprise Miner Product Managers:
  Wayne Thompson (Wayne.Thompson@sas.com) for the Americas
  Sascha Schubert (Sascha.Schubert@eur.sas.com) for Europe, Middle East, Africa

SAS was founded in 1976 and launched its text mining initiative in 2002, with the support of its text mining partner (Inxight Corporation, Sunnyvale, California).

2.2 Product

SAS Text Miner offers a rich list of text mining features:

Performing predictive text mining: Automatic classification and predictive model-ling.
Hierarchical and expectation-maximization clustering for disclosure of topical contents.
Dimension reduction of vocabulary using singular value decomposition.
Rank ordering terms by importance using various statistical weight functions.
Compact representation of documents by singular values and/or term indices.
Optional filtering of terms by syntactic category.
Locating documents similar to a selected set of documents.
Locating terms similar to a selected set of terms.
Surfing among clusters, documents, or terms and propagating their effects.
Use of a dynamic hierarchical cluster browser.
Customizing term start lists, stop lists, and synonym lists.
Creating a semantic map of central terms in a document collection through link analysis and concept graphs.
Exploiting the full integration with data mining and applying data mining algorithms (such as Kohonen clustering, association, sequences, neural nets, regressions, decision trees, comparative model assessments) to text.
generation of scoring code for production deployments.
Java interface with use of Enterprise Miner 5.1.
Use in interactive and batch mode.
Support for remote collaboration over the web.

2.3 Incorporation of domain knowledge

Domain specific terminology can be entered via a start list and/or synonym list; Domain specific descriptions (such as an ontology) can be loaded into SAS and a classification model can be built in order to map documents to ontology levels.

2.4 Supported languages

Entity extraction support is offered for English, French, German, and Spanish. Text parsing is offered for English, French, German, Italian, Portuguese, Spanish, Dutch, Swedish, and Danish.

2.5 IT requirements

SAS Text Miner is currently Windows based. Unix version is expected to be available soon.

Memory requirements depend on number of documents and terms; minimum recommendation is 512 MB for server.

SAS Text Miner requires SAS Enterprise Miner, enabling both, the classical client server architecture as well as the new N-tier architecture.

3 SPSS

3.1 Company description

Website: www.spss.com
Contact Names: Peter Caron, pcaron@spss.com; Olivier Jouve, ojouve@spss.com.
SPSS Inc., founded in 1968, is a provider of predictive analytics software (*e.g.* data and text mining) and services, with more than 250,000 commercial, academic, and public sector customers.

In February 2002, SPSS acquired LexiQuest, Inc., a France-based developer of linguistics-based information management software. LexiQuest text mining technology continues to be a component of SPSS product offerings. For additional information, visit www.spss.com.Text Mining software revenues are not available.

3.2 Product functionality

Database access and other file formats (SPSS files for open-ended survey response analysis) are available with Text Mining for Clementine, a LexiQuest text-mining add-on to SPSS' data mining workbench.

Data Pre-Processing Facilities: LexiQuest products have pre-processing facilities to convert document formats and prepare a corpus for linguistic con-cept extraction. In addition, post-processing facilities are available for each LexiQuest product, depending on its use. Text Mining for Clementine is part of an offering with a full complement of structured data pre- and post-processing techniques.

LexiQuest techniques include linguistic extraction, concept organization and categorization. A full complement of data mining techniques are available for use on textual data with Text Mining for Clementine.

3.3 Incorporation of domain knowledge

Domain knowledge in the form of dictionaries and patterns, while not required, can be added to improve extraction.

3.4 Exporting discovered knowledge

Discovered knowledge can be browsed, exported as lists or reports, used to drill down into source documents, integrated with an application or website for categorization or used to combine with structured data.

3.5 Supported languages

English, French, German, Spanish, Dutch, Italian and Japanese.

3.6 IT requirements

Summary System Requirements: Hardware Intel and Sun SPARC, Operating system Microsoft Windows and Solaris.
Software Architecture:
LexiQuest Mine is a Web-based application; LexiQuest Categorize is a Java application with supporting SDK; Text Mining for Clementine is a plug-in module for the Clementine data mining workbench. All products have a client-server architecture.

3.7 Marketing information

3.7.1 LexiQuest Mine—a text mining application
LexiQuest Mine 'reads' thousands of documents in a matter of minutes (more than 1GB an hour) and extracts the concepts held within the text. The software enables organizations to extract concepts such as company names, product names, terms, patent numbers, people's names and locations. These concepts are presented in a color-coded map for easy identification, and analysis of the connections and trends. Find new research as soon as it is published, identify up and coming competitors before they start stealing market share, track and capitalize on emerging customer preferences.

3.7.2 LexiQuest Categorize—a categorization engine
LexiQuest Categorize can 'read' hundreds of documents in a matter of minutes (more than 1GB an hour) and, based on strategies determined during an initial training, assign documents to the correct category based on their content. Whether your delivery vehicle is your intranet, an electronic portal, an application, or your website, the software enables you to keep ahead of the rising information tide while maintaining order and ease of navigation for your customers and employees.

3.7.3 Text Mining for Clementine—an add-on to the data mining suite
Text Mining for Clementine provides integrated text mining, adding a new and rich source of data for data mining applications, such as detecting fraud and improving customer relationships. By capitalizing on the wealth of unstructured data contained in e-mail, database note fields and other text-based records, Text Mining for Clementine helps organizations achieve the most reliable, comprehensive results by increasing the predictive power of Clementine.

3.8 Customer base

Peugeot SA, BMS, CIM'O1, France Telecom, Children's Memorial Hospital Chicago, Groupe de Boeck, CNES, INIST, NTT.

4 Synthema

4.1 Company description

Synthema (http://www.synthema.it), small enterprise established by former computer researchers of the IBM Research Center in Pisa in 1993, provides Human Language Tecnologies, including both Machine Translation Systems. Synthema operates in the field of Text Mining since 1997.

Contact Name: Claudio Cirilli (CEO), Remo Raffaelli (R&D director), Federico Neri (Text Mining manager).

4.2 Product

Synthema – jointly with TEMIS – has developed TEMIS Online Miner Light, 4006-LXS. Synthema has been in charge for the Knowledge Extractor development, the full product packaging and its technical support, while TEMIS developed the Search and Clustering Engine. Released in October 2002.

4.3 Incorporation of domain knowledge

General dictionaries for English (76,000 lemmas), French (53,000 lemmas), Italian (52,000 lemmas), German (64,000 lemmas), Spanish (30,000 lemmas), Portuguese-Brazilian (41,000 lemmas).

4.4 Exporting discovered knowledge

The analysis results can be exported into a Microsoft Excel compatible file – so as to allow users to create their own charts and graphs – or to an XML file.

4.5 Supported languages

English, French, German, Italian, Spanish, Portuguese-Brazilian.

4.6 IT requirements

Hardware Platform: Personal Computer, PIII 800 MHz, 300 MB HDD, SVGA 800×600 64k colors, network card. Memory Requirements: 1024 MB.
Software Requirements: Windows NT 4.0/2000 and Windows XP as OS, MySQL as database manager.
Software Architecture:
TEMIS Online Miner Light has a client/server architecture consisting of the following main components:

1. The Knowledge Extractor, the back-end application, extracts the relevant information from a set of documents, indexing them by the concepts they contain and storing them into the application database by ODBC.
2. The application database is managed by MySQL.

3. The Search & Clustering Engine, the front-end application, allows to do the following:

    - search for documents by performing simple queries,
    - navigate query results by concepts,
    - make time based projections of concepts
    - classify results according to the concepts they share.

Being the front-end application available as a servlet or JSP pages, Apache Tomcat is used as the servlet container. This choice offers a way to shift processes from clients to server. In fact, Java servlets JSPs are more efficient, powerful and portable than traditional programs. They handle each request through a lightweight Java thread, not a heavyweight operating system process, and they cache previous computations, keeping database connections always open.

4.7 Customer base

- AREA Science Park (the largest science park in Italy, located in Trieste), to provide Technology Watch services to SMEs.
- FirenzeTecnologia (Firenze), to increase knowledge and stimulate interaction between enterprises and technological research centres, setting up technological and strategic partnerships to enlarge scenarios and opportunities. Demo availabale accessing: http://www.spi-rit.net: 8080/OM/index.jsp
- TechNapoli (Napoli), to support the access of SMEs to innovative circuits and promote the rising of innovative suppliers and new companies.

5 TEMIS

5.1 Company description

TEMIS Text Mining Solutions SA (www.temis-group.com), was founded in 2000 by ex-IBM text mining experts. It is directly present in France, Germany, Italy, UK and USA. In 2003 it acquired the Xerox linguistics operations (MKMS).
Contact: info@temis-group.com

5.2 Products

5.2.1 Insight discoverer clusterer (IDC)

IDC reads each document present in the dataset, compares documents pair ways and looks for similarities between each other, using a statistical model.

For each document, it generates a list of the closest documents within the document database and then compares all generated lists and regroups together the most similar lists, hence regrouping together the most similar documents.

The user can navigate using the topics to go to a subset of the search result.

5.2.2 Insight discoverer categorizer (IDK)

IDK sorts documents according to a pre-defined set of categories chosen by the user.

The user defines the categories into which he wants to classify the documents (*e.g.* human resources/sales and marketing/finance or product X/product Y/product Z) and assigns a limited set of documents to each category.

IDK reads each document within each category and creates a statistical model by which it automatically assigns any new document to one or several of the pre-defined categories.

5.2.3 Insight discoverer extractor (IDE) and skills cartridges (SC)

IDE extracts knowledge from unstructured texts. To achieve that, the software identifies concepts and relationships between several texts, or within the same text. Skill Cartridges complement Temis IDE, by bringing specialized knowledge databases that will help IDE to identify key concepts and extract them. Skills Cartridges are already available for Competitive Intelligence, Customer Rela-tionship Management, Pharmaceutical Industry. Other Skills Cartridges will be soon available.

IDE begins by reading each document (whatever its format) and identifies the language of the document, using a statistical algorithm that calculates the re-partition of the letters and of the words. IDE then works phrase after phrase, with the following process:

(a) Morpho-syntactic analysis. IDE:
- assigns a grammatical function to each element of the phrase.
- then summarizes each element of the phrase and reduces them to their canonical form (verbs are converted to their infinitive form, while names are converted to their masculine, singular form).

(b) Rule based analysis:
- IDE analyzes each element of the phrase and tries to assign roles to each of them.

(c) Concept centred analysis: The presence of the concept 'competitive intelligence – buying – acquisition' triggers the 'right and left actor identification'. The software tries to identify proper names around the concept of 'competitive intelligence – buying – acquisition' within the sentence. It then finds Vivendi Universal, already identified as a 'potential company name', and understands that, since it is located before the verb, this 'potential company name' becomes:
'Sure company name' (the software validates its initial analysis),
'Subject of the 'competitive intelligence – buying – acquisition' relationship previously detected.
IDE then looks for the acquired company, and understands that USA Networks is a 'sure company name', hence the object of the 'competitive intelligence – buying – acquisition' relationship.

(d) Database feeding: the software has now validated the phrase 'Vivendi Universal said it would buy USA Networks' as a sentence expressing

the rumour of an acquisition led by the company Vivendi Universal, towards the company USA Networks. It stores the phrase in the database used with IDE, and, with the phrase, the elements it has understood regarding the phrase (who is the acquirer, which company will be bought, it is only a rumour, and it relates to a company acquisition).

5.2.4 Online miner (OM)

Online Miner is an application born to search and analyze web sites, eventually merging them with documents coming from the company (proprietary documents), CDs, online databanks, chats, emails, instant messages and extracting key facts.

Temis ID Online Miner brings together IDE, IDC, and IDK, coupling them with Skills Cartridges and analyzing simultaneously documents written in different languages.

Online Miner was firstly born for competitive intelligence applications, but is currently used in CRM, R&D and other applications.

5.2.5 Xelda

The linguistic engine XeLDA® is a multilingual linguistic engine which models and standardizes unstructured documents in order to automatically exploit their content. It is the base of other Temis products, based on a technology developed through 20 years of research and development in Xerox Scientific Centers, and bought by Temis in 2003. XeLDA® offers a scalable range of services based on natural language processing components that may be integrated in business applications:

- Language identification: automatically recognizes the language used by each document
- Segmentation: divides a text into sentences.
- Tokenization: splits a text into basic lexical units
- Morphological analysis: returns the normalized form (the lemma) and the potential grammatical categories for all the words identified during the tokenization. stage.
- Morpho-syntactic disambiguation: determines the exact grammatical category of a word according to its context.
- Extraction of noun phrases: extracts sequences of words that form noun phrases.
- Dictionary lookup: identifies the context of a word to find the corresponding dictionary entry.
- Recognition of idiomatic expressions: recognizes the expressions found in a text
- Relational morphology: groups together the words in a text that belong to the same family of derivatives, *i.e.* they have the same morphological root.

5.3 Incorporation of domain knowledge

Domain specific terminology (as Strategic Intelligence, CRM, Biology *etc.etc.*) is assured by specialistic Skills Cartridges which, as ontologies, may be loaded into Temis model.

5.4 Supported languages

English, French, German, Spanish, Portuguese, Italian, Dutch, Czech, Greek, Hungarian, Polish, Russian.
Danish, Swedish, Norwegian (2 versions), Finnish, Arabic are available at special conditions.

5.5 IT requirements

Hardware: Pentium-class PC or Server (Pentium III 1 Ghz). 256 MB RAM. Required space on the hard disc for software installation: 100 MB.
Operating System: Windows NT, Windows 2000, Windows XP, Linux.
Software: Java virtual machine: Java 1.3.1.

5.6 Customer base

In a couple of years Temis managed to build a solid customer base in Europe and America, in different sectors. Its main customers are:

Novartis, Roche, Dresdner Bank, Credit Lyonnais, Telecom Italia Mobile, France Telecom, Conoco, Total, DaimlerChrysler, Renault, Peugeot, Fiat, Hachette, European Parliament, Regione Calabria local government.

5.7 Partners

Several partners have signed agreements with Temis. The most known are IBM, AskMe, BCT.

6 Others

6.1 Autonomy

Autonomy (www.autonomy.com) was founded in 1996, with HQs in Cambridge, UK.

Autonomy use a combination of technologies that employs advanced pattern matching techniques using Bayesian Inference and Shannon Information theory. It identifies the patterns that occur in text, based on the usage and frequency of words or terms that corresponds to specific ideas or concepts. Autonomy is generally referred to as a key player in information retrieval market (*i.e.* well placed in search and categorization), appearing weaker in information extraction and clustering, well positioning *vs.* other players such as Verity.

6.2 Clearforest

Clearforest (www.clearforest.com) was founded in 1998, formerly called Instinct Software, present in Israel and USA.

The technology is based on pattern matching, and is considered particularly good in information extraction. Limitations in utilization with new languages.

6.3 Convera

Convera (www.convera.com) is the result of the merging between Excalibur and Intel Interactive Media Services Division.

Convera offers a high-performance, cross and multilingual search system which relies on semantic networks technology, with rich word dictionary and word relationship database.

6.4 Entrieva

Entrieva (www.entrieva.com), previously Semio Corp, was founded in 1966. It is based on computational linguistics, and has a wide range of products in text mining and information retrieval.

6.5 Fast

Fast Search & Transfer (www.fastsearch.com) was founded in 1997 in Norway. Its search and real time filtering technologies, its compression products and services are based on a modular technology platform. A consistent attribute of its technology is its scalability and efficiency.

6.6 IBM

IBM (www.ibm.com), the largest player in information technology worldwide, offers TMS –Text Mining Server (http://www.developer.ibm.com/solutions /isv/igssg.nsf/list/bycompanyname/86256B7B0003EBBF86256DF600491B20?OpenDocument) developed jointly with Temis.

The IBM offering in the unstructured data management is large, listing products as Intelligent Miner for Text, Lotus Discovery Server, Enterprise Information Portal, Web Fountain and several partnership.

6.7 Insightful

Insightful (www.insightful.com), founded in 1986, known worldwide for its statistics suite S-PLUS, enlarged its offering with its InFact, a stand-alone text mining solution for information retrieval based on semantic and syntactic methods. It understands questions, so the user doesn't need to be a subject expert to find the needed information.

6.8 Inxight

Inxight (www.inxight.com) was founded in 1997 as a spin off of Xerox PARC. Inxight is a key player in the text mining sector, thanks to large product mix and technical competence. Their offering is strong in information extraction and categorization, but also has summarization and visualization.

6.9 Verity

Verity (www.verity.com) is a player in information retrieval market, competing with Autonomy to be defined the market leader. It provides business portal infrastructure software, with a particularly good technology for full-text search, retrieval and categorization (by rules).