A multilanguage platform for Open Source Intelligence

N. Baldini[1], F. Neri[2] & M. Pettoni[3]
[1]*Focuseek, Italy*
[2]*Synthema, Italy*
[3]*CIFIGE, II Information and Security Department (RIS),*
STATO MAGGIORE DIFESA, Italy

Abstract

Open Source Intelligence (OSINT) is an intelligence gathering discipline that involves collecting information from open sources and analyzing it to produce usable intelligence. The revolution in information technology is making open sources more accessible, ubiquitous, and valuable, making open intelligence at less cost than ever before. The explosion in OSINT is transforming the intelligence world with the emergence of open versions of the specialistic arts of human intelligence (HUMINT), overhead imagery intelligence (IMINT), and signals intelligence (SIGINT). The international Intelligence Communities have seen open sources grow increasingly easier and cheaper to acquire in recent years. However, up to 80% of electronic data is textual and most valuable information is often hidden and encoded in pages which are neither structured, nor classified. The process of accessing all these raw data, heterogeneous in terms of source and language, and transforming them into information is therefore strongly linked to automatic textual analysis and synthesis, which are greatly related to the ability to master the problems of multilinguality. This paper describes a multilingual indexing, searching and clustering system, designed to manage huge sets of data collected from different and geographically distributed information sources, which provides language independent search and dynamic classification features. The Joint Intelligence and EW Training Centre (CIFIGE) is a military institute, which has adopted this system in order to train the military and civilian personnel of Defence in the OSINT discipline.

Keywords: open source intelligence, focused crawling, natural language processing, morphological analysis, syntactic analysis, functional analysis, unsupervised clustering.

WIT Transactions on Information and Communication Technologies, Vol 38, © 2007 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA070321

1 Open Source Intelligence and information overload

Open Source Intelligence (OSINT) is an intelligence gathering discipline that involves collecting information from open sources and analyzing it to produce usable intelligence. The specific term “open” refers to publicly available sources, as opposed to classified sources. OSINT includes a wide variety of information and sources. With the Internet, the bulk of predictive intelligence can be obtained from public, unclassified sources. The revolution in information technology is making open sources more accessible, ubiquitous, and valuable, making open intelligence at less cost than ever before. In fact, monitors no longer need an expensive infrastructure of antennas to listen to radio, watch television or gather textual data from Internet newspapers and magazines.

The availability of huge amount of data available in the open sources leads to the well-identified nowadays paradox: an overload of information means no usable knowledge. Besides, open source texts are - and will be - written in various native languages, but these documents are relevant even to non-native speakers. Independent information sources can balance the limited information normally available, particularly if related to non-cooperative targets. The process of accessing all these raw data, heterogeneous both for type (scientific article, patent, free textual document), source (Internet/Intranet, database, etc), protocol (HTTP/HTTPS, FTP, GOPHER, IRC, NNTP, etc) and language used, and transforming them into information, is therefore inextricably linked to the concepts of focused crawling, textual analysis and synthesis, hinging greatly on the ability to master the problems of multilinguality. This task can require undoubtedly remarkable efforts.

The Joint Intelligence and EW Training Centre (CIFIGE) is a military institute, which depends from the II Information and Security Department (RIS) of the Italian Defence General Staff. It constitutes a reference point for the ongoing military national doctrine and the pertaining Intelligence aspects. The center is responsible for the training in the field of Military Intelligence by courses and seminaries. The Intelligence disciplines are managed and examined in depth during courses, doctrines and on the Job trainings.

2 The Logical components

2.1 The crawler

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 hard to be accomplished. Exporting data from an existing database means that all the people providing and using the content has 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.

Searchbox is a multimedia content gathering and indexing system, 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, the *gatherer* is the coordinator of a pool of 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 - the 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.

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 then processed to produce a set of features using appropriate algorithms. For instance, the features extracted from a portion of text might be a list of keywords/lemmas/concepts, while the extraction of features from a bitmap image might be extremely sophisticated. Even more 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 provide 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, wading through millions of documents. 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.

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 which is available in the user interface. 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: 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 and alert concepts, when managing tons of documents.

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 the 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 the focused crawling algorithm builds a model for the context within which topically relevant pages occur on the Web. This algorithm shows significant performance improvements in crawling efficiency over standard focused crawling. In fact, the credit assignment 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. 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. The general framework and the specific implementation of such a context model are called Context Graph. It has a rapid and efficient initialization phase, being 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 far 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.

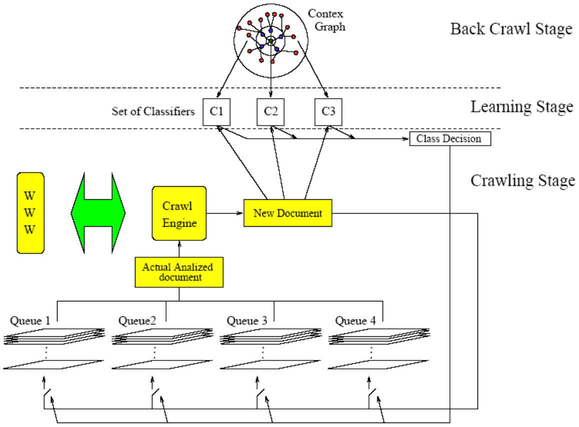


Figure 1: Graphical representation of the Context Focused Crawler.

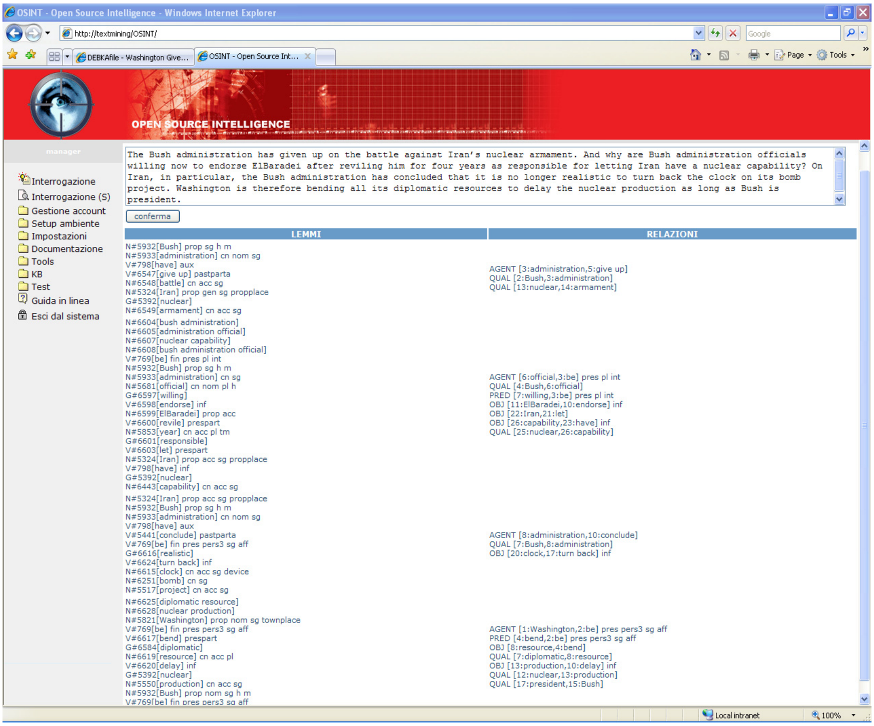
2.2 The lexical system

The automatic linguistic analysis of the 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.

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, it recognises 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 (the Linguistic Preprocessing extracts bilingual lexicons from comparable and parallel corpora, enriching existing bilingual dictionaries and helping overcome the language barrier for cross-language information classification. The major problem consists in the different syntactic structure and words definition these two languages may have. So a direct phrasal alignment is often needed. The following bilingual morphological analysis recognises as relevant terminology only those terms or phrases, that exceed a threshold of significance. This morphological analysis detects significant Simple Word Terms (SWT) and Multi Word Terms (MWT), annotating their headwords, their relative and absolute positions. SYNTHEMA strategy on multilingual dictionary construction consists in the assumption that, having taken in account a specific term S and its phrasal occurrences, its translation T can be automatically detected by analysing the correspondent translated sentences. Thus, semi-automatic lexicon extraction and storage of multilingual relevant descriptors become possible), they are used as descriptors for documents [7–10]. 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].



The Bush administration has given up on the battle against Iran's nuclear armament. And why are Bush administration officials willing now to endorse ElBaradei after reviling him for four years as responsible for letting Iran have a nuclear capability? On Iran, in particular, the Bush administration has concluded that it is no longer realistic to turn back the clock on its bomb project. Washington is therefore bending all its diplomatic resources to delay the nuclear production as long as Bush is president.

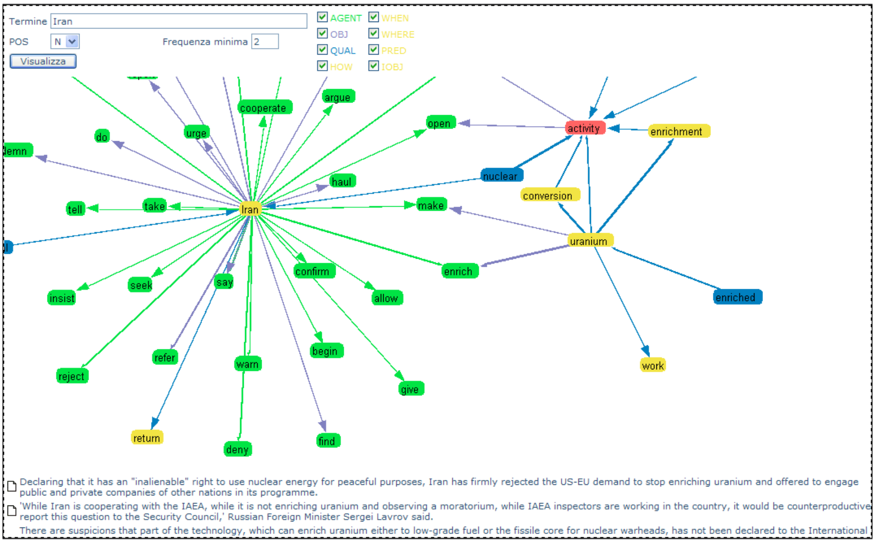
LEMMI	RELAZIONI
N#5932[Bush] prop sg h m N#5933[administration] cn nom sg V#798[have] aux V#6547[give up] pastparta N#6548[battle] cn acc sg N#5324[Iran] prop gen sg propplace G#5392[nuclear] N#6549[armament] cn acc sg	AGENT [3:administration,5:give up] QUAL [2:Bush,3:administration] QUAL [13:nuclear,14:armament]
N#6604[bush administration] N#6605[administration official] N#6607[nuclear capability] N#6608[bush administration official] V#769[be] fin pres pl int N#5932[Bush] prop sg h m N#5933[administration] cn sg N#5681[official] cn nom pl h G#6597[willing] V#6598[endorse] inf N#6599[ElBaradei] prop acc V#6600[revile] prespart N#5853[year] cn acc pl tm G#6601[responsible] V#6603[let] prespart N#5324[Iran] prop acc sg propplace V#798[have] inf G#5392[nuclear] N#6443[capability] cn acc sg	AGENT [6:official,3:be] pres pl int QUAL [4:Bush,6:official] PRED [7:willing,3:be] pres pl int OBJ [11:ElBaradei,10:endorse] inf OBJ [22:Iran,21:let] OBJ [26:capability,23:have] inf QUAL [25:nuclear,26:capability]
N#5324[Iran] prop acc sg propplace N#5932[Bush] prop sg h m N#5933[administration] cn nom sg V#798[have] aux V#5441[conclude] pastparta V#769[be] fin pres pers3 sg aff G#6616[realistic] V#6624[turn back] inf N#6615[clock] cn acc sg device N#6251[bomb] cn sg N#5517[project] cn acc sg	AGENT [8:administration,10:conclude] QUAL [7:Bush,8:administration] OBJ [20:clock,17:turn back] inf
N#6625[diplomatic resource] N#6628[nuclear production] N#5821[Washington] prop nom sg townplace V#769[be] fin pres pers3 sg aff V#6617[bend] prespart G#6584[diplomatic] N#6619[resource] cn acc pl V#6620[delay] inf G#5392[nuclear] N#5550[production] cn acc sg N#5932[Bush] prop nom sg h m V#769[be] fin pres pers3 sg aff	AGENT [1:Washington,2:be] pres pers3 sg aff PRED [4:bend,2:be] pres pers3 sg aff OBJ [8:resource,4:bend] QUAL [7:diplomatic,8:resource] OBJ [13:production,10:delay] inf QUAL [12:nuclear,13:production] QUAL [17:president,15:Bush]

Figure 2: Lexical analysis.

2.3 Functional navigation

Users can search and navigate by roles, exploring sentences and documents by the functional role played by each concept/lemma, as shown in Figure 3. Users can navigate on the relations chart by simply clicking on nodes or arches, expanding them and taking a look of sentences/documents characterized by the selected criterion.

This can be considered a visual investigative analysis component specifically designed to bring clarity to complex investigations. It automatically enables investigative information to be represented as visual elements that can be easily analyzed and interpreted. Functional relationships - AGENT, ACTION, OBJECT, QUALIFIER, WHEN, WHERE, HOW - among human beings and organizations can be searched for and highlighted, pattern and hidden connections can be instantly revealed to help investigations, promoting efficiency into investigative teams. Should human beings be cited, their photos can be shown by simple clicking on the related icon.



Termine: Iran

POS: N

Frequenza minima: 2

☑ AGENT ☑ WHEN
☑ OBJ ☑ WHERE
☑ QUAL ☑ PRED
☑ HOW ☑ IOBJ

Declaring that it has an "inalienable" right to use nuclear energy for peaceful purposes, Iran has firmly rejected the US-EU demand to stop enriching uranium and offered to engage public and private companies of other nations in its programme.

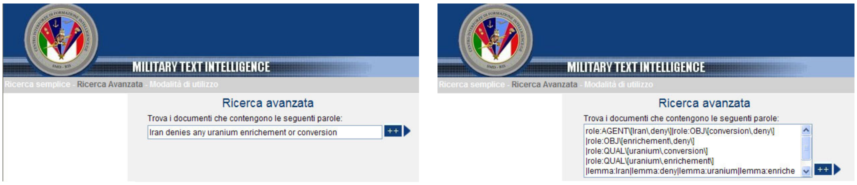
'While Iran is cooperating with the IAEA, while it is not enriching uranium and observing a moratorium, while IAEA inspectors are working in the country, it would be counterproductive report this question to the Security Council,' Russian Foreign Minister Sergei Lavrov said.

There are suspicions that part of the technology, which can enrich uranium either to low-grade fuel or the fissile core for nuclear warheads, has not been declared to the International

Figure 3: Functional search and navigation.

2.4 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 concepts/lemmas in the same functional role as in the query, trying to retrieve all the texts which constitute a real answer to the query.



MILITARY TEXT INTELLIGENCE
Ricerca semplice - Ricerca Avanzata - Modalità di utilizzo
++

Ricerca avanzata

Trova i documenti che contengono le seguenti parole:

Iran denies any uranium enrichement or conversion

MILITARY TEXT INTELLIGENCE
Ricerca semplice - Ricerca Avanzata - Modalità di utilizzo
++

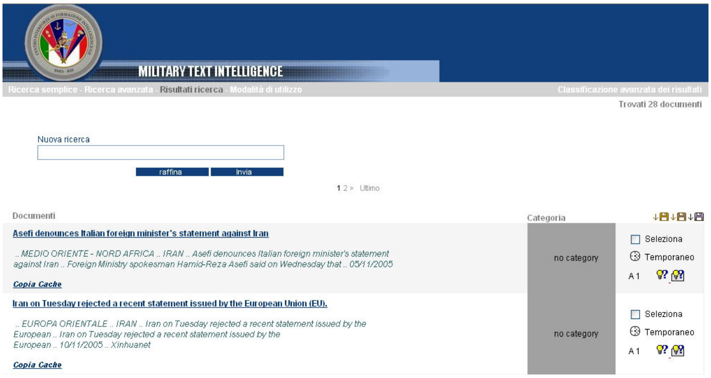
Ricerca avanzata

Trova i documenti che contengono le seguenti parole:

```
role:AGENT\[Iran\,deny\]|role:OBJ\[conversion\,deny\]
|role:OBJ\[enrichement\,deny\]
|role:QUAL\[uranium\,conversion\]
|role:QUAL\[uranium\,enrichement\]
|lemma:Iran|lemma:deny|lemma:uranium|lemma:enriche
```

Figure 4: Natural language query and its functional and conceptual expansion.

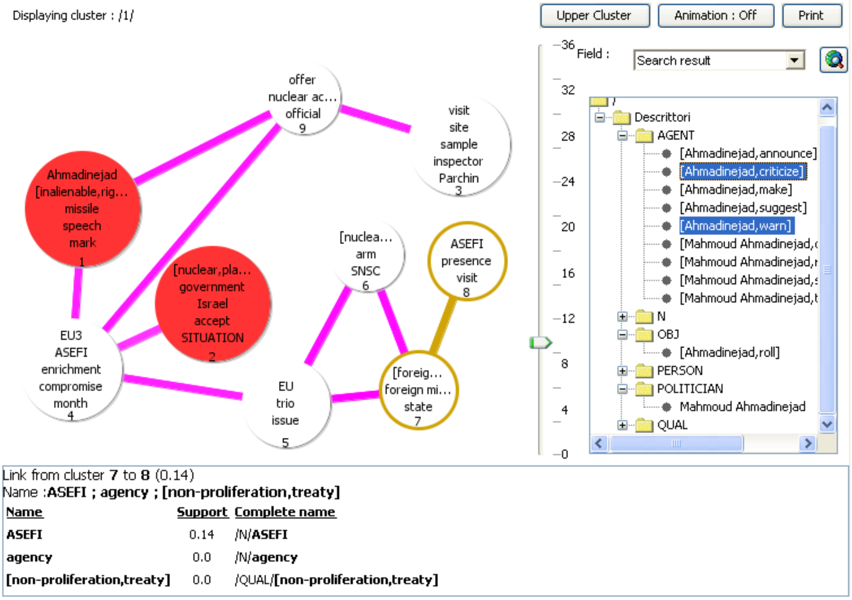
Results are then displayed and ranked by relevance, reliability and credibility, as expressed by the A1-F6 schema (source reliability ratings range from A to F: A=Reliable, B=Usually Reliable, C=Fairly Reliable, D=Not usually Reliable, E=Unreliable, F=Cannot Be Judged. An F rating does not necessarily mean the source is unreliable, but that the processing personnel have no previous experience upon which to base a determination. Information credibility ratings instead range from 1 (Confirmed) to 6 (Cannot Be Judged)).


MILITARY TEXT INTELLIGENCE
Ricerca semplice - Ricerca avanzata - Risultati ricerca - Modalità di utilizzo
Classificazione avanzata dei risultati
Nuova ricerca
raffina
Invia
1 2 > Ultimo

Trovati 28 documenti

Documenti	Categoria	
Asefi denounces Italian foreign minister's statement against Iran *.. MEDIO ORIENTE - NORD AFRICA .. IRAN .. Asefi denounces Italian foreign minister's statement against Iran .. Foreign Ministry spokesman Hamid-Reza Asefi said on Wednesday that .. 05/11/2005* ***Copia Cache***	no category	□ Seleziona Temporaneo A 1
Iran on Tuesday rejected a recent statement issued by the European Union (EU). *.. EUROPA ORIENTALE .. IRAN .. Iran on Tuesday rejected a recent statement issued by the European .. Iran on Tuesday rejected a recent statement issued by the European .. 10/11/2005 .. Xinhuanet* ***Copia Cache***	no category	□ Seleziona Temporaneo A 1

Figure 5: Search results.



Link from cluster **7** to **8** (0.14)
Name :**ASEFI ; agency ; [non-proliferation,treaty]**

Name	Support	Complete name
ASEFI	0.14	/N/**ASEFI**
agency	0.0	/N/**agency**
[non-proliferation,treaty]	0.0	/QUAL/**[non-proliferation,treaty]**

Figure 6: Thematic map, functional search and projection inside topics.

Conceptual and lexical descriptors can be exported to I2 Analyst's Notebook®, or to Microsoft® Excel.

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. The application dynamically discovers the thematic groups that best describe the detected documents, according to the K-Mean approach and using functional relationships and lemmas as descriptors. This phase allows users to access documents by topics, not by keywords. The application provides a visual summary of the analysis. 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. Then users can project clusters and documents on lemmas and their functional relations. The output results can be viewed by a simple Web browser.

3 Conclusions

This paper describes a Multilingual Text Mining platform for Open Source Intelligence, adopted by Joint Intelligence and EW Training Centre (CIFIGE) to train the military and civilian personnel of Italian Defence in the OSINT discipline.

Multilanguage Lexical analysis permits to overcome linguistic barriers, allowing the automatic indexation, simple navigation and classification of documents, whatever it might be their language, or the source they are collected from. This approach enables the research, the analysis, the classification of great volumes of heterogeneous documents, helping intelligence analysts to cut through the information labyrinth.

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, September 2000, 10-12.

[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-layered 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, Finland, August 1990, 24.

[7] Raffaelli, R., *ABCD – A Basic Computer Dictionary*, Proceedings of ELS Conference on Computational Linguistics, Kolbotn, Norway, August 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 principale, 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 (Greece), Proceedings, Management Information Systems, Vol 11, A. Zanasi Ed., ISBN: 1-84564-017-9, May 2005, 25-27.

[13] Neri, F., Raffaelli, R., *Text Mining applied to Multilingual Corpora*, Knowledge Mining: Proceedings of the NEMIS 2004 Final Conference, Springer Verlag Pub., Spiros Sirmakessis Ed., ISBN-13: 978-3540250708.

[14] Baldini, N., Neri, F., *A Multilingual Text Mining based content gathering system for Open Source Intelligence*, IAEA International Atomic Energy Agency, Symposium on International Safeguards: Addressing Verification Challenges, Wien, Austria, IAEA-CN-148/192P, Book of Extended Synopses, pp. 368-369, October 2006, 16-20.