2011 3rd Conference on Data Mining and Optimization (DMO)
28-29 June 2011, Selangor, Malaysia

Data Mining Technique for Expertise Search in a Special Interest Group Knowledge Portal

Wan Muhammad Zulhafizsyam Wan Ahmad, Shahida Sulaiman, Umi Kalsom Yusof
School of Computer Sciences
Universiti Sains Malaysia
11800 USM, Penang, Malaysia
zulhafizsyam.com08@student.usm.my, {shahida, umiyusof}@cs.usm.my

Abstract—**The Internet contributes to the development of electronic community (e-community) portals. Such portals become an indispensable platform for members especially for a Special Interest Groups (SIG) to share knowledge and expertise in their respective fields. Finding expertise over the e-community portal will help interested people and researchers to identify other experts, working in the same area. However, it is quite a cumbersome task to search such expertise in the portal. In order to find an expert, expertise data mining could be a solution to ease the search of experts. Performing effective data mining technique will help to analyze and measure expertise level accurately in a SIG portal. This paper proposes a method called Expertise Data Mining (EDM) that comprises a few techniques for expertise search in a SIG portal. It expects to improve the finding of experts among the members of a SIG e-community.**

Keywords- Data mining; information retrieval; knowledge discovery; Web mining; expertise search

I. INTRODUCTION

A Special Interest Group (SIG) normally refers to a research community who shares the same interest in their respective fields. Nowadays, such SIG members adopt an electronic community (e-community) to share their information and knowledge among them. Finding "the real expert" in a SIG portal is a cumbersome task. In an e-community domain especially in a SIG knowledge portal, some criteria such as the number of publications, user contribution to the portal and interaction among portal members must be measured carefully in order to know members' expertise level in a portal.

Introduction of the Web enables researchers to broadcast their publication faster and at a lower cost [1]. Massive amount of data in the Web requires a filtering-system to select only relevant data to be extracted and processed to become knowledge that can support related decision making. Some information in the Web is irrelevant to an individual researcher. These characteristics make the Web a difficult domain for knowledge discovery [1]. Information of diverse research expertise is hidden among the massive amount of information and cannot be extracted automatically from the Web [2].

Scientific publications available online tend to be poorly organized, causing the search for relevant research publications to be difficult and time consuming [3]. In order to collect useful knowledge quickly and easily from a Web database, users need the help of an information-filtering system that can automatically extract only relevant records as they appear in a stream of incoming records [4].

Commercial search engines such as Yahoo! [5], Lycos [6], and Google [7] are ineffective for searching scientific publications accurately due to the fact that most research documents are mainly available in Post Script or PDF (Portable Document Format) and they are not properly indexed by commercial search engines [3]. Other technique includes Term Frequency x Inverse Document Frequency (TFIDF) that mines Web pages and analyse author co-citations in order to find expertise within an organization or a small domain area [2].

Existing data mining techniques such as CiteSeer [8], PubSearch [3] and DBLP [9] perform well in mining citations information and finding expertise from the Web. However, they still have weaknesses as they do not provide information on researchers who work in the same research area. These techniques only perform the mining of citations of papers without including researchers' projects which are also an important source of information for SIG members.

Finding expertise over the Web helps researchers in knowing the experts who work in the same area. Thus, we propose an Expertise Data Mining (EDM) technique to mine research publications and projects that will be used to measure the expertise level. The proposed technique is different from that of existing techniques as it also considers research projects of each member in the calculation of a researcher's expertise level. We anticipate the proposed technique will increase the accuracy of expertise search among members of SIG portals regardless whether they are from the academic or industrial background.

This paper deals about motivation of research in Section II, followed by the details of related work in Section III. Section IV describes about expertise data mining. Finally, we conclude the work in Section V.

II. MOTIVATION

Some of the SIG portals such as Daniweb.com [10] and Forums.sun.com [11] measure their members to identify their expertise level. The current measurement of members is solely

978-1-61284-212-7/11/$26.00 ©2011 IEEE

based on rating by other members, and interaction with the portal. This kind of measurement is not accurate to determine the expertise level of each member in the portal. Other important factors such as member publications and projects must be included to measure the expertise level of members. Previous studies show that expertise is recommended on the basis of publications and human resource records [12][13]. Another study shows that classification of members' expertise in SIG portal is important to measure members' expertise level [14].

Few researchers have encountered some limitations and problems in the process of evaluating experts [15][16][17]. Evaluation of experts' performance tends to face external influences mainly due to individual subjective judgment and discrimination on the result of expert selection [15]. For example, National Science Foundation of China (NSFC) uses a method to evaluate experts, in which it uses the "word of mouth" method for conducting an evaluation. This method is not efficient because it is mainly based on a subjective judgment from division managers [16]. This method is incomplete, inefficient and unfair to other qualified experts. Besides, it is difficult for an organization to predict the success of the project reviewed by experts [17]. Another problem faced by researchers is on how to estimate the level of knowledge of each expert, to show the creativity of experts, and to display how experts work together in a group [18]. Thus, existing problems motivate us to provide an effective method for expertise data mining called EDM that can measure expert members and can accurately find 'real expert' mainly in a SIG e-community.

III. RELATED WORK

An important component to measure members' expertise level is through members' publications. Several methods and techniques [1][2][11][18][19][20][21][22] are available to mine citations and publications from the Web and databases in order to find research expertise. However, these methods and techniques still can be improved to produce better results. Table 1 shows the comparison of current data mining techniques used to retrieve publications from the Web.

A citation-based retrieval system known as PubSearch [3] adopts data mining techniques such as document clustering, author clustering, author co-citation analysis and multi-clustering that are applied to the Web Citation Database in order to cite keywords and authors. PubSearch generates Web Citation Database from online scientific publications and support the retrieval of publications based on Web Citation Database. The Web Citation Database is a data warehouse used for storing citation indices, which contain the references that the publications cite [3].

Another current citation-based retrieval system includes Institute for Science Information (ISI) [24] that offers two types of search; general search and cite reference search. General search works to find publications by keyword terms, author names, journal titles and author affiliations. On the other hand, cite reference search works to find publications that cite authors or publications specified by users. Besides, Open Citation (OpCit) [19] links all cited work as archives and provides a mechanism that enables users to retrieve publications from these archives. CiteSeer is known as Research Index uses Autonomous Citation Indexing (ACI) to build citation database [1].

TABLE I. COMPARISON OF DATA MINING TECHNIQUES TO IDENTIFY RESEARCH EXPERTISE

Tools	Data Mining Technique	Type of Data Extracted	Type of Search	User Update
CiteSeer [8]	Sophisticated template matching	PostScript, PDF, citation indices	Keyword and author search	√
PubSearch [3]	Multi clustering	PostScript, PDF, citation indices	Keyword and author search	X
Harzing's PoP [21]	Google Scholar	Citation indices	Keyword and author search	X
DBLP [9]	Machine learning algorithm	Citation indices, user sumission	Keyword and author search	√ (by admin)

In identifying a research expertise, Agglomerative Hierarchical Clustering (AHC) [25] is performed on scholarly citation database to generate keyword or document cluster, author clusters, journal clusters, date clusters and organization clusters [3]. Most systems are developed based on social and collaborative aspects in making expertise recommendations [26]. However, this approach has difficulty in creating user profiles. Expertise Recommender [22] and Expertise Finder [23] use data mining for mining experts which include 'keyword-based mining' and Author Co-Citation Analysis (ACA). Both rely on simple keyword retrieval for finding information on expertise and such technique is only applicable to find expertise within an organization or a small domain area. CiteSeer approach for mining publication based on citation information and it can identify authors of research papers that are related to query keywords [1]. However, CiteSeer does not provide information about researchers who work in the same research areas.

Digital bibliography has become an important source of research citation from various fields. Many tools have been developed to retrieve selected information for certain purposes such as to know how many people have cited the researcher's publication. Harzing's Publish or Perish (PoP) retrieves and analyzes academic citations [21]. It uses Google Scholar (GS) to obtain raw citations, then analyzed and converted to a number of statistics. PoP calculates the citation metrics that include the total number of papers, the total number of citations, the average number of citation per paper, the average number of citation per author, the average number of paper per author, Hirsch's h-index, Egghe's g-index, Age-weighted citation rate and analysis of the number of authors per paper. Digital Bibliography and Library Project (DBLP) provide bibliographic information on major Journal and proceedings in the field of computer science [20]. DBLP started in 1993 as a pure HTML application, but later essential parts were converted to XML. The idea is to enter the table of contents (TOCs) into a format. DBLP extracts the bibliographic information (title, authors, page numbers and DOIs) from HTML pages with a script. These data mining techniques use different techniques to mine citations from the Web. However, they have a similar purpose and target that is to do Web based citation effectively and accurately.

IV. EXPERTISE DATA MINING (EDM)

The architecture of the proposed EDM method is illustrated in Fig. 1. It consists of an indexer, a Web citation database, publication and project extractor and expert deep analyzer that should be adopted in a SIG portal. Indexer searches for research publications and research projects in related websites and parse the bibliographic section by searching the keywords. Then, it extracts the citation information to be stored in the Web citation database. This paper focuses on the improvement of h-index to measure the cumulative impact of a researcher's output.

A. Indexer

The task of indexer is to search the scientific publications and projects from existing websites. For publications, it identifies the bibliographic segment via searching keywords or authors. It parses the bibliographic segment to extract citation information from the Web citation database. Indexer will help in searching relevant publications easily and faster. The indexer involves the following steps to search scientific publications:

1. Search members' scientific publications from identified databases such as Google Scholar [27] and CiteSeer [8].
2. Identify the bibliographic section by searching authors' name.
3. Parse bibliographic section to extract the citation information.
4. Remove duplicate records by matching titles of publication from existing publications.
5. Store the records into the Web citation database.

Indexer has been tested to extract a member's publications from selected publication portals such as Google Scholar [27] and CiteSeer [8]. Fig. 2 shows an example of raw data extracted using the indexer for the member named "Shahida Sulaiman" and stores the extracted data into the Web citation database. In this research, MySEIG portal [28] has been used as the problem domain. The relevant information will be filtered out from the Web citation database using the extractor. The indexer is further improved to search projects from existing websites besides publications.

The search and extraction of data triggered when members access their profile from the SIG portal via a Web browser. After a member's profile is accessed, the indexer performs a search, and then it identifies, filters and extracts the data to store in the Web citation database. In order to increase its efficiency, search and extraction of data; it also can be triggered using a temporal event such as by weekly, monthly or yearly. Another option is to trigger an event to search and extract the data. It extracts relevant data simultaneously for all members in the portal.

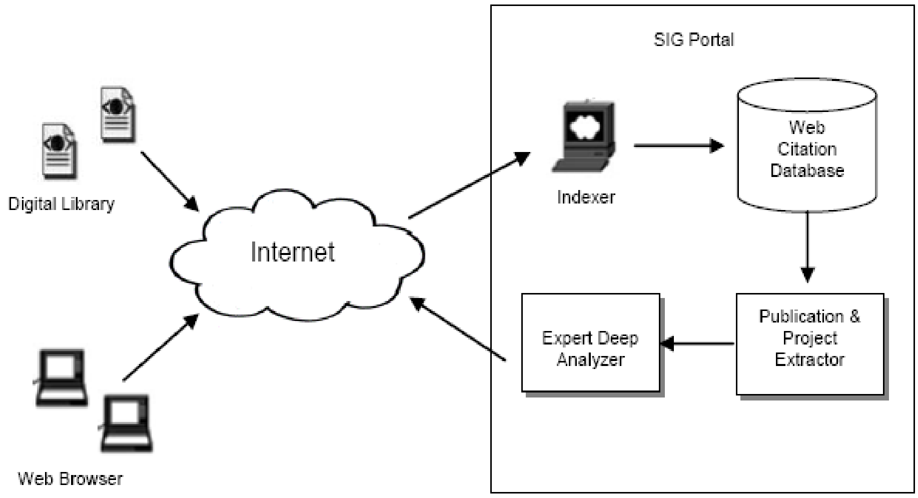
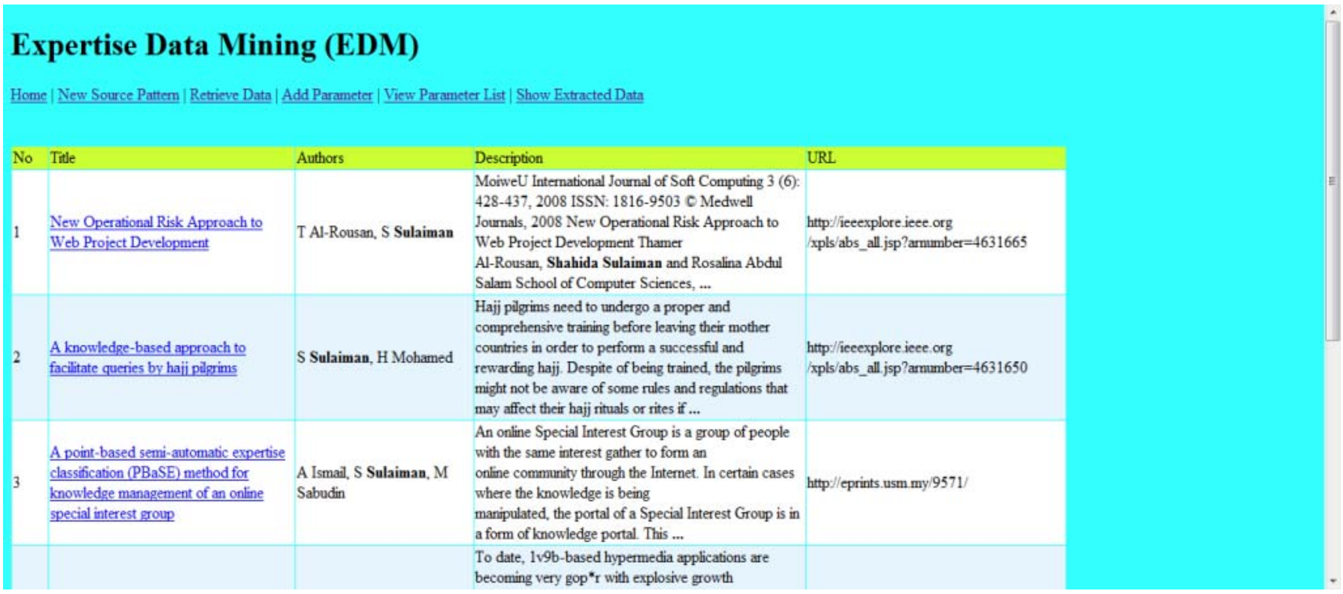


Figure 1. The architecture to implement the proposed EDM in a SIG portal

978-1-61284-212-7/11/$26.00 ©2011 IEEE



Expertise Data Mining (EDM)
Home | New Source Pattern | Retrieve Data | Add Parameter | View Parameter List | Show Extracted Data

No	Title	Authors	Description	URL
1	New Operational Risk Approach to Web Project Development	T Al-Rousan, S **Sulaiman**	MoiweU International Journal of Soft Computing 3 (6): 428-437, 2008 ISSN: 1816-9503 © Medwell Journals, 2008 New Operational Risk Approach to Web Project Development Thamer Al-Rousan, **Shahida Sulaiman** and Rosalina Abdul Salam School of Computer Sciences, ...	http://ieeexplore.ieee.org /xpls/abs_all.jsp?arnumber=4631665
2	A knowledge-based approach to facilitate queries by hajj pilgrims	S **Sulaiman**, H Mohamed	Hajj pilgrims need to undergo a proper and comprehensive training before leaving their mother countries in order to perform a successful and rewarding hajj. Despite of being trained, the pilgrims might not be aware of some rules and regulations that may affect their hajj rituals or rites if ...	http://ieeexplore.ieee.org /xpls/abs_all.jsp?arnumber=4631650
3	A point-based semi-automatic expertise classification (PBaSE) method for knowledge management of an online special interest group	A Ismail, S **Sulaiman**, M Sabudin	An online Special Interest Group is a group of people with the same interest gather to form an online community through the Internet. In certain cases where the knowledge is being manipulated, the portal of a Special Interest Group is in a form of knowledge portal. This ...	http://eprints.usm.my/9571/
			To date, 1v9b-based hypermedia applications are becoming very gop*r with explosive growth	

Figure 2. Example of raw data extracted using the indexer

B. Publication and Project Extractor

Text extraction is used to filter out relevant information from massive amount of texts available online [3]. Specific fields or attributes like author's name, title, description, date of publication and URL are extracted. During the data cleansing step, it removes duplicated records and consolidates similar terms. This function becomes very important to prevent duplicated data and to combine data from multiple databases. Apart from that, fuzzy matching techniques are used to identify and combine similar entities among authors, keywords, and affiliations. Fuzzy matching techniques are not within the scope of this paper.

During the extraction process, the citation format of a publication such as author-first or title-first format is omitted because the important data required is only number of publications for each member. Thus, it increases number of publications for each member that possibly increases their expertise level besides the projects they are involved in.

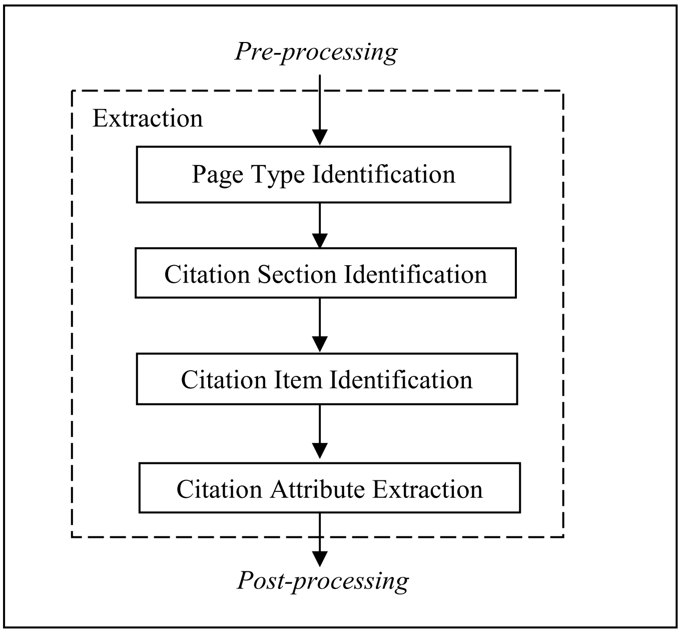
Publication and project extraction starts when target pages are identified, extracted, and relevant information is stored in the Web citation database for further retrieval. Project extraction is quite direct. However, publication extraction consists of three processes that are pre-processing, extraction and post-processing.

1) Pre-processing

In this process, the extracted pages will be analyzed to construct several patterns for subsequent retrieval using Regular Expression. For publications, several patterns constructed which consist of publication attributes such as title, authors, year of publication, description and URL. These patterns can be constructed manually based on source code retrieved in publication and project pages from selected databases such as Google Scholar [27] and CiteSeer [8]. The patterns are constructed to extract specific and relevant information from these pages. The subsequent extraction is performed in the next process.

2) Extraction

The steps involved in the extraction process are shown in Fig. 3. This process focuses on publication extraction.



Pre-processing
Extraction
Page Type Identification
Citation Section Identification
Citation Item Identification
Citation Attribute Extraction
Post-processing

Figure 3: Extraction process for publications

The Page Type Identification step identifies the extracted page whether it is ideal page or mixed page. Ideal page contains only publication and project information and used by major databases such as CiteSeer [8] and DBLP [9] while mixed page contain publication and project information together with additional information such as author's biography. Page Type Identification uses keyword search technique to identify page type.

Citation Section Identification identifies exact location of data to be retrieved using functional tag embedded in page layout. Page layout normally uses the HTML tag to display the page content in the website. If publication and project page uses different tag than HTML, it will automatically convert the page into HTML tag structure. Next step is Citation Item Identification.

Citation Item Identification is implemented using Regular Expression pattern. Pattern matching string that is constructed will identify and match the pattern string of item to be extracted. If the matching is correct, then citation attribute extraction will be executed.

In the last step, Citation Attribute Extraction is performed. Regular Expression pattern is divided into five components that are title, authors, year of publication, description and URL link of online available publications. It only extracts the attribute that matches the string pattern defined in Citation Item Identification and removes irrelevant data.

3) Post-Processing

Since extraction process depends heavily on heuristic and pattern matching, different page structure can affect the result of extraction. Citation section in the publication page may contain different structure for each citation and a few irrelevant citations could be misidentified as a true citation. As a result, a number of citations for certain members could be higher than it is supposed to be and affect the accuracy of the information. Therefore, post-processing step attempts to identify and remove wrong citations that have been extracted. Four steps introduced to perform this task include:

1. Publication title becomes a mandatory attribute to be extracted. Citation without an attribute of title will be removed.
2. Each citation must contain at least two attributes. Citation that contains a single attibute will be removed.
3. Author attribute must be similar or partially similar with existing members in the portal.
4. Publication date is important attribute to be classified as true citation. Therefore, citation extracted must contain publication date. Citation without publication date will be removed.

C. Expert Deep Analyzer

Expert Deep Analyzer is a technique to analyze and measure an expert's level of knowledge in a portal based on a modified h-index. The h-index is defined as follows: "A researcher has index h if h of their Np papers have at least h citations each, and the other ($Np - h$) papers have no more than h citations each" [29]. Its objectives are to measure the cumulative impact of a researcher's output by looking at the number of citations of their work. Hirsch argues that the h-index is preferable to other single-numbered criteria, such as the total number of papers, the total number of citations and citations per paper [30].

The benefit of the h-index is that it combines an assessment of both quantity (number of papers) and quality (impact, or citations to these papers) [29]. An academic researcher cannot have a high h-index without publishing a substantial number of papers. However, this is not enough. These papers need to be cited by other academics in order to count for the h-index. It is also preferable over the number of papers as it takes into consideration papers that are not cited. Hence, the h-index favours academics that publish a continuous stream of papers with a lasting and above-average impact factor [31].

In order to measure the expertise level more accurately, a new factor that is number of research projects of each member has been included in the calculation of h-index. Existing h-index is based on the total number of papers, the total number of citations and citations per paper. This may not be accurate enough to measure members' expertise level in a SIG portal. Some SIG members have an industrial background. Thus they only depend on research projects but not research papers. Therefore, measurement of expertise level using the modified h-index by adding a new factor that is the number of research projects of each member in the calculation of h-index will increase the accuracy of all members' expertise level regardless whether they are from the academic or the industrial background.

V. CONCLUSION

This paper has discussed the study of using data mining technique for expertise search in a SIG knowledge portal. The proposed EDM method consists of a number of techniques to mine expertise information based on scientific publications and research projects using an indexer, publication and project extractor, and also analyze the extracted information using expert deep analyzer. The expertise measurement is based on the modification of h-index by adding a new factor that is the number of research projects of each member. The proposed EDM is expected to increase the accuracy in mining the expertise, and improve the effectiveness of finding the "real expert" mainly in a SIG e-community knowledge portal.

ACKNOWLEDGMENT

This research is partially supported by short-term grant of EPISoDe project 304/PKOMP/6310044 and APEX incentive grant 1002/PKOMP/ARSP13100.

REFERENCES

[1] K. D. Bollacker, S. Lawrence, and C. L. Giles, "Discovering relevant scientific literature on the Web," *Intelligent Systems and their Applications, IEEE*, vol. 15, pp. 42-47, 2000.

[2] W. Hu, J., Yuan, and S., Yuantao, "The research of a Web mining method in research areas," *The Sixth Wuhan Interenational Conference on E-Business,* pp. 314-319, 2005.

[3] Y. He, S. C. Hui, and A. C. M. Fong, "Mining a Web citation database for document clustering," *Applied Artificial Intelligence: An International Journal*, vol. 16, pp. 283 - 302, 2002.

[4] F. Christos and W. O. Douglas, "A survey of information retrieval and filtering methods," University of Maryland at College Park, 1995.

[5] Yahoo, [Online]. Available: http://www.yahoo.com, 2011.

[6] Lycos, [Online]. Available: http://www.lycos.com, 2011.

[7] Google, [Online]. Available: http://www.google.com, 2011.

[8] CiteSeer, [Online]. Available: http://citeseer.ist.psu.edu/, 2011.

[9] DBLP, [Online]. Available: http://*dblp.uni-trier.de*/, 2011.

[10] Daniweb.com, [Online]. Available: http://www.daniweb.com/, 2011.

[11] Forums.sun.com, [Online]. Available: http://forums.sun.com/index.jspa, 2011.

[12] Q. T. Tho, S. C. Hui, and A. C. M. Fong, "A Web mining approach for finding expertise in research areas," in *Proceedings of the 2003 International Conference on Cyberworlds*, IEEE CS, 2003.

[13] Q. T. Tho, S. C. Hui, and A. C. M. Fong, "A citation-based document retrieval system for finding research expertise," *Information Processing & Management*, vol. 43, pp. 248-264, 2007.

[14] A. Ismail, S. Sulaiman, M. Sabudin, and S. Sulaiman, "A point-based semi-automatic expertise classification (PBaSE) method for knowledge management of an online Special Interest Group," in *Proceedings of International Symposium on Information Technology*, ITSIM'08, IEEE, vol. 2, pp. 794-800, 2008.

[15] Y. Tianbiao, Z. Jing, Z. Kai, W. Wei, and W. Wanshan, "Study on project experts' evaluation based on analytic hierarchy process and fuzzy comprehensive evaluation," in *Proceedings of the 2008 International Conference on Intelligent Computation Technology and Automation - Volume 01*: IEEE CS, 2008.

[16] S. Yong-Hong, M. Jian, F. Zhi-Ping, and W. Jun, "A group decision support approach to evaluate experts for R&D project selection," *Engineering Management, IEEE Transactions on*, vol. 55, pp. 158-170, 2008.

[17] T. Qijia, M. Jian, L. Jiazhi, C. W. K. Ron, and L. Ou, "An organizational decision support system for effective R&D project selection," *Decis. Support Syst.*, vol. 39, pp. 403-413, 2005.

[18] O. V. Stukach, "Teaching computer science using visualization of the evaluation process," presented at Computational Technologies in Electrical and Electronics Engineering, 2008. SIBIRCON 2008. IEEE Region 8 International Conference on, 2008.

[19] S. Harnad and L. Carr, "Integrating, navigating and analyzing eprint archives through Open Citation linking (the OpCit project)," vol. 79, pp. 629-638, 2000.

[20] M. Ley and P. Reuther, "Maintaining an online bibliographical database: the problem of data quality," presented at Extraction et gestion des connaissances (EGC'2006), Actes des sixiÃ¨mes journÃ©es Extraction et Gestion des Connaissances, Lille, France, vol. 2, RNTI-E-6, 2006.

[21] A. W. K. Harzing and R. van der Wal, "Google Scholar as a new source for citation analysis," *Ethics in Science and Environmental Politics*, vol. 8, pp. 61-73, 2008.

[22] C. Purnima, J. Anupam, Y. Michelle Shu, and R. Ramya, "An expertise recommender using Web mining," in *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference*: AAAI Press, 2001.

[23] C. Richard, V. H. Gareth, and H. Wendy, "An agent based approach to finding expertise," in *Proceedings of the 4th International Conference on Practical Aspects of Knowledge Management*: Springer-Verlag, 2002.

[24] Thomson ISI, "Institute for Scientific Information (ISI) Web Page", Available: http://www.isinet.com, 2011.

[25] B. Everitt, *Cluster Analysis*, 3rd ed. London: Edward Arnold, 1993.

[26] W. M. David and S. A. Mark, "Expertise recommender: a flexible recommendation system and architecture," in *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, ACM, 2000.

[27] Google Scholar, [Online]. Available: http://scholar.google.com, 2011.

[28] MySEIG.org, [Online]. Available: http://www.myseig.org, 2011.

[29] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, pp. 16569-16572, 2005.

[30] W. Glänzel, "On the opportunities and limitations of the h-index," Library of Chinese Academy of Sciences, 2006.

[31] L. Bornmann and H.-D. Daniel, "What do we know about the h-index?," *Journal of the American Society for Information Science and Technology*, vol. 58, pp. 1381-1385, 2007.