A Text Mining based content gathering system as strategic support for SMEs

N. Baldini[1], F. Neri[2] & M. Perrone[3]
[1]*Focuseek, Italy*
[2]*Synthema, Italy*
[3]*FirenzeTecnologia, Italy*

Abstract

The diffusion of innovation culture is an essential condition to guarantee entrepreneurial progress and improvement to the social fabric. Innovation processes in enterprises need constant Research and Development. The majority of Italian Small and Medium Enterprises (SMEs) experience difficulties in keeping in touch with innovations in the research world, which hinders competitiveness growth in actual global markets. FirenzeTecnologia together with Synthema-TEMIS-Focuseek has developed an integrated platform named SPI-RIT ("Services for Enterprises - Research and Technological Innovation"), which is directly oriented to facilitate structured and life-long relationships between the regional industrial world and scientific one. SPI-RIT collects heterogeneous and distributed data, integrating focused crawling, multilingual natural language processing indexing and searching, unsupervised clustering. SPI-RIT advocates and encourages integration between research and industry; promotes the industrial use of research results; increases the technological development of enterprises to support their competitiveness, relying upon a network of contacts with institutions, research institutes and consulting organisations.
Keywords: technological innovation, technological transfer, focused crawling, natural language processing, morphological analysis, syntactic analysis, functional analysis, unsupervised clustering.

WIT Transactions on Information and Communication Technologies, Vol 37, © 2006 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA060361

1 Introduction

FirenzeTecnologia is the Special Agency established by the Chamber of Commerce of Florence to promote technological innovation and foster the competitive development of the local productive system.

Since it was founded, FirenzeTecnologia has paid an increasing attention to the removal of obstacles between research and industry, encouraging the creation of a network of firms and local universities, supporting them in their research activities. To achieve its goals, FirenzeTecnologia has created an integrated platform named SPI-RIT ("Services for Enterprises - Research and Technological Innovation"), which is directly oriented to facilitate structured and life-long relationships between regional industrial and scientific areas, promoting information and innovation awareness among the SMEs. In fact, SPI-RIT supports spontaneous innovation processes, reacting to market needs and promoting the transfer of technology and know how; facilitating relations between the institutions, the scientific environment and the business sector. Then it suggests useful ways of anticipating needs, highlighting the risks due to rapid changes in the markets and in technology, creating more awareness and opportunities for company emancipation;

SPI-RIT consists in a package of on-line services that encourage the use of the Internet to search for information, keep updated and discover new solutions to increase business competitiveness. SPI-RIT is a complete Internet application, which collects and analyzes large sets of textual data according to morphological-syntactic-functional and statistical criteria, indexing them by their most significant lemmas and noun phrases they contain. It allows users to search for information and dynamically cluster results by Text Mining techniques.

2 The infrastructure

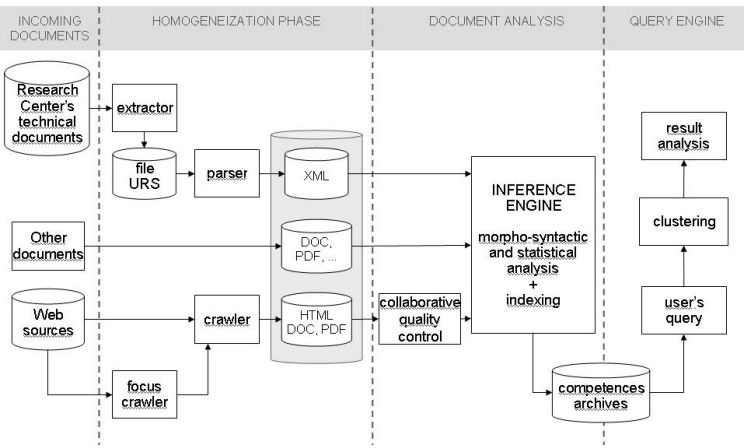

Figure 1: Overall system infrastructure.

2.1 The gathering system: Searchbox

In any large company or public administration the goal of aggregating contents from different and heterogeneous sources (even if they are located and managed by the company itself) is really a difficult task to accomplish. Exporting data from an existing database means that all the people providing and using the content have to obtain the necessary authorizations, or some human resources have to be allocated in order to write the sw procedures needed to get the data. In this scenario, a crawling technology can enormously simplify the integration task, because the crawler acts exactly like any other authorized user whose accessing procedures are already defined and accepted by all departments inside the organization.

The crawler aggregates many different information sources and provides some standard application services to access them (Figure 2).

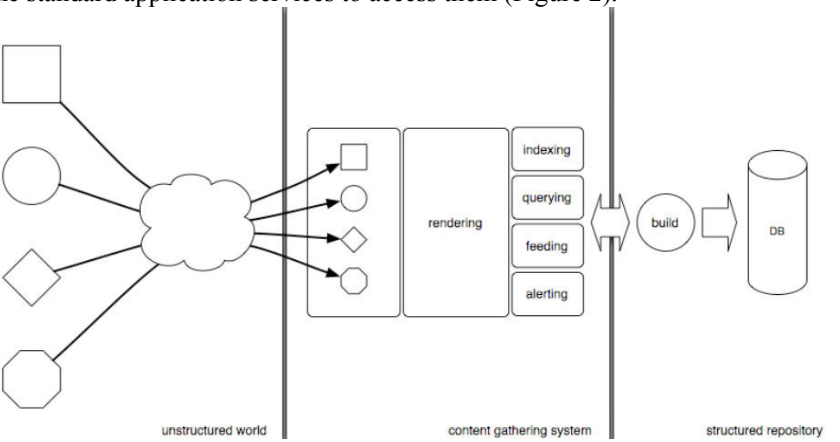

Figure 2: The content gathering component.

On the left side of Figure 2, the heterogeneous world of contents providers is sketched: the different shapes represent the different protocols and formats used to access documents. Then the content gathering module – in the middle – chooses the right adapter to gather information from any content provider, filling the structured repository on the right.

Searchbox is a multimedia content gathering and indexing system [3], whose main goal is managing huge collections of data coming from different and geographically distributed information sources. Searchbox, whose architecture has been conceived as a layer for information retrieval services in large enterprises, government institution, and Internet vertical portals, provides a very flexible and high performance dynamic indexing for content retrieval.

In Searchbox (see Figure 2), the *gatherer* is the coordinator of a pool of gathering agents whose task is to acquire new data from an information source, as soon as it is available. For instance, a noticeable example of a gathering agent

is the focused Web crawler, which starts form a set of initial Web pages (seeds) and performs intelligent navigation on the basis of appropriate classifiers. The gathering activities of the Searchbox, however, are not limited to the standard Web, but operate also with other sources like remote databases by ODBC, Web sources by FTP-Gopher, Usenet news by NNTP, WebDav and SMB shares, mailboxes by POP3-POP3/S-IMAP-IMAP/S, file systems and other proprietary sources [2].

The *renderer* is a central component in the Searchbox architecture. Searchbox indexing and retrieval system does not work on the original version of data, but on the "rendered version". Any piece of information (e.g. a document) is processed to produce a set of features using an appropriate algorithm. For instance, the features extracted from a portion of text might be a list of keywords/lemmas, while the extraction of features from a bitmap image might be extremely sophisticated. Even complex sources, like video, might be suitably processed so as to extract a textual-based labeling, which can be based on both the recognition of speech and sounds. All extracted features are then compiled in an internal XML format and passed to the indexing module. The extraction process of the renderer component is done by a pipeline of plug-ins, which provides the compilation of the final XML representation.

The *indexer* creates the index of the collection of information gathered from multiple sources, while the querying module offers a complete query language for retrieving original contents. The index is fully dynamic in the sense that any indexed content is almost-immediately available for queries. This is a crucial feature when the system is used on highly dynamic sources.

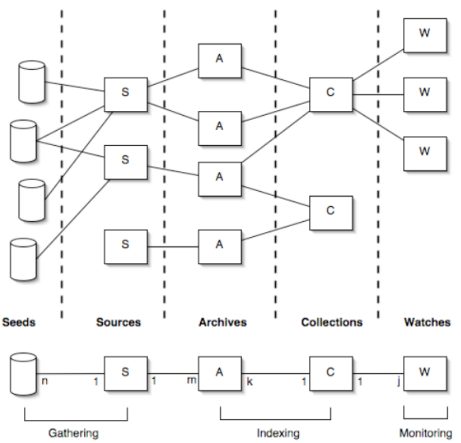

Figure 3: Searchbox main concepts: seeds, sources, archives, collections, watches.

The Searchbox indexer module can manage any feature that a specific renderer plug-in is able to extract from the original raw content. All of the extracted and indexed features can be combined in the query language made

available by the query interface of the indexer module. Searchbox provides default plug-ins to extract text from most common types of documents, like HTML, XML, TXT, PDF, PS and DOC. Other formats can be supported using specific plugins. Finally, a multilevel cache is available, which can be used to store and index multiple versions of the same content. The possibility to "historicize" different versions of the same document is a relevant practical feature, which turns out to be especially interesting for the implementation of the watch Searchbox concept.

2.1.1 Focused crawling

Focused crawling aims to crawl only the subset of the Web pages related to a specific category. The major problem in focused crawling is performing appropriate credit assignment to different documents along a crawl path, such that short-term gains are not pursued at the expense of less-obvious crawl paths that ultimately yield larger sets of valuable pages. To address this problem we present a focused crawling algorithm that builds a model for the context within which topically relevant pages occur on the Web. This context model can capture typical link hierarchies within which valuable pages occur, as well as model content on documents that frequently co-occur with relevant pages. This algorithm further leverages the existing capability of large search engines to provide partial reverse crawling capabilities. The algorithm shows significant performance improvements in crawling efficiency over standard focused crawling. Credit assignment for focused crawlers can be significantly improved by equipping the crawler with the capability of modelling the context within which the topical materials is usually found on the Web [1]. Such a context model has to capture typical link hierarchies within which valuable pages occur, as well as describe off-topic content that co-occurs in documents that are frequently closely associated with relevant pages. It is presented a general framework and a specific implementation of such a context model, which we call a Context Graph. Algorithm further differs from existing focused crawlers in that it leverages the capability of existing exhaustive search engines to provide partial reverse crawling capabilities. As a result it has a rapid and efficient initialization phase, and is suitable for real-time services. The Context Focused Crawler (CFC), uses the limited capability of search engines like AltaVista or Google to allow users to query for pages linking to a specified document. This data can be used to construct a representation of pages that occur within a certain link distance (defined as the minimum number of link traversals necessary to move from one page to another) of the target documents. This representation is used to train a set of classifiers, which are optimized to detect and assign documents to different categories based on the expected link distance from the document to the target document. During the crawling stage the classifiers are used to predict how many steps away from a target document the current retrieved document is likely to be. This information is then used to optimize the search. There are two distinct stages to using the algorithm when performing a focused crawl session:

1. An initialization phase when a set of context graphs and associated classifiers are constructed for each of the seed documents

2. A crawling phase that uses the classifiers to guide the search, and performs online updating of the context graphs.

The complete system is shown in Figure 5.

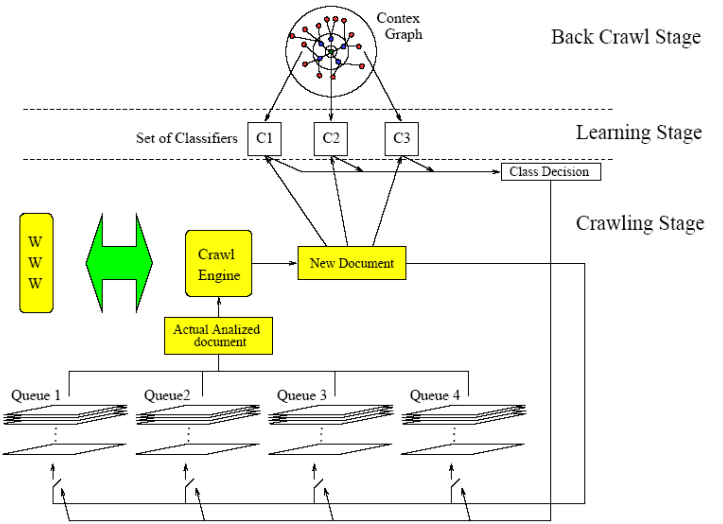

Figure 4: Graphical representation of the context focused crawler.

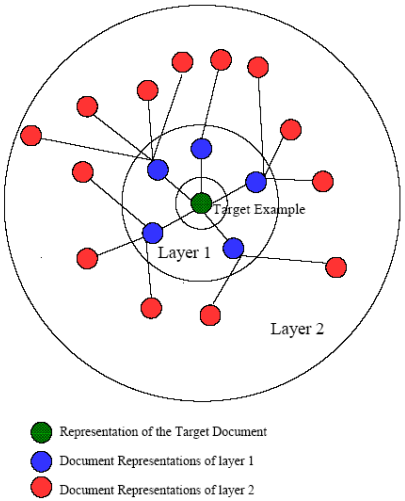

Figure 5: A context graph represents how a target document can be accessed from the Web.

3 The lexical system

The automatic linguistic analysis of free textual documents is based on Morphological, Syntactic, Functional and Statistical criteria. This phase is intended to identify only the significant expressions from the whole raw text. At the heart of the lexical system is a theory of McCord's, known as Slot Grammar [4]. A slot, explains McCord, is a placeholder for the different parts of a sentence associated with a word. A word may have several slots associated with it, and these form a *slot frame* for the word. In order to identify the most relevant terms in a sentence, the system analyzes it and, for each word, the Slot Grammar parser draws on the word's slot frames to cycle through the possible sentence constructions. Using a series of word relationship tests to establish context, the system tries to determine the meaning of the sentence. Each slot structure can be partially or fully instantiated and it can be filled with representations from one or more statements to incrementally build the meaning of a statement. This includes most of the treatment of coordination, which uses a method of 'factoring out' unfilled slots from elliptical coordinated phrases. The parser - a bottom-up chart parser - employs a parse evaluation scheme used for pruning away unlikely analyses during parsing as well as for ranking final analyses.

Figure 6: Lexical analysis.

By including semantic information directly in the dependency grammar structures, the system relies on the lexical semantic information combined with functional application rules.

Shouldn't the lexical system be able to detect the proper functional role of each word and recognise as relevant information only those terms or phrases that comply with a set of pre-defined morphological patterns (i.e.: `noun+noun` and `noun+preposition+noun` sequences) and whose frequency exceeds a threshold of significance? The Information Quotient is calculated taking in account the term, its *Part Of Speech* tag, its relative and absolute frequency, its distribution on documents [7].

The detected terms and phrases are then extracted, reduced to their *Part Of Speech* (NOUN, VERB, ADJECTIVE, ADVERB, etc.) *and Functional* (AGENT, OBJECT, WHERE, CAUSE, etc.) tagged base form [5]. Once referred to their language independent entry inside the sectorial multilingual dictionary, they are used as descriptors for documents [7–9]. In multilingual dictionaries, each lemma is referenced to syntax or domain dependent translated terms, so that each entry can represent multiple senses. Besides, the multilingual dictionaries contain lemmas together with simple binary features, as well as sophisticated tree-to-tree translation models, which map - node by node - whole sub-trees [7].

4 The search and clustering system

Users can search document by keywords combined by Boolean operators, or by typing their own query in Natural Language, expressed using normal conversational syntax. Traditional Boolean queries, while precise, require strict interpretation that can often exclude information that is relevant to user interests. The system analyzes the query, identifying the most relevant terms contained, their semantic and functional interpretation, expanding terms and concepts to all the languages supported by the system (English, French, German, Italian, Spanish, Portuguese). The search engine returns as result all the documents which contain the query lemmas, having the same functional role.

The automatic classification of results is made by Online Miner Light, which is an application developed by TEMIS jointly with SYNTHEMA, and fulfils the Unsupervised Classification schema. (TEMIS was established in 2000 as a Technology & Consulting Company, specialized in Text Intelligence and Advanced Computational Linguistics to develop applications related to Competitive Intelligence, Customer Relationship Management and Knowledge Management.) The application dynamically discovers the thematic groups that best describe the detected documents, according to the K-Mean approach. This phase allows users to access documents by topics, not by keywords. The application provides a visual summary of the analysis (See Figure 7). A map shows the different groups of documents as differently sized bubbles (the size depends on the number of documents the bubble contains) and the meaningful correlation among them as lines drawn with different thickness (that is level of correlation). Users can search inside topics and have a look of the documents populating the clusters. The output results can be viewed by a simple Web browser.

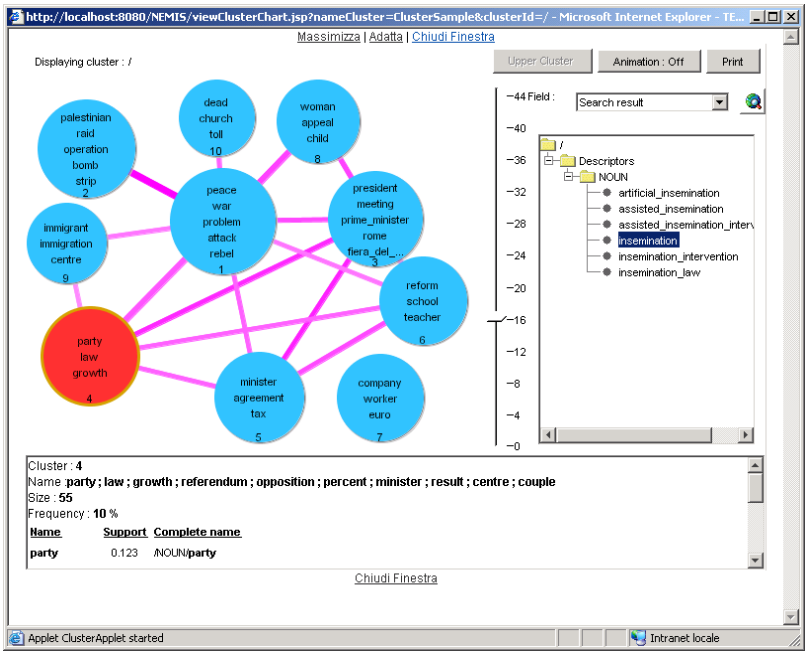

Figure 7: Thematic map and search in topics.

5 Conclusions

This paper describes a Multilingual Text Mining application, which promotes information and innovation awareness among the SMEs. Lexical analysis and Translation Memories permit to overcome linguistic barriers, allowing the automatic indexation and classification of documents, whatever it might be their language, or the source they are collected from. This new approach enables the research, the analysis, the classification of great volumes of heterogeneous documents, helping people to cut through the information labyrinth. Being multilinguality an important part of this globalised society, Multilingual Text Mining is a major step forward in keeping pace with the relevant developments in the challenging and rapidly changing world.

References

[1] Diligenti, M., Coetzee, F. M., Lawrence, S., Giles, C. L., Gori, M., *Focused Crawling Using Context Graphs*, Proceedings of 26th International Conference on Very Large Databases, VLDB, pp. 527-534, 10/9 - 12/9 2000.

[2] Baldini, N., Gori, M., Maggini, M., *Mumblesearch: Extraction of high quality Web information for SME*, 2004 IEEE/WIC/ACM International Conference on Web Intelligence.

[3] Baldini, N., Bini, M., *Focuseek searchbox for digital content gathering*, AXMEDIS 2005 - 1st International Conference on Automated Production of Cross Media Content for Multi-channel Distribution, Proceedings Workshop and Industrial pp. 24-28.

[4] McCord, M. C., *Slot Grammar: A System for Simpler Construction of Practical Natural Language Grammars* Natural Language and Logic 1989: 118-145
McCord, M. C., *Design of LMT: A Prolog-Based Machine Translation System*, Computational Linguistics 15(1): 33-52 (1989)
McCord, M. C., *Using Slots and Modifiers in Logic Grammars for Natural Language*. Artif. Intell. 18(3): 327-367 (1982)
McCord, M. C., *Slot Grammars*, American Journal of Computational Linguistics 6(1): 31-43 (1980)

[5] Raffaelli, R., *An inverse parallel parser using multi-layerd grammars*, IBM Technical Disclosure Bulletin, 2Q, 1992.

[6] Marinai, E., Raffaelli, R., *The design and architecture of a lexical data base system*, COLING'90, Workshop on advanced tools for Natural Language Processing, Helsinki, Sweden, Aug 1990, 24.

[7] Raffaelli, R., *ABCD – A Basic Computer Dictionary*, Proceedings of ELS Conference on Computational Linguistics, Kolbotn, Norway, Aug 1988, 30-31.

[8] Galli, G., Raffaelli, R., Saviozzi, G., *Il trattamento delle espressioni composte nel trattamento del linguaggio naturale*, IBM Research Center, internal report, Pisa, Italy, pp. 1-19, 1992.

[9] Cascini, G., Neri, F., *Natural Language Processing for Patents Analysis and Classification*, ETRIA World Conference, TRIZ Future 2004, Florence, Italy.

[10] Neri, F., Raffaelli, R., *Una nuova procedura multilingua Text Mining, basata sulla rilevazione della terminologia principali, delle memorie di traduzione e sul Clustering*, Text Mining, uno strumento strategico per imprese ed istituzioni, di S.Bolasco, A. Canzonetti, F.Capo, pp. 71-74, CISU Ed., ISBN: 88-7975-341-X.

[11] Neri, F., *Multilingual Text Mining*, Data Mining VI, *Sixth International Conference on Data Mining, Text Mining and their Business Applications*, Skiathos (Grecia), Proceedings, Management Information Systems, Vol 11, A. Zanasi Ed., ISBN: 1-84564-017-9, 25-27/5/2005.

[13] Neri, F, Raffaelli, R., *Text Mining applied to Multilingual Corpora*, NEMIS Final Conference, Network of Excellence in Text Mining and its Applications in Statistics, Knowledge Mining, Atene (Grecia), Proceedings, Springer Verlag Ed., 25/10/2004.