Automatic Content Extraction on the Web with Intelligent Algorithms

Pablo Cababie, Alvaro Zweig, Gabriel Barrera and Daniela Lopéz De Luise

Abstract-**Since the INTERNET outburst, consumer perception turned into a complex issue to be measured. Non-traditional advertising methods and new product exhibition alternatives emerged. Forums and review sites allow end users to suggest, recommend or rate products according to their experiences. This gave raise to the study of such data collections. After analyze, store and process them properly, they are used to make reports used to assist in middle to high staff decision making. This research aims to implement concepts and approaches of artificial intelligence to this area. The framework proposed here (named GDARIM), is able to be parameterized and handled to other similar problems in different fields. To do that it first performs deep problem analysis to determine the specific domain variables and attributes. Then, it implements specific functionality for the current data collection and available storage. Next, data is analyzed and processed, using Genetic Algorithms to retro feed the keywords initially loaded. Finally, properly reports of the results are displayed to stakeholders.**

Index Terms: Opinion mining, crawling, Genetic Algorithms

I. Introduction

The paper represents the result of research carried out in the ITLab University of Palermo. Within this context, emerged the proposal for a system to collect and process information in a particular topic and to show the results in report form for analysis and decision making process.
The problem arises from the need for a pharmaceutical company to obtain the perceptions of consumers available on the web about their products and competitors. Opinions are subjective expressions that reflect the perceptions or feelings of people about events or entities. When someone needs to make a decision, one factor that can cooperate to take it wisely is the opinion of others. Before the explosion of the Web, when an individual needed to make a decision, he used to consult with the family and alleges. When a company needed to know the opinion of the general public about a product or service, it used to send polls and interest groups.
With the emergence of the Web, information started to appear online and available to everybody in public forums, discussion groups, or bogs. This new sites are defined as part of the concept called Web 2.0 user-generated content. The world wide web , having over 350 million pages, continues to grow rapidly at a million pages per day [1]. About 600 GB of text changes every month [2]
This available information becomes an essential tool for decision-making process based on a new paradigm called "crowdsourcing". [3]
Thus, these opinions and debates on the Web become highly relevant for companies or people to make decisions. However, it must be pointed that these views are sometimes not as easily identifiable and are hidden in different users' personal pages or forums. Therefore, the main challenge of the project is based on the collection, identification, processing and reporting of results of this "crawling". Throughout the work, covering topics such as background and previous and related research of the topic, the structure of the system, the process and analysis of the collected information and finally conclude with the future work proposed.

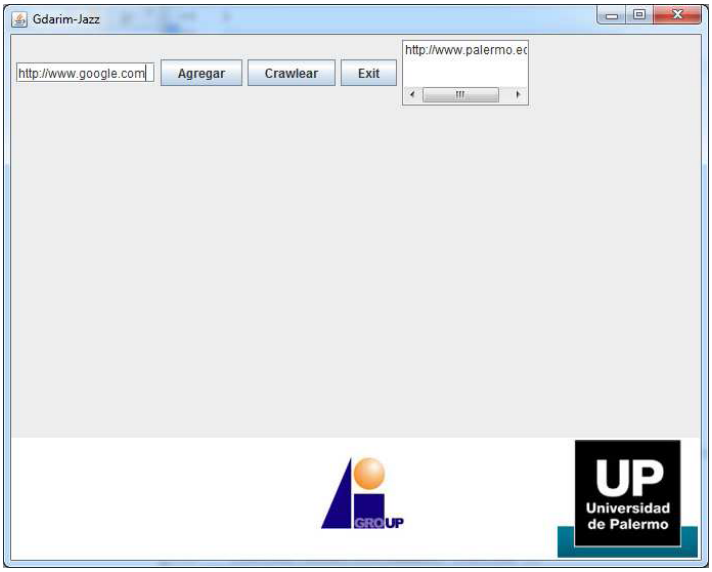

FIGURE 1 – Gdarim application screenshot

Manuscript received Aug 19, 2010. This work was supported in part by the University of Palermo in Buenos Aires.

Pablo Cababie is a systems engineer. He works as data manager in the petroleum industry and as a researcher since 2005 specialized in Genetic Algorithms implementations (phone: +5411 5199- 4520; fax: +5411 4963-1560; e-mail: pcabab@palermo.edu).

Alvaro Zweig, is an advanced student at University of Palermo in Buenos Aires. He is now a free lance developer for Java, and .Net technologies (e-mail: alvarozweig@gmail.com).

Gabriel Barrera is a systems engineer and coordinator for research lines at University of Palermo (e-mail: gmbarrera@gmail.com).

Daniela Lopez de Luise is a Ph.D in Computer Science at the University of La Plata. She is the research team manager at the University of Palermo.

II. BACKGROUND AND RELATED RESEARCH

There are research papers relating to opinion mining. Most of them approach the problem from the point of view of semantic interpretation of the wording found on the Web. [4] [5] [6] For example, crawl a forum and analyzing the text for product reviews, recommendations and complaints.[7] [8] [9] Several papers focuses in the identification of opinion. If it is critical, comparation, complains, praise or sites directly sense devoted to the exchange of views. [5] [10] Also, there are studies that are based on a list of words classified as good or bad, used to catalog the mention of the view as positive, negative or neutral. The lists are defined with a lot of words dictionary as well, excellent, spectacular, bad, poor, etc. [11] [12] [13] [14] [15] Also, to this list are incorporated opinion phrases such as "cost an arm and a leg" or "you need to rob a bank" or "is to pull the money."

Related work can also be found in a research which it is proposed to differentiate the genuine opinion of the "opinion Spam" or "unwanted view". [16] In this research it is analyzed opinion spam's factors and proposes methodologies to identify and isolate it.

Among the background of the topic have found many works and established a theoretical framework on the subject quite extraction of opinions on the Web, Opinion mining and adjacent tracks, however, still have not been documented implementations of these concepts applied to any industry or a non-scientific or academic purpose. Existing studies focus upon the discovery and conceptualization of new terms and modeling of the new reality brought by the advent of the Internet and new communication technologies, but not in use for practical purposes.

Keep in mind that the work and research more relevant and committed to this issue are recent ones, since about 20 years ago, these concepts were unthinkable or difficult to conceive even with technologies that were unknown. Moreover, there was a published paper in which the objective is to detect trends in electoral campaigns using existing technology and information collected on social Web sites. [17]. In the latter study it identifies different strategies for collecting information to analyze:

1. Comprehensive tracking: collecting all possible information in a given period of time
2. Incremental Crawl: We revisit the pages already stored for changes and if changes, these are re-done.
3. Tracking focused: looking for information on a topic based on a ranking algorithm that filters the results that are not relevant
4. Deep Tracking: Collect important information about a particular issue. Unlike the focused crawling, it has the ability to complete forms on the web to store and access the pages returned a completed form.

In addition, there were found reserch papers approaching crawling from different point of views. Crawlers and agent have grown more sophisticated [18] Topical crawler have been studied extensively the last years[19,20,21,22, 23]

Some interesting methods proposed in recent years are those of fish search [24] and focused crawling [25].

Focused crawling concept was implemented using a "clasifier that evaluates the relevance of hypertext document with respect to the focus topics and a distiller that identifies hypertext nodes that are great access points to many relevant sources[19]. There shuoldn't be forgotten to analyze the likage sociology, locating specialty sites and community culture [19]. The focused crawling is different in using a topic taxonomy, learning from example and using graph distillation to track topical hubs. After this reasearch, it was found a lot of anecdotal evidence that bicycle pages are not refer a lot of other biclycle pages, but also refer more significantly more than one moght expect to rd cross and first aid pages. Similarily, HIV / AIDS pages often do not directly refer to other HIV/ AIDS pages but refer to hospital hoe pages.

AI implementations for crawling was proposed begining with a basic exposure to search algorithms and then to be extended in a number of directions to include information retrieval, Bayesian learning, unsupervised learning, natural language rocessing, and knowledge representation.[26]

III. PROPOSED STRUCTURE AND MODEL

The system that supports the research consists on a set of three modules:

1) Crawler

The crawler is in charge of Internet searching and text by storing in a database for further processing. This module has the following input components (input minimum):

* Parameter to search
* Pages where to look
* Deep level navigation links (if there is no limit would be sought through the Internet and never end this stage)
* Restrictions (eg search only in a domain)
* Parameters to function as a filter (words that should not contain the text)

The operator enters the start point pages and then navigates the system for their "children" (linked) to the depth defined in the configuration. Is relevant to mention that the average number of outlinks on web pages is 7 [27] This module basically follows the following behavior:

1. Loading a page
2. Debug the code and convert it to plain ASCII text
3. Read the HTML code in search of the parameter and if does not contain the filter words
4. Search on the same code links to other pages (which must not exceed the maximum level of depth, no restrictions skip) to form a list of URLs to keep searching.
5. If step 3 was yes, the code goes to the analyzer.

The module generates text files with different information. Among them, there will be a metadata file, one with the title of the page, one with a header and the contents of the text in the body. In this way, can be isolated and properly process each part of the page separately.

Defined directory structure

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

It has been defined a structure to store the necessary files with information gathered after the sweep of the sites. The structure consists on a directory for each type of file stored. All those listed and indexed in a flat file. (Bd.txt). In this file each destination will have an ID followed by the URL. The other files will have the ID as a name and an extension that indicates their content, for example:
www.pagina.com ---> 12345
Then the files will emerge from this page:
12345.src (source)
12345.bdy (text body of the page)
12345.lin (links page)
12345.tit (title tag information)
122345.mta (information from meta data page)
12345.etc. (Additional information varies)
12345.ima (images listed on page)
12345.ifr (information contained in the Iframe tag)
12345.hrf (information inside the href tag)
It should be noted that for testing purposes, it was chosen to use ANMAT (National Administration of Drugs, Food and Medical Technology) as a start point. (http://www.anmat.gov.ar/). Later new URLs will be added to the scan list.
The scanned websites are analyzed to make sure that they are written in Spanish. The procedure that we use to define their language is based on the amount of times the letter "e" appeared in the text

As is illustrated in Figure 2, the architecture was implemented using a standard model-view-controller. Figure 3 shows the defined scheme for the content of the module "Model."

In this module, it is implemented Genetic Algorithms to explore and expand the scope of the search criteria.
Basically the system analyzes word by word and associates it with an ID which later is linked to its frequency. The implementation use genetic algorithms operations such as mutation to infer and deduce new words to crawl and amplify the range of search, providing more opportunities to find the information needed.
The system on each crawling operation estimates metrics about the loading page performance, its pagerank value, to later ponderation and scoring.

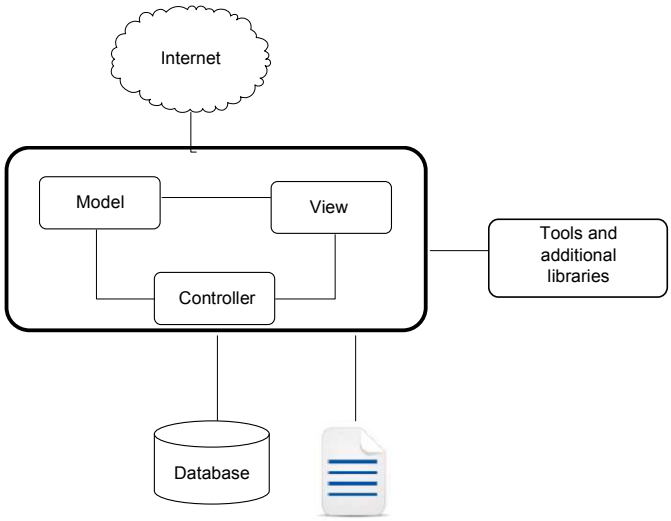

FIGURE 2 – Architecture

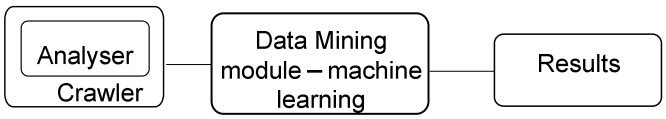

FIGURE 3 – Module "Model"

i) Sub analyzer module
This sub module is part of the crawler or search engine.

The analyzer (see Figure 4) was implemented using the Composite design pattern. It crawls the directory structure that hosts the pages provided by the Crawler.

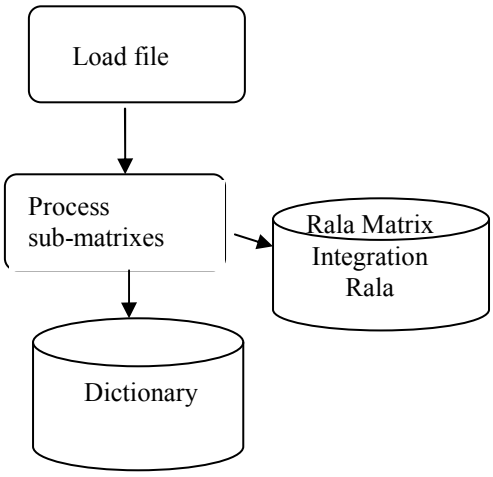

FIGURE 4 – Sub module Parser Analyzer

The following is the description of the behavior of the algorithm to process the files:
For each plain text file:

1. Take the next line L in the text.
2. Apply Porter's algorithm, obtaining the root R of each word.

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

3. For each R:
    3.1. Save the dictionary Dictionary.dct
    3.2. Save <file> matrix. MTRX
4. If there are more lines in A, then go to 1.

After generating the dictionary and the array of frequencies for each file, proceed to the integration of all partial matrices to a single array called <integration>. MTRX. In this stage it was implemented the following algorithm: [18]

Generate sparse matrix <integration>. MTRX empty:
1. Take a matrix <file>. MTRX
2. Integrate content in <integration>. MTRX
3. If more <file>. MTRX then go to 1

It should be noted that each word in the dictionary Dictionary.dct is unique and its records have the following format:

WEIGHT-ON ID + APPEARANCES

At the same time in the <file>. MTRX there are records with the following structure.

ID + QTY-occurrence

Where CANT-occurrence is a counter from 1 (indicating the first appearance of the word identification ID) to n (indicating the total number of times the same word that appears in row) Finally, in the matrix <integration>. MTRX records with the following structure.

ID-FILE + ID + OCCURRENCE

Where ID-Archive, is the unique identifier for each file processed (usually associated with a single URL) and can OCCURRENCE 1 (indicating the occurrence of the word with ID within the file A) or 0 (indicating the absence of such same word in A).
As an example, suppose the following ej.txt file with the contents:

"The practices are complicated
There is a practice file
they claim that the situation is complicated"

The resulting matrix for ej.txt.mtrx file will contain:

1,1
2,1
3,1
4,1
5,1
1,1
6,1
7,1
8,1
9,1
3,1
1,1
10,1
11,1
12,1
13,1
14,1
15,1

Also, the generated entries in the dictionary Dictionary.dct are:

Afirm,10,0.0714285746216774,1
Las,2,0.0714285746216774,1
son,4,0.0714285746216774,1
Hay,6,0.0714285746216774,1
un,7,0.0714285746216774,1
de,9,0.0714285746216774,1
situ,13,0.0714285746216774,1
archiv,8,0.0714285746216774,1
practic,3,0.1428571492433548,2
la,12,0.0714285746216774,1
que,11,0.0714285746216774,1
complic,5,0.1428571492433548,2
es,14,0.0714285746216774,1

2) Data mining module

This module processes the views stored in the sparse matrix contained in text files using advanced techniques of "machine learning" [29][8].

3) Presenter module:

This module is the last in the system and is responsible for exposing the user's search results completed and all information processed in the previous modules. The results are presented through pie charts and reports with all the needed information for analysis and decision making.

IV. Scope

In the initial analysis of the research it was required to define the scope of the system developed to constrain the domain of the problem. This simplification provides the possibility of facilitating the conceptualization and development in a maintainable and orderly system. Moreover, these limitations on the system will let verify the results of the research and the application developed. The restrictions are:

1. The system will process only is Spanish pages
2. The depth level is part of the system configuration
3. The module only processes HTML, Excel, PDF and Word files
4. The application uses a dictionary to identify opinions

V. Conclusion & Future Work

The investigation as it progresses seems even more exciting and viable. The publications so far do not provide relevant information to solve the specific problem. The project will represent a significant improvement for the collection and administration of specific information in an efficient and automatic way.
The next few months the project will focus on refining the relations in the database, model and improve the development and system design that enable collecting data wherever they

are.

After test this development deeply, will proceed to implement the concept of genetic algorithms to optimize the information search task. Tests will be required and adjustments on fitness functions to improve performance and to consider all the alternative answers.

In particular, for the crawler module, the keywords must be defined to find the type of pages to go in case of multilingual sites. It will be added check boxes (checkboxes) to the user interface to setup the options such as learning threshold and histogram pruning.

REFERENCES

[1] A technique for measuring the relative sizeand overlap of the public web search engines, K. Bharat, A Broder, 1998
[2] Preserving the internet, B. Kahle, March 1997
[3] The rise of Crowdsourcing, Wired, Vol 14, No. 6 , Howe, J, 2006
[4] Entity Discovery and Assignment for Opinion Mining Applications, Xiaowen Ding, Bing Liu, Lei Zhang
[5] Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews , Dave, D., Lawrence, A., and Pennock, D., WWW'03, 2003.
[6] A Holistic Lexicon-Based Approach to Opinion Mining Ding, X., Liu, B., and Yu, P., WSDM'08, 2008.
[7] Mining and summarizing customer reviews, Hu. M, and Liu, B. KDD-2004.
[8] Classification Using Machine Learning Techniques , Pang, B., Lee, L., and Vaithyanathan, S. Thumbs up Sentiment. EMNLP'02, 2002.
[9] Generalizing Dependency Features for Opinion Mining, Malesh Joshi, Carolyn Penstein-Rose, 2009
[10] Opinion Mining, Bing Liu
[11] Determining Term Subjectivity and Term Orientation for Opinion Mining, Esuli, A., and Sebastiani, F, EACL'06, 2006.
[12] Predicting the Semantic Orientation of Adjectives, Hatzivassiloglou, V., and McKeown, K., ACL-EACL'97, 1997.
[13] Mining Comparative Sentences and Relations, Jindal, N., and Liu, B., AAAI'06, 2006.
[14] Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents, Kaji, N., and Kitsuregawa, M., EMNLP'07, 2007.
[15] Fully automatic lexicon expansion for domain-oriented sentiment analysis, Kanayama, H., and Nasukawa, T., EMNLP'06, 2006.
[16] Opinion Spam and analysis, Nlitin Jindal, Bing Liu , 2008
[17] The design of OPTIMIST, an Opinion Mining System for Portuguese Politics, Mario J. Silva Carvalho, Luis Sarmiento
[18] Searching the world wide web, S. Lawrence and C.L. Giles, April 1998
[19] Focused crawling, a new approach to topic-specific Web resource discovery, Soumen Chakrabarti, 1999
[20] Acelerated focused crawling through online relevance feedback, S. Chakrabarti, Hawaii, 2002
[21] Information retrieval in the world wide web: making client-based searching feasible, P.M.E. De Bra and R.J.D. Post, 1994
[22] The shark search algorithm – An application: tailored web site mapping, M. Hersovici,1998
[23] Adaptative retrieval agents: internalizing local context and scaling up to the we web. Machine learning, F. Menczer and R.K. Belew, 2000
[24] Improved algorithms for topic distillation in hyperlinked environment, K. Bharat, M. Henzinger, ACM SIGIR conference, 1998
[25] Intelligent crawling on the world wide web with arbitrary predicates, Charu C. Aggarwal, 2001
[26] Web crawling as an AI project, Christopher h. Brooks
[27] Stochastic models for the web graph, S. RaviKumar P. Raghavan, Nov 2000
[28] New models in probabilistic information retrieval, C.J. van Rijsbergen, S.E. Robertson and M.F. Porter, 1980, London: British Library. (British Library Research and Development Report, no. 5587).
[29] Machine learning definition, http://en.wikipedia.org/wiki/Machine_learning

ISBN: 978-988-17012-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)