Annals of Library and Information Studies
Vol. 61, June 2014, pp. 93-101

A generic algorithm for extraction, analysis and presentation of important journal information from online journals

Virendra Kumar[1], Gaurav Gupta[2], H K Sardana[3] and Akash Deep[4]

Central Scientific Instruments Organisation (CSIR-CSIO), Sector-30C, Chandigarh – 160030, India,
E-mail: virendrakumar@csio.res.in, gauravgupta@csio.res.in, hk_sardana@csio.res.in, dr.akashdeep@csio.res.in

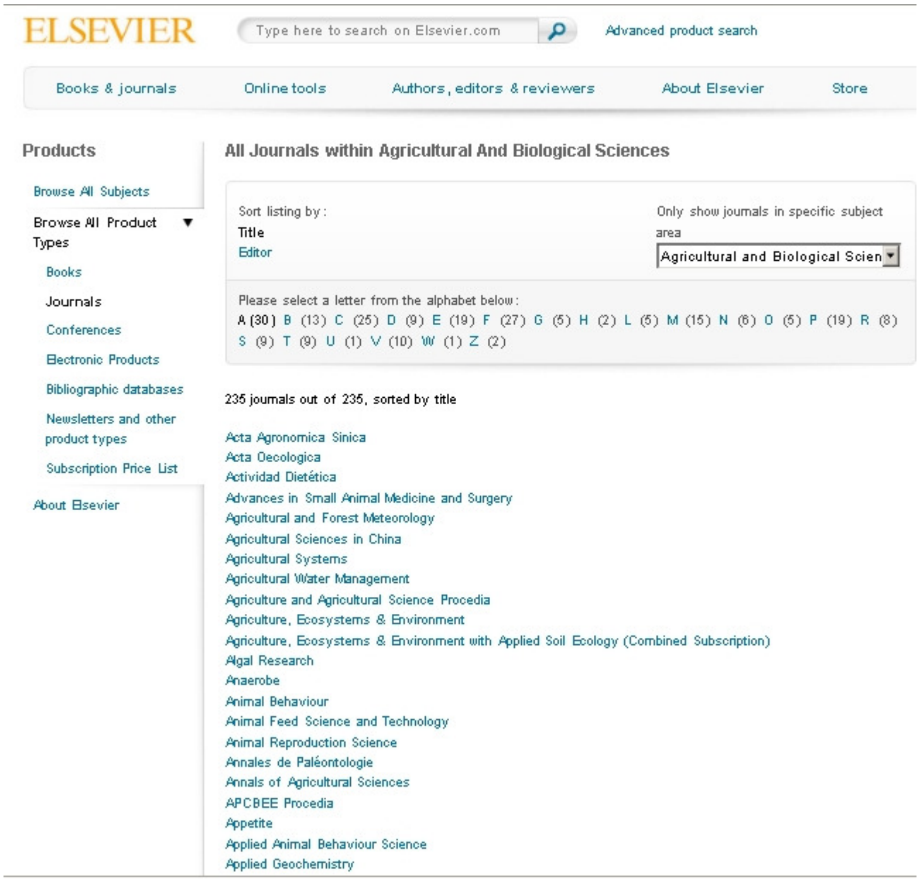
Received: 17 September 2013; revised: 5 May 2014; accepted: 28 May 2014

A generic algorithm for the extraction, analysis and presentation of summarised information from any individually searched journal is described. The algorithm currently analyses from a list of more than 7000 online available journals and displays desired information about the impact factor, publication numbers etc. It will help the user in sorting suitable alternatives for the dissemination of his/her research findings.

Keywords: Impact factor; Online journals; Data extraction; URL

Introduction

In today's web world, a large and quickly mounting amount of information is continuously produced, shared and consumed. Web data extraction systems represent a broad class of software applications, which offers the extraction of information from web sources[1,2] with a limited human effort and then stores the collected information for further processing. The processed information is converted to the desired structured format for targeted applications[3,4]. The web data extraction systems are useful for a wide range of applications, including the analysis of text documents (like e-mails, support forum, technical and legal documentation etc), business and competitive intelligence[5], trends of social web platforms[6,7] and bio-informatics[8]. The state-of-the-art of the web data extraction has been evaluated in a number of reviews. Laender et al. proposed the classification of web data extraction systems by introducing a set of criteria and a qualitative analysis of various such tools[9]. Kushmerick proposed the profiling of the web data extraction through finite state approaches, including the wrapper induction approach (automatic generation of wrappers) and the maintenance approach (updating of wrapper each time the structure of the web source changes)[10]. Natural Language Processing and Hidden Markov Models were also discussed. Some other informative articles on the wrapper induction approach include the work of Flesca et al.[11] and Kaiser and Miksch[12]. The papers from Chang et al.[13] and Fiumara[14] discuss the tri-dimensional categorization of web data extraction systems on the basis of task difficulties, techniques used and degree of automation. Sarawagi[15] has also summarized some useful information in the related context. The work from Arasu and Garcia-Molina[16] and Elmeleegy et al[17] may deserve special mention. The former authors presented an algorithm 'ExALG' for extracting the structured data from web pages by employing a two-step concept, involving the discovery of sets of tokens associated with the same type constructor in the (unknown) template used to create the input pages, and using the above sets to deduce the template that extracts the values encoded in the pages. Elmeleegy et al[17] proposed the 'ListExtract' technique to extract data from the webpages that are arranged in the form of list. The technique executes as a sequence operations over the input list which can be divided into three steps, an independent splitting phase, an alignment phase, and a final refinement phase. Both the above reports targeted the structured web pages. The unstructured data management systems (UDMSs) are the software systems that analyze raw text data, extract and integrate structures from them, and build



Products

Browse All Subjects
Browse All Product Types
Books
Journals
Conferences
Electronic Products
Bibliographic databases
Newsletters and other product types
Subscription Price List
About Elsevier

All Journals within Agricultural And Biological Sciences

Sort listing by :
Title
Editor

Only show journals in specific subject area
Agricultural and Biological Scien

Please select a letter from the alphabet below :
A (30) B (13) C (25) D (9) E (19) F (27) G (5) H (2) L (5) M (15) N (6) O (5) P (19) R (8) S (9) T (9) U (1) V (10) W (1) Z (2)

235 journals out of 235, sorted by title

Acta Agronomica Sinica
Acta Oecologica
Actividad Dietética
Advances in Small Animal Medicine and Surgery
Agricultural and Forest Meteorology
Agricultural Sciences in China
Agricultural Systems
Agricultural Water Management
Agriculture and Agricultural Science Procedia
Agriculture, Ecosystems & Environment
Agriculture, Ecosystems & Environment with Applied Soil Ecology (Combined Subscription)
Algal Research
Anaerobe
Animal Behaviour
Animal Feed Science and Technology
Animal Reproduction Science
Annales de Paléontologie
Annals of Agricultural Sciences
APCBEE Procedia
Appetite
Applied Animal Behaviour Science
Applied Geochemistry

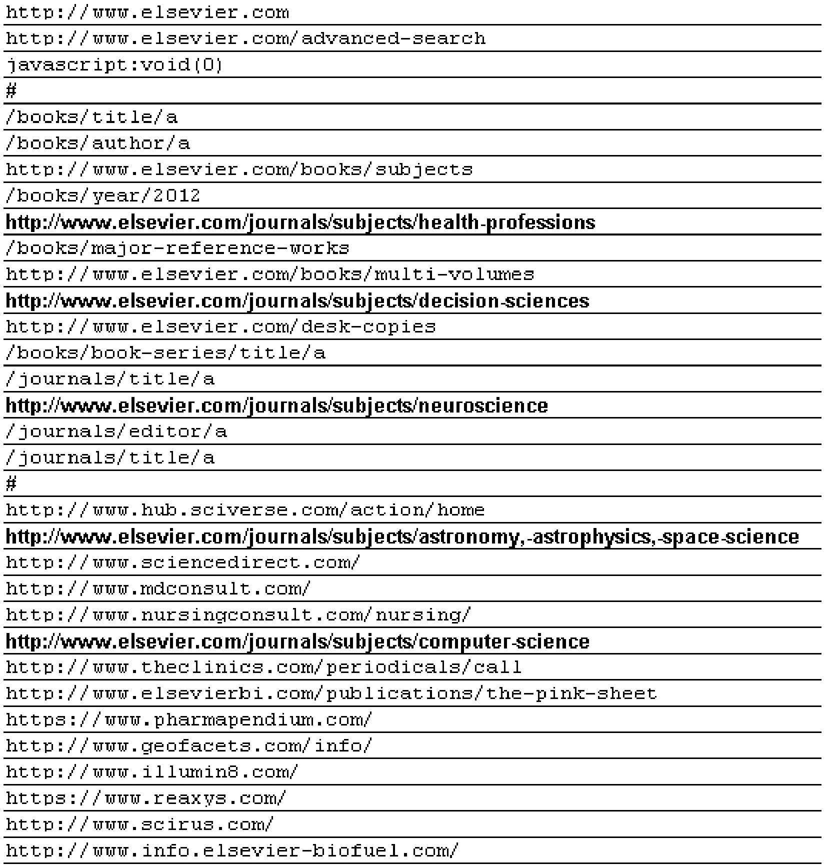
Fig. 1—List of journals under agricultural and biological sciences with each journal having corresponding URL

the database. UDMSs are a relevant and challenging example of web data extraction systems. The contributions from Doan et al.[18], describing Cimple (UDMSs developed at the University of Wisconsin) and Baumgartner et al.[19] are the two most relevant surveys on the state-of-the-art of the UDMs. A recent survey from Ferrara et. al[20] also covers the main developments. The development of UDMSs may be targeted for various important applications; however, not many reports are available on the management of unstructured data at this time.

A researcher is always interested in submitting his research work in a domain specific journal with wide readership and reputation. In most of the scientific and technical areas, a researcher would make a decision with respect to the topic profile and impact factor of the journal. The user generally has to visit individual journal's home page to access the desired information. The ever-growing trend of introducing new journals by different big and small publishers, along with the emergence of so many open-access options make it a difficult for an individual to identify the most suitable journals for submission of research article.

In this paper, we have proposed an algorithm for the selective extraction and analysis of the specific information from various journals. The user can access the relevant information about the included set of journals through a simple search string. More than 7000 journals have been so far added in the locally stored server and efforts are being made to include many more journals. As an example, all the Elsevier journals (around 2700 journals), with list of subject-wise journals has been included in the database and user can access the vital information from any journal through a single search query. (Fig. 1).

The proposed generic algorithm may provide quicker and much desired information about any journal's impact factor, publisher, ISSN etc. The tool may be useful to researchers who have to find better options for his/her research work's submission. In the process,



http://www.elsevier.com
http://www.elsevier.com/advanced-search
javascript:void(0)
#
/books/title/a
/books/author/a
http://www.elsevier.com/books/subjects
/books/year/2012
http://www.elsevier.com/journals/subjects/health-professions
/books/major-reference-works
http://www.elsevier.com/books/multi-volumes
http://www.elsevier.com/journals/subjects/decision-sciences
http://www.elsevier.com/desk-copies
/books/book-series/title/a
/journals/title/a
http://www.elsevier.com/journals/subjects/neuroscience
/journals/editor/a
/journals/title/a
#
http://www.hub.sciverse.com/action/home
http://www.elsevier.com/journals/subjects/astronomy,-astrophysics,-space-science
http://www.sciencedirect.com/
http://www.mdconsult.com/
http://www.nursingconsult.com/nursing/
http://www.elsevier.com/journals/subjects/computer-science
http://www.theclinics.com/periodicals/call
http://www.elsevierbi.com/publications/the-pink-sheet
https://www.pharmapendium.com/
http://www.geofacets.com/info/
http://www.illumin8.com/
https://www.reaxys.com/
http://www.scirus.com/
http://www.info.elsevier-biofuel.com/

Fig. 2—Pre-Phase of RetrieveURLs with all extracted URLs

large amount of unstructured data from web pages are extracted, analyzed, cleaned and organized in relational tables. The suggested method for the data extraction does not dependent upon the lists or pre-formatted templates.

Problem definition and strategy

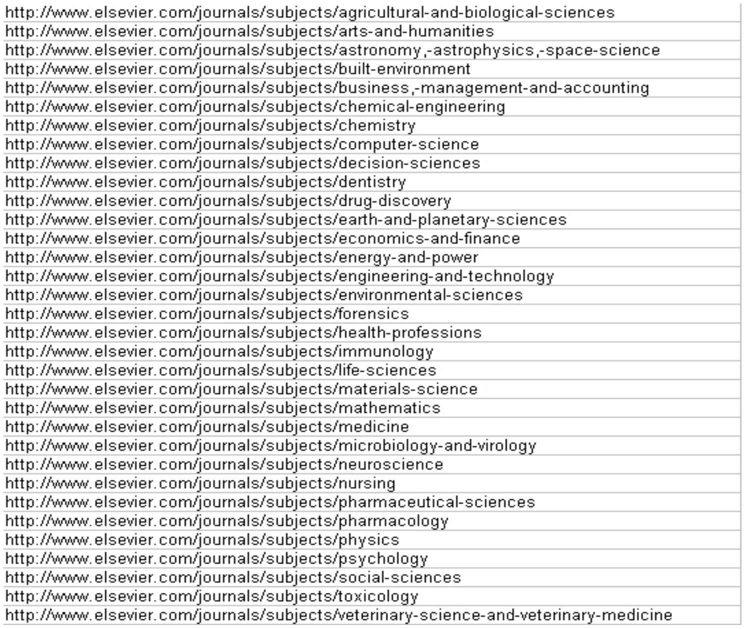
Consider an unstructured web page consisting of data values of heterogeneous nature: The data values would be extracted from attributes in order to form a relation so that the same can be grouped into tuples. The strategy for the data extraction include: Identification of the key data elements and grouping on the basis of their attributes; Pre-processing of the data elements for the identification of the noise contained in the tuples; and Processing of the elements by recursive algorithm to remove the noisy data along with tabulation of the tuples. The problem can be briefly described as “the extraction of relevant url’s from an unstructured webpage of *‘n’* html elements with *‘u’* numbers of linked url’s having associated label *‘l’*.

Algorithm overview

Herein applied algorithm performs URL extraction and filtering operations, followed by repeated iteration on the relevant URLs and finally storing the results in the database along with their descriptions. Each stored URL page is read to identify the subject URL, journal name, journal URL, which are stored in the database. Subsequently, for each stored URL, the whole webpage (including ‘about the journal’ page) is read and the impact factor is updated in the corresponding tuple. The different steps can be divided into the following phases:

Phase 1 (Extraction)

This phase extracts all the URLs from the journal domain name and filters the records based on reserved keyword (Fig 2). Each record is read and its URL is explored to search for impact factor. If found, the information is updated in the table (Fig 3), else another reserved keyword’s URL is accessed to extract the desired impact factor (Fig. 4).



```
http://www.elsevier.com/journals/subjects/agricultural-and-biological-sciences
http://www.elsevier.com/journals/subjects/arts-and-humanities
http://www.elsevier.com/journals/subjects/astronomy,-astrophysics,-space-science
http://www.elsevier.com/journals/subjects/built-environment
http://www.elsevier.com/journals/subjects/business,-management-and-accounting
http://www.elsevier.com/journals/subjects/chemical-engineering
http://www.elsevier.com/journals/subjects/chemistry
http://www.elsevier.com/journals/subjects/computer-science
http://www.elsevier.com/journals/subjects/decision-sciences
http://www.elsevier.com/journals/subjects/dentistry
http://www.elsevier.com/journals/subjects/drug-discovery
http://www.elsevier.com/journals/subjects/earth-and-planetary-sciences
http://www.elsevier.com/journals/subjects/economics-and-finance
http://www.elsevier.com/journals/subjects/energy-and-power
http://www.elsevier.com/journals/subjects/engineering-and-technology
http://www.elsevier.com/journals/subjects/environmental-sciences
http://www.elsevier.com/journals/subjects/forensics
http://www.elsevier.com/journals/subjects/health-professions
http://www.elsevier.com/journals/subjects/immunology
http://www.elsevier.com/journals/subjects/life-sciences
http://www.elsevier.com/journals/subjects/materials-science
http://www.elsevier.com/journals/subjects/mathematics
http://www.elsevier.com/journals/subjects/medicine
http://www.elsevier.com/journals/subjects/microbiology-and-virology
http://www.elsevier.com/journals/subjects/neuroscience
http://www.elsevier.com/journals/subjects/nursing
http://www.elsevier.com/journals/subjects/pharmaceutical-sciences
http://www.elsevier.com/journals/subjects/pharmacology
http://www.elsevier.com/journals/subjects/physics
http://www.elsevier.com/journals/subjects/psychology
http://www.elsevier.com/journals/subjects/social-sciences
http://www.elsevier.com/journals/subjects/toxicology
http://www.elsevier.com/journals/subjects/veterinary-science-and-veterinary-medicine
```

Fig. 3—Pre-Phase of RetrieveURLs with journals URLs

Phase 2 (Cleaning)

A table is created with attributes- Publisher, Subject, Journal Name, Journal URL and Impact Factor (Fig. 2) with candidate keys as

Publisher, Subject, Journal Name -> Journal URL

Publisher, Subject, Journal Name -> Impact Factor

A sub-set is then calculated for each tuple of the impact factor attributes and undesired entries (i.e. HTML tags) are truncated.

Phase 3 (Classification)

The resulted data is updated in the table containing information about the impact factor of each journal.

Phase 4 (Presentation)

The search parameter is passed on to the table having journal name and impact factor to retrieve the resultant impact factor along with the information about the citation index.

Extraction phase

This phase is sub-divided into two phases:-

Pre-Phase (URL Extraction)

A URL of a publisher's website is passed as parameter and the total numbers of URLs are calculated. All URLs and their corresponding labels are retrieved. When a search request is received, each URL is checked with respect to the entered query and only the relevant URLs are considered. The herein used algorithm 'RetrieveURL' extracts the URLs from the publishers' domain and then filters them according to the search string. Figure 2 shows the total number of URLs extracted and Figure 3 shows the URLs with "journal" keyword. After employing RetrieveURLs algorithm, 33 rows are added to the list out of 193 extracted URLs when passing Elsevier URL (a demonstrative parameter in the present case).

Algorithm 1 RetrieveURLs:

```
1. retriveurls (journaldomainname)
2. retrieve all URLs
3. if URL contains "journal"
4. add URLto list
5. return URL list
```

Post-Phase (Impact Factor Extraction):

Total numbers of URLs are calculated and each URL is read. The URLs without "#" in the last and satisfying the keyword "journal" are stored in the database table T1 (URL |URL Label| Publisher's Name). Subsequently, each record of the database table T1 is read to fetch further details from the individual URL and the attributes about the Journal name, Journal URL are stored in the database table T2 (URL, Journal Name, Journal URL) (Fig. 4).

<u>Algorithm 2: ReadImpactFactor</u>

```
read_impactfactor(urllist)
while(lurllist has url)
readpage(url)
end while
```

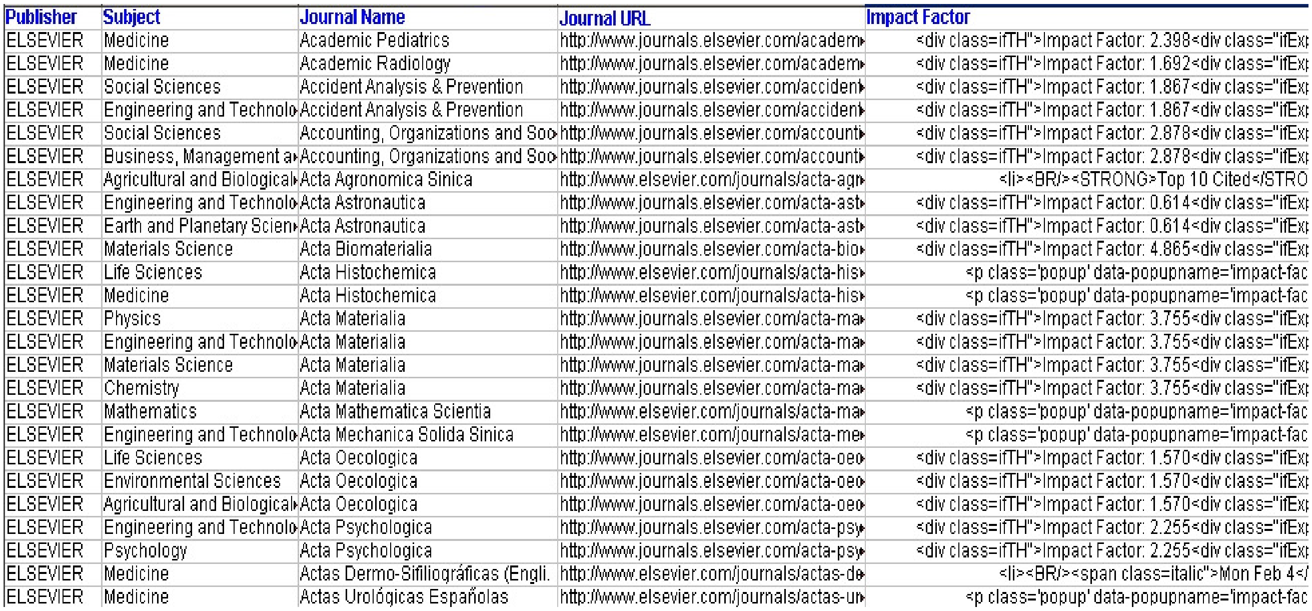
The database table T2 is read for each URL and is then updated with the attribute Impact factor, if found; Else the URL page is read for the keyword "about this journal" and the found information is updated in table T2.

<u>Algorithm 3: Read Page</u>:

```
readpage(url)
if page contains "impact factor"
boolean ischaractorfound;
int last_index=0;
int first_index= impact_factor_contain_data.first
   iidnexof("impact factor");
while(ischaractorfound)
char ch=impact_factor_contain_data.read(first_
   index)
if(ch conatin '<')
last_index=first_index;
break;
else
first_index++;
end while;
impact_factor=sub_string(impact_factor_contain_
   data(first_index,last_index)
update database impact_factor;
end if
else
readallurl(url)
if ulrlabel contains "about this journal"
readpage(url)
end if
end else
```

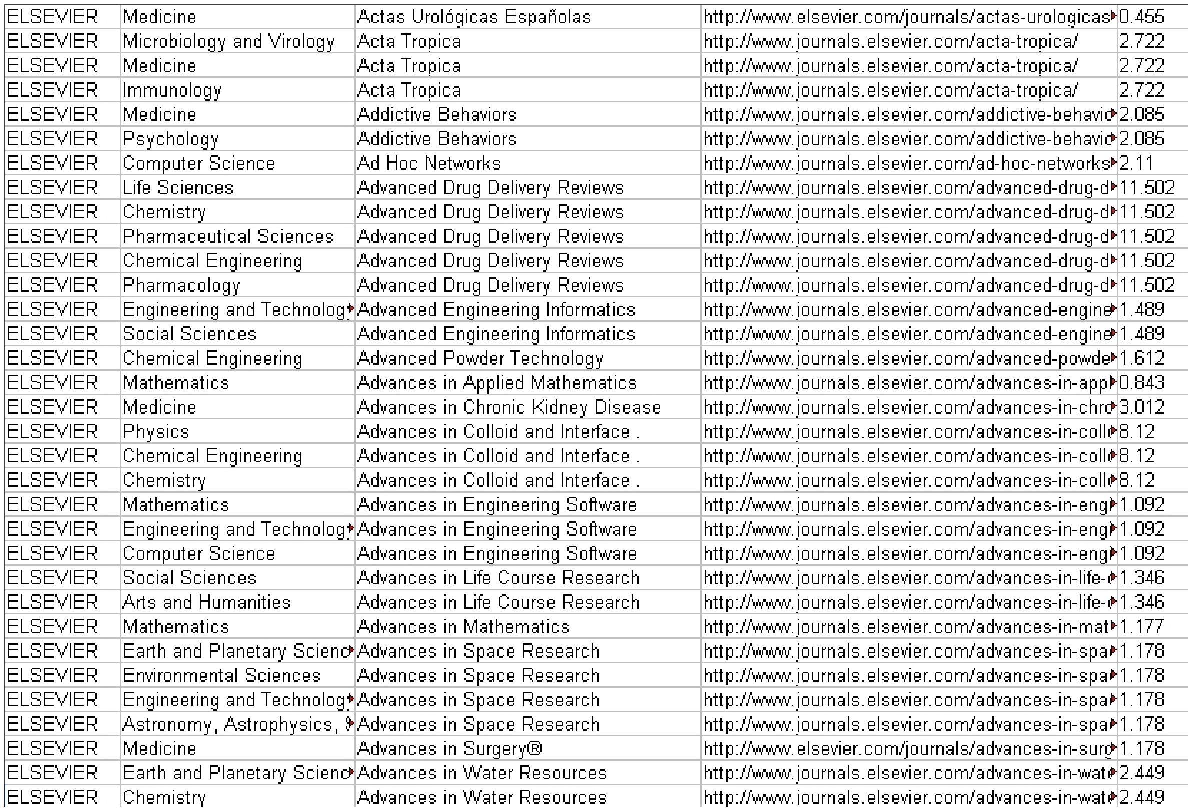
Cleaning phase

The extracted table (Fig. 5) may contain undesired noisy tuples in Impact Factor attributes, which needs to be eliminated before the table is processed for classification. In this section, we describe the method to filter out the dirty and erroneous data. It is usually a two-step process, involving the detection and correction of errors in a data set. To detect the errors, three procedures have been employed: Descriptive Statistics, Scatter Plot and Histogram. In context of the present work, the accuracy of the impact factor



Publisher	Subject	Journal Name	Journal URL	Impact Factor
ELSEVIER	Medicine	Academic Pediatrics	http://www.journals.elsevier.com/academ	<div class=ifTH">Impact Factor: 2.398<div class="ifExp
ELSEVIER	Medicine	Academic Radiology	http://www.journals.elsevier.com/academ	<div class=ifTH">Impact Factor: 1.692<div class="ifExp
ELSEVIER	Social Sciences	Accident Analysis & Prevention	http://www.journals.elsevier.com/acciden	<div class=ifTH">Impact Factor: 1.867<div class="ifExp
ELSEVIER	Engineering and Technolo	Accident Analysis & Prevention	http://www.journals.elsevier.com/acciden	<div class=ifTH">Impact Factor: 1.867<div class="ifExp
ELSEVIER	Social Sciences	Accounting, Organizations and Soc	http://www.journals.elsevier.com/account	<div class=ifTH">Impact Factor: 2.878<div class="ifExp
ELSEVIER	Business, Management a	Accounting, Organizations and Soc	http://www.journals.elsevier.com/account	<div class=ifTH">Impact Factor: 2.878<div class="ifExp
ELSEVIER	Agricultural and Biological	Acta Agronomica Sinica	http://www.elsevier.com/journals/acta-agr	<li><BR/><STRONG>Top 10 Cited</STRO
ELSEVIER	Engineering and Technolo	Acta Astronautica	http://www.journals.elsevier.com/acta-ast	<div class=ifTH">Impact Factor: 0.614<div class="ifExp
ELSEVIER	Earth and Planetary Scien	Acta Astronautica	http://www.journals.elsevier.com/acta-ast	<div class=ifTH">Impact Factor: 0.614<div class="ifExp
ELSEVIER	Materials Science	Acta Biomaterialia	http://www.journals.elsevier.com/acta-bio	<div class=ifTH">Impact Factor: 4.865<div class="ifExp
ELSEVIER	Life Sciences	Acta Histochemica	http://www.elsevier.com/journals/acta-his	<p class='popup' data-popupname='impact-fac
ELSEVIER	Medicine	Acta Histochemica	http://www.elsevier.com/journals/acta-his	<p class='popup' data-popupname='impact-fac
ELSEVIER	Physics	Acta Materialia	http://www.journals.elsevier.com/acta-ma	<div class=ifTH">Impact Factor: 3.755<div class="ifExp
ELSEVIER	Engineering and Technolo	Acta Materialia	http://www.journals.elsevier.com/acta-ma	<div class=ifTH">Impact Factor: 3.755<div class="ifExp
ELSEVIER	Materials Science	Acta Materialia	http://www.journals.elsevier.com/acta-ma	<div class=ifTH">Impact Factor: 3.755<div class="ifExp
ELSEVIER	Chemistry	Acta Materialia	http://www.journals.elsevier.com/acta-ma	<div class=ifTH">Impact Factor: 3.755<div class="ifExp
ELSEVIER	Mathematics	Acta Mathematica Scientia	http://www.elsevier.com/journals/acta-ma	<p class='popup' data-popupname='impact-fac
ELSEVIER	Engineering and Technolo	Acta Mechanica Solida Sinica	http://www.elsevier.com/journals/acta-me	<p class='popup' data-popupname='impact-fac
ELSEVIER	Life Sciences	Acta Oecologica	http://www.journals.elsevier.com/acta-oeo	<div class=ifTH">Impact Factor: 1.570<div class="ifExp
ELSEVIER	Environmental Sciences	Acta Oecologica	http://www.journals.elsevier.com/acta-oeo	<div class=ifTH">Impact Factor: 1.570<div class="ifExp
ELSEVIER	Agricultural and Biological	Acta Oecologica	http://www.journals.elsevier.com/acta-oeo	<div class=ifTH">Impact Factor: 1.570<div class="ifExp
ELSEVIER	Engineering and Technolo	Acta Psychologica	http://www.journals.elsevier.com/acta-psy	<div class=ifTH">Impact Factor: 2.255<div class="ifExp
ELSEVIER	Psychology	Acta Psychologica	http://www.journals.elsevier.com/acta-psy	<div class=ifTH">Impact Factor: 2.255<div class="ifExp
ELSEVIER	Medicine	Actas Dermo-Sifiliográficas (Engli.	http://www.elsevier.com/journals/actas-de	<li><BR/><span class=italic">Mon Feb 4</
ELSEVIER	Medicine	Actas Urológicas Españolas	http://www.elsevier.com/journals/actas-ur	<p class='popup' data-popupname='impact-fac

Fig. 4—Retrieval of impact factor from journals' URLs



ELSEVIER	Medicine	Actas Urológicas Españolas	http://www.elsevier.com/journals/actas-urologicas	0.455
ELSEVIER	Microbiology and Virology	Acta Tropica	http://www.journals.elsevier.com/acta-tropica/	2.722
ELSEVIER	Medicine	Acta Tropica	http://www.journals.elsevier.com/acta-tropica/	2.722
ELSEVIER	Immunology	Acta Tropica	http://www.journals.elsevier.com/acta-tropica/	2.722
ELSEVIER	Medicine	Addictive Behaviors	http://www.journals.elsevier.com/addictive-behavio	2.085
ELSEVIER	Psychology	Addictive Behaviors	http://www.journals.elsevier.com/addictive-behavio	2.085
ELSEVIER	Computer Science	Ad Hoc Networks	http://www.journals.elsevier.com/ad-hoc-networks	2.11
ELSEVIER	Life Sciences	Advanced Drug Delivery Reviews	http://www.journals.elsevier.com/advanced-drug-d	11.502
ELSEVIER	Chemistry	Advanced Drug Delivery Reviews	http://www.journals.elsevier.com/advanced-drug-d	11.502
ELSEVIER	Pharmaceutical Sciences	Advanced Drug Delivery Reviews	http://www.journals.elsevier.com/advanced-drug-d	11.502
ELSEVIER	Chemical Engineering	Advanced Drug Delivery Reviews	http://www.journals.elsevier.com/advanced-drug-d	11.502
ELSEVIER	Pharmacology	Advanced Drug Delivery Reviews	http://www.journals.elsevier.com/advanced-drug-d	11.502
ELSEVIER	Engineering and Technolog	Advanced Engineering Informatics	http://www.journals.elsevier.com/advanced-engine	1.489
ELSEVIER	Social Sciences	Advanced Engineering Informatics	http://www.journals.elsevier.com/advanced-engine	1.489
ELSEVIER	Chemical Engineering	Advanced Powder Technology	http://www.journals.elsevier.com/advanced-powde	1.612
ELSEVIER	Mathematics	Advances in Applied Mathematics	http://www.journals.elsevier.com/advances-in-app	0.843
ELSEVIER	Medicine	Advances in Chronic Kidney Disease	http://www.journals.elsevier.com/advances-in-chro	3.012
ELSEVIER	Physics	Advances in Colloid and Interface .	http://www.journals.elsevier.com/advances-in-coll	8.12
ELSEVIER	Chemical Engineering	Advances in Colloid and Interface .	http://www.journals.elsevier.com/advances-in-coll	8.12
ELSEVIER	Chemistry	Advances in Colloid and Interface .	http://www.journals.elsevier.com/advances-in-coll	8.12
ELSEVIER	Mathematics	Advances in Engineering Software	http://www.journals.elsevier.com/advances-in-eng	1.092
ELSEVIER	Engineering and Technolog	Advances in Engineering Software	http://www.journals.elsevier.com/advances-in-eng	1.092
ELSEVIER	Computer Science	Advances in Engineering Software	http://www.journals.elsevier.com/advances-in-eng	1.092
ELSEVIER	Social Sciences	Advances in Life Course Research	http://www.journals.elsevier.com/advances-in-life-	1.346
ELSEVIER	Arts and Humanities	Advances in Life Course Research	http://www.journals.elsevier.com/advances-in-life-	1.346
ELSEVIER	Mathematics	Advances in Mathematics	http://www.journals.elsevier.com/advances-in-mat	1.177
ELSEVIER	Earth and Planetary Scienc	Advances in Space Research	http://www.journals.elsevier.com/advances-in-spa	1.178
ELSEVIER	Environmental Sciences	Advances in Space Research	http://www.journals.elsevier.com/advances-in-spa	1.178
ELSEVIER	Engineering and Technolog	Advances in Space Research	http://www.journals.elsevier.com/advances-in-spa	1.178
ELSEVIER	Astronomy, Astrophysics,	Advances in Space Research	http://www.journals.elsevier.com/advances-in-spa	1.178
ELSEVIER	Medicine	Advances in Surgery®	http://www.elsevier.com/journals/advances-in-surg	1.178
ELSEVIER	Earth and Planetary Scienc	Advances in Water Resources	http://www.journals.elsevier.com/advances-in-wat	2.449
ELSEVIER	Chemistry	Advances in Water Resources	http://www.journals.elsevier.com/advances-in-wat	2.449

Fig. 5—Extracted table with impact factor

(which can range from 0 to maximum) needs to be verified. The standard deviation has been employed to detect the erroneous data. Fig. 6 gives the information about the scattering of the desired 'impact factor' data with respect to the different journals searched. It can be deduced that the desired values are within the reasonable and acceptable ranges. Some manual checking also confirmed that the extracted impact factor data were in coherence with the actual values.

Presentation

Fig. 5 contains the attributes viz. Publisher, Subject, Journal Name, Journal URL and Impact Factor. Extraction and Presentation algorithm were implemented in Java. The interface for accessing the table and presentation of the desired result provides the convenient space for entering the search parameter. The web request is passed to the table and the corresponding impact factors are displayed (Fig. 7).

Experiments

Necessary experiments have been performed to study the accuracy and performance of the three extraction algorithms i.e RetrieveURLs, ReadImpactFactor and ReadPage algorithm. After applying the said algorithms, the record set is cleansed. Passing of a search parameter on the web server through HTTP service presents the result of the desired query. Finally, the obtained result is compared with the manually searched documents on the actual Elsevier domain.

Setup

The whole experimental setup for the Elsevier publications is shown as Fig. 8.

We consider the JList, that holds Journal URLs along with their labels, and IFList, that holds the details about publisher (in this case Elsevier), journal, subject and impact factor. A URL address of the publisher's domain (http://www.elsevier.com/journals/subjects#) works as a parameter to retrieve all the contained URLs from the web page. These retrieved URLs are added to JList and the algorithm *RetrieveURLs* then filters the obtained record sets which match with the desired journal hyperlink. Simultaneously, the URLs

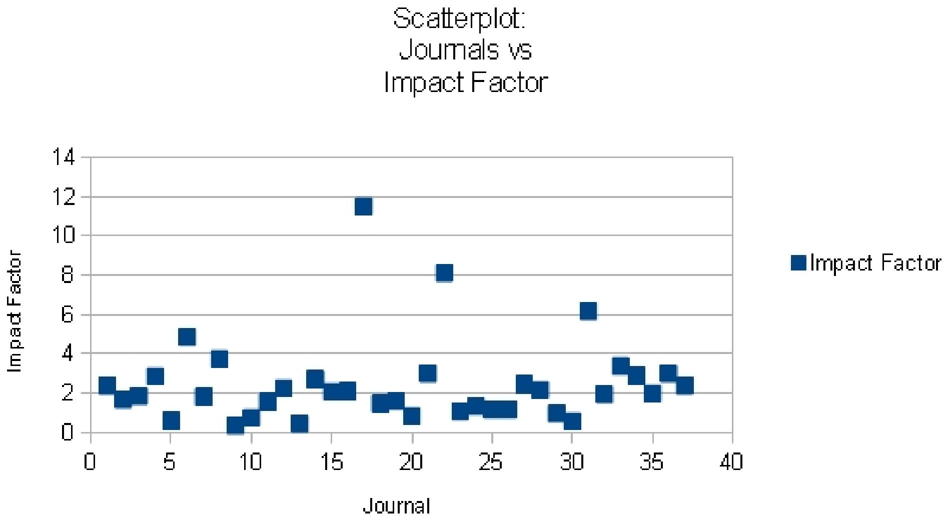
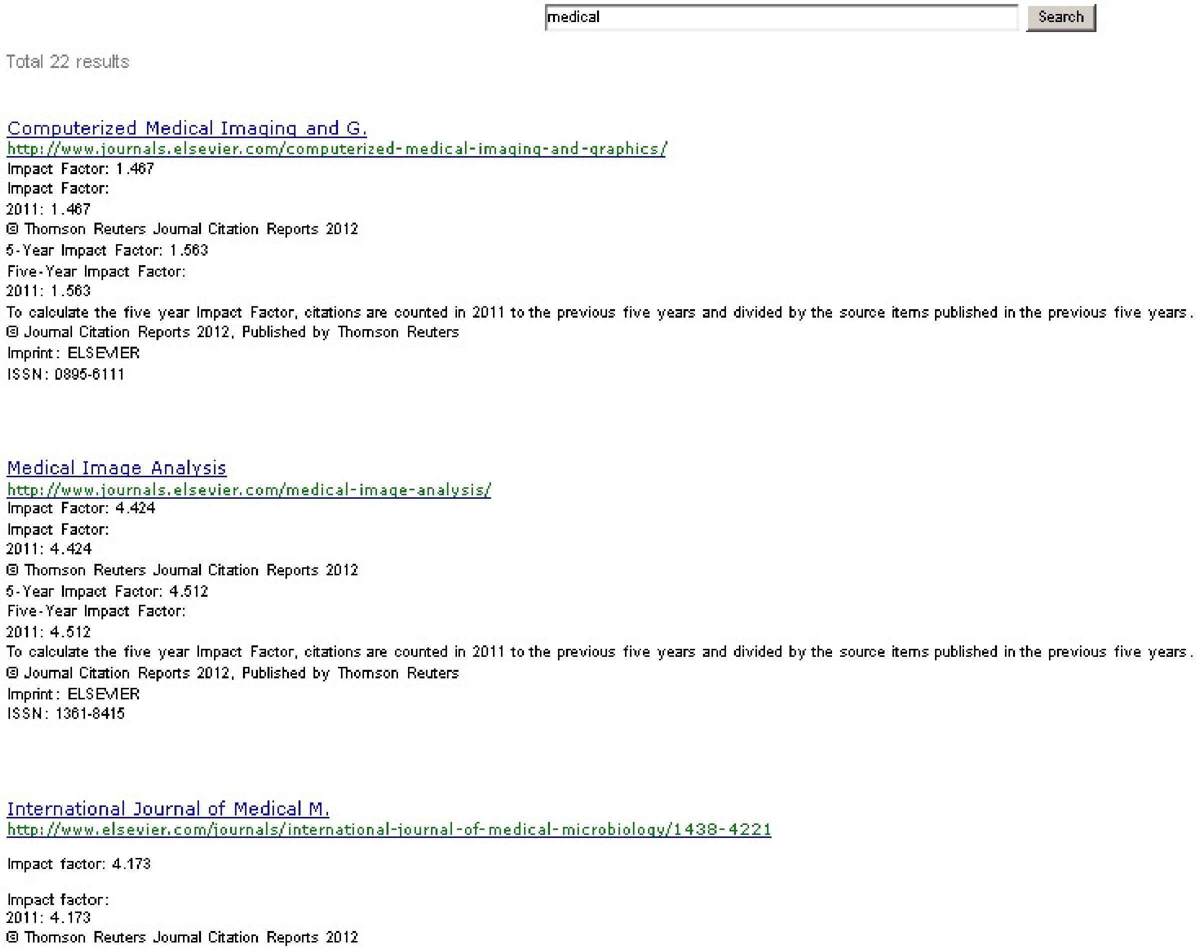


Fig. 6—Scatter plot of Journal vs. impact factor



Total 22 results

Computerized Medical Imaging and G.
http://www.journals.elsevier.com/computerized-medical-imaging-and-graphics/
Impact Factor: 1.467
Impact Factor:
2011: 1.467
© Thomson Reuters Journal Citation Reports 2012
5-Year Impact Factor: 1.563
Five-Year Impact Factor:
2011: 1.563
To calculate the five year Impact Factor, citations are counted in 2011 to the previous five years and divided by the source items published in the previous five years.
© Journal Citation Reports 2012, Published by Thomson Reuters
Imprint: ELSEVIER
ISSN: 0895-6111

Medical Image Analysis
http://www.journals.elsevier.com/medical-image-analysis/
Impact Factor: 4.424
Impact Factor:
2011: 4.424
© Thomson Reuters Journal Citation Reports 2012
5-Year Impact Factor: 4.512
Five-Year Impact Factor:
2011: 4.512
To calculate the five year Impact Factor, citations are counted in 2011 to the previous five years and divided by the source items published in the previous five years.
© Journal Citation Reports 2012, Published by Thomson Reuters
Imprint: ELSEVIER
ISSN: 1361-8415

International Journal of Medical M.
http://www.elsevier.com/journals/international-journal-of-medical-microbiology/1438-4221

Impact factor: 4.173

Impact factor:
2011: 4.173
© Thomson Reuters Journal Citation Reports 2012

Fig 7—Demonstration of user interface for the search of impact factors

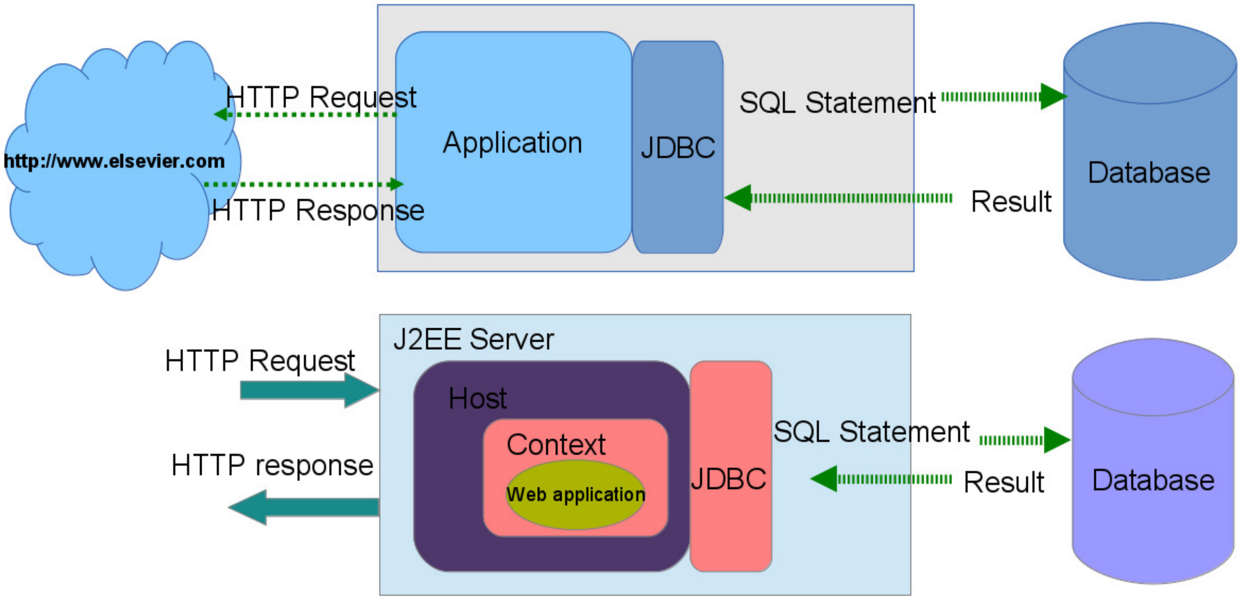


Fig. 8—Experimental setup for data retrieval and presentation

not having the keyword "journal" are dropped. The total number of hyperlinks on the domain web page is 193 and only 33 relevant record sets are added to the JList. These 33 records denote the actual number of subjects in which journals are grouped. JList contains the URL, caption of the hyperlink and attributes. The same can be presented as JList(Publisher, Subjects, SubjectURL).

The IFList (Publisher, Subject, Journal Name, Journal URL, Impact Factor) is then populated by importing the data from the JList URLs. The collected information contains Journal name and URL, publisher's name and subject. At this stage the attribute Impact factor remains NULL. The total number of tuples (including duplicates) extracted from web pages and added to the IFList is 4438 out of 7492 tuples. Finally, each web page from the IFList is parsed with the individual Journal URLs and the attribute Impact factor is updated in the list. The total number of journals from which the impact factors could be extracted is 3052. This is 68.76% of the total number of journals available.

Overall performance

- Performance ratio vis-à-vis the total number of subjects listed on the Elsevier and the extraction performed herein in JList: 1
- Performance ratio vis-à-vis the total number of Journals on the Elsevier domain and the extraction performed herein in IFList: 0.98
- Performance ratio vis-à-vis the total number of Journals on the Elsevier domain and the Impact factor extraction: 0.68.

The seemingly incomplete extraction of Impact factor values is mainly due to the fact a large numbers of journals might not have been indexed for this particular parameter, and thus not having the values on the web pages.

Application of extraction algorithm

Extraction algorithm is a generic algorithm to extract unstructured data from publisher's domain such as Elsevier. The presented impact factor data in this paper has been retrieved only for the journals available in the Elsevier domain. However, the said algorithm can also extract the impact factors from other publisher's domain. One can also use the proposed algorithm for the indexing of the followings:

Citation Index: As with the case of impact factor, we can extract the total citation index of the journals from unstructured data by parsing the web pages available in the publisher's domain.

Intelligent analysis for researcher: A researcher can analyze the journals along with their impact factor subject-wise that helps in seeking the right journal to publish the paper without visiting each publisher's domain and journals.

Conclusions

The developed extraction algorithm is able to extract unstructured data from multiple publishers' domain. The proposed generic algorithm would not require major changes for the comprehensive analysis of various journals. This has been experimentally validated by applying the extraction algorithm on Elsevier domain. The authors have also successfully tested the process for other publishers' domain, such as Springer, Nature, American Chemical Society, Wiley, etc. It has been possible to create a database of more than 10,000 journals with filtered information on the impact factor and Journal subject. The extraction algorithm can be further extended to deduce other web information such as average impact factor of previous years and citation index that may further extend the utility of the proposed algorithm.

Acknowledgements

Authors are thankful to the Director, CSIR-CSIO for providing necessary facilities.

References

1. Baumgartner R, Gatterbauer W and Gottlob G, *Encyclopedia of Database Systems*, Springer (2009), pp. 3465-3471.
2. Alberto H F Laender, Berthier A Ribeiro-Neto, Altigran S and Juliana S, *A brief survey of web data extraction tools*. SIGMOD Record, 31 (2002) 84-93.
3. Irmak U and Suel T, *Interactive wrapper generation with minimal user effort*. In Proc. 15th International Conference on World Wide Web, May 23-36, 2006, Edinburgh, Scotland, (2006) 553-563, doi:10.1145/1135777.1135859.
4. Zhao H, *Automatic wrapper generation for the extraction of search result records from search engines*. PhD thesis, State University of New York at Binghamton, Binghamton, NY, USA (2007)
5. Baumgartner R, Frolich O, Gottlob G, Harz P, Herzog, Herzog M, Lehmann P and Wien T, *Web data extraction for business intelligence*, In Proc. 12th Conference on Datenbanksysteme in Business, Technologie and Web, March 2-4, 2005, Karlsruhe, Germany, (2005) 48-65.
6. Salvatore A Catanese, De Meo P, Ferrara E, Fiumara G and Provetti A, *Crawling facebook for social network analysis purposes*. In Proc. International Conference on Web Intelligence, Mining and Semantics, Sogndal, Norway, May 25-27, (2011) 52.
7. Gjoka M, Kurant M, Butts C T and Markopoulou A, *Walking in Facebook: a case study of unbiased sampling of OSNs*. In Proc. 29th Conference on Information Communications, March 15-19, 2010, San Diego, CAA, USA, (2010) 2498-2506.
8. Plake C, Schiemann T, Pankalla M, Hakenberg J and Leser U, *Alibaba: Pubmed as a graph*. Bioinformatics, 22(19) (2006) 2444-2445.
9. Laender A H F, Ribeiro-Neto BA, da Silva A S and Teixeira J S, A brief survey of web data extraction tools. *SIGMOD Record*, 31(2) (2002) 84-93,
10. Kushmerick N, Finite-state approaches to web information extraction. In: Pazienza, M.T. (Eds.) *Information Extraction in the Web Era*, Springer (2003) pp. 77-91,
11. Flesca S, Manco G, Masciari E, Rende E and Tagarelli A, Web wrapper induction: a brief survey. *AI Communications*, 17(2) (2004) 57-61.
12. Kaiser K, Miksch S, *Information extraction a survey Technical report*, E188 - Institute of Software Technology and Interactive System (2003-2013).
13. *Softwaretechnik und Interaktive Systeme*; Technische Universitat. Wien, Germany (2005).
14. Chang C, Kayed M, Girgis M R and Shaalan K F, A survey of web information extraction systems. *IEEE Transactions on Knowledge and Data Engineering*, 18 (10) (2006) 1411-1428.
15. Fiumara G, *Automated information extraction from web sources: a survey*. In Proc. of Between Ontologies and Folksonomies Workshop, June 28, 2007, Michigan, USA, (2007) 1-9.
16. Sarawagi S, Information extraction. *Foundations and Trends in Databases*, 1(3) (2008) 261-377.
17. Arasu A and Garcia-Molina H, *Extracting structured data from web pages*, SIGMOD, (2003) 337-348.
18. Elmeleegy H, Madhavan J and Halevy A, Harvesting relational tables from lists on the web. *The VLDB Journal*, 20 (2) (2011) 209-226.
19. Doan A H, Naughton J F, Ramakrishnan R, Baid A, Chai X, Chen F, Chen T, Chu E, DeRose P and Gao B, Information extraction challenges in managing unstructured data, *ACM SIGMOD Record*, 37(4) (2009) 14-20.
20. Ferrara E, De Meo P, Fiumara G and Baumgartner R, Web data extraction, applications and techniques: a survey, Cornell University Library, arXiv: (2013), 1207.0246v2.