Proceedings of the 5th WSEAS Int. Conf. on SIMULATION, MODELING AND OPTIMIZATION, Corfu, Greece, August 17-19, 2005 (pp237-242)

A Simple Web Interface for Citation Mining

H. D. CORTÉS, J. A. DEL RÍO, E. O. GARCÍA.
Centro de Investigación en Energía
Universidad Nacional Autónoma de México
A.P. 34, 62580 Temixco, Morelos
MEXICO

http://www.cie.unam.mx/

Abstract: - Citation Mining is an integration of citation bibliometrics and text mining. It can be used to measure the impact, both direct and indirect, of scientific and technological research. In this work we present a Simple Web Interface for a set of command line oriented programs for Citation Mining. The separation of the Web Application and the Program Application allow us an easier and faster development cycle. This system is used by the Government of the State of Morelos, Mexico, to identify users, applications and impact of the Research Institutions located in the State.

Key-Words: - Citation Mining, Text Mining, Science Impact.

1 Introduction

Recently, scientists have addressed the problem of citation in scientific research from different perspectives: looking for topological description of citation 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 cites of the scientific article. To overcome the above mentioned limitations of these techniques, recently it has developed a phenomenological approach to 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.

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 for citation mining. In order to do this we modified some algorithms. This web interface will help in obtaining a more complete profile of the citing papers, and thereby get a more complete representation of the impact of science.

The organization of the paper is as follows: in section 2 we explain the methodology and the specific data we are dealing with. In this section we also include the explanations of the moduli we have used to separate the complete program. In section 3 we present illustrative examples of the outputs of our web interface. Finally we close the paper with some comments.

2 Method

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

2.1 Program 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 mining: it is optimized for scanning arbitrary text files, extracting information from those text files, and printing reports based on that information [1]. Perl has no arbitrary limits for data size, 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].

We have divided the software in three moduli. The first obtain the counts required in a bibliometric analysis. The second uses a statistical physics algorithm to extract the relevant words, and the last one measures the similarities. These two last moduli are based on entropy arguments. In the following, we explain in detail these three moduli and the format format of the data source.

2.2 Data Source.

From the database Web of Science (`http://isiknowledge.com/`) we obtain a field tagged file, bounded by a search criteria. In the specific example worked here, we use SCI-EXPANDED citation database, search criteria is (SOURCE *nature* OR *science*) AND (ADDRESS *mexico* NOT *new*). The complete analysis of this example has been performed and published [11].

A full data record includes title, authors, source, abstract, language, document type, keywords, addresses, cited references, times cited, publisher information, ISSN, source abbreviation, page count, and subject category. A sample of this record is shown in Appendix A.

Being the field tagged file a text file, fields and records are variable size. Some fields are single line and single data, others fields are single line and multi data, others fields are multi line and multi data, and others span in several lines, but these lines represent 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 all available previously, such as e-mail addresses. The software developed for mining this specific format must deal with all these issues.

2.3 bibliometrics.pl.

In citation mining the first tool is the bibliometric analysis [6,7]. The following fields are directly counted: source (SO), document type (DT), language (LA), published year (PY), source abbreviation (J9) and times cited (TC). From the field authors (AU) we obtain information about the authors themselves, and number of authors per article (NAU). From field CR we get the most cited authors (CAU). We extract information about research institutions and countries from field C1. A more elaborated analysis can tell us the collaborations among research institutions and countries. For each field of interest, a CSV text file is generated. Theses files can be post-processed with a graphic tool like OpenCalc.

2.4 wordsdev.pl.

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 ref. [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, 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 run length of words from 1 to 3. For each length a CSV text file is generated and then interpreted. Here it is important to mention that with

[1] perl(1)

[2] http://www.cpan.org/

this method it is not necessary to have a dictionary with meaningless words (as the, a, is by, are, etc.) or foreknowledge about the topic of the text. This is important because eliminate preprosesing.

2.5 relentropy.pl.

In order to compare the similarities between the abstract, we have used a compression algorithm. Recently a zipping method to recognize the subject treated in a text was proposed [10]. This method uses that the entropy of a string can be measured when this string is zipped (compressed). The main idea is that when one compresses two 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. Then, this algorithm considers that two close papers will have a relative informational entropy close to zero. This algorithm is based on the statistical basis, 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 analysis.

2.6 Web Application

In order to make easier the use of the citation mining software, we have developed a simple web interface. It is simple from both the point of view of the user, and the developer. We avoid the use of unnecessary features, like fancy javascripting or complex layouts, but still using widely accepted features like CSS.

We have divided the software into several moduli closed related with their command line citation mining program. This web interface is written in Perl using the CGI.pm module[3], and it is hosted by a PC running Fedora Core[4]. The web server is Apache. In the following a typical session is described.

The user enters to the system, and can select a previously loaded project, or start a new one. When a new project is started, the user has to provide a short name for the project. All subsequents operations will be realized on the selected project.

The first process is to upload a data file. Several integrity checks are realized to assure a valid field tagged file. If the uploaded file is found valid it is stored on the web file system. We perform a field count for cross-references among the other programs.

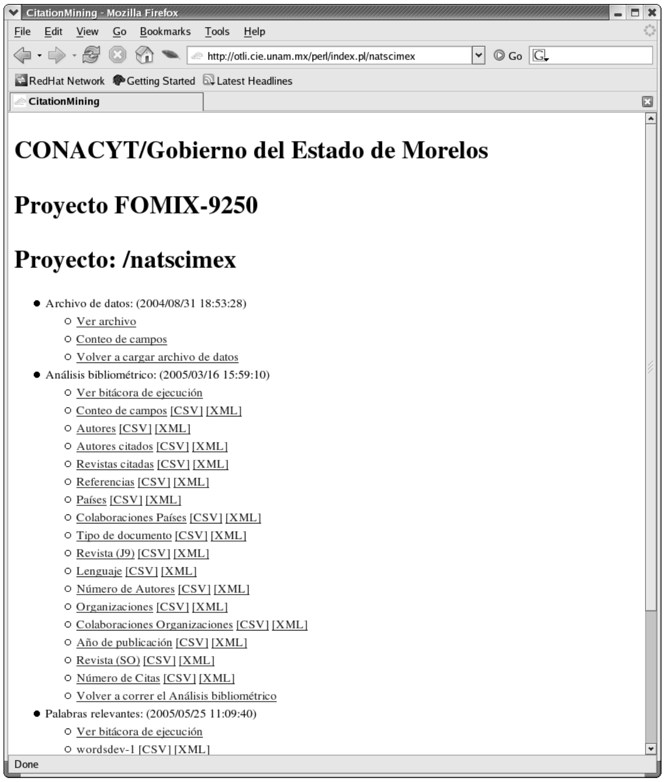
Next, the user can perform the bibliometric analysis. The web interface invokes the previously described bibliometric program with the appropriated arguments. For each file generated, a link is showed.

Later, the user can apply the relevant words search. Also the web interface invokes the corresponding program, and for each generated file a link is showed.

Finally, the user can perform the similarity analysis. In the next section we present some examples.

3 Examples

Here we present screenshots of the simple open source web interface for citation mining. In Fig. 1 we present the main page when the complete analysis has been performed. In it we can see the first two moduli with the complete set of outputs, also there are shown the time stamp of the analyzes.


CitationMining - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://otli.cie.unam.mx/perl/index.pl/natscimex
Go
RedHat Network
Getting Started
Latest Headlines
CitationMining
Done

CONACYT/Gobierno del Estado de Morelos

Proyecto FOMIX-9250

Proyecto: /natscimex

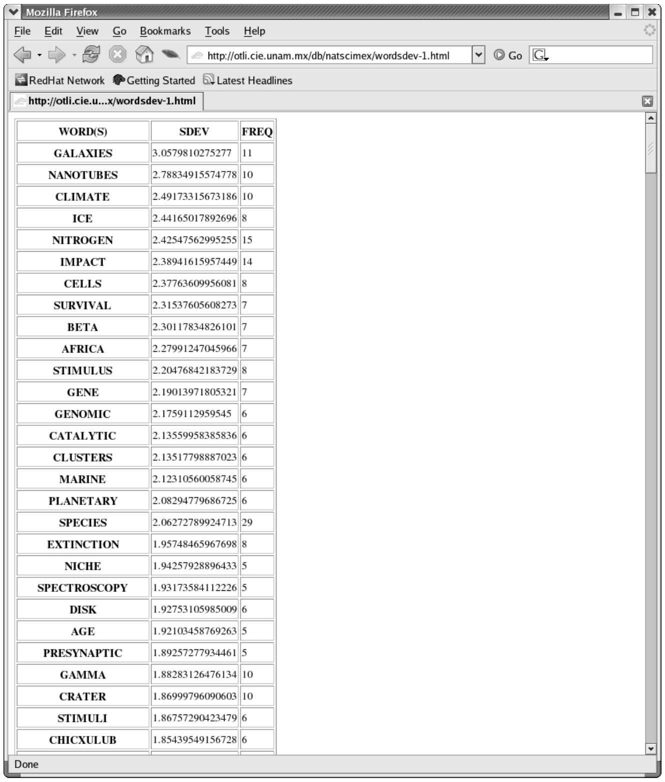
- Archivo de datos: (2004/08/31 18:53:28)
  - Ver archivo
  - Conteo de campos
  - Volver a cargar archivo de datos
- Análisis bibliométrico: (2005/03/16 15:59:10)
  - Ver bitácora de ejecución
  - Conteo de campos [CSV] [XML]
  - Autores [CSV] [XML]
  - Autores citados [CSV] [XML]
  - Revistas citadas [CSV] [XML]
  - Referencias [CSV] [XML]
  - Países [CSV] [XML]
  - Colaboraciones Países [CSV] [XML]
  - Tipo de documento [CSV] [XML]
  - Revista (J9) [CSV] [XML]
  - Lenguaje [CSV] [XML]
  - Número de Autores [CSV] [XML]
  - Organizaciones [CSV] [XML]
  - Colaboraciones Organizaciones [CSV] [XML]
  - Año de publicación [CSV] [XML]
  - Revista (SO) [CSV] [XML]
  - Número de Citas [CSV] [XML]
  - Volver a correr el Análisis bibliométrico
- Palabras relevantes: (2005/05/25 11:09:40)
  - Ver bitácora de ejecución
  - wordsdev-1 [CSV] [XML]

Fig. 1. Display of the main result page

[3] CGI(3pm)

[4] http://fedora.redhat.com/

In Fig. 2 we present the output of the more relevant single words. As it can be seen there are not meaningless words and can also be observed a very indicative words like *chicxhulub* (a crater created by a meteor and possible cause of dinosaur extinction), and other words remarking that astronomy is a relevant field in Mexican Science.

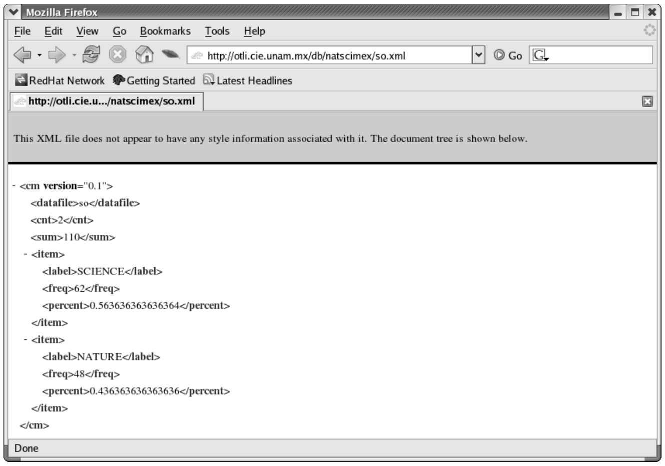


http://otli.cie.unam.mx/db/natscimex/wordsdev-1.html

http://otli.cie.u...x/wordsdev-1.html

WORD(S)	SDEV	FREQ
GALAXIES	3.0579810275277	11
NANOTUBES	2.78834915574778	10
CLIMATE	2.49173315673186	10
ICE	2.44165017892696	8
NITROGEN	2.42547562995255	15
IMPACT	2.38941615957449	14
CELLS	2.37763609956081	8
SURVIVAL	2.31537605608273	7
BETA	2.30117834826101	7
AFRICA	2.27991247045966	7
STIMULUS	2.20476842183729	8
GENE	2.19013971805321	7
GENOMIC	2.1759112959545	6
CATALYTIC	2.13559958385836	6
CLUSTERS	2.13517798887023	6
MARINE	2.12310560058745	6
PLANETARY	2.08294779686725	6
SPECIES	2.06272789924713	29
EXTINCTION	1.95748465967698	8
NICHE	1.94257928896433	5
SPECTROSCOPY	1.93173584112226	5
DISK	1.92753105985009	6
AGE	1.92103458769263	5
PRESYNAPTIC	1.89257277934461	5
GAMMA	1.88283126476134	10
CRATER	1.86999796090603	10
STIMULI	1.86757290423479	6
CHICXULUB	1.85439549156728	6

Fig. 2. Display of the single relevant words



http://otli.cie.unam.mx/db/natscimex/so.xml

http://otli.cie.u.../natscimex/so.xml

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
- <cm version="0.1">
    <datafile>so</datafile>
    <cnt>2</cnt>
    <sum>110</sum>
  - <item>
      <label>SCIENCE</label>
      <freq>62</freq>
      <percent>0.563636363636364</percent>
    </item>
  - <item>
      <label>NATURE</label>
      <freq>48</freq>
      <percent>0.436363636363636</percent>
    </item>
  </cm>
```

Fig. 3. Display of an XML output

In Fig. 3 we illustrate a preliminary XML output showing the display of source journals. Here we observe that Mexican scientists sort his/her work in both journals almost by equal.

4 Conclusions

In this paper we present a simple open source web interface for citation mining. This interface has been implemented using PERL to deal with the text file obtained from the Web of Science database of Institute for Scientific Information. In the development of the interface it has been paid special attention to avoid creating databases from the extracted data from ISI and of course the interface asks for the file from ISI which can be obtained only for authorized users. Thus we are close to ISI's acceptable use policy. Of course the software can be modified in order to deal with any other database of scientific journals with an output in text (CVS) format. This software has been use for the analysis of the scientific production of Mexico in the main stream journals as Nature and Science during last decade [11] and also has been use to deal with several thousand of papers allowing immediately analysis of the results.

5 Acknowledgements

This paper has been partial supported by FOMIX CONACyT-Morelos under project 9250.

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 RN, del Río JA. 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., Muñoz, 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 Río, J.A. and Cortés, H.D. La ciencia mexicana en Nature y Science: La última década, Ciencia, (journal of the Mexican Academy of Sciences AMC) in press (2005).

Appendix A: Sample ISI Record

In this appendix we shown a standard ISI record

```
FN ISI Export Format
VR 1.0
PT J
AU Kostoff, RN
   Bedford, CD
   del Rio, JA
   Cortes, HD
   Karypis, G
TI Macromolecule mass spectrometry:
   Citation mining of user documents
SO JOURNAL OF THE AMERICAN SOCIETY FOR
   MASS SPECTROMETRY
LA English
DT Article
ID ULTRAVIOLET-LASER DESORPTION;
   ELECTROSPRAY ION-SOURCE; IONIZATION;
   PROTEINS; BIBLIOMETRICS; INFORMATION;
   SCIENCE; PHYSICS; TRENDS
AB Identifying research users,
   applications, and impact is important
   for research performers, managers,
   evaluators, and sponsors.
   Identification of the user audience
   and the research impact is complex and
   time consuming due to the many
   indirect pathways through which
   fundamental research can impact
   applications. This paper identified
   the literature pathways through which
   two highly-cited papers of 2002
   Chemistry Nobel Laureates Fenn and
   Tanaka impacted research, technology
   development, and applications.
   Citation Mining, an integration of
   citation bibliometrics and text
   mining, was applied to the >1600 first
   generation Science Citation Index
   (SCI) citing papers to Fenn's 1989
   Science paper on Electrospray
   Ionization for Mass Spectrometry, and
   to the >400 first generation SCI
   citing papers to Tanaka's 1988 Rapid
   Communications in Mass Spectrometry
   paper on Laser Ionization Time-of-
   Flight Mass Spectrometry.
   Bibliometrics was performed on the
   citing papers to profile the user
   characteristics. Text mining was
   performed on the citing papers to
   identify the technical areas impacted
   by the research, and the relationships
   among these technical areas. (C) 2004
   American Society for Mass
   Spectrometry.
C1 Off Naval Res, Arlington, VA 22217
   USA.
   Univ Nacl Mexico, Ctr Invest Energia,
   Mexico City, DF, Mexico.
   Univ Minnesota, Minneapolis, MN USA.
RP Kostoff, RN, Off Naval Res, 800 N
   Quincy St, Arlington, VA 22217 USA.
EM kostofr@onr.navy.mil
CR BEAVIS RC, 1989, RAPID COMMUN MASS SP,
   V3, P233
   BEAVIS RC, 1989, RAPID COMMUN MASS SP,
   V3, P432
   BEAVIS RC, 1990, ANAL CHEM, V62, P1836
   CUTTING DR, 1992, P 15 ANN INT ACM
   SIG, P318
   DAVIDSE RJ, 1997, SCIENTOMETRICS, V40,
   P171
   DELRIO JA, 2002, ADV COMPLEX SYST, V5,
   P19
   FENN JB, 1989, SCIENCE, V246, P64
   FENN JB, 1990, MASS SPECTROM REV, V9,
   P37
   GARFIELD E, 1985, J CHEM INF COMP SCI,
   V25, P170
   GOLDMAN JA, 1999, METHOD INFORM MED,
   V38, P96
```

GORDON JS, 1998, AM HERITAGE, V49, P8
GREENGRASS E, 1997, TRR520296 NAT SEC AG
GUHA S, 1998, P ACM SIGMOD INT C M, P73
HEARST MA, 1998, NAT LANGUAGE INFORMA
HEARST MA, 1999, P ACL 99 37 ANN M AS
JAEGER HM, 1992, SCIENCE, V255, P1523
KARAS M, 1985, ANAL CHEM, V57, P2935
KARAS M, 1987, INT J MASS SPECTROM, V78, P53
KARAS M, 1988, ANAL CHEM, V60, P2299
KARAS M, 1989, INT J MASS SPECTROM, V92, P231
KARYPIS G, 1999, COMPUTER, V32, P68
KARYPIS G, 2002, CLUTO CLUSTERING TOO
KOSTOFF RN, SCI TECHNOLOGY TEXT
KOSTOFF RN, 1997, J INF SCI, V23, P4
KOSTOFF RN, 1998, SCIENTOMETRICS, V43, P27
KOSTOFF RN, 2000, J AIRCRAFT, V37, P727
KOSTOFF RN, 2001, J AM SOC INF SCI TEC, V52, P1148
KOSTOFF RN, 2003, ENCY LIB INFORMATION, V4, P2789
KOSTOFF RN, 2003, INT HDB INNOVATION, P388
KOSTOFF RN, 2003, MED HYPOTHESES, V61, P265
KOSTOFF RN, 2004, FRACTALS, V12, P1
KOSTOFF RN, 2004, INT J BIFURCAT CHAOS
LOO JA, 1989, ANAL BIOCHEM, V179, P404
LOO JA, 1990, SCIENCE, V248, P201
LOSIEWICZ P, 2000, J INTELL INF SYST, V15, P99
MACROBERTS MH, 1996, SCIENTOMETRICS, V36, P435
MANN M, 1989, ANAL CHEM, V61, P1702
NARIN F, 1976, MONOGRAPH NSF
NARIN F, 1994, EVALUATION REV, V18, P65
PRECHELT L, 2002, J UNIVERS COMPUT SCI, V8, P1016
RASMUSSEN E, 1992, INFORMATION RETRIEVA
SCHUBERT A, 1987, SCIENTOMETRICS, V12, P267
SMITH RD, 1990, ANAL CHEM, V62, P882
STEINBACH M, 2000, 00034 U MINN DEP COM
SWANSON DR, 1986, PERSPECT BIOL MED, V30, P7
SWANSON DR, 1997, ARTIF INTELL, V91, P183
TANAKA K, 1987, P 2 JAP CHIN JOINT S, P185
TANAKA K, 1988, RAPID COMMUN MASS SP, V2, P151
VIATOR JA, 2001, J ACOUST SOC AM 1, V109, P1779
WHITEHOUSE CM, 1985, ANAL CHEM, V57, P675
WILLETT P, 1988, INFORMATION PROCESSI, V24, P577
WISE MJ, 1992, STRING SIMILARITY GR
WONG SF, 1988, J PHYS CHEM-US, V92, P546
YAMASHITA M, 1984, J PHYS CHEM-US, V88, P4451
YAMASHITA M, 1984, J PHYS CHEM-US, V88, P4671
YOSHIDA T, 1988, MASS SPECTROSC JAPAN, P36
ZAMIR O, 1998, P 21 ANN INT ACM SIG, P46
ZHAO Y, 2003, IN PRESS CRITERION F
ZHU DH, 2002, TECHNOL FORECAST SOC, V69, P495

NR 59
TC 0
J9 J AMER SOC MASS SPECTROM
JI J. Am. Soc. Mass Spectrom.
PD MAR
PY 2004
VL 15
IS 3
BP 281
EP 287
PG 7
SC CHEMISTRY, ANALYTICAL; CHEMISTRY, PHYSICAL; SPECTROSCOPY
ER

EF