International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:2, No:9, 2008

Web Application to Profiling Scientific Institutions through Citation Mining

Héctor D. Cortés, Jesús A. del Río, Esther O. García, Miguel Robles.

Abstract**—** Recently the use of data mining to scientific bibliographic data bases has been implemented to analyze the pathways of the knowledge or the core scientific relevances of a laureated novel or a country. This specific case of data mining has been named citation mining, and it is the integration of citation bibliometrics and text mining. In this paper we present an improved WEB implementation of statistical physics algorithms to perform the text mining component of citation mining. In particular we use an entropic like distance between the compression of text as an indicator of the similarity between them. Finally, we have included the recently proposed index h to characterize the scientific production. We have used this web implementation to identify users, applications and impact of the Mexican scientific institutions located in the State of Morelos.

Keywords**—** Citation Mining, Text Mining, Science Impact

I. Introduction

RECENTLY, scientists have addressed the phenomena of citation in scientific research from different perspectives: looking for topological descriptions of citations [1], or for power laws in citation networks [2], or obtaining power laws in number of cites received by journals according to their number of published papers [3]. Aggregation of citation number counts is characteristic of almost all published citation studies [1], [3], [4]; this approach identifies R&D units that have had (and have not had) gross impact on the user community.

The detailed analysis of all the available data of the citing community is required to obtain more information and knowledge [5]. Until recently there has been no comprehensive systematic methodology to deal with the information available through the cites of scientific articles. To overcome the above mentioned limitations of these techniques, there have been recent developments through phenomenological approaches that deal with all the citation information available, and obtain a more detailed description of this complex system [6], [7]. The phenomenological methodology was implemented with the use of commercial software [7].

In all those studies the data source was the Web version of the Science Citation Index (SCI 5300 leading research journals), that allows a broad variety of bibliometric analyzes of R&D units (papers, researchers, journals, institutions, countries, technical areas) to be performed.

Manuscript recieved July 1st, 2006; revised July 17, 2006; This work was partially supported by FOMIX CONACyT Morelos under project 9250 and Macroproyecto Tecnologías de la Informática y Computación para la Universidad, U.N.A.M.

Authors are with Centro de Investigacion en Energia, Universidad Nacional Autonoma de Mexico, A.P. 34, 62580 Temixco, Morelos, Mexico (e-mails: hdcg@cie.unam.mx, arp@cie.unam.mx, eogm@cie.unam.mx, mrp@cie.unam.mx; URL: http://www.cie.unam.mx).

The aim of this paper is to show how the phenomenological approach for citation mining can be implemented in a simple open source web interface. This web interface will help in obtaining a more complete profile of the citing papers, and therby get a more complete representation of the impact of science.

This paper is organized as follows: in section 2 we explain the methodology and the specific data we are dealing with. Also we have included in this section the explanations of the moduli we have used to implement the main program. In section 3 we present illustrative examples of the outputs of the web interface. Finally, in section 4 we close the paper with some comments.

II. Method

In this section we describe the data source, the applications for citation mining, the used algorithms, and the web interface.

A. Citation Mining Application

The citation mining application is a set of programs written in Perl. It is well known that Perl provides a powerful set of features for text processing: it is optimized for scanning arbitrary text files, extracting information from those text files, and printing reports based on that information. The language is intended to be practical (easy to use, efficient, complete) rather than beautiful (tiny, elegant, minimal) [1]. Perl has no arbitrary limits for data size, the native associative arrays (hashes) can grow without loosing performance and regular expression pattern matching techniques are fast and efficient. Moreover, Perl supports both procedural and object-oriented (OO) programming, and has one of the world's most impressive collections of third-party modules at CPAN [2].

The software is modularized in five moduli. The first is an administrative module to mantain data files. The second obtains the counts required in a bibliometric analysis. The third uses a statistical physics algorithm to extract the relevant words, Fourth measures the similarities. These two last moduli are based on entropy-like arguments. Finally, the fifth module measures the coherence of the scientific production based on the citation number.

In the following, we explain in detail these five moduli and the format of the data source.

[1] perl(1)
[2] http://www.cpan.org/

B. Data Source

The data source is a field tagged file from the Thomson ISI Web of Knowledge database `http://isiknowledge.com/`, meeting a search criteria, thus the software and algorithms here presented can be easily used in many different study-case.

A full data record includes authors, titles, journal source, abstract, language, document type, keywords, addresses, cited references, cited references count, times cited, publisher information, ISSN, source abbreviation, page count, and recently e-mail addresses and subject category.

The field tagged file is in nature a text file. Each record is composed with serveral text lines which are the record's fields. Both records and fields are of variable size. Some fields fit in a single line and contains single data. Others fields fit in a single line but represent multiple data. There are fields that span several lines, and represent multiple data. Some fields span in several lines, but contain one single data. Moreover, some fields can be missed because they are not included within the document type, or because an incomplete search criteria. Also, new fields are introduced with information not yet available previously, such as e-mail addresses, or subject category. The software developed for mining this specific format must deal with all these issues.

C. Bibliometric Analysis

The first tool in citation mining is the bibliometric analysis [6], [7]. There are some fields that can be counted directly. From the field *authors* (AU) we obtain information about the authors themselves, and number of authors per article (NAU). From the field *Cross References* (CR) we get the most Cited Authors (CAU). From field *Addresses* (C1) we extract information about research institutions and countries. Collaborations among research institutions and countries can be obtained with a more elaborated analysis.

For each field of interest, an XML file is generated by the program. An HTML file and a CSV file can be obtained using XSLT. Theses files can be post-processed with a graphic tool like OpenCalc.

D. Relevant words

This program extracts the relevant words within the abstracts of the papers in the field tagged file. In order to do this, we follow the procedure indicated in reference [8]. This method uses the standard deviation of the distance between successive occurrences of a word in a text as an indicator of the relevance of the words in the analyzed text. For completeness a frecuency analysis is performed to improve accuracy. The standard deviation is actually close to the entropy [9] in such a way that random distance between same words/phrases indicates a non-relevant word/phrases.

We follow this algorithm and we select the words with normalized standard deviation of the distance between successive occurrences higher than 1 as relevant words in the corresponding text. The algorithm is applied with word lengths from 1 to 3. For each word length a set of XML, HTML and CVS files is generated.

TABLE I
PROPOSED SIMILARITY CRITERIA.

$E(AB) \leq 0.11$	$E(BA) \leq 0.11$	Very similar
$E(AB) \leq 0.11$	$0.11 < E(BA) \leq 0.22$	Similar
$0.11 < E(AB) \leq 0.22$	$0.11 < E(BA) \leq 0.22$	Related
$E(AB) \leq 0.22$	$E(BA) > 0.22$	Related but different

Here it is important to emphasize that with this method the use of stop words it's not necessary, neither foreknowledge about the topic of the text nor the language of text.

E. Similarities

In order to compare the similarities between the abstracts, we have used a compression algorithm. Recently a zipping method to recognize the subject treated in a text was proposed [10]. This method uses the entropy of a string measured when this is zipped (compressed).

The main idea is that when one compresses two text strings, one after another, the compression rate will increase if the second string is similar to the first one, and then the zipped string will have less disorder (entropy) than the previous two strings. In the other side, if the second string is not similar to the first one, the compression ratio will decrease. This algorithm considers that two close papers will have a relative informational entropy close to zero. This algorithm is based on statistical basics, this means that it works better with long files. In our case the abstracts are not actually long files, however this method works properly as we have seen in previous analyzes [11], [12].

In order to classify the degree of similarity we proposed the criteria shown in Table I. Those have been obtained through the comparison of an statistical analysis and manual classification of various given real cases.

It is important to mention that if $E(AB) = E(BA) = 0$ implies that the abstracts are identical.

F. Index h

Recently an index for characterize the scientific production of a researcher has been proposed [13]. The index h of a scientist is defined as the number of papers with citation number higher or equal to h. The index h is an easily computable index which gives an estimate of the importance, significance and broad impact of a scientist's cumulative research contributions. Specifically, this index has been one of the newest ideas in the field of scientific indexes

This module calculates the index h for each author easily, by ordering papers by the field *times cited* (TC). Also, the index h for each institution can be calculated using all the corresponding papers. For each case, the program generates an XML file, and a couple of XSLT transformations generate the HTML and CSV files.

G. Web Application

In order to join these applications we have developed a web interface to capture information and store it to a database. This web interface is written in Perl using the CGI.pm module [3],

[3]CGI(3pm)

and it is hosted in a PC running Fedora Core [4]. The web server is Apache[5].

The user enters to the system providing its username and password. We currently use the basic authentication scheme according to the HTTP specification. This means that the browser sends the username and password with every request to the server. In that sense the user is never logged into the server as in telnet or ftp connections. Every request is made independently from all previous requests and in that sense there is no concept of logging out. However, the user usually perceives the dialog as a session that is equivalent to a telnet or ftp session. But the written code does nothing to support the session concept.

The requested data are organized in the following topics: general information, teaching, industry links, relationships between academic institutions and private sector and technology development.

Once the information is collected we achieve an institutional profile useful to find the impact of the scientific institution on each given topic. The data bases acquired are analyzed using the citation mining process, beginning from the institution name, to obtain the universe of articles belonging to it; form these data the bibliometric analysis is performed.

Finally it is worth mentioning that the system is built with a restricted user level to access only the records and an administrator user to maintain the complete data base.

III. Examples

Here we present screenshots of the web interface.

In Fig. 1 we show the main menu of the web application accessing as a restricted user. The user can select any of the sections and capture the information related with its own institution. In the *análisis bibliométrico y de minería* section, user can perform the citation mining analyzes according with the captured addresses information.

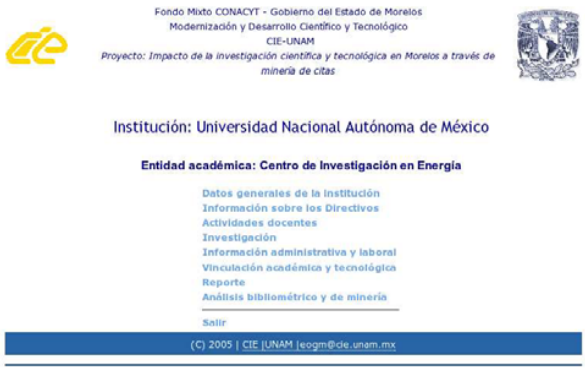

Fig. 1. Display of the main result page

In Fig. 2 we present the main page when the complete analysis has been performed. In it we can see the five moduli with the complete set of outputs (HTML, XML, CSV), also time stamp of the analyzes is shown.

[4] http://fedora.redhat.com/
[5] http://www.apache.org/

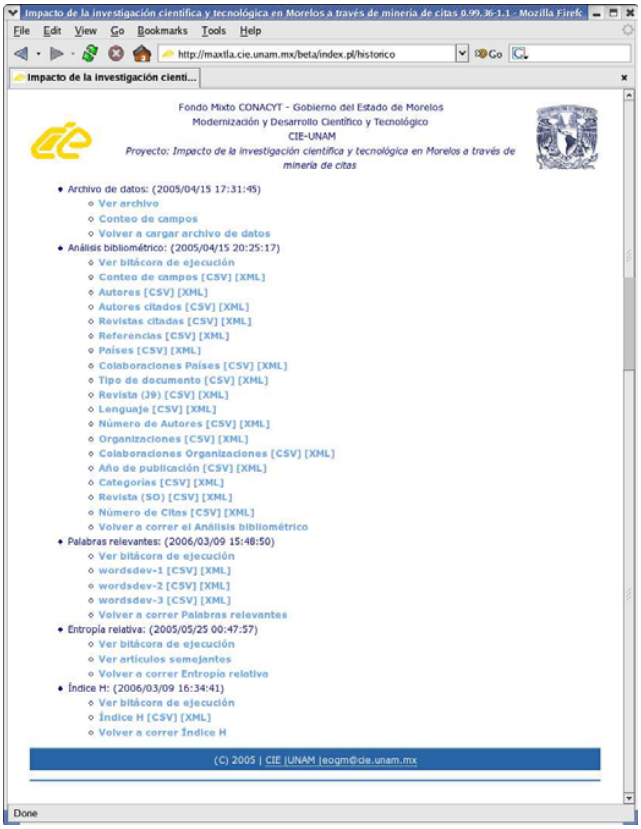

Fig. 2. Display of the main result page

In Fig. 3 we present the output of the more relevant single words. As it can be seen there are no meaningless words, it can be observed in a very indicative word like *atole*, a porridge of maize meal and water, or milk; an important component of the rural Mexican diet.

In Fig. 4 we illustrate a sample output of similar papers. When the abstracts of a pair of papers differ by very few words, or none, we have a pair of identical papers. A very similar set of papers means that the relative entropic distance is small, and should require visual inspection to discard identical papers. Similar papers are the case of one paper published on congress proceedings (short paper) and the other one on a jornal (long paper). In the case of related papers we have clearly different papers in the same area, or technique.

In Fig. 5 we illustrate the output of the Index h. We confirm that the highers indexes corresponds to the successful scientistics with long career.

IV. Conclusions

In this paper we present an implementation of citation mining to profiling scientific institutions. This system has been implemented using Perl and it is a clear sample of different data mining technics applied to text data in order to obtain valuable formated information that can be easily interpreted.

The example developed here is based on the ISI database and very close to ISI's acceptable use policy. Of course the

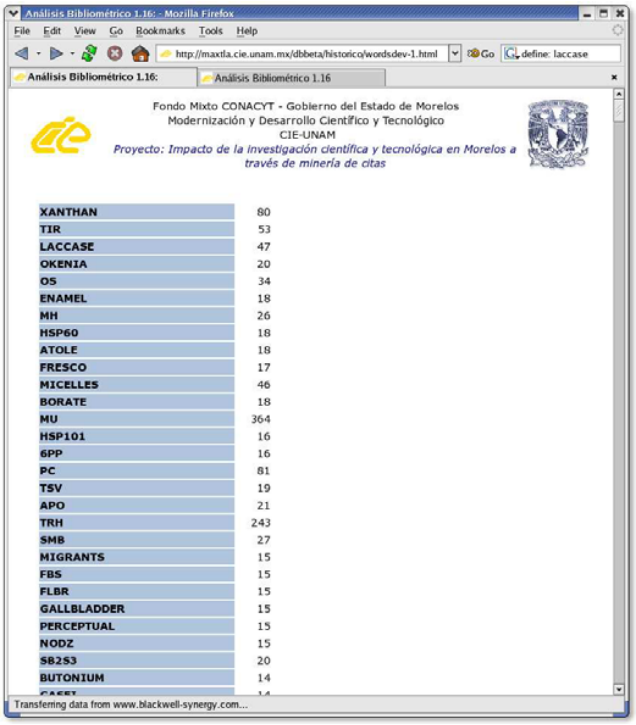

Fig. 3. Display of the most relevant words

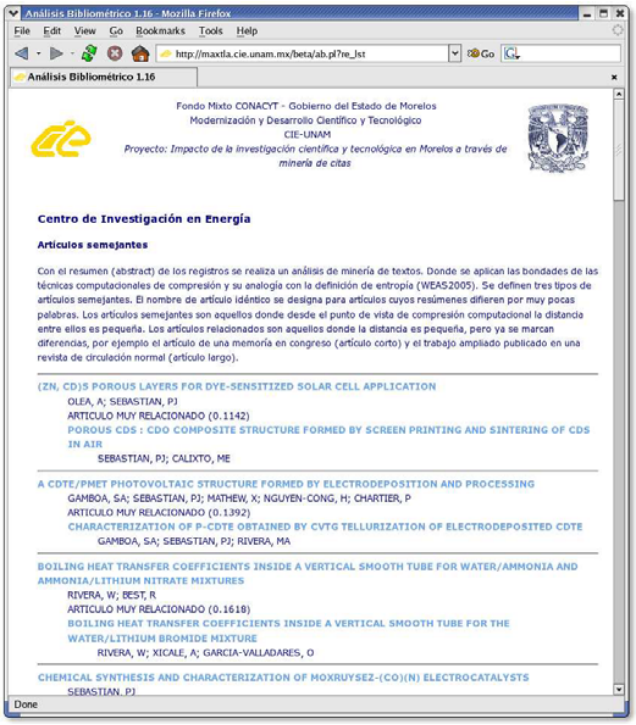

Fig. 4. Similar papers

software can be modified in order to deal with any other database of scientific journals. A very important achieve is that the application may be used to obtain a picture of the profile of science on an institution, a region or even a country, this rely only on the choice of the subset of the ISI's data base selected. At present the profiling is being done for different institutions in the Mexican state of Morelos, and is planned to be applied in other Mexican institutions in near future.

Finally, an important part to be stressed is that the module containing the procedure for obtain the index h is a novel feature for citation mining.

Appendix. Sample ISI Record

In this appendix we shown a standard ISI record.

```
FN ISI Export Format
VR 1.0
PT J
AU Kostoff, RN
   del Rio, JA
   Humenik, JA
   Garcia, EO
   Ramirez, AM
TI Citation mining: Integrating text mining and
   bibliometrics for research user profiling
SO JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION
   SCIENCE AND TECHNOLOGY
LA English
DT Article
ID DATABASE TOMOGRAPHY; SCIENCE; IMPACT
AB Identifying the users and impact of research is
   important for research performers, managers,
   evaluators, and sponsors. It is important to know
   whether the audience reached is the audience desired.
   It is useful to understand the technical
   characteristics of the other research/development/
   applications impacted by the originating research,
   and to understand other characteristics (names,
   organizations, countries) of the users impacted by
   the research. Because of the many indirect pathways
   through which fundamental research can impact
   applications, identifying the user audience and the
   research impacts can be very complex and time
   consuming. The purpose of this article is to
   describe a novel approach for identifying the
   pathways through which research can impact other
   research, technology development, and applications,
   and to identify the technical and infrastructure
   characteristics of the user population. A novel
   literature-based approach was developed to identify
   the user community and its characteristics. The
   research performed is characterized by one or more
   articles accessed by the Science Citation Index
   (SCI) database, because the SCI's citation-based
   structure enables the capability to perform
   citation studies easily.
C1 Off Naval Res, Arlington, VA 22217 USA.
   UNAM, Ctr Invest Energia, Temixco, Morelos, Mexico.
   NOESIS Inc, Manassas, VA USA.
RP Kostoff, RN, Off Naval Res, Arlington, VA 22217 USA.
CR *DOD, 1969, AD495905 DTIC DOD
   *DOE, 1983, DOEER0194
   *DOE, 1986, DOEER0275 DOE
   *IDA, 1991, P2192 IDA, V1
   *IITRI, 1968, TECHN RETR CRIT EV S
   AVERCH HA, 1994, EVALUATION REV, V18, P77
   BRAUN T, 1987, LIT ANAL CHEM SCIENT
   DAVIDSE RJ, 1997, SCIENTOMETRICS, V40, P171
   DELRIO JA, 2000, CITATION MINING CITI
   EGGHE L, 1990, INTRO INFORMETRICS
   HERRING SD, 1999, J AM SOC INFORM SCI, V50, P358
   JAEGER HM, 1992, SCIENCE, V255, P1523
   KOSTOFF RN, 1994, SCI PUBLIC POLICY, V2
   KOSTOFF RN, 1997, ADA296021 DTIC
```

Fig. 5. Index h

```
   KOSTOFF RN, 2000, J AIRCRAFT, V37, P727
   KOSTOFF RN, 2000, J CHEM INF COMP SCI, V40, P19
   LOSIEWICZ P, 2000, J INTELL INF SYST, V15, P99
   NARIN F, 1989, EVALUATION SCI RES, P120
   NARIN F, 1994, EVALUATION REV, V18, P65
   PRICE DJD, 1986, LITTLE SCI BIG SCI
   STEELE TW, 2000, JASIS, V15, P476
   TASSEY G, 1999, EVAL PROGRAM PLANN, V22, P113
NR 22
TC 11
PU JOHN WILEY & SONS INC
PI NEW YORK
PA 605 THIRD AVE, NEW YORK, NY 10158-0012 USA
SN 1532-2882
J9 J AM SOC INF SCI TECHNOL
JI J. Am. Soc. Inf. Sci. Technol.
PD NOV
PY 2001
VL 52
IS 13
BP 1148
EP 1156
PG 9
SC Computer Science, Information Systems; Information
   Science & Library Science
GA 485KG
ER

EF
```

REFERENCES

[1] Amaral, L.A.N., Gopikrishnan, P., Matia, K., Plerou, V. and Stanley E.H. Application of statistical physics methods and concepts to the study of science & technology systems, Scientometrics, Vol. 51, No. 1, 2001, pp 9-36.
[2] Bilke, S. and Peterson, C. Topological properties of citation and metabolic networks, Phys. Rev. E Vol. 64, 2001. 036106
[3] Katz, J.S. The self-similar science systems, Research Policy, Vol 28, 1999, pp. 501-517
[4] Redner, S. How popular is your paper? An empirical study of the citation distribution, Eur. Phys. J., Vol 4, 1998, 131.
[5] Kostoff R.N., del Río J.A. The impact of physics research. Phys World. Vol. 14, 2001, pp 47-51.
[6] Kostoff, R.N., del Río, J.A., Humenik, J.A, García, E.O. and Ramírez, A.M., Citation Mining: Integrating Text Mining and Bibliometrics for Research Users Profile. J. Am. Soc. Inform. Scien. & Tech. Vol. 52, 2001, pp. 1148-1156.
[7] J.A del Río, R.N. Kostoff, E.O. García, A.M. Ramírez and J.A. Humenik, Phenomenological approach to profile impact of scientific research: citation mining, Adv. Complex Syst. Vol. 5, 2002. pp. 19-42.
[8] Ortuno, M., Carpena, P., Bernaola-Galvan, P., Muoz, E. and Somoza, A.M., Keyword detection in natural languages and DNA. Europhysics Letters, Vol. 57, 2002, pp. 759-764.
[9] Montrol, E.W, About the Physics of no-physical systems. J. Stat Phys, Vol. 42, 1986, 647.
[10] Benedeto, D., Caglioti E., Loreto V., Language Trees and Zipping, Physical Review Letters, Vol. 88, 2002, 048702.
[11] del Ro, J.A. and Corts, H.D. La ciencia mexicana en Nature y Science: La ltima dcada, Ciencia, (journal of the Mexican Academy of Sciences AMC). In press (2006).
[12] Cortes H.D., del Rio J. A., Garcia E.O., Web Implementation of Entropy-like Algorithms for Citation Mining, WSEAS Transactions on Information Science and Applications, Vol. 2, No. 9, Pp. 1430 - 1437, 2005
[13] J. E. Hirsch. An index to quantify an individual's scientific research output. PROC. NAT. ACAD. SCI. 102, 16569-16572 (2005), also available in http://arXiv.org/physics/0508025.