Volume 5, No. 6, June 2014
Journal of Global Research in Computer Science
ISSN-2229-371X
RESEARCH PAPER
Available Online at www.jgrcs.info

A METHOD TO IDENTIFY THE INFLUENTIAL RESEARCHER FROM RESEARCH PAPERS

Nilesh Jain[1]

Research Scholar, Mewar University,

NH - 79 Gangrar, Chittorgarh, Rajasthan-312 901

nileshjainmca@gmail.com

Dr. Vijay Singh Rathore[2]

Professor & Director,

Shree Karni College Jaipur(Raj.)

Abstract: The research works are published by the researchers in various journals and conferences, which are made publically available by the publishers. Though such research is spread all over the world, there should be a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research. In this paper, we have plotted a method to find the influential researcher for a particular input document. The proposed approach has series of steps to process out the influential researcher. Initially; we extract the reference from the text document which is given as input. The basic text extraction algorithms are utilized for the process. Later on the authors of those particular references are identified for further processing. Then, a clustering process is subjected for identifying the similar referral documents. The clustering processes in defined with a semantic similarity measure that will help to extract most similar documents and their information. This will enhance the clustering process. Finally, the queries regarding intended research is subjected and compared with the clusters and parameter score is extracted. 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 analysis showed that the proposed approach is efficient processing the influential researcher.

Key words: research work, influential researcher, similarity, clustering.

1.INTRODUCTION

Usually, many scientific publications are made available on the Web supported by individual researchers, research institutions and database (such as EI, SCI, SSCI, etc.) to share their research findings. However, it is apparent that information on expertise on research areas is hidden in the huge sea of information and cannot be extracted automatically from the Web. Hence research on finding expertise from Internet has become a meaningful task, attracting much interest from the academic community [1]. An improved understanding of the customer's habits, needs, and interests can allow the business to profit by, for instance, "cross selling" or selling items related to the ones that the customer wants to purchase. Hence, reliable knowledge about the customers' preferences and needs forms the basis for effective methodology [3]. The fast pace and large amounts of data available in these online settings have recently made it imperative to use automated data mining or knowledge discovery techniques to discover Web user profiles. These different modes of usage or the so-called mass user profiles can be discovered using Web usage mining techniques that can automatically extract frequent access patterns from the history of previous user click streams stored in Web log files.

Web mining is the application of data mining techniques to extract knowledge from Web. Though such research is scattered all over the world, there is a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research [5]. Government and the other state organizations periodically need adaptation of certain new technology or process for improving or implementing certain policies. Such new requirements are generally met by calling for researchers to join in hand with the organization with their proposal. Filtering such proposals also takes a hectic schedule and thorough understanding of the researchers profile and to gauge his ability to complete the work [6]. In order to solve this problem we emphasize on extracting meaningful information from the web through web mining techniques that helps understanding the region wise trends in research activities and further extract more meaningful information like patterns that suggest the progress in a particular area and prominent contributors in the area.

© JGRCS 2010, All Rights Reserved

Recently many researchers have emerged with idea of selecting most relevant research domain from the existing research journals. Nilesh Jain and Dr. Vijay Singh Rathore [1] have proposed a method to measure research related domains through web mining techniques. The proposed approach is based on a specific web mining techniques to identify relevant documents. A semantic similarity based web usage mining is represented in the paper [3] for identifying the user interest or user profiles. In the proposed approach, we utilizes the information from these researches to identify the influence researcher to enhance the credibility of the research.The proposed method includes three major steps, the initial step is extract the reference from the text document which is given as input. The basic text extraction algorithms are utilized for the process. Later on the authors of those particular references are identified for further processing. In the second step, a clustering process is subjected for identifying the similar referral documents. The clustering processes in defined with a semantic similarity measure that will help to extract most similar documents and their information. This will enhance the clustering process. In the third step, the queries regarding intended research is subjected and compared with the clusters and parameter score is extracted. Then we list the authors with higher parameter score. The relevant authors are selected based on the parameters score.

The contributions of the proposed approach are,

- Analyzed various web mining techniques for relevant information extraction
- Design and develop a method to identify influential researcher from research papers through web mining techniques
- Conduct different performance and comparative analysis for the assessment of the proposed approach.

The rest of the paper is organized as, the second section plots the literature review and third section plots the motivation behind the proposed approach. Thefourth section plots the proposed approach and fifth section includes the experimental analysis. The conclusion of the proposed approach is plotted in the sixth section.

2.LITERATURE REVIEW

Nilesh Jain and Dr. Vijay Singh Rathore [1] have proposed a method to measure research related domains through web mining techniques. The research works are published by the researchers in various journals and conferences, which are made publically available by the publishers. Though such research is spread all over the world, there should be a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research. Government and the other state organizations periodically need adaptation of certain new technology or process for improving or implementing certain policies. Such new requirements are generally met by calling for researchers to join in hand with the organization with their proposal. Filtering such proposals also takes a hectic schedule and thorough understanding of the researchers profile and to gauge his ability to complete the work. In order to solve this problem we emphasize on extracting meaningful information from the web through web mining techniques that helps understanding the region wise trends in research domain activities and further extract more meaningful information like patterns that suggest the progress in a particular area and prominent contributors in the area.

GokceBanuLaleci, Mustafa Yuksel, and AsumanDogac [2] have described an initial implementation of the Semantic Framework developed within the scope of SALUS project to achieve interoperability between the clinical research and the clinical care domains. In their Semantic Framework, the core ontology developed for semantic Mediation is based on the shared conceptual model of both of these domains provided by the Biomedical Research Integrated Domain Group (BRIDG) initiative. The core ontology is then aligned with the extracted semantic models of the existing clinical care and research standards as well as with the ontological representations of the terminology systems to create a "model of meaning" for enabling semantic mediation. Although SALUS is a research and development effort rather than a product, the current SALUS knowledge base contains around 4.7 million triples representing BRIDG DAM, HL7 CDA model, Clinical Data Interchange Standards Consortium standards, and several terminology ontologies. In order to keep the reasoning process within acceptable limits without sacrificing the quality of mediation, we took an engineering approach by developing a number of heuristic mechanisms. The results indicate that it is possible to build a robust and scalable semantic framework with a solid theoretical foundation for achieving interoperability between the clinical research and clinical care domains.

OlfaNasraoui, MahaSoliman, EsinSaka, Antonio Badia, Richard Germain [3] have presented a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing an ontology of the Web content. Even though the Web site under study is part of a nonprofit organization that does not "sell" any products, it was crucial to understand "who" the users were, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, we present an approach for discovering and tracking evolving user profiles. We also describe how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior.

Chuck Litecky, Andrew Aken, Altaf Ahmad, and H. James Nelson [4] have proposed a Web content data mining application that extracted almost a quarter million unique IT

© JGRCS 2010, All Rights Reserved

job descriptions from various job search engines and distilled each to its required skill sets. We statistically examined these, revealing 20 clusters of similar skill sets that map to specific job definitions. The results allow software engineering professionals to tune their skills portfolio to match those in demand from real computing jobs across the US (see the sidebar "Computing Jobs in the US") to attain more lucrative salaries and more mobility in a chaotic environment

3.MOTIVATION BEHIND THE APPROACH

The research works are published by the researchers in various journals and conferences, which are made publically available by the publishers. Though such research is spread all over the world, there should be a distinct pattern for this, which suggest that different geographical areas observes distinct trend towards particular direction of research. Recently, Nilesh Jain and Dr. Vijay Singh Rathore [1] proposed a method for measuring research related domain through web mining techniques. The method intended to provide idea to research scholars in selecting their research area. Inspired from the above research, weproposed a method for finding the influential researcher from the documents through web mining techniques. The influential researcher is considered as to be the author, whose research has provided dominant motivation to an intended research. The proposed approach uses a series of steps to identify the influential researcher from the input document. The details of the proposed approach are furnished in the following section.

4. THE PROPOSED APPROACH FOR IDENTIFYING THE INFLUENTIAL RESEARCHER

The method proposed is intended to provide idea to research scholars in selecting their research area. Inspired from the above research, I have intended to propose a method for finding the influential researcher from the documents through web mining techniques. The influential researcher is considered as to be the author, whose research has provided dominant motivation to an intended research. The major steps regarding the proposed approach is listed as,

- Extracting author information
- Parameter score calculation
- Identifying influential researcher

The proposed method includes three major steps, the initial step is extract the reference from the text document which is given as input. The basic text extraction algorithms are utilized for the process. Later on the authors of that particular references are identified for further processing. In the second step, a clustering process is subjected for identifying the similar referral documents. The clustering processes in defined with a semantic similarity measure that will help to extract most similar documents and their information. This will enhance the clustering process. In the third step, the queries regarding intended research is subjected and compared with the clusters and parameter score is extracted. Then we list the authors with higher parameter score. The relevant authors are selected based on the parameters score.

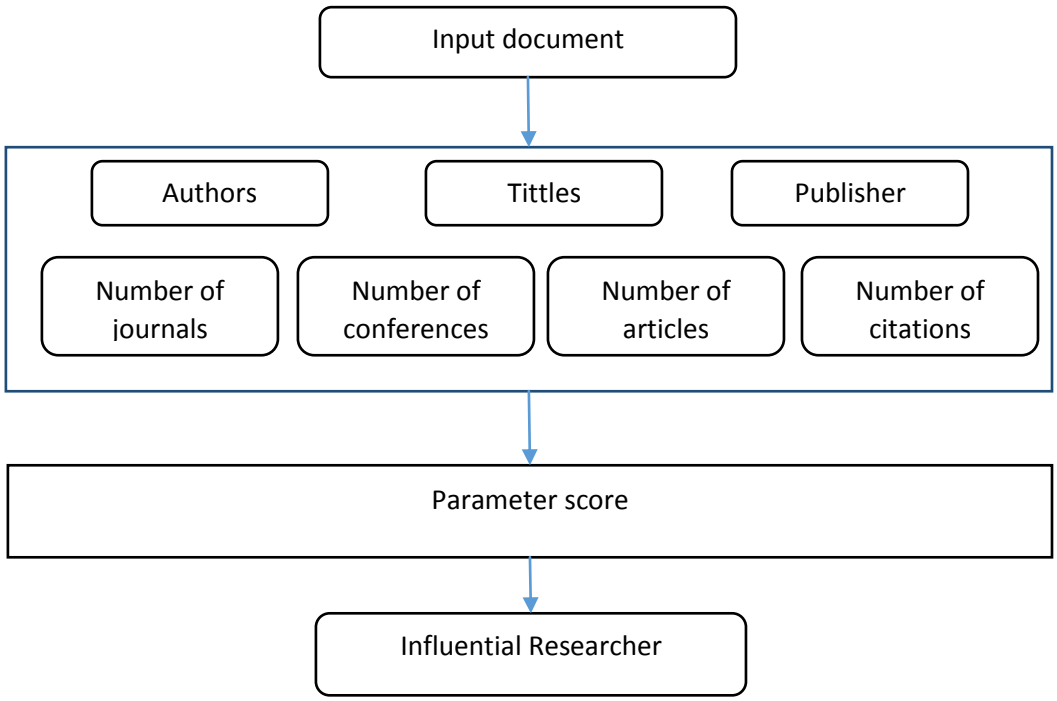

Fig.1. Block diagram

© JGRCS 2010, All Rights Reserved

The figure 1 represents the block diagram of the proposed approach. The second block of the diagram represent the author's information extraction phase, which mostly deals with the parameter selection and their value calculation. The third block of the diagram represent the parameter score calculation phase, through that the influential researcher is identified. The following section is a plot of detailed discussion about the proposed approach.

a. EXTRACTING AUTHOR INFORMATION

The main idea of the proposed approach is to find the influential researcher from the given input document. There is a series of steps to process the input document in order to effectively identifying the influential researcher. The basic text processing methods are carried out to process the document to retrieve the content from the documents. Initially, the reference section of the document is selected for processing. The total number of references are identified by selecting the numbers of each references. If the references are not provided with number, the full stop "." is considered as the identifier or the line separation. Once all the reference number are identified, we move on to identify the author and tittle of the documents in the references. We store the author names in reference to a reference in set A,

Where, $A = [a_1, a_2, ..., a_n]$

"A" represents the set of autors and a_1 represents the author names of document 1, a_2 represents the author names of document 2 and so on. Once all the author names corresponding each reference are extracted, we subject a search on document tittles. The tittle of each reference of the document is saved on a set called, T,

$$T = [t_1, t_2, ..., t_n]$$

Where, T is the set of tittles and $t_1$represents the tittle of reference one, t_2 represents the tittle f reference 2 and so on. In the similar way, the publication details of each of the references are also collected in a set defined as P.

$$P = [p_1, p_2, ..., p_n]$$

Once all the details regarding an author are collected by the text processing methods. A table is created to lest all the parameters. The table includes three columns for author, tittle and publications. The data to the table are collected from the sets A, T and P.

Serial NO	Author	Tittle	Publisher
1	a_1	t_1	p_1
2	a_2	t_2	p_2
…	…	…	…
N	a_n	t_n	p_n

Table.1. parameter table

The table 1 represents the parameter table for the input document. The parameter table list out all the authors and their details. The parameter table will help the program to easily identify the authors and to fetch details regarding them. The parameter table is a major resource of the proposed approach in the coming steps. The parameters tables indicate an author then corresponding tittles and publish of that particular author. The proposed approach primarily concentrates on the parameters like author name, title of the document and references. Based on these features, we extract four mathematical calculations from each document, which are considered as weighted parameters. These weighted parameters are used to calculate the influential researchers from the document.

1. Number of journals

The weighted parameter, number of journals is calculated by selecting each author name from the references. The author names from the reference is selected and given to the internet for fetching the details regarding the particular author. The results are selected and the journal published by the author is selected into account. We define the weighted parameter, number of journal as,

$$N(j) = sum(journal\,by\,author)$$

Here, the value N(j) represents the number of journals. The sum of journals is represented in the right side of the equation. In this context, we consider only journals published by the author. The value N(j) is calculated for all authors selected from the references. Consider the following example,

© JGRCS 2010, All Rights Reserved

Author Name	N(j)
Lin Li	37
Guandong Xu	638

Table.2. N(j) values

The above table 2 represents the reference of the page "Random walk based rank aggregation to improving web search". The author of the paper is listed in the table with their number of journals published.

2. **Number of conferences**

In this section, we select the number of conferences possessed by the author as the weighted parameter. In similar to the number of journal calculation, for numbe of conference calculation, we select the total number of conference attended by the author. The sum of conference attended represented as N(conf) and are calculated as,

$$N(conf) = sum(conferences\ by\ author)$$

On considering the same example as above, we can see that the same authors possess N(conf) value as,

Author Name	N(conf)
Lin Li	15
Guandong Xu	18

Table.3. N (conf) value

The table 3 represents the value of N(conf) calculated for the author Lin Li and Guandong Xu for the selected input document of the context.

3. **Number of Articles**

The parameter, the number of articles possessed by the author represents total number journals and conferences. The parameter is calculated as the sum of number of journals and number of conferences possessed by the author. i.e., if the author has total of 21 journals and a total of 11 conference, then the value of number of articles possessed by the author becomes 33. The parameter is represented as,

$$N(art) = sum(N(j), N(conf))$$

The value N(art) represents the number of articles and as discussed above it is the sum of N(j) and N(conf) of the author in reference. The value N(art) helps the proposed method to identify the influence of the particular author in the overall domain.

Author Name	N(art)
Lin Li	52
Guandong Xu	656

Table.4. N(art) values

The above table 4 represents the demonstration of N(art) calculation of the proposed approach. The authors listed above are selected from a reference of the document titled as, "Random walk based rank aggregation to improving web search". The value of N (art) helps to identify the contribution of the selected authors to the domain of web search and hence can help the proposed method to identify how influential the authors are to the selected input document.

4. Number of citations

The proposed method discusses a fourth parameter in this section. The parameter is known as number of citations about the author. The number of citations shows the contribution of the author's works in other works. The number of citations are represented using N(cit), which is different from the other three parameter. For the parameters, N(j), N (conf) and N(art), the calculations are entirely based on the author name's perspective. On the other hand, the parameter N (art), is calculated based on the articles, which uses the author's publications as references.The values of N (art) can be calculated as,

$$N(art) = sum\_of\_citations(authors's\_name)$$

The value of N(art) help us to find out the contribution of the author in reference to the domain under discussion. So the value also helps in finding the influence of the particular author in the current input document also. The number of article is calculated in relevant the input document and total citations of the author in the contributing domain. On considering the example, which is discussed above,

Author Name	N(cit)
Lin Li	15
Guandong Xu	19

Table.5. N(cit) values

The table 5 depicted in the above shows the values of number citations by the authors Lin Li and Guandong Xu for the paper tittle as "Random walk based rank aggregation to improving web search". The table represents that, the author Lin Li's references are cited 15 times in other articles and latter name has cited 19 times in the other article. The value of N(art) is used along with other three parameter value to calculate the influential researcher.

b. Parameter score calculation

In the parameter calculation, we select the title from the parameter table and subjected for search in internet. This search will reveal details relevant the authors regarding the particular title. The search results from the tittles produce the abstracts of the tittles that are subjected to search. Every abstracts are selected and stored for the further process. We apply a clustering algorithm to groups the abstracts. The K-means clustering algorithm is used for the clustering process and this clustering process groups the abstracts that are similar characteristics. In order to process with the k means clustering, the abstracts are converted to data points by comparing the abstract of the input document. To generate the data value, we use a mutual similarity measure between

© JGRCS 2010, All Rights Reserved

the sentences of input document abstract and abstract extracted. This will produce a data value for each of the abstract and that data points are used for the clustering process.

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

Here, the numerator represents the joint probability and denominator represents the probability of sentences in abstract from the reference document. The joint probability is calculated between the sentences from the abstract of input document and from the abstract of reference document. The calculation produces data value for each of the abstracts. The data points are then stored in the set named as D,

$$D = [d_1, d_2, ..., d_n]$$

The k means clustering is applied on the data points in set D. thus k number of clusters are formed from the data. Once the clusters are generated, the centroid corresponding to each cluster is selected. The selected centroid are compared with centroid extracted from the input document's abstract. The distance between both centroids are calculated for the purpose. The cluster with least distance to the centroid of the input document is selected for further processing.

$$dist(c_{cluster}, c_{input}) = eucledeiandis\tan ce(c_{cluster}, c_{input})$$

We use the Euclidian distance for calculating the distance between the centroids. Once the least distinct cluster has been identified, the references regarding that cluster are retrieved. The title of the reference are again selected from the parameter table and subjected for another internet search. The search is used to reveal another parameters like, Number of journals, Number of conferences, number of articles and number of citations of the particular tittle. On the basis of these there parameters a scores is generated for the extracted references from the cluster. The scores is added to the right side of the parameter table in reference to the author and tittle. The parameter score is calculated as,

$$parameter\_score = \frac{N(j)}{J} + \frac{N(conf)}{C} + \frac{N(art)}{Art} + \frac{N(cit)}{Cit}$$

Here, N(j) represents the number of journals, J represents the maximum number of journals, N(conf) represents the number of conferences, C represents the maximum number of conferences, N(art) represents the number of articles, Art represents the maximum number of articles and N(cit) represents the number of citations and Cit represents the maximum number of citations. We can list the parameter score as follows,

Serial NO	Author	Tittle	Publisher	Score
1	a_1	t_1	p_1	s_1
2	a_2	t_2	p_2	s_2
…	…	…	…	…
k	a_k	t_k	p_k	s_k

Table.6. parameter table with score

The table 6 represents the parameter table with parameter score. The parameter scores are calculated to the references which are present in the selected cluster. That is the reason behind limiting the entry to the table by k.

c. Identifying influential researcher

The process of identifying the influential researcher is the final step of the proposed approach. The parameter table with parameter score is selected for the purpose of identifying the influential researcher. The parameter score on the authors are selected form the parameter table. The parameter scores are arranged in their decreasing order. We select top n number of authors for identifying the influential researcher. The abstracts of the n selected reference document are selected and their abstract are again compared with the input documents abstract. The references that possess high similarity value is selected as the influential researcher.

5. Experimental results and analysis

In this section experimental analysis for the proposed approach is presented. The experimental results of the proposed approach is plotted to evaluate the performance of the proposed approach. A detailed analysis subjected to evaluate the performance of the proposed method for identifying the influential researcher.

Experimental setup and dataset

© JGRCS 2010, All Rights Reserved

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, doc 2 and doc 3for the ease of use. Each of the documents contains more than 20 references and among those one is influential researcher.

Responses regarding Doc 1.

The document 1 is published by the author Hung Yi Lin [8] 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.7. parameter table for doc 1.

The table 7 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

© JGRCS 2010, All Rights Reserved

		spatial searching						

Table.8. updated parameter table

The table 8 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.

Responses regarding doc 2.

The document 2 is presented by Shingo Mabu et al.[9], the paper deals with intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming. The doc2 is supplied to the proposed approach as input. The responses on processing the document 2 is also encouraging similar to the document 1. Let us consider the 2 parameter tables considered for the document 2.

Sl. No	Authors	Tittle	Publication
1	J. G.-P. A. El Semaray , J. Edmonds and M. Papa	Applying data mining of fuzzy association rules to network intrusion detection	IEEE
2	A. S. S. Forrest , S. A. Hofmeyr and T. A. Longstaff	sense of self for unix processes	IEEE
3	J. Luo	Integrating fuzzy logic with data mining methods for intrusion detection	IEEE
4	J. Zhang , M. Zulkernine and A. Haque	Random-forests-based network intrusion detection systems	IEEE
5	Z. Yu , J. J. P. Tsai and T. Weigert	An automatically tuning intrusion detection system	IEEE

Table.9. parameter table

The table 9 represents the parameter table for the document 2. In this table, we presented the top 5 relevant references regarding the input document. According to the proposed approach, the tittles from the tables are selected for the further processing.

Sl No	Author	Tittle	publisher	N(j)	N(conf)	N(art)	N(cit)	Parameter score
1	J. G.-P. A. El Semaray , J. Edmonds and M. Papa	Applying data mining of fuzzy association rules to network intrusion detection	IEEE	332	56	4	10	1.2210
2	A. S. S. Forrest , S. A. Hofmeyr and T. A. Longstaff	sense of self for unix processes	IEEE	89	81	9	15	0.91230
3	J. Luo	Integrating fuzzy logic with data mining methods for intrusion detection	IEEE	90	56	12	30	1.10112
4	J. Zhang , M. Zulkernine and A. Haque	Random-forests-based network intrusion detection systems	IEEE	38	41	11	36	1.04212
5	Z. Yu , J. J. P. Tsai and T. Weigert	An automatically tuning intrusion detection system	IEEE	42	66	8	12	0.94323

Table.10. parameter table with parameter score

© JGRCS 2010, All Rights Reserved

The parameter score for the document is calculated based on the equation presented in the proposed approach. The analysis from the tables shows that all the authors are highly influenced for the preparation of document 1. Since the reference 1 has higher parameter score according to the calculations, we set the authors, G.P. A. El Semaray , J. Edmonds and M. Papa as the influential researcher to the input document 1.

Responses regarding doc 3

The document 3 is referenced to the topic artificial neural network-based merging score for Meta-search engine. The similar process described in the above are repeated for the document 3 also. The parameter table of the doc 3 is represented in the table 11.

Sl. No	Authors	Tittle	Publication
1	Amir Hosei, Keyhanipour, BehzadMoshiri, Majid Kazemian, Maryam Piroozmand and Caro Lucas	Aggregation of web search engines based on users' preferences in Web Fusion	ELSEVIER
2	Lin Li, Guandong Xu, Yanchun Zhang and Masaru Kitsuregawa	Random walk based rank aggregation to improving web search	ELSEVIER
3	Sugiura A., Etzioni O	Query routing for Web search engines: architecture and experiments	CiteSeerX

Table.11. parameter table of doc 3

Sl No	Author	Tittle	publisher	N(j)	N(conf)	N(art)	N(cit)	Parameter score
1	Majid Kazemian	Aggregation of web search engines based on users' preferences in Web Fusion	Elsevier	631	28	1	29	0.79725
2	Lin Li	Random walk based rank aggregation to improving web search		37	15	0	15	0.27773
3	Guandong Xu	Random walk based rank aggregation to improving web search		638	18	1	19	0.62450
4	Sugiura A	Query routing for Web search engines: architecture and experiments		75	13	5	18	0.55063

Table.12. updated parameter table

The analysis from the table 12 states that the author Majid Kazemianacquired the highest parameter score and thus the influential researcher to the doc 3 is Majid Kazemian.

6. CONCLUSION

In this paper, we have plotted a method to find the influential researcher for a particular input document. The proposed approach has series of steps to process out the influential researcher. Initially; we extract the reference from the text document which is given as input. The basic text extraction algorithms are utilized for the process. Later on the authors of those particular references are identified for further processing. Then, a clustering process is subjected for identifying the similar referral documents. The clustering processes in defined with a semantic similarity measure that will help to extract most similar documents and their information. This will enhance the clustering process. Finally, the queries regarding intended research is subjected and compared with the clusters and parameter score is extracted. 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

© JGRCS 2010, All Rights Reserved

analysis showed that the proposed approach is efficient processing the influential researcher.

REFERENCES

1] Nilesh Jain and Dr. Vijay Singh Rathore, " an approach to measure research related domains using web-mining techniques", Journal of Global Research in Computer Science, Vol. 2, No. 8, pp: 04-08, 2011.

[2] GokceBanuLaleci, Mustafa Yuksel, and AsumanDogac, "Providing Semantic Interoperability between Clinical Care and Clinical Research Domains", IEEE journal of biomedical and health informatics, VOL. 17, NO. 2, pp: 356 -369 , 2013.

[3] OlfaNasraoui, MahaSoliman, EsinSaka, Antonio Badia, Richard Germain, "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE transactions on knowledge and data engineering, VOL. 20, NO. 2, PP: 202 – 215, 2008.

[4] Chuck Litecky, Andrew Aken, Altaf Ahmad, and H. James Nelson, "Mining for Computing Jobs", Transactions on IEEE software, pp: 78 – 85, 2010.

[5] Ching-Seh Wu, Ibrahim Khoury, and Hemant Shah, "Optimizing Medical Data Quality Based on Multiagent Web Service Framework", IEEE transactions on information technology in biomedicine, VOL. 16, NO. 4, PP: 746 – 757, 2012.

[6] Valentina Tamma, "Semantic Web Support for Intelligent Search and Retrieval of Business Knowledge", Transactions on IEEE intelligent systems, PP: 84 -88, 2010.

[8] Hung Yi Lin, "Efficient and compact indexing structure for processingof spatial queries in line-based databases", Data and Knowledge engineering, ELSEVIER, Vol. 64, pp: 365- 380, 2008.

[9] Shingo Mabu, Nannan Lu, Kaoru Shimada,KotaroHirasawa, " An Intrusion-Detection Model Based on Fuzzy Class-Association-Rule Mining Using Genetic Network Programming", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART C: APPLICATIONS AND REVIEWS, VOL. 41, NO. 1, PP: 130-139 , 2011.

© JGRCS 2010, All Rights Reserved