Nilesh Jain et al. International Journal of Recent Research Aspects ISSN: 2349-7688, Vol. 5, Issue 2, June 2018, pp. 36-42

A Dynamic Approach to Identify Page Score for Research Papers to Improve Page Ranking

Nilesh Jain[1] and Dr. Vijay Singh Rathore[2]

[1]Research Scholar, Mewar University, NH - 79 Gangrar, Chittorgarh, Rajasthan-312 901
[2]Professor & Director, Shree Karni College Jaipur (Raj.)

Abstract: **With a growing amount of data in multiple domains, users have access to loads of data. Retrieving information from the data is a challenge, when data is un/semi-structured. An additional issue is the relation between information in various domains, given domain restrictions. The collection of information becomes very hard to find, extract, filter or evaluate the relevant information for the users. In this paper, we have studied the basic concepts of web mining, classification, processes and issues. In addition, this paper works analyzed that are published through researchers in numerous conferences and journals, which are complete publicaly presented by the publishers. Although such research is spread everywhere world, there must be a targeted sample for this, which endorse that quite a lot of areas observes particular trend in the direction of specified research direction. In this paper, we have plotted a technique to discover influential researcher for a specific input document. The proposed method has steps series to procedure out influential researcher. Clustering procedure is subjected for recognizing similar referral documents. Clustering procedures in describe with semantic similarity measure that will help to similar documents extract and their data. This will increase clustering procedure. Ultimately, queries regarding meant research are subjected and compared with the parameter score and clusters is removed. Then we list the authors with higher parameter score. The relevant authors are selected based on the parameters score. A number of documents are selected for the experimentation process, in order to evaluate the efficiency of the proposed approach. The experimental evaluation showed that the proposed process is effective processing the influential researcher. Study shows that the accuracy of recommendation system has improved significantly with the use of the proposed technique.**

General Terms: **Web Mining, Page Rank, Score, Research, Authors.**

I. INTRODUCTION

The web mining technology is an important data mining branch. The Web contains huge quantity of information, applying data mining technology on Web, namely the Web mining technology, becomes the most important research along with fast internet development [1].

Research is an on-going process which results in writing Research publications that are shared with the world in different conferences. Research publication header information that includes Conference name in which paper was published, Title of paper, Author(s) of paper, Affiliations of author(s), Email of Author(s), Keywords in paper and Abstract of paper. This knowledge, when removed can be very valuable in numerous data mining scenarios such as one can find collaboration among various universities through looking for publication which have authors from two or more than two different universities.

Another scenario can be one can find trends of research done in particular field or area. The typical issue faced in removing data from research papers is that they do not have usual format describe in which research publications are written. Typically all journal or conference has its own format for writing research publication. Information extraction from research paper is complete through three methods. First approach is to do structural analysis of PDF along with pattern matching. Such as, matching for words like Keyword and Abstract and looking for font size to extract Title as title is typically of larger font size. This approach overall is not very efficient due to large number of paper structural formats in which papers are published. Another approach used is Web based lookup from a knowledge base.

II. RELATED WORK

Most existing keyphrase extraction techniques used only the textual content of the target document [2]. Recently, Wan and Xiao [3] addressed this simplification using a model that incorporates a local neighborhood of a document for extracting keyphrases. We posit that, in adding to a document's textual content and textually -similar neighbors, other useful neighborhoods exist in research document collections that have the potential to improve keyphrase extraction.

In a citation network, information flows from one paper to another via the citation relation[4]. This information flow as well as the influence of one paper on another is specifically captured by means of citation contexts. Keyphrase removal was previously studied applying both unsupervised and supervised methods for various kinds of documents concluding scientific abstracts, newswire documents, webpages and meeting transcripts.

The SemEval 2010 Shared Task was focused on comparing keyphrase removal for scientific articles [4], indicating once again significance of this problem. Supervised methods use annotated documents with "correct" keyphrases to train classifiers for discriminating keyphrases extracted from a document [5,6]. In unsupervised keyphrase removal, domain-

specific knowledge and numerous measures for example term frequencies, inverse document frequencies, topic proportions, etc. are used to score terms in a document that are later aggregated to find scores for phrases [7,8].

The PageRank algorithm is widely-used in keyphrase extraction models. Other centrality measures such as betweenness and degree centrality were also previously studied for keyphrase extraction. However, based on current experiments, PageRank family of approaches and tf-idf based scoring can be considered the state-of-the-art for unsupervised keyphrase removal [9].

TextRank was made for scoring key phrases using the PageRank values obtained on a word graph built from the adjacent words in a document. Wan and Xiao extended the TextRank approach to Single Rank by adding edges between words within a window size greater than 2 and edge weights in the graph based on co-occurrence between words. Unlike the TextRank and Single Rank models, where only the content of the target document are used for key phrase extraction, textually-similar neighboring documents are included in the scoring procedure in Expand Rank. In contrast to the approaches above, we present a model for key phrase extraction from research papers that are embedded in citation networks. The underlying algorithm of our model is PageRank applied to word graphs constructed from target papers and their local neighborhood in a citation network. In addition, unlike the approaches so far, our model incorporates multiple neighborhoods and includes a flexible way to incorporate different weights for each neighborhood [10]. Similar to key phrase extraction, tag recommendation systems are designed to predict descriptive terms or tags for organizing and sharing Web resources.

Yao et.,al.[11] identified GROBID, ParsCit, Mendeley, HeaderParserService, PDFSSA4MET, PDFMEAT, Zotero and PaperPile. Generation of Bibliographic Data(GROBID)[12] has used Conditional Random Fields machine learning algorithm that is implemented using MALLET[13] to extracts the bibliographical data corresponding to the header data and references. Parse Citation (ParsCit) [14] achieves Reference string parsing and logical structure scientific documents parsing. Mendeley [15] uses SVM and Web based lookup for embedded metadata and citation extraction removal details from research publication.

Mendeley provides an application based on windows that helps in organizing and research publication collaboration. Mendeley classifies Author, Title, Journal, Year and keywords. PDF Structure and Syntactic Analysis for Metadata Extraction and Tagging [8] (PDFSSA4MET) provide metadata extraction and tagging based on the syntactic and structural analysis of research publication. PDFFSSA4MET extracts and tag Title, Author, Section headings and references. PDFMEAT[16], Zotero[17] and PaperPile[18] do web based lookup for getting Research publication information. Yao et.,al[11] proposed a framework that will call PDFSSA4MET, ParsCit, HeaderParserService and PDFMEAT from command line and generate a uniform result set from it.

III. PROPOSED WORK

We have proposed an algorithm which works on pattern discovery and pattern analysis phases of data processing. The proposed work gives the user-based idea about the web og that is used. The whole web log is studied and the results are obtained on user basis. The user- based calculation makes it more precise and usable for the users. The whole proposed process is explained here with the help of a flowchart.

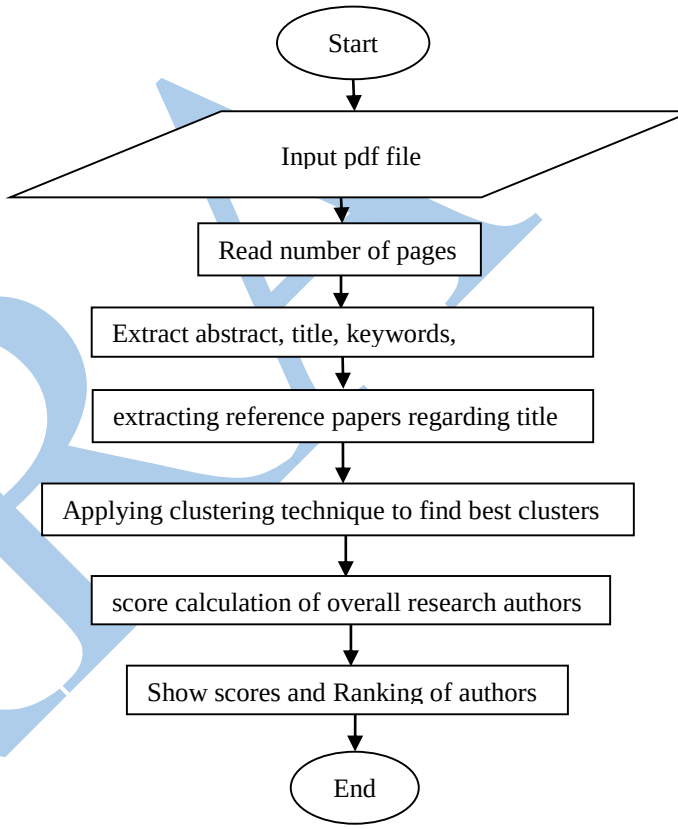

Fig-1: Flowchart of Proposed Algorithm

This flowchart explains the whole process of the proposed algorithm. The pseudocode for this algorithm has been explained in detail below:

1. Input research file name.
2. Read number of pages in the pdf file using itextpdf.
    a. To get the number of pages in the research paper
    No_pages = getNumberOfPages()
3. Extract abstract, title, keywords, references from the pfd file.
4. For i = 1:length(references)
    a. Count reference numbers.
    b. Segregate all references with their names and titles.
    c. Extract paper for each reference name.
    End.
5. for each author= 1:50
    a. Authors' names and their papers are shown in the output window.

b. Search author name = "author name"+ "google scholar citations".
c. Store the data in a text file as snippet.txt.
d. read(snippet.txt) and find authors' names and their links .

End for.

6. for i=1: length(title)
   a. read title of each paper.
   b. if(number_words_matching >= 3)
      i. choose the title and the research paper as the matching paper of the research topic.
      ii. Extract the papers of these titles and authors.
      iii. Extract the url of the paper.

      end.
   c. Use the url to find papers title, conference/journal, citations and year.
   d. Read abstract of such papers.

   end.
7. Applying k-means for clustering the papers.
8. for i=1:count(papers)
   a. no.of clusters = 5.
   b. Select data points for cluster formation which is given by

$$datavalue = \frac{P(ab_{input}|ab_{reference})}{P(ab_{reference})}$$

   c. Cluster the papers based on the centroid and eucledian distance

$$dist(c_{cluster}, c_{input}) = eucledian distance(c_{cluster}, c_{input})$$

   d. Best clusters are formed for the research papers to find the best influential researcher.

   end.
9. Score calculation is done based on: citations, journals, conferences, articles.
10. for j=1:length(papers)
    a. N(j)=sum(journalbyauthor)
    b. N(conf)= sum(confbyauthor)
    c. N(cit)= sum(N(j),N(conf))
    d. N(art)= sum_of_citations(author's_name)

    end.
11. $parameter_{score} = \frac{N(j)}{J} + \frac{N(conf)}{C} + \frac{N(art)}{Art} + \frac{N(cit)}{cit}$
12. Thus parameter score are calculated for various papers.
13. End.

IV. RESULT ANALYSIS

The proposed approach is implemented in java programming language with JDK 1.7.0. The program is implemented in a system with processor; inter core i5, a RAM of 4GB and hard disk space of 500 GB. The experimental section plotted in the following section discusses the responses of 5 input documents. The documents will processed through the different phases of the proposed approach and their responses are discussed. The documents are labeled as doc1, doc2 and doc3 for the ease of use. Each of the documents contains more than 20 references and among those one is influential researcher.

The results are generated from the code by adding a pdf file and start reading the file. The process flows like:

1. On clicking the run button, the "INFLUENTIAL RESEARCH FINDER" window, which is the first screen for the input.

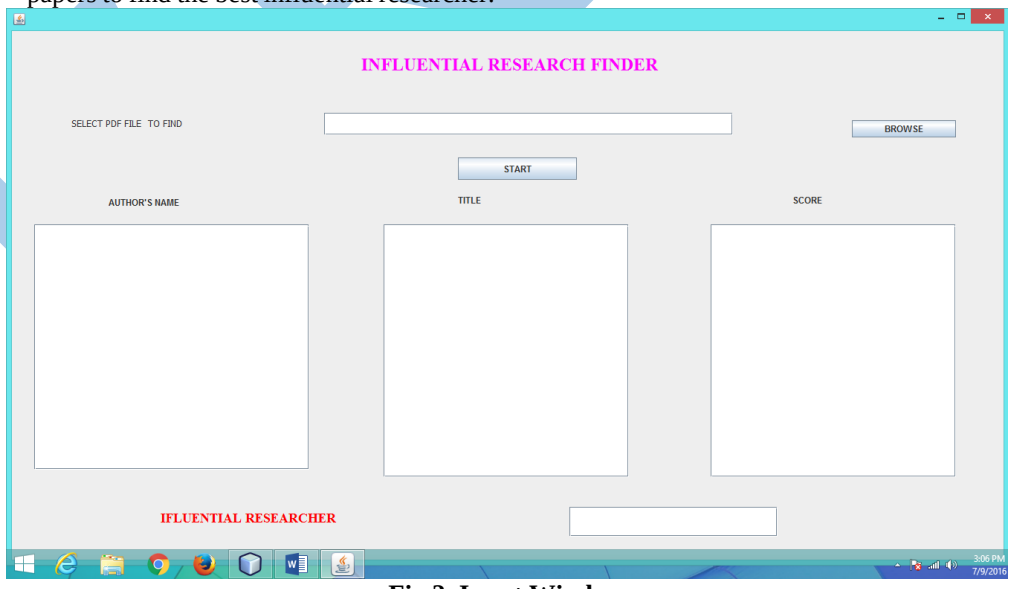

Fig 2. Input Window

2. Next step is to select the research paper in pdf format and upload it for the processing. The windows shows the paper selected:

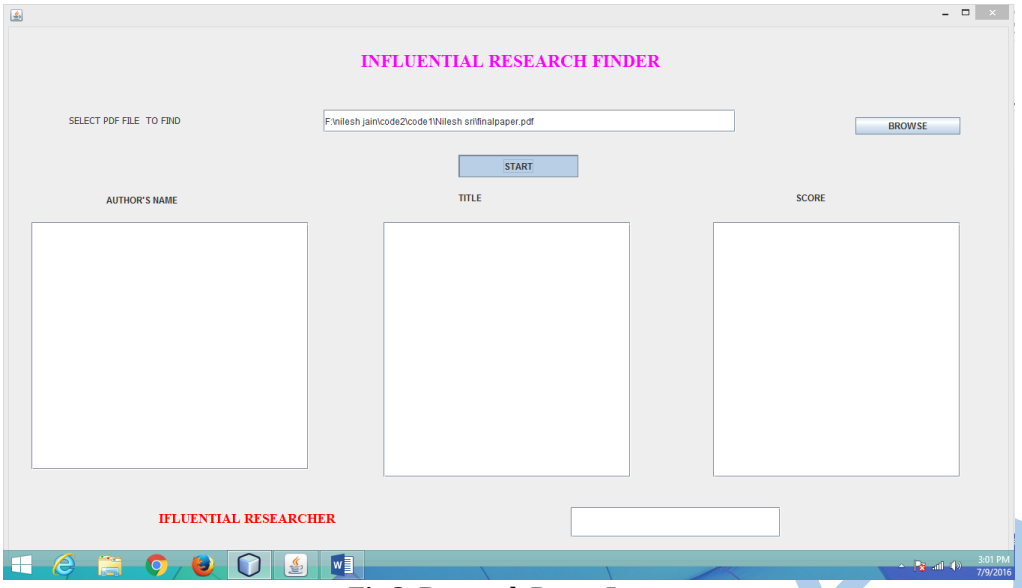

Fig 3. Research Paper Input

3. The processing of the paper runs in background which is prompted on the command window. The result shows the title name of the paper, total no. of pages and the references of the paper extracted. The results are displayed like:

This PDF has 26 pages.

Title ARTIFICIAL NEURAL NETWORK-BASED MERGING SCORE FOR META SEARCH ENGINE

Metasearch engine, neural network, retrieval of documents, ranking list

****

Leonidas Akritidis, Dimitrios Katsaros and Panayiotis Bozanis, Effective rank aggregation for metasearching

Hideaki Ishii, Roberto Tempo and Er-Wei Bai, A Web Aggregation Approach for Distributed Randomized PageRank Algorithms

Amir Hosei, Keyhanipour, Behzad Moshiri, Majid Kazemian, Maryam Piroozmand and Caro Lucas, Aggregation of web search engines based on users' preferences in WebFusion

Gholam R, Amin and Ali Emrouznejad, Optimizing search engines results using linear programming

Lin Li, Guandong Xu, Yanchun Zhang and Masaru Kitsuregawa, Random walk based rank aggregation to improving web search

Qizhi Fang, Han Xiao and Shanfeng Zhu, Top-d Rank Aggregation in Web Meta-search Engine

Lin Li, Zhenglu Yang and Masaru Kitsuregawa, Using Ontology-Based User Preferences to Aggregate Rank Lists in Web Search

Sugiura A., Etzioni O., Query routing for Web search engines: architecture and experiments

Manning C.D., Raghavan P., Schutze H., Introduction to Information Retrieval

Meng W., Yu C., Liu K.-L., Building efficient and effective metasearch engines

Spink A., Jansen B.J., Blakely C., Koshman, S., Overlap among major Web search engines

Aslam J.A., Montague M.H., Metasearch consistency

Vogt C.C., Adaptive combination of evidence for information retrieval

Dwork C., Kumar R., Naor M., Sivakumar D., Rank aggregation methods for the Web

Ailon N., Charikar M., Newman A., Aggregating inconsistent information: ranking and clustering

Fig 4. Processing of Research Paper

4. The next step shows the total references that the research paper consists and related research papers of the topic of the authors which match them.

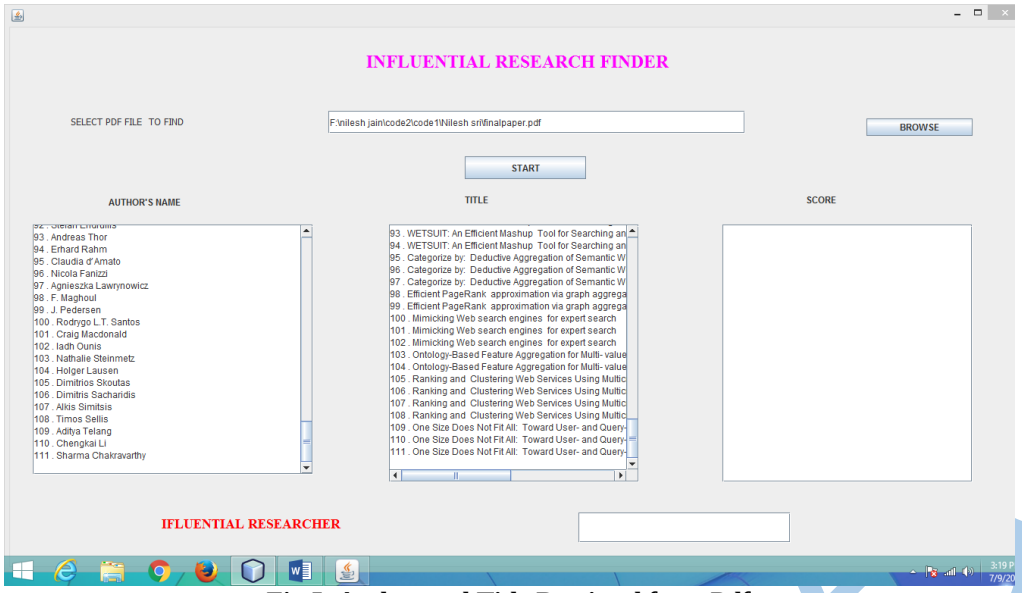

Fig 5. Author and Title Retrieval from Pdf

5. The last window after processing displays the final result which is score and the influential researcher i.e. the researcher having the highest score value.

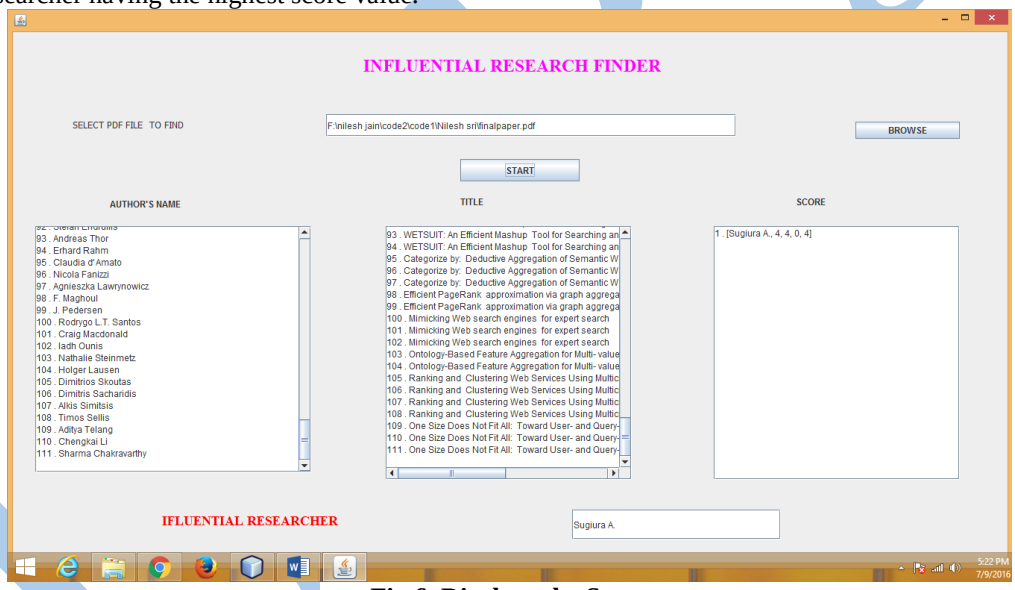

Fig 6. Displays the Score

ILLUSTRATION USING EXAMPLE

Explaining the proposed work using an example will make it better to understand.

Responses regarding Doc 1.

The document 1 is published by the author Hung Yi Lin [224] on the topic, Efficient and compact indexing structure for processing of spatial queries in line-based databases. The proposed approaches accept the doc 1 as input and initiate the text processing methods. The initial step generates the parameter table for the doc 1. The parameter table for the doc 1 is presented below and we have listed top 5 references from the list.

Sl. No	Authors	Tittle	Publication
1	J.L. Bentley	Multidimensional binary search trees used for associative searching	ACM
2	H. Blanken, A. Ijbema, P. Meek, B. Akker	The generalized grid file: description and performance aspects	IEEE
3	E.I. Chong, J. Srinivasan, S. Das, C. Freiwald, A. Yalamanchi, M. Jagannath, A.T. Tran, R. Krishnan, R. Jiang	E.I. Chong, J. Srinivasan, S. Das, C. Freiwald, A. Yalamanchi, M. Jagannath, A.T. Tran, R. Krishnan, R. Jiang	ACM
4	V. Gaede, O. Gunther	Multidimensional access methods	ACM
5	A. Guttman	R-trees: a dynamic index structure for spatial searching	ACM

Table 1. Table Parameter for Doc 1.

The table 1 represents the parameter table for the document 1. We have listed the top 5 reference of the document for saving the space. The document possesses a total of 22 references and the proposed approach has processed all 22 references in the parameter table for experimentation. Later the tittles from the parameter table are selected and second phase of the proposed approach is executed. This will reveal the parameter score regarding the authors listed in the parameter table. The program will generate much larger parameter table with more fields in it. The updated parameter table can be presented as,

Sl No	Author	Tittle	publisher	N(j)	N(conf)	N(art)	N(cit)	Parameter score
1	J.L. Bentley	Multidimensional binary search trees used for associative searching	ACM	24	37	1	3	0.89210
2	H. Blanken, A. Ijbema, P. Meek, B. Akker	The generalized grid file: description and performance aspects	IEEE	66	81	2	4	0.71230
3	E.I. Chong, J. Srinivasan, S. Das, C. Freiwald, A. Yalamanchi, M. Jagannath, A.T. Tran, R. Krishnan, R. Jiang	E.I. Chong, J. Srinivasan, S. Das, C. Freiwald, A. Yalamanchi, M. Jagannath, A.T. Tran, R. Krishnan, R. Jiang	ACM	37	41	2	3	0.70112
4	V. Gaede, O. Gunther	Multidimensional access methods	ACM	40		2	6	0.64212
5	A. Guttman	R-trees: a dynamic index structure for	ACM	12	10	1	1	0.54323

Table 2. Updated Parameter Value

The table 2 represents the parameter table with parameter score. The analysis from the table shows that, the author J.L. Bentley has the higher parameter score as compared to other authors. Thus, we select the author as the influential researcher for the doc 1 as per out approach. Through the manual checking, it is found that the doc 1 is inspired from the approaches detailed in J.L Bentley's articles.

V. CONCLUSION

This research has given an efficient method for finding an influential researcher in the field of research and development. The proposed approach uses a series of steps to process the research paper and applied clustering algorithm to cluster the authors based on the similarity measure. Finally, score calculation user helps to know about author who is influential researcher. The relevant authors are particular based on parameters score. A number of documents are selected for the experimentation process, in order to evaluate the efficiency of the proposed approach. The experimental analysis present that proposed method is effective processing influential researcher.

References

[1] Fuchun Peng and Andrew McCallum. 2004. "Accurate Information Extraction from Research Papers using Conditional Random Fields".

[2] Mihalcea, R., and Tarau, P. 2004. Textrank: Bringing order into text. In EMNLP.

[3] Wan, X., and Xiao, J. 2008. Single document keyphrase extraction using neighborhood knowledge. In AAAI.

[4] Shi, X.; Leskovec, J.; and McFarland, D. A. 2010. Citing for high impact. In JCDL.

[5] Frank, E.; Paynter, G. W.; Witten, I. H.; Gutwin, C.; and Nevill-Manning, C. G. 1999. Domain-specific keyphrase extraction. In IJCAI.
[6] Hulth, A. 2003. Improved automatic keyword extraction given more linguistic knowledge. EMNLP 216–223.
[7] Nguyen, T., and Kan, M.-Y. 2007. Keyphrase extraction in scientific publications. In Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers, volume 4822.
[8] Liu, F.; Pennell, D.; Liu, F.; and Liu, Y. 2009. Unsupervised approaches for automatic keyword extraction using meeting transcripts. In Proceedings of NAACL '09, 620–628.
[9] Marujo, L.; Ribeiro, R.; de Matos, D. M.; Neto, J. P.; Gershman, A.; and Carbonell, J. G. 2013. Key phrase extraction of lightly filtered broadcast news. CoRR.
[10] Boudin, F. 2013. A comparison of centrality measures for graph-based keyphrase extraction. In IJCNLP.
[11] Kevin Yao, Mario Lipinski, Bela Gipp and Jim Pitman. "Header Extraction from Scientific Documents".
[12] Patrice Lopez, "GROBID: Combining Automatic Bibliographic Data Recognition and Term Extraction For Scholarship Publications".
[13] A. McCallum and A. Kachites.2002. MALLET: "A Machine Learning for Language Toolkit".
[14] Isaac G. Councill, C. Lee Giles and Min-Yen Kan, "ParsCit: An opensource CRF reference string parsing package".
[15] Mendeley is a desktop and web program for managing and sharing research papers, discovering research data and collaborating online. Available: http://www.mendeley.com/.
[16] PDFMEAT is Metadata acquisition and embedding tool for papers in PDF format. Available: http://code.google.com/p/pdfmeat/
[17] Zotero is a powerful, easy-to-use research tool that helps to gather, organize, and analyze sources and then share the results of research. Available: http://www.zotero.org/
[18] Paperpile is a tool to find, organize, cite and share scientific papers. Available: http://paperpile.com/

