International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:8, 2007

A Web Text Mining Flexible Architecture

M. Castellano, G. Mastronardi, A. Aprile, and G. Tarricone

***Abstract*—**Text Mining is an important step of Knowledge Discovery process. It is used to extract hidden information from not-structured o semi-structured data. This aspect is fundamental because much of the Web information is semi-structured due to the nested structure of HTML code, much of the Web information is linked, much of the Web information is redundant. Web Text Mining helps whole knowledge mining process to mining, extraction and integration of useful data, information and knowledge from Web page contents.

In this paper, we present a Web Text Mining process able to discover knowledge in a distributed and heterogeneous multi-organization environment. The Web Text Mining process is based on flexible architecture and is implemented by four steps able to examine web content and to extract useful hidden information through mining techniques. Our Web Text Mining prototype starts from the recovery of Web job offers in which, through a Text Mining process, useful information for fast classification of the same are drawn out, these information are, essentially, job offer place and skills.

***Keywords*—**Web text mining, flexible architecture, knowledge discovery.

I. Introduction

KNOWLEDGE Discovery Process is a creative process which requires a number of different skills and knowledge. Currently there is no standard framework in which to carry out knowledge discovery process projects. This means that the success or failure of a knowledge discovery process project is highly dependent on the particular person or team carrying it out and successful practice can not necessarily be repeated across the enterprise. Knowledge Discovery process needs a standard approach which will help translate business problems into mining tasks, suggest appropriate transformations and mining techniques, and provide means for evaluating the effectiveness of the results and documenting the experience[1,2,3,4,5,6]. There have been a number of techniques to realize a Knowledge Discovery Process using different approaches, the most recent effective strategy being Mining Techniques[7].

Data Mining is concerned with finding patterns in data which are interesting, according to some user-defined measure of interestingness, and valid, according to some user defined measure of validate. This area has recently gained much attention of industry, due to the existence of large collections of data in different formats, including large data warehouses, and the increasing need of data analysis and comprehension. In addition to mining of data stored in data warehouses, there has recently been also increased interest in Text and Web mining. Web mining is a application of data mining techniques to discover patterns from the Web. According to analysis targets, web mining can be divided into three different types, which are Web usage mining, Web content mining and Web structure mining. Web content mining is the process to discover useful information from the content of a web page. The type of the web content may consist of text, image, audio or video data in the web. Web content mining sometimes is called Web Text Mining because the text content is the most widely researched area. Web Text Mining belongs to the area of web content mining and refers to the process of extracting interesting and non-trivial patterns or knowledge from web texts. Usually it is viewed as an extension of text mining and data mining. The technologies that are used in Web Text Mining are NLP, natural language processing and IR, information retrieval [8].

Web Text Mining is an increasingly important area of research driven by the dramatic growth in access to very large text stores such as the Web, email and domain-specific reports. At the heart of Web Text Mining is the need for tools that discover and exploit re-occurring patterns in text. Typically advances in machine learning and statistics are exploited for pattern discovery. Knowledge acquisition and representation techniques help transform discovered patterns into new knowledge structures. These structures are more transparent and manageable. Importantly, are able to provide higher level abstractions that facilitate the solution of text-related tasks [9].

In this work we describe a web text mining process based on flexible architecture able to discover knowledge in a distributed and heterogeneous multi-organization environment. This paper is composed as it follows: in the second section we introduce a generic web text mining lifecycle. In the third section we describe a web text mining process through the description of four steps: Crawling, Pre-Processing, Text Mining and Presentation of Results. In the fourth section we present the service oriented architecture and our flexible web text mining architecture. Finally we present a prototyping of web text mining process.

II. The Web Text Mining Lifecycle

Text Mining phase is preceded from Text Refining phase that transforms input raw text in structured information. Text Refining input are not-structured data as texts or semi-structured data as HTML pages coming from Web. These data are submitted to the opportune process composed by recovery of documents, cleaning of the draw out documents and extraction of the useful information to the Text Mining phase. Text Refining output is stored in database, XML file or any

other structured form that is called Intermediate Form [10].

Text Mining techniques are applied to the Intermediate Form. These techniques can be also very different between them like pure and simple transposition of Data Mining techniques, like reworking of these techniques for the source data nature or like the search of original techniques [11].

The range is very ample and the choice depends from the type of waited knowledge and from Intermediate Form type produced in Text Refining phase. Central element of process is understanding the issue, rather than the refinement of the involved technologies. Text Mining process doesn't consider only the presence or the frequency of a word or a concept inside a document but aims to find the relationship between. these words or concepts and others inside document. In this way it aims to find information contained in text.

Text Mining phases are: document clustering, document categorization, pattern extraction.

Document clustering is the assignment of entity multivariate to few categories (classes, groups) not previously defines. The criterion is to gather among them similar entity. It is necessary to establish the variable of classification, the measure of the proximities among the entities, the number of groups in which to leave again the entities and finally to pass to an integration phase with other methods of analysis. Principal techniques are distinguished in base to the type of analysis that can be hierarchical or not hierarchical type. Hierarchical analysis is distinguished by agglomerative and divisive techniques. Not hierarchical analysis is like the possibility of subdivision in mutually exclusive classes (partitions) or subdivision in overlapped classes. In this type of analysis is fundamental to establish previously the number of cluster to produce or to plan the analysis so that to get solutions which foresee a varying number of cluster. Textual clustering is used as automatic process to divide a documents collection in groups. Inside groups documents are similar among them on base of selected characteristics: author, length, dates, keywords. Textual clustering can be used for provide a panning of the contents of documents collection, for identify hidden similarity, for facilitate the process of browsing and for find correlated or similar information. If it works with keywords or with features that represents the semantics of the documents, the individualized groups will be distinguished in base to the different themes treated in the corpus. Algorithms of hierarchical clustering are used for textual data.

In Document categorization, the objects must be attribute to one or more classes or categories already notes in start. In text reference, this technique of mining is said Text Categorization (TC). Text Categorization is often developed through algorithms of Machine Learning. [12]. Classification is the process in which meaningful correlations among frequent whole frequent data are found. This is a typical process of Data Mining. There are association rules for Text Categorization. All algorithms for produce association rules operate in two phases. Gives keywords and a document collection, the extraction of associations, which satisfy given ties of support and confidence, is effected in two passages: before producing all the whole keywords with greater or equal support respect to the reference which are called frequent set, then producing all the association rules that can be derived from the produced frequent set and that satisfy the given confidence.

In Pattern Extraction some patterns are identified following the analysis of associations and of tendencies. The discovery of the associations is the process in which meaningful correlations among frequent whole data are found. This process can be applied to textual data if these data are been transformed previously in a structured form. Pattern Extraction needs particular attention in Web Text Mining process and is carried out through following techniques:

- Predictive Text Mining which is used for identifying the tendencies in collected documents in a time period.
- Association Analysis which identifies the relationships among the attributes, like if the presence of a pattern implicates the presence of another pattern in documents.

Finally, presented techniques are used for introduce the patterns and for visualize the results.

Fig. 1 shows the web text mining lifecycle.

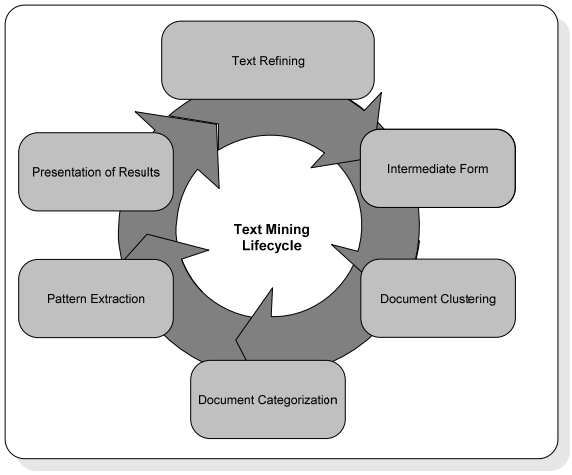

Fig. 1 Text Mining Lifecycle

III. The Web Text Mining Process

The information coming from Web represents an important role in Knowledge Discovery Process. To examine web contents and to extract useful information to a purpose through techniques of Mining is scientifically called Web Text Mining [13].

Fig. 2 shows the Web Text Mining Process taken into account.

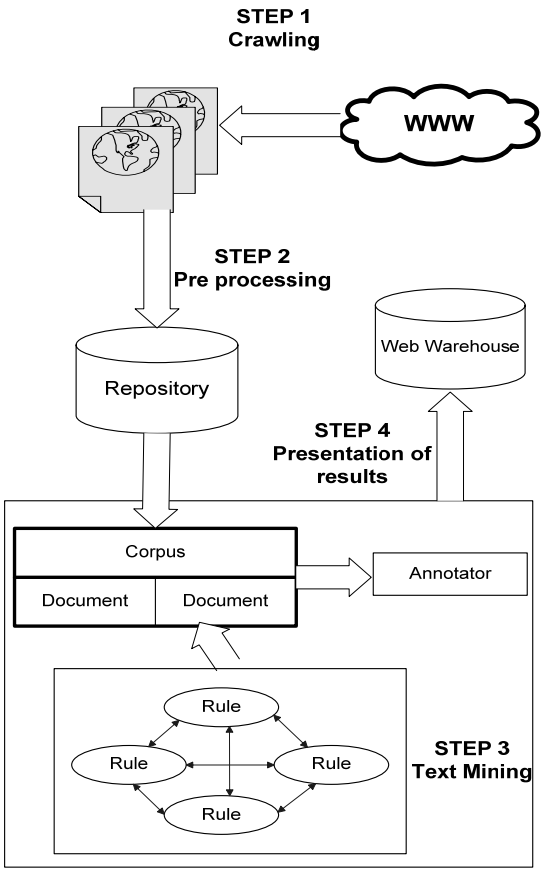

Fig. 2 The Web Text Mining Process

The Web Text Mining Process is implemented through four steps.

The first step foresees the recovery of the useful information from Web. This information consists in textual contents present in web pages. The references solution covers the first step of Web Text Mining process [14,15], dealing with automatic retrieval of all relevant documents and ensuring at the same time that the not-relevant ones macaws fetched as few as possible. The recovery of the useful information is effected through crawling departing from one or more URL. The crawling is implemented by a Web Crawler, as shown in the Fig. 3. Web Crawler's architecture fits the guides lines of to Focused Crawler[16], as it is designed to only gather documents on to specific topic, thus reducing the amount of network traffic and downloads. It is composed from four components: Master, Slaves, Scribes, H-Information. Every component has a specific assignment. H-Information (Human Information) is the start of whole process of crawling. It allows the system administrator to interact with Web Crawler inserting a list of sites Web believed of particular importance for the service that the system must realize. Master is the kernel of Web Crawler that allows the start and the management of the whole process of recovery of the information. Master calls Slaves and passed them URLs of the pages in accord with the built list by the H-Information. Slaves accede to the Web for crawling the pages which are associated to URLs indicated by Master. Different Slaves can operate at the same time rendering more efficient, reducing the answer times and avoiding simultaneous accesses to the same resources [24]. Slaves call Scribes [22,23]. Scribes receive Web pages crawled from Slaves. For each Scribe there is a Slave. Scribe parses entire page and captures different information according to the type of the examined document.

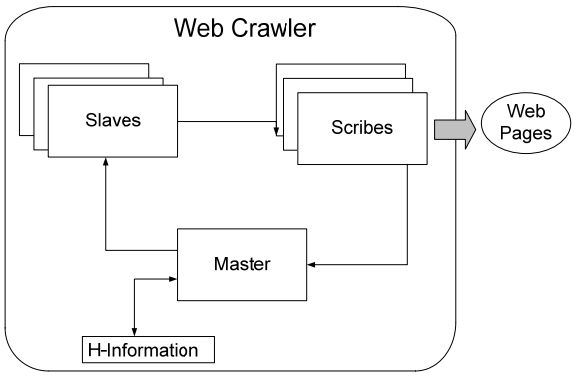

Fig. 3 Web Crawler Architecture

The second step is the pre-processing. It foresees the creation of a repository of the information from the web during the execution of the first step. During this phase it is had to provide to effect a control and a cleaning (Text Cleaning) of the pages reached during the first step. The definition of Template, which contains keywords of the universe, allows a choice on the single content of the examined web page. If it contains information that correspond to one of the Template defined for the system in matter, it is accepted and on it a cleaning of the content is effected to eliminate all what to be held "garbage", that is to say tag, banner, information not tightly connected to the content of the same page. Otherwise the page is discarded. If the page has found correspondence with one of the Template defined for the system, the "polished up" information of the page are memorized in a repository.

The third step is the fundamental step on which the whole Web Mining process founds him. In this step a first phase of Information Extraction and the following phase of Text Mining are effected. It founds on the use of Text Mining tools that receiving in input the data which are contained in the repository of the second step and they are able to extract "intrinsic information" contained in the document or in all documents.

A general Text Mining scheme founds its functioning on Document, Corpus, Rules and Annotators. A Document is defined as a single aggregation of text coming from a same source, a Corpus is all Documents, that is, a Document Clustering, coming from one or more web sources, the Rules are the most important part of Text Mining and their definition from the programmer will allow to get or not a good result of the whole Web Text Mining process. Before applying the

rules, Text Mining software handles the phase of Information Extraction effecting a tokenization and lemmatization of the text that it will allow an accurate analysis of the document during the Text Mining. Tokenization is a in which a text is split into a stream of words by removing all punctuation marks and by replacing tabs and other non-text characters by single white spaces. This tokenized representation is then used for further processing. The set of different words obtained by merging all text documents of a collection is called the dictionary of a document collection. Lemmatization try to map verb forms to the infinite tense and nouns to the singular form. In order to achieve this, the word form has to be known, that is. the part of speech of every word in the text document has to be assigned[17].

Subsequently to the Information Extraction, the application of one or more rules on a Document or on a Corpus will provide the extraction of "hidden information" that will be made available to the system in XML format, the useful format to fourth step. The rules can interact between them to create a more general rule. The accuracy in the discovery of intrinsic information of the document, is better if more rules are defined. But it is important to not define a lot of rules for avoiding the problem of "too much accurate", because in this case the process results could not useful. The discovery information in this phase is momentarily memorized in the Annotators which have the assignment of memorize all useful information in a Document or in a Corpus. The further analysis on the Annotators, that is to say, elimination of the duplicates, elimination of the errors, etc. are the start of the discovery of the real information called Skill. The Skills are keywords which allows to the system to have a general vision of the semantic content of the Document or the Corpus [18].

Fig.4 shown as the application of Rules on the Keywords (nodes) allows the individualization of Annotators.

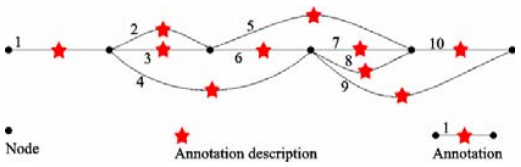

Fig. 4 Rules and Annotators scheme

The fourth is last step of the process; it foresees the phase of presentation of the obtained results. In our architecture, the information drawn out during the execution of the third step is stored in a second repository. This repository created during this phase will follow the rules of the Web Warehousing according to the Whoweda schemas. The repository is composed by URLs and nodes. URLs are the reference URLs of the document, that is to say, “referenced from" and "it references who" whereas the nodes are the skills found in the third step. To fill this repository is necessary to effect a parsing of the result in XML supplied by software during the third phase for recover the necessary information. Use of Whoweda schemas is fundamental in the implementation of Web Warehouse because our system is directed to the execution of Data Mining on the data produced by the whole Web Text Mining process. Following Whoweda schemas is possible realize more views and more schemas on the whole Web Warehouse. This allows fast interrogation of Web Warehouse in the Data Mining phase [19]. The assignment to effect Data Mining and therefore to define views and schemas for the interrogation is submitted to the opportune programming of a Miner as discussed in the next section.

The whole process of Web Text Mining has been conceived and projected according to the Service Oriented Architecture (SOA) as discussed in the next section.

IV. The Web Text Mining Architecture based on Service Oriented Architecture

The Service-Oriented Architecture (SOA) approach is the latest of a long series of attempts in software engineering trying to encourage the reuse of software components; the SOA provides a scheme to design distributed and orchestrated applications. The SOA is a strategy and a flexible and scalable technical framework for supporting open standards and open application interfaces. Open applications that are built on an open architecture promise to be significantly less expensive to implement and to operate.

A service in SOA is an application or function with well-defined interfaces that is packaged as a reusable component for use in a business process. Simply stated, in an SOA, business processes appear as a set of separate components that can be joined and choreographed to create composite applications and processes and so there are following benefits:

- Flexibility, a service can be located on any server, and relocated as necessary, as long as it maintains its registry entry, prospective clients will be able to find it;
- Scalability, services can be added and removed as demand varies;
- Replaceability, provided that the original interfaces are preserved, a new or updated implementation of a service can be introduced, and outdated implementations can be retired, without disruption to users [20].

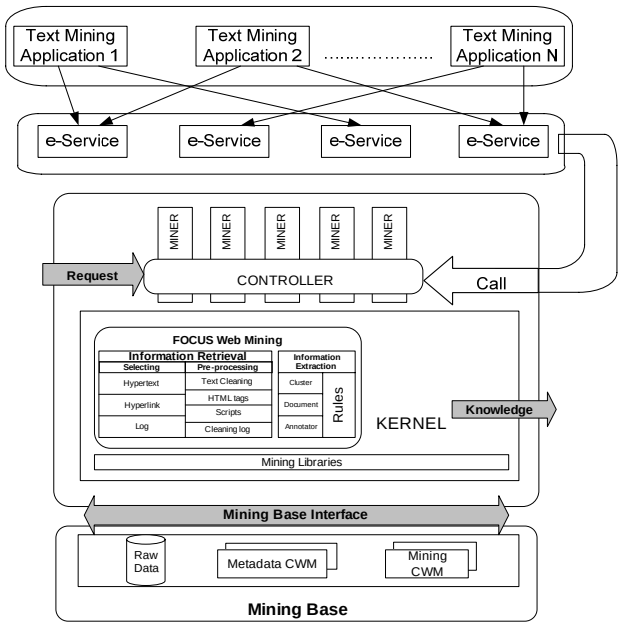

Fig. 5 The Flexible Web Text Mining Architecture

As shown in Fig. 5 the specialized architecture is composed by the Web Mining functional component which effects the information retrieval from the WWW and extracts intrinsic information from a textual analysis of the content of documents coming from web through a phase of Text Content Mining [21].

The Controller handles the process flow, calling the appropriate Miners and determining the next steps to be complete. The Kernel is the core of our architecture where the new mining models has built. Finally, the Mining Base is the knowledge repository of the whole architecture and its functions are those of repository for raw data, knowledge metadata and mining mode [5,6].

In Fig. 5, we have shown a flexible architecture that modularizes the full lifecycle of web text mining process into reusable components in order to create the interaction among these different components of a mining architecture and the assembling of their logic into a manageable application, these components are called “Miners” which play an important role in our architecture for flexibility. The Miners are building blocks that can be used to build a complex application based on e-services. Every Miners has a well defined mining task.. The system realizes a decoupling between the services and the knowledge discovery process in text application adopting the miners reusable components [1,2,3,4,5,6]. In this sense each of four steps of the web text mining process corresponds to an e-service and every e-service can be defined in terms of these elementary building blocks, as shown in Fig. 6.

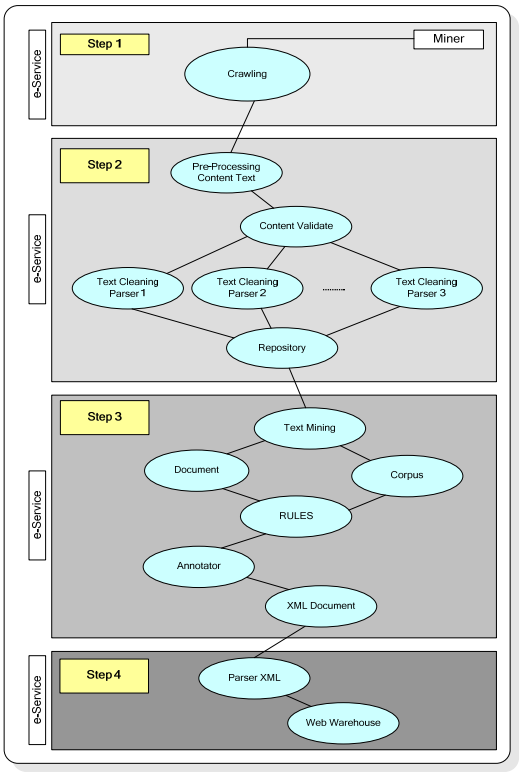

Fig. 6 E-service Workflow in Web Text Mining Process

Miner Orchestration is defined as the process able to create a workflow of building blocks and in this sense Miner Orchestration provides more complex functionality in the Knowledge Discovery Process. Miner Orchestration confers dynamic, flexible, and adaptable to meet the changing needs of a organization [5,6].

V. Experimental Results

A Web Text Mining prototype has been developed to prove the proposed architecture. It is decided to experiment studies through the realization of a prototype for evidencing that all the phases of knowledge extraction process in text, starting from not-structured information, like those present in the web, can concretely be applied to various issues. The prototype has been realized to extract useful information from job advert description through a Text Mining phase. It extracts both *places* and *skills* respect to the considered job adverts according to the scheme shown in Fig. 7.

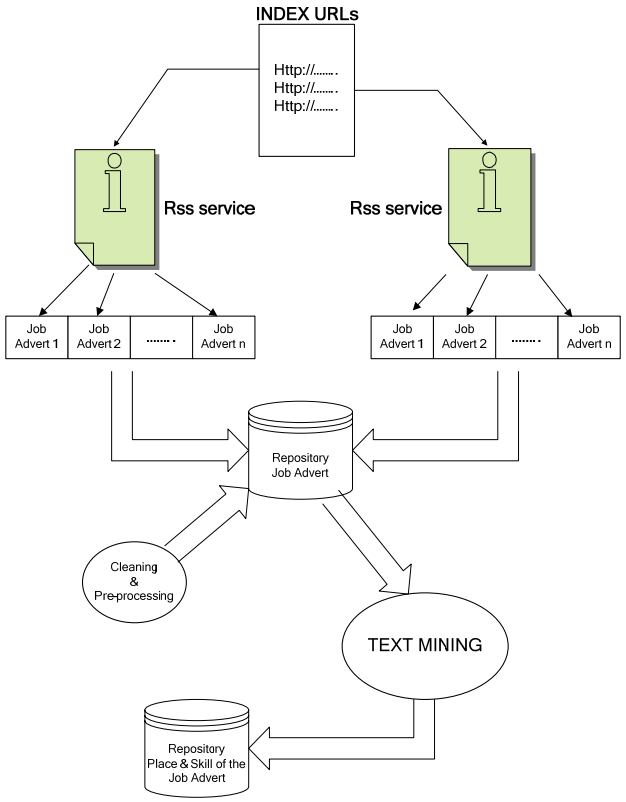

Fig. 7 Prototype Model

The process is composed by following phases: Crawling, Pre-processing, Text Mining and Results Presentation.

The Crawling consists in recovery of the job adverts from Web. In this phase, the prototype reads the content of XML configuration file in which there are URLs corresponding to RSS services of job adverts. The considered example of XML file, called SystemInfo.xml is shown in Fig. 8.

```
<SYSTEMS>
  <SUBSYSTEM URL="http://www.cercolavoro.com/rss/ciao.xml" name="OfferteLavoro" />
  <SUBSYSTEM URL="http://www.cercolavoro.com/rss/offerte+lavoro+rss.jsp" name="OfferteLavoro"/>
  <SUBSYSTEM URL="http://milano.bakeca.it/rss20-offerte-di-lavoro.xml" name="OfferteLavoro"/>
  <SUBSYSTEM URL="C:/RSS/offerte-lavoro-rss-test-URLNULL.xml" name="OfferteLavoro" />
  .
  .
  .
</SYSTEMS>
```

Fig. 8 SystemInfo.xml file

When URL is read, the prototype enters to Internet to recover the RSS file that contains job adverts. The recovered RSS file is submitted to a validation that consists to verify if the file is well formed. If the file is good the useful information for the Text Mining phase is extracted through a parsing of the file, otherwise the file is discarded. The prototype is based on a multithread algorithm that allows parallel appeals. This allows to analyze more pages to optimize the run-time of the whole process of extraction.

The Fig. 9 shows as in information extraction in Crawling phase there are special characters as &lt, &gt which must be replaced with the corresponding characters.

Description
LA DITTA MELANU' SRL (FABBRICA DI LIQUORI) CON SEDE IN SALERNO ·
Cerchiamo: ·Agente di commercio ·
L'Istituto Italiano di Scienze Umane indice una selezione per curriculum e colloquio al fine
Work & Trade seleziona Agenti, Consulenti Commerciali, Venditori, Distribu
Work & Trade seleziona addetti di telemarketing per le sedi di Casoria (N
Montini Agenti, Consulenti Commerciali, Venditori, Distributori e Promotori per le

Fig. 9 Text before Pre-processing phase: to be note the presence of useless special characters

The Pre-processing consists in the preparation of the information extracted by RSS file for the next Text Mining phase. In this phase information come "cleaning up" of the whole useless content. Pre-processing is important to increase the effectiveness and the efficiency of Text Mining because the presence of useless content can bring to altered results. In the prototype the Pre-processing phase is able to eliminate and to replace characters which create problems in Text Mining phase. The Fig. 10 shows the text after Pre-processing phase.

Description
LA DITTA MELANU' SRL (FABBRICA DI LIQUORI) CON SEDE IN SALERNO ·CERCA
Cerchiamo:··Agente di commercio/··Requisiti del candidato:/·· Spiccate doti relazionali e
L'Istituto Italiano di Scienze Umane indice una selezione per curriculum e colloquio al fine
Work & Trade seleziona/ Agenti, Consulenti Commerciali, Venditori, Distributori e Pro
Work & Trade seleziona/ addetti di telemarketing per le sedi di Casoria (Na), Modug
Montini/ Agenti, Consulenti Commerciali, Venditori, Distributori e Promotori per le divisioni I

Fig. 10 Text after Pre-processing phase

The result of Pre-processing phase is stored in a database.

The Text Mining phase, the prototype accedes to the database and draws out the URL and the description of job adverts. An analysis is done on description by a tool called GATE through API Java. The text contained in the description is identified as Document inside a Corpus. On the Document three operations are executed: tokenization and lemmatization, which are descript in second section, and mining rules application. To have a good tokenization and lemmatization it is important to get a clean and correct text by pre-processing.

After tokenization and lemmatization, the application of mining rules to the Document is executed. It is very important to apply the rules in the correct sequence that is suitable in external file called main.jape as shown in Fig. 11. Changing the rules sequence it is gotten different result. The rules are based on keywords lists referred to precise universe, like the keywords city list contains only city names. The rule execution consists in a matching between the tokenized Document and the rule reference lists. If the token is present in more lists, the rule analyzes the lemmatized text and decides to list the token belongs. This decision is taken on percentage.

```
MultiPhase:    TestTheGrammars
Phases:
city
organization
jobtitles
province
skill
```

Fig. 11 Main.jape file

The rules are stored in external files with jape extension like city.jape shown in Fig. 12.

```
Phase: City
Input: Token Lookup
Options: control = appelt

Rule: City

(
 {Lookup.majorType == citta}
):city
-->
{
  gate.AnnotationSet city = (gate.AnnotationSet)bindings.get("city");
  gate.Annotation cityAnn = (gate.Annotation)city.iterator().next();
  gate.FeatureMap features = Factory.newFeatureMap();
  features.put("rule", "City");
  outputAS.add(city.firstNode(), city.lastNode(), "City", features);
}
```

Fig. 12 City.jape file: to be note as the file is divided in two parts until arrow is GATE code, after arrow is Java code

This rule, called City, individualizes cities which are contained in a text through a matching on all tokens of the same document through the list. This list is called citta.

The matching result is stored in Annotators in Document in which there are information on the individualized tokens.

The fourth phase of the process is Results presentation, in this phase a XML file is created through the information stored in the Annotators, it is used to fill a database through a file parsing. The database contains URLs, places, skills, jobtitle.

We bring back experimental results obtained from a job adverts site that we use like RSS Web service source.

Number of job adverts made available by the service: 100.

Number of job adverts taut out by prototype in crawling phase: 100.

Number of job adverts on which the rules of Text Mining have been applied with happened: 82.

Number of job adverts in which the job center has been individualized: 63 (55 municipalities and 8 regions).

Number of job adverts in which skills has been individualized: 65.

Crawling phase has recovered the 100% of Job adverts. Text Mining phase applies same rules on all adverts having 18% loss limited on the total of the database, because, job adverts do not contain searched information or because, limited number of rules defined in this prototype are inadequate. On 39 job adverts it has been possible to characterize both searched information, on 26 only job center, on 17 only skills. As shown in Fig. 13, errors are minimal percentage and, in particular, there is a skill, called Perito, which is confuses with a city.

Link	City
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Milano
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	San Gimignano
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Cesena
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Genova
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Roma
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Como
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Bologna
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Catania
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Como
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Perito
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Varese
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Noto
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Roma
http://www.cercolavoro.com/offerta_personalizzata.jsp?ID...	Roma

Fig. 13 Results

VI. Conclusion

In this paper has been presented a Flexible Web Text Mining Architecture. To this purpose, the guide lines of the SOA and the Miners model have been considered as a way to manage a workflow of reusable building blocks with well defined tasks and able to interoperate with one another for the creation of new services. The main advantages offered by this architecture are the quick designing of a process according to one's own business needs and the creation of new flexible services without resorting to substantial changes.

Acknowledgment

The authors acknowledge the financial support provided by the Italian Ministry of Education, University and Research which has made possible the realization of this work as result of our research activities.

References

[1] M. Castellano, G. Mastronardi, A. Aprile, G. Bellone de Grecis, F. Fiorino, "Applying a Flexible Mining Architecture to Intrusion detection", ARES 2007, *Second International workshop Data Warehousing and Data Mining*, DAWAM 2007, Vienna, April, 2007.

[2] M. Castellano, N. Pastore, F. Arcieri, V. Summo, and G. Bellone de Grecis, "A Knowledge Center for a Social and Economic Growth of the Territory", *IEEE Computer Society Press*, International Conference On System Sciences, Big Island Hawaii, 3-6 January 2005.

[3] M.Castellano, N. Pastore, F. Arcieri, V. Summo, and G. Bellone de Grecis, "An e-Government Cooperative Framework for Government Agencies", *IEEE Computer Society Press,* International Conference On System Sciences, Big Island Hawaii, 3-6 January 2005.

[4] M.Castellano, N.Pastore, F.Arcieri, V. Summo, and G. Bellone de Grecis, "A Flexible Mining Architecture for Providing New E-Knowledge Services", *IEEE Computer Society Press,* International Conference On System Sciences, Big Island Hawaii, 3-6 January 2005.

[5] M. Castellano, N. Pastore, F. Arcieri, V. Summo, and G. Bellone de Grecis, "Orchestrating Knowledge Discovery Process", *E-Service Intelligence: Methodologies, Technologies and Application*, Springer, pp 447-496.

[6] M. Castellano, F. Fiorino, F. Arcieri, V. Summo, and G. Bellone de Grecis, "A Web Mining Process for e-Knowledge Service", *E-Service Intelligence: Methodologies, Technologies and Application*, Springer, pp 447-496. A Web Mining.
[7] W. Lee, SJ. Stolfo, KW. Mok, "Data Mining Approaches for Intrusion Detection", *Proceeding of the 7th USENIX Security Symposium*, 1998.
[8] W. Zhong, X. Tang, "Web Text Mining on XSSC" Institute of System Science, Academy of Mathematics and System Science.
[9] Knowledge Discovery for Text, RGU: school of Computing, California.
[10] A.H. Tan, *Text Mining: The State of the Art and the Challenges*, in PAKDD99 Whorkshop on Knowledge Discovery from advanced Databases, Beijing, China, April 1999.
[11] Nahm U.Y. e Mooney R.J., ***Using Information Extraction to Aid the Discovery of Prediction Rules from Text***, in KDD2000 Workshop on Text Mining, Boston, Massachusetts, USA, August 2000.
[12] B. Mobasher, R. Cooley, and J. Srivastava: *Creating Adaptive Web Sites Through Usage-Based Clustering of URLs*(1999), In Proc. of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999.
[13] R. Kimball and R. Merz: "The Data Webhouse Toolkit, Building the Web-Enabled Data Warehouse", John Wiley & Sons, January 2000.
[14] Cooley, R. et al, "Web Mining: Information and Pattern Discovery on the World Wide Web", *In Proceeding of IEEE International Conference Tools* with AI. Newport Beach, California, USA, pp. 558-567, (1997).
[15] Etzioni, O., "The World Wide Web: Quagmire or GoldMine", Communication of the ACM, Vol. 39, No. 11, pp. 65-68, (1996).
[16] Chakrabarti, S. et al, Focused Crawling, "A New Approach to Topic-Specific Web Resource Discovery", *In Proceeding on the 8th International Word Wide Web Conference,*. Toronto, Canada, pp. 1623-1640, (1999).
[17] A. Hotho, A. Numberger, G. Paab, *A brief Survey of Text Mining*, University of Kassel, School of Computer Science, Knowledge Discovery Group, 13 May, 2005.
[18] GATE – General Architetcture for Text Engineering, http://gate.ac.uk/
[19] Saurav S. Bhowmick, Wee Keong Mg, Sanjay Madria, "Web Schemas in WHOWEDA", Data Warehousing and OLAP. McLean, Virginia, United States. Year 2000, pp. 17 – 24, ISBN:1-58113-323-5.
[20] Service Oriented Architecture, SOA, White Paper.
[21] Brin, S. and Page, L., "The Anatomy of a Large Scale Hypertextual Web Search Engine", *In Proceeding of the 7th International World Wide Web Conference*, Brisbane, Australia, pp. 107-117, (1998).
[22] MARINHO, Leandro Balby Marinho, Girardi Rosario: "Mineração da Web", *Revista Eletrônica de Iniciação Cientfica*, São Luiz, Jun. 2003.
[23] Adriana Marotta, Regina Motz, Raul Ruggia, "Managing Source Schema Evolution in Web Warehouses", Instituto de Computación, Facultad de Ingeniería Universidad de la República. Montevideo, Uruguay, 2001.
[24] O.Etzioni, "The world wide web: Quagmire or gold mine", *Comm.of the ACM*,39(11):6568,1996.

Marcello Castellano was born in 1961. He received "Laurea cum Laude" in Computer Science in 1985 from University of Bari (Italy). Currently he is Assistante Professor at the Department of Electrical and Electronic Engineering of the Polytechnic of Bari, Italy. Previously, he has been staff member researcher at National Institute of Nuclear Physics, and computer specialist at Italian National Council of Researches. He received a scientific associate contract from Center European of Nuclear Researcher and Visiting Researcher at New Mexico State University and Gran Sasso International Laboratory (Italy). He serves as reviewer in several scientific international journals and conferences. Dr Castellano's main research interests are in machine learning, data analysis and mining. In these fields, he authored high-quality scientific papers in international journals.