Process of Retrieval of Knowledge Starting From Data Texts

Mouhcine El Hassani[1], Belaid Bouikhalene[2], Mohamed Naimi[3]

[1](Laboratory of Flows and Transfers Modeling (LAMET), Graduate Student, Sultan Moulay Slimane university, Morocco)

[2](Laboratory of Mathematics Innovation and Information Technology (LIMATI), Phd Professor, Sultan Moulay Slimane university, Morocco)

[3](Laboratory of Flows and Transfers Modeling (LAMET), Phd Professor, Sultan Moulay Slimane university, Morocco)

Corresponding Author: Mouhcine El Hassani

Abstract: *Nowadays, large importance is given to publications of numerical information via the Net, witch necessitates a Knowledge Extraction System from the Text (KEST).*

The basic principle consist to identify the needs of a person to do, a search in a text, which can be not structured, to recover the terms and the information related to the research subject and to structure them as classes of useful information.

Keywords: *Big data, Data mining, data warehouse, Extraction of knowledge, Text mining.*

Date of Submission: 20-08-2018 Date of acceptance: 03-09-2018

I. Introduction

Currently, given being the role, how important is, of New Information and Communication Technologies (NICT), leading to a considerable information number, available in numerical forms, such as electronic reviews, books published on Internet, structured information disseminated in social networks like Facebook, Twitter, and many others. It is necessary to extract the relevant and reliable ones. Therefore, it appears necessary to preconize a credible and performing system dealing with all textual information, in order to deduce structured and useful knowledge.

The basic principle of textual searching is to organize information in entities and classes of words, while maintaining the associations between these classes and the mutual interactions of these objects.

Thus, all new found information will enrich a structured database represented in tables depending on each type of data.

II. Text Mining [1]

There are many technical definitions of text mining both in Internet and textbooks, but the definition adopted in the current paper is the whole of process of searching models related to the artificial intelligence allowing to find rules of association starting from unstructured text. There exist several methods which are based on sorting, regrouping (starting from requests SQL: Structured Query Language) of words and counting the number of repeated words to identify their importance.

In general, a process of text extraction (text mining) passes by four stages:

1. The first one consists in preparing the data for treatment by transforming the raw data format to another format in order to subject them to adequate operations.
2. The second is the search of the frequent reasons in the extracted text and the extraction rules of association.
3. The Third aims at presenting the data in visual form using graphs or diagrams. In this part, a data-processing tool like 2D or 3D visualization software would be necessary for recognizing relevant and useful information.
4. The fourth consists of applying both cleaning and optimization operations to reduce the size of found information.

The technics of data processing do not only depend on these four stages but also on the location of the information being exploited and, particularly on the algorithms and methods used.

III. Information Extraction

The operation of extraction of the information is the first most significant stage in the pretreatment of the text.

One extracts information [2] to find a text structured in natural language. The DARPA (Defense Advanced Research Projects Agency) initiated and financed a series of specific annual workshops, called "Message Understanding Conferences" (MUC), which assembled researchers that focus on the methods of extracting unstructured text. For each MUC, participating groups were given sample messages and instructions on the type of information to be extracted, and developed a system to process such messages. The performance of each participant's system was evaluated against its competitors. The data to be extracted were presented in the form of a model specifying a list of sites of under-chains extracted from the document.

In general, the information to be extracted is defined by a model representing a list of exits to be filled, although sometimes it is represented by annotations in the file. The fillers of vacuum can be a group of specified values or of chains instantaneously extracted from the document.

The following example visualizes the constraints related to the pretreatment of the text:

Annonce : Besoin, de Développeur Web
Emplacement: Rabat . Maroc
Cette personne est responsable de la conception et de la mise en œuvre des composants d'interface Web du serveur ABC et des tâches générales de développement de l'arrière-plan.

Un candidat retenu doit posséder une expérience qui comprend :

- Un ou plusieurs des éléments suivants : Solaris, Linux, Windows / NT
- Programmation en C / C ++, Java
- Accès à la base de données et intégration : Oracle, ODBC
- CGI et scripting: un ou plusieurs de JavaScript,
- Perl, PHP, ASP
- L'exposition à ce qui suit est un plus: JDBC, FrontPage et / ou Fusion de cuivre.

Une expérience de 2 ans (ou équivalent) est nécessaire.
Modèle rempli
• Réf : \ Développeur Web "
• Lieu : Rabat. Maroc
\ Langages : \ C / C ++ ", \ Java", \ JavaScript ", \ Perl", \ PHP ", \ ASP"
• plates-formes : \ Solaris ", \ Linux", \ Windows / NT "
• applications : \ Oracle ", \ ODBC", \ JDBC ", \ FrontPage", \ Cold Fusion "
• zones: \ Database ", \ CGI", \ scripting "
• degré requis: expert
• années d'expérience: \ 2+ ans "

Figure 1: Sample text and template completed for job posting

Figure 1 displays a model of simplified document, which is recovered from an operation of information extraction obtained from a social network, which publishes job offers. The text in this example contains only sectors charged by chains extracted directly from the document. Several sectors can have multiple fillers for the use of the advertisement of activities as languages (of programming), environment of development, applications and fields. Text mining may yield interesting and important inputs for predictive modeling and can provide insightful recommendation for decision makers. Practical applications has been shown in analyzing hotel reviews, banner pages of courses, conference advertisements , job offers, job advertisements ,company newspapers, and many more. Furthermore, the extraction of knowledge starting from text is still an adequate method to automate the update of the static and dynamic webpages. Technics of the artificial intelligence using automatic learning were presented to extract information from files and texts documents in order to easily generate databases starting from information, which makes the text hence more accessible in line. For example, the information extracted from the working stations on the Web can be used to build a consultable database, for the inventory of computer equipment, in order to define its needs clearly.

IV. Outline On Text Mining (Excavation Of Text)

The exploration of rough information estimates that information "to be left" is already in a local database. By misfortune, for several applications, information in numerical form is available only in the shape of files in free natural language rather than well-organized databases. Since the extraction of information treats the difficulty in changing a set of textual files into a more structured database, the extracted data built by a module Text Mining can facilitate the KDD process (Knowledge Discovery in Databases). Artificial Intelligence also supports KDD for more pushed use of extracting knowledge from data as illustrated in Figure 2. The extraction of information can play an obvious role in text mining.

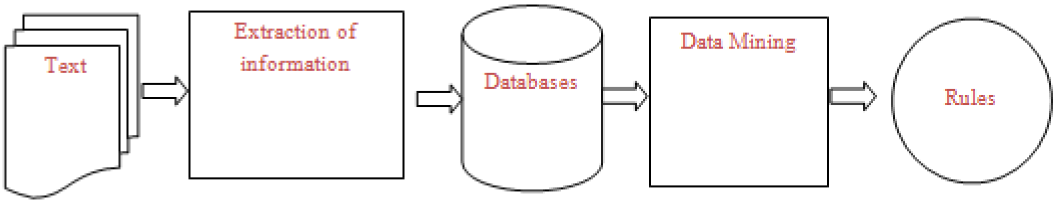

Figure 2: Example of Text Mining process

Admittedly, the construction of a Text Mining system is a difficult operation, however, there were new significant improvements in the processes use of machine learning to help to automate the realization of systems Text Mining. By manually handling a small number of documents with the goal of extracting data, Text Mining systems system can be useful, for long text, for building a database. Nevertheless, the exactitude of the current Text Mining systems is limited and, consequently, a database deduced automatically will comprise surely a significant number of faults. However, the question that may be of interest to us is whether the knowledge acquired from this database is clearly less credible than in the case of a cleaner database. This article presents examples showing that the rules noticed, starting from a database automatically extracted, are an inaccuracy close to that discovered starting from a manually built database.

V. Process Of Extracting Information

Several originators set out the process of extraction of information in stages of distinct granularity, using specific systems of extraction created for this purpose. The analysis of the various approaches, used for information extraction, helps to identify six main stages of this process:

- Pretreatment.
- The discovery of the proper names.
- Syntactic analysis.
- The extraction of the events and the relations.
- Resolution of the anaphora (the anaphora (female substantive) is a stylistic device which consists in beginning worms, sentences ,sets of sentences or worms by the same word or the same syntagm.)
- Production of operating results.

1. Pretreatment:

In this stage, we divide the document text into several portions that constitute phases, segments, empty zones, etc. This operation can be carried out by various components related to the programming language like the string Tokenizer of java, the splitters, the segmenters, etc. Note that the tokenization consists of dividing the text into several Tokenas delimited by a character or spaces preset in advance. This technic is effective for most of texts written by using various languages, except for Chinese, Japanese, and other specific languages.

The phase, following the treatment is the lexical and morphological analysis of the text. It consists in locating the words and sentences representing the exceptions and ambiguities, and then specialized dictionaries of several languages are used to resolve these constraints. These dictionaries can gather the names of countries, cities of the scientific terms, etc. A simple example consists in progressively locating the words of a text editor, like Word, with the seizure, and to give suggestions of the faults made during the seizure. Users can always introduce new terms and enrich the dictionary.

2. The discovery of the proper names:

It is one of the most significant tasks in the process of extraction of information. It makes it possible to locate the whole of classes and of entities representing the proper names, such as people, companies, monuments, countries. This information can be easily identified, since they are written in the form of text and even by the availability of the tools of control of the programming languages "the regular expressions of java".

3. Syntactic analysis:

During this stage, a syntactic analysis of the sentences in the documents is carried out. After having identified the basic entities and classes in the preceding phase, the sentences are analyzed to identify the group of names of some of these entities and the groups of verbs. In this stage of analysis, the work is prepared for the next stage of extraction of the events and the relations in which they collaborate. The groups of nouns and verbs are used like sections to start to work at the stage of correspondence of reason. The identification of these groups is performed by the application of a set of especially built regular expressions. However, the complete analysis is not an easy task; it thus requires expensive calculations, which, slow down all the process of information extraction. Since we deal with such a difficult problem, the complete analysis is likely to introduce errors. On the other hand, sometimes, the total syntactic analysis is not necessary. So groups of research on the extraction of information tend to use what is called the partial or surface analysis instead of integrality. By using only local information, they observed that the not very deep analysis creates partial syntactic fragments, which overlap only with higher degree of confidence. Following the application of the partial syntactic analysis, they showed a better performance than that of the sites, which tried to create complete syntactic structures [4].

4. Extraction of events and relations:

This process is carried out by the creation and the application of rules of extraction, which specify different reasons. The text is adjusted with these reasons and, if a correspondence is found, the element of the text is labelled and extracted afterwards. The formalism of this writing differs from a system of extraction of information to another [5].

5. Resolution of the ankaphora:

Any class given in a text can be revoked on several occasions and each time that could be returned differently. In order to recognize all the manners used to give a name to this entity, one carries out a documented resolution of reference. There are several types of correlation, but the best current types are the proper names and the pronominal nomenclature, when a name is replaced in the first case by a pronoun and by another name or a noun phrase in the second [6].

6. Production of operating results:

This phase contains the modification of the structures, which were arisen during preceding operations in the models of exit according to formats specified by a process. It can include different operations of standardization for the dates, the hour, the currencies, etc. For example, a method of district for the percentages can be carried out and the number of surface 60.96 will be entirety converted 61.

Because a particular process of information extraction cannot have all the possible components, all the operations should not obligatorily be completed in just one extraction of information. According to Appelt and Israel (MUC-5 1993), there are several factors affecting the choice of the components of the systems, such as:

- the language: as previously said for the text processing in the case of Chinese or Japanese. These languages do not comprise clear words and limits of sentences. Texts in German language contain words of difficult morphological structures; witch requires a comparison to English documents in such cases.
- the kind of text and properties: in the transcriptions of abstract speeches, for example, misspellings can occur in addition to the implicit limits of the sentence. If the information must be extracted from these texts, these questions must be taken into account and be addressed when designing system by adding the corresponding modules.
- the process of extraction: for the recognition of the names, the modules of analysis and resolution of anaphora cannot be necessary.

VI. Evaluation Of The Extraction Of Information

By taking into account the text entry or a block of texts, the awaited exit of a system of extraction of information can be defined in a precise way. To facilitate the evaluation of various systems and approaches of the extraction of information the parameters of precision and recall were adopted by "the IR research community" in this respect. They measure the system effectiveness from the point of view of the user, i.e. the extent to which the system produces all the suitable output (recall) and only the suitable exit (precision). Thus, the recall and the precision can be regarded respectively as the measurement of exhaustiveness and exactitude. To define them in a formal way, this one allots to # key the total number of slots, which must be filled according to an annotated corpus of reference, representing a degree of reliability or a Gold Standard, and # correct (# incorrect) the number of slots correctly filled (incorrectly) the answer of the system. A slit known as is filled correctly if it is not aligned with a slit in the Gold Standard (slit parasitic) or if a invalid value were allotted to it. Then, the precision and the recall can be defined as follows:

$$Précision = \frac{\#correct}{\#correct + \#incorrect}$$ (equation 1)

$$recall = \frac{\#correct}{\#key}$$ (equation 2)

In order to obtain a finer image of the performance of information extraction system, the precision and the recall are often measured separately for each type of site.

Measurement F is used as balanced harmonic mean of precision and recall, which is defined as follows:

$$F = \frac{(\beta^2+1) * Précision * recall}{(\beta^2 * Précision) + recall}$$ (equation 3)

In the definition above, β is a nonnegative value, used to adjust their relative weighting

(when $\beta^2 = 1.0$ the recall and the precision are considered to have the same weight, for lower values of β more weight is given to the precision).

Other parameters are also used in the literature, for example the error rate of slit, SER [7] [8], which is given by:

$$SER = \frac{\#incorrect + \#missing}{\#key}$$ (equation 4)

Where $\#missing$ indicates the number of sites in the reference, which are not aligned with any site in the response of the system. It reflects the relationship between the total number of erroneous slot and the total number of slots in the reference. According to particular cases, unquestionable needs standard for errors (for example of the parasitic slits) can be balanced in order to consider them more or less significant than others.

VII. Conclusion

In this paper, it is shown that text-mining systems can be developed relatively rapidly and evaluated easily on existing Information Extraction (IE) corpora by using existing information extraction and data mining technology. General steps are presented for better information extraction performance, shedding light on key elements that should be considered when extracting data information, particularly from unstructured texts. Also, Factors, such as accuracy, recall, F-measurement and the error rate of the slot, are used to better improve the evaluation in the extraction of information.

.

References

[1]. Un Yong Nahm and Raymond J. Mooney Department of Computer Sciences, University of Texas, Austin, TX 78712-1188 fpebronia,mooneyg@cs.utexas.edu , *" Text Mining with Information Extraction ",p 2*

[2]. Un Yong Nahm 2004 ,*Doctoral Dissertation :Text mining with information extraction*,The University of Texas at Austin ©2004 ISBN:0-496-01283-5

[3]. Lehnert, W., Cardie, C., Fisher, D., Riloff, E., & Williams,R. 1991 a. University of Massachusetts, *Description of the CIRCUS ,System as Used for MUC-3. Proceedings, Third Message Understanding Conference (MUC-3). San Diego, CA, Morgan Kaufrnann, pp. 223-233.*

[4]. Ralph Grishman,1997. *Information Extraction: Techniques and Challenges* ,Computer Science Department, New York University New York NY10003 U.S.A pp. 7-9.

[5]. Alexandre Saidi , *Textual Information Extraction Using Structure Induction*, LIRIS-CNRS (UMR 5205)Ecole Centrale de Lyon, Mathematics and Computer Science Department,p 4-9

[6]. Ronen Feldman, *THE TEXT MINING HANDBOOK Advanced Approaches in Analyzing Unstructured Data*,

[7]. Thierry Poibeau,Horacio Saggion,Jakub Piskorski,Roman Yangarber, *Multi-source, Multilingual Information Extraction and Summarization, p 27-28*

[8]. Gokhan Tur,Renato De Mori, *Spoken Language Understanding: Systems for Extracting Semantic Information from speetch*, John Wiley & Sons,Ltd 2011

* Mouhcine El Hassani. Process of Retrieval of Knowledge Starting From Data Texts ." IOSR Journal of Computer Engineering (IOSR-JCE) 20.4 (2018): 33-37.