MfWMA: A Novel Web Mining Architecture for Expert Discovery

Muhammad Naeem[1], Saira Gillani[2] and Sheneela Naz[2]

[1]*Department of Computer Science, Mohammad Ali Jinnah University Islamabad, Pakistan*
[2]*Centre of Research in Networks & Telecom, Mohammad Ali Jinnah University University Islamabad, Pakistan*
naeems.naeem@gmail.com, sairagilani@yahoo.com, shahneela.cs@gmail.com

Abstract

Identification of expert to domain knowledge in any field of interest is essential for consulting in industry, academia and scientific community. The objective of this study is to address the expert-finding task in contemporary communities. We proposed Multifaceted Web Mining Architecture (MfWMA) and implemented a tool with data extracted from Growbag, dblpXML and web authors home page resource to identify personnel with specific expertise. We mined two thousand and five hundred author's personal web pages with the underlying criteria of a dozen of key parameters; while parsing on each page in pursuit of 8 thousand topics. This study corroborate this quantification in terms of a measure of expertise. The prototype provides its users to distinguish the level of expertise in a particular area; thus resulting in the capability to mark people with broader expertise. Through this extension to the web enabling technique, we have demonstrated that the proposed architecture presents a novel web mining approximation with realistic results.

Keywords: *Text mining; multifaceted; high profile; expertise finding, information seeking*

1. Introduction

In any corporate entity, the knowledge of expertise is a non-trivial resource. Although, critical projects in corporate sector have been observed with focus on design and implementation issues, the success of any project and research problem also involves careful selection of right experts. Collaboration can't be effective unless one can identify the person with whom communication might be required. Previous research has helped clarify the amount of engineering effort devoted to communication. Particularly in engineering, one classic study spent around 16% of their time in communicating with experts [1]. Interestingly, Allen [1] reported a tendency for high-performing engineers to consult much more with experts outside their own discipline as compared to low-performing engineers, although both groups spent almost the same proportion of time for communication.

People with work locations separated by 30 meters have been observed to communicate about as infrequently as people whose workplaces are located in different continents [2], which shows importance of consultancy with concerned experts. So, if any organization expects projects with members spread across multiple floors of a single building, it might experience much-reduced communication among more widely separated members. Previous work suggested an approach for solving the expertise-finding problem. In an empirical study of finding experts in a software development organization, Ackerman, *et al.*, [3] pointed out that experience was the primary criterion engineers ordinarily used to determine expertise. In

fact, developers often used change history to identify those who had experience with a particular file, generally assuming that the last person to change it was most likely to be "the expert". This strategy had several shortcomings, including the inability to determine if s/he is the person who carried out the change had made a large or small change, and whether the person had made many or only a few alterations in the relevant code. Additionally, when someone with capabilities in depth was deemed, it was quite difficult to identify such person from the changing information stored in individual files.

There are also expertise detection systems based entirely on an analysis of user activity and behaviour while being engaged in an electronic environment. Krulwich, *et al.*, [4] have analyzed the number of interactions of an individual within a discussion forum as a means of online structuring an expert's profile. Although such an approach is useful in monitoring user participation, measures such as number of interactions on a particular topic, which in itself is not reflective of knowledge levels of individuals. Knowledge can be categorized into two classes: tacit and explicit knowledge. Management systems focus on explicit knowledge because it can be articulated in written language. However, according to the Delphi Group's study carried out on more than 700 US companies, a large portion of corporate knowledge (42%) was revealed out to be tacit knowledge.

Expertise, a major component of tacit knowledge, is the most valuable knowledge because it defines an organization's unique capabilities and core competencies [5]. The great value of expertise can only be exploited when an individual's expertise can be shared with others [5]. Recently Li, *et al.*, [6] discussed the importance of expert reviewer in the field of marketing. They argued to find the potential influential nodes for effectively and quickly spreading product impressions within a marketing network. However, codifying expertise is difficult and expensive [7]. One effective method of sharing expertise is to enhance people to communicate with each other. Expertise matching – the process of finding experts with a specific expertise – plays an important role in connecting people.

The topic facets efficiently organize one particular facet, using such metadata with respect to user-provided keywords [8]. The main difference to existing (static) facet organizations is that this topic facet is sensitive with respect to time and user community. This provides a motivation for exploiting the currently available metadata from *Growbag* and *dblpXML* collection for computer science. The great value of expertise can be exploited only when an individual's expertise can be shared with others. Generalized processes to find experts are expensive whereas automatic expert finding systems already have delivered ambiguous results. Manual approaches are limited to specific projects only because of costly resources.

However, the fundamental question still remains. How can a person be identified as an expert in a domain? Kajikawa, *et al.*, [9] pointed out that deluge of publications has raised a problem of achieving a comprehensive view even on a topic with limited scope. Zainab et al. [10] argued over the objectivity and functionality of the research publications by showing a detailed statistical data about research publications. In this study, we have relied on two reference models. The first model is Academia Europaea [11]. They have focused on prestigious awards, especially Nobel Laureates, in their membership nomination form. The importance of publications and citations was no doubt considered but it was not the only criteria. The second reference model is Pakistan Academy of Sciences [12]. The page for the fellowship has again focused on numerous local/international awards. We can draw conclusion from careful examination of the two reference models that number of publications and citations does play a role but still there are other factors that organizations consider while selecting an individual as an eminent expert in a specific field.

This study explores the discovery of expertise within the context of a digital electronic journal; the *Growbag* an updated *dblpXML* has very large number of articles covering all

topics of computer science. A reference work related to the journal-ranking problem has been drawn recently [13]. Our work handles the problems of finding experts using automatic multifaceted approach, which handles automation errors using multi-feature extraction. We justified results by multiple facets using different metrics and find appropriate intensive experts. Presented work mines different metrics from *Growbag* dataset resulting in weighted constrains while calculating expert score. Facets offer different dimensions. Such facets can be considered a way to categorize content or document collections for intuitive user interaction. We shall summarize the main contributions of this work as below.

- To the best of our knowledge, the proposed technique to dig out the web-based faceted ranks is first of its kind in the area of finding experts in academics.
- Our main contributions center around a context-sensitive web mining based approach heuristic inspired by the concept of finding automated and manual approach as described by Afzal, *et al.*, [7].

The technique is aimed at rendering help to journal editors and conference organizers to assign score to mark any author for his/her potential role in capacity of reviewer-ship.

2. Related Work

Discovery of expertise is not a trivial task. Many people and organizations are working on it to fairly find an expert. Both autonomous system and manual efforts have been exercised. In manual approach persons have to perform huge amount of effort but in the end quality of output is very fine. Many measuring factors are used to find the pertinent information in finding experts. An expert is a major member (either a software agent or a human expert), with the knowledge of the agent world in a complex multi-agent domain but with focused expertise for a particular problem solver in a special field [14].

Finding an expert may vary from field to field such as: for academia profile, projects, publication and many other factors (herein called weights) could be used to find the exact expert. If we talk about finding reviewer expert work, then *Most Expert Finder* systems are based on highly localized, privatized and specialized datasets, beneficial only in narrow margin with small settings [15]. By facilitating the task of finding suitable reviewers, we anticipate that the quality of an overall conference could improve, since both the number of reviewers available for consideration would be larger and the extent of their expertise would be determined and useful in the selection process. If we delve into the application of expert discovery, then there is a potential possibility to fulfil the requirement of fair distribution of staff in an enterprise and all together the same can be applied into projects, awards, publications, etc. Unfortunately, active experts do not have enough time to preserve sufficient descriptions of their continuously changing and specialized skills [16]. One notable example is MITRE database where it was pointed out that quickly maintaining and updating previous experience databases is not a trivial job.

ExpertFinder fills this gap by mining information and activities related to experts while providing it in an intuitive fashion to end-users [16]. A specific example is university that is considered a well knowledge-based organization. The authorities at universities have also realized that effective development and management of their organizational knowledge base is critical for survival in today's competitive service industry. The knowledge and expertise of a university staff involved in teaching and research in various areas is the major asset that a university holds [17]. When the user searches using a specific term, the system ranks employees by the mentioned term or phrase and its statistical association with the employee name resulting into the realization that one of the most important problems in developing

expert systems is knowledge acquisition from experts [18, 16]. In order to mechanize this problem, many techniques and inductive learning methods, such as induction of decision trees [19, 20], rule induction methods [19, 21, 22] and rough set theory [23, 24] were introduced and performed which have been shown reasonable suitability to extract knowledge from databases. Other researchers investigated the discovery of communities of practicing experts via a prototype called XperNet [16]. XperNet is designed to extract expertise networks. It uses statistical clustering techniques and social network analysis to glean networks or affinity groups consisting of people having related skills and interests [16, 25].

Mockus, *et al.*, [26] applied a technique over data from a software project's change management records to find people with desired expertise in a large organization [26]. In literature, some other systems have been reported which detect experts entirely on an analysis of user activity, behaviour, likes and dislikes while being engaged in an electronic environment. A notable example in past decade is the analysis of number of interactions of an individual within a discussion forum as a means of constructing an expert's profile [4]. Even though this kind of approach is helpful in monitoring user contribution, the measures such as number of interactions on a particular topic in itself requires significant insightful knowledge levels of individuals. Another approach discussed in literature was related to use of semantic structure expert/expert-locator (EEL) pair requests for technical information in a large research and development company [27]. The system automatically constructs a semantic space of organizations and terms, using a statistical matrix decomposition technique (singular value decomposition) to represent semantic similarity present in large text sources. McDonald, *et al.*, [28] reported on a system that uses various files organizationally closest to the requester, and how well the requester knows the expert (based on a previous analysis of the social network in the organization). The problem of finding experts is not limited to widely distributed teams, however. In fact, people whose offices are separated by 30 meters communicate about as infrequently as people who are located on different continents [29].

S.D Neil, *et al.*, [30] analysed quality filter in scientific communication process and proposed information analyst be used as a filter to identify quality research papers, especially using the validity criterion, fact lead to our author research work quality phenomena to extract legendary in field. Awang Ngah Zainab, *et al.*, [31] measured trends for expert systems in library and information services based upon authorship patterns and expressiveness of published titles. He identified the total, trends, focus of studies, authorship pattern and expressive quality of publications covering ES (expert system) applications in the broad or sub-domain of LIS (library information system). Robert P. Vecchio, *et al.*, [32] raised issue of particularistic bias, agreement, and predictive validity in manuscript Review Process. He applied his study process on 853 manuscripts and initial study shown majority of the reviewed papers were rejected after initial review (603, or 81.6%), whereas the remainder (136, or 18.4%) received an invitation to revise and resubmit, which leads research quality. Anne S. Tsui, *et al.*, and John R. Hollenbeck, *et al.*, [33] suggest that conversation should be about addressing the large gap between the demand for effective reviewers and the supply of individuals who are both successful authors and effective reviewers. Towards parallelism in a structural scientific discovery Gehad M. Galal, *et al.*, [34] investigated approaches for scaling particular knowledge discovery in databases (KDD) system to discover interesting and repetitive concepts in graph-based databases from a variety of domains.

A similar approach proposed by Mockus, *et al.*, [26] could be adopted to compute expertise for researchers across different topics. Studies indicate that engineers and scientists instinctively do not communicate much with colleagues whose offices are distant to each other, so there are fewer opportunities to find out whoever holds expertise in various areas when teams are distributed [26]. Cameron, *et al.*, [15] collected the expertise of a subset of

researchers who have published papers in World Wide Web and Semantic Web Conferences. This dataset includes more than 1,200 researchers and 1,504 relationships to about 100 unique topics. Expertise, a major component of tacit knowledge, is the most valuable knowledge because it defines an organization's unique capabilities and core competencies [5]. The most widely used approach for expertise matching within academia is to build an expertise database where individuals specify their expertise using several keywords or short sentences resulting in empowering the users to search these databases to find an expert [17]. A prototype system has been implemented based on the architecture with the aim to help PhD applicants find potential supervisors [17]. The literature details a number of systems that undertake a fully automatic approach to locate experts, including, Who Knows [27], Agent Amplified Communications [35], Contact Finder [4], Yenta [2], MEMOIR [35], Expertise Recommender [28], Expert Finder [36], SAGE [37] and the KCSR Expert Finder [38]. This is reflected by wide variety of expertise evidence, such as emails [35], electronic messages on bulletin boards [4], program codes [39, 40], Web pages [2, 35], and technical reports [41, 38] used in expert finder systems. Sim, *et al.*, [41] proposed that the heterogeneity of information sources should be used as an indicator for reflecting experts' competencies. Expert finder systems can be integrated into other organizational systems, such as information retrieval systems, recommender systems and Computer Supported Cooperative Work systems [41].

XML is accepted as the standard for data interchange [42]. Heterogeneous data structures can be represented in a uniform syntax (nested tagged elements). On the other hand in XML user can add tags and the same information can be represented differently by different XML structures. Recently Razikin, *et al.*, [43] carried out an important work investigating the effectiveness of tags in facilitating resource discovery by means of machine learning and user-centric approaches. They showed that all of the tags are not useful for content discovery. Their research was limited to only 100 frequent tags extracted from a corpus of 2,000 documents. Lu, *et al.*, [44] reported importance of tagging in social computing within the domain of digital library science. They highlighted the difference and connections between expert-assigned subject terms and social tags in order to uncover the potential obstacles for implementation of social tagging in the domain of digital libraries. Researchers as well as organization design different systems, tools for expert discovery whether their techniques are different but aim is same to find the expert in nick of time so that time could be saved. Chen, *et al.*, [13] argued that previous studies addressed the problem of journal ranking through expert survey metrics, or using an objective approach such as citation-based metrics. They suggested integrating both of these approaches. However, their work focused only on journal ranking problem [13].

By virtue of the complexity of temporary nature of transient information available on the web, it has been a challenge to find out the right actor in mixed service-oriented systems. [45]. Daniel, *et al.*, [45] presented an approach Human-Provided Services (HPS) with the argument of necessitating automated inference of knowledge and trust in an environment of distributed collaboration. They illustrated that the skill and capabilities of experts be treated as service. Recently, Lopez, *et al.*, [46] has reported the importance of coordination of expertise based upon crowd sourcing so that the corporate services, including IT Service Delivery, IT Inventory Management and End-User Support, can benefit from the knowledge network.

3. Problem Statement

We shall formulate the problem statement into two sections:

1. Can an expert (*E*) in an academic environment be ranked (*R*) by its web weights (*W*) alongside the conventional ranking scores (*S*) such as citations, co-author network and publication count?

$$E^R <- \bigcup_{s_i \in S} s_i + \bigcup_{w_i \in W} w_i$$

2. Is any correlation is found between web weights and non-web weights?

$$\bigcup_{s_i \in S} s_i <-> \bigcup_{w_i \in W} w_i$$

To solve this problem, we need to find expert weight from *Growbag* dataset, *dblpXml* and author's homepage. This leads to our focus on mining web for author's homepages to identify multifaceted parameters to rank and build expert profile. Authors profile required building with highly concerned parameters to identify highly ranked authors on specific domain.

4. Web Mining for Expert Discovery

In order to achieve the optimized utilization of the expertise held by individuals within an organization, various organizations have reportedly adopted the searching system: Expert Recommender Systems (ERS). Usually, the prime interest of an inquirer is to find out an expert to address a specific problem [47]. Albeit ERS permits quick searching of experts, the inquirers may notice the absence of capability of system for informing accurate usefulness. Fully automated systems have been reported as an alternative to these self-reporting recommender systems such as SAGE [37], bulletin boards [4], systems with email as input [48], Web pages [2], software coding system [28, 36], technical reports [38] and the artefacts of social software systems such as Wikis and Weblogs and also social networks, *e.g.*, Lin and Griffiths-Fisher, *et al.*, [48]. However, Crowder, *et al.*, [38] found that systems mimic like ERSs have been found to be prone to problems of concerning expertise analysis support, heterogeneous information sources, reusability and interoperability. Ehrlich, *et al.*, [48] illustrated the social impact of finding and contacting domain experts. They discussed SmallBlue and ERS developed for IBM for mapping each staff member's social network for providing the information of "who is connected to whom and where social networks overlap". Competent expert discovery systems in the past have been innovatively applied in helping PhD scholars and research community in finding germane supervisors [17]. Peer-reviewers identification for conferences and the former made use of a manually derived expertise profile database and employed reference mining for all papers submitted to a conference [49]. Later on, co-authorship network was constructed for each submitted paper making use of a measure of conflict-of-interest to ensure that associates did not review papers. Manually constructed taxonomy in which manually crafted taxonomy employed for 100 topics in DBLP covering the research areas of a small sample of researchers appearing in DBLP [15].

We enhanced the work towards topics identification and considered co-occurred keywords as well as general term as Topics for Growbag dataset. Our technique efficiently finds credible results for which we developed a tool. We retrieved authors from each topic with their publication analysis. Moreover, we employed technique of web mining for author's homepages to get their profiles in different aspects. Our proposed work and implemented tool

considerably delivered results of more than 2,500 experts' homepages analysis on behalf of multifaceted parameters.

Algorithm 1. Expert Profile Algorithm

Input: Topic T. Year x. WebfactorCount k
Output: Collection of Authors with their ranks

for *each Topic do*
get 'authors'
for each *'author' do*
get *'author's co-author network size'*
get *'author's publication'*
get *'publication'* in last x years
get_auth_home-Page
for each *"Home Page" do*
get *Bool P_Score*
{" 'Project' , 'Awards' , 'Honorarium' , 'Affiliations' , 'RFCs' , 'Supervision' , 'Collaboration' , 'Relevance' , 'Keynote_Speaker' , 'Reviewer' , 'Protocol Design' , 'Distinctions' "
}
end

non_web_wt ← (citations/size(publications));
non_web_wt ← non_web_wt +(size(publications/size(co.auth.net));
non_web_wt ← non_web_wt +(size(publications/last_x_years_publication);
non_web_wt ← non_web_wt +(size(publications in relevant field /publication);

$$web\_wt \leftarrow \sum_{i=1}^{k} web\_factor$$

Expert Profile ← non_web_wt + web_wt;
end
return *expert profile;*
end

5. Proposed Methodology

We employed and focused our work on dblpXML for mining home pages to build expert profile in MfWMA. In this aspect we sorted out different facets like contribution of a particular domain expert, authors project contribution. In this view we parsed her/his online homepage to search out whether s/he majorly contributed in well-known project or supervised? Whether s/he received awards and other achievements? *Growbag* database provided by DBLP has been reported as very imperfect database for researcher in the domain of computer science [50]. In this study, we have endeavored to identify the reviewers behind the research papers in the margin of qualifying scoring weights. We have precisely classified the weights into two categories, *Grwobag* weights (or non-web weights) and web weights. Incomplete as well as inconsistent information were not treated at all. The model in which we acquired different weights to fill expert profile building blocks has been shown in the Figure 1.

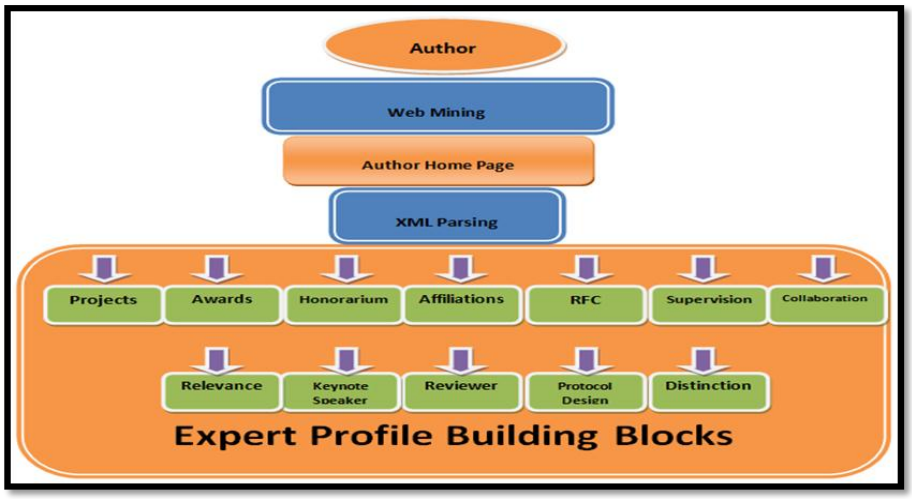

Figure 1. Expert Profile Building Blocks (MfWMA)

We shall describe each of them as below:

- **Projects:** We employed parameter of Project on Authors homepage while using text mining techniques and XML parsers to find whether there is any role of author in any technical project. At first level we used Boolean values to take decision of this parameter.
- **Awards:** An award is a key weight to find out an author's credibility. This leads to our examination whether there is any reputed award won by authors.
- **Honorarium:** Honorariums deliver the benchmark values of author's contribution showing his/her contribution in his domain in well-formed way.
- **Affiliations:** It shows an author's significant influential role in his field because multiple affiliations build the portfolio. This indicates the versatile proficiency of authors in various domains of knowledge.
- **RFCs:** Request for Comments (RFCs) is popular in the domain of computer networks and communication. RFC is produced as the result of a large number of experimentation in a specific field. Usually, RFC is not ranked, but we were impressed by the reality that author's practical experimentation knowledge in a specific area based on a large-scale handshaking methodologies demands lot of expertise. Consequently, if an author has profile with contribution in RFC, then it is a positive and loud indication of his/her expertise in particular domain.
- **Supervision:** A PhD scholar needs supervisor and a researcher needs guidance in project supervision, etc. Supervisor plays vital role in the success of any project or scholar's research deliverables, which was included as weight in our expert profile.
- **Collaboration:** Experts in every field play role as collaboration in different versatile features, which impact better on community considered as weight.
- **Relevance**: Basically for the domain expert it is necessary to find expert relevant to the field. So an expert belonging to the B-Topic is not meant for C-Topic within scope of consideration for B-Topic, so we evaluated relevance.

- **Keynote Speaker**: A keynote speaker in any domain of knowledge demonstrates the gist of a theme. Not only in corporate but also in commercial environments, a keynote speaker enjoys a significant importance. Prime functionality of the keynote speaker is to lay down the framework associated with the central dogma of a theory or discussion. In other words, we can say that a keynote speaker can play a role in the capacity of convention moderator whether it is the process of reviewing research articles or examining any experimental evaluation. The crucial importance of keynote speaker has motivated us to include this status in our web weights.
- **Reviewer:** A reviewer is an expert who evaluates a product. The product may be scholarly publication or an industrial/commercial service or hardware. In an academic journal or conference, a reviewer decides and measures the strength of contributed knowledge diffusion. Any person who is already involved in the capacity of reviewer indicates that s/he is trusted by an organization. So we consider this measure considerably for building blocks of expert profile.
- **Protocol Design:** Protocol standards are the patent resource of communication and processing within heterogeneous environment which necessarily build upon an intelligence strategy of handshaking or other protocol requirements demonstrating an author's value and hands on expertise in relevant domain. These reasons were sufficient to consider it in one of web weights in this research work.
- **Distinctions:** If author A has significant distinction among his/her peers, then it indicated his/her credibility towards expertise profile building.

6. Experimental validation

This section will elaborate our results with their validation in detail. The performance of the system is measured on standard statistical measures including sensitivity, specificity and selectivity. The performance measures of implemented system are given by equation 1 to 5. These measures are defined formally as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

$$Error = \frac{FN + FP}{TP + TN + FP + FN} \quad (2)$$

$$Sensitivy(\mathrm{Re}call) = \frac{TP}{TP + FN} \quad (3)$$

$$Selectivity(\mathrm{Pr}ecision) = \frac{TP}{TP + FP} \quad (4)$$

$$F - measure = 2 \times \frac{\mathrm{Pr}ecision \times \mathrm{Re}call}{\mathrm{Pr}ecision + \mathrm{Re}call} \quad (5)$$

In the domain of information retrieval, the accuracy is described as the degree of closeness of measurements towards its real quantitative value. Conventionally, experts are measured in terms of number of publications and citations, *etc.* S.D. Neil [30] pointed out that judgment of quality of the produced research articles is of great importance. They proposed that the information analysis be used as a gauge filter of a research paper's quality. As shown in

Figure 2, the error rate for all of the web weights ranges from 5% to 14%. The highest error rate we encountered is in RFC. The precision which is equivalent to selectivity is also described as the degree of closeness but with repeatability experiment. It was discussed in the literature that accuracy-cum-error rate alone is not sufficient to describe any measurement values but presision is also a mandatory requirement. In the literature, two kinds of erros have been reported: error of accuracy or error of precision.

A close examination of Figure 2 and 3 shows that the errors encountered in retrieving the results are of precison. This statement can be validated by the fact that the error of accuracy is always biased in some specific direction and usually delivers a speicif pattern. However, this is not true in our case where no significant pattern is observed conforming our statement that this is error of precison in nature. Yet again it was pointed out that precision alone is not enough rather recall is also an important measure for the presentation of the estimation of the results. Another measure which encompasses both precison and recall is known in the name of F-measure. It has been exploited significantly in scientific experiments for the validation of the results. Figure 3 illustrates the detail of the F-measure of each of the web weights. A careful examination highlights that "RFC", "Protocol Design" received very low F-meaure followed by "Honorarium". On the other hand, "Awards", "Affiliations", "Project", "Distinction" and "Collaboration" exhibit high F-measure value which shows the strenth of results used in our methodology. The rest of the web weights deliver intermediate values of F-measure. This analysis shows that more than 50% web weights yield reliable results.

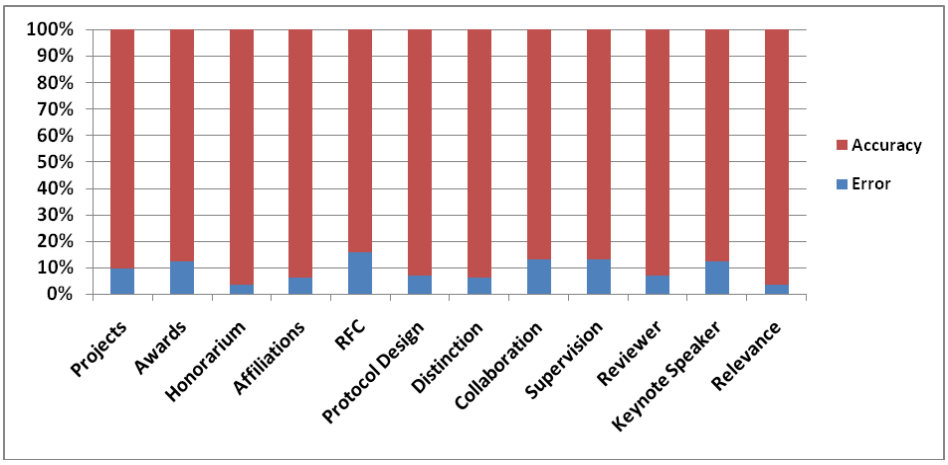

Figure 2. Web Weights Accuracy and Error obtained comparison

An expected relation R between S and W is a subset of Cartesian product (S x W). If $(s, w) \in R$ or $s\,R\,w$. When R holds a relationship on a set S it means that R is a subset of SxS. This arise the investigation into the reflexive, symmetric and transitive relationship.

Lemma-1

Web Scores S and web scores W both does not hold reflexive, symmetric and transitive relation such that the relationship $R \subseteq S \times S$ exists if $s\,R\,w$ such that $s \in S$.

Proof:

It is evident from the experimental validation depicted from Figure 3 that for every member of the conventional non-web score, a positive or negative relationship exists. The figure shows that a positive relationship exists for every member of S towards every member

of web scores. This indicates that no strong relationship exists between members of both of the sets. In general only some of the factors have tight relationship towards the web scores. But nevertheless a monotonic relationship is found. It corroborates that class of empty pair-wise disjoint sets are found. Hence, it is proven that both of the sets have no reflexive, symmetric or transitive relationship.

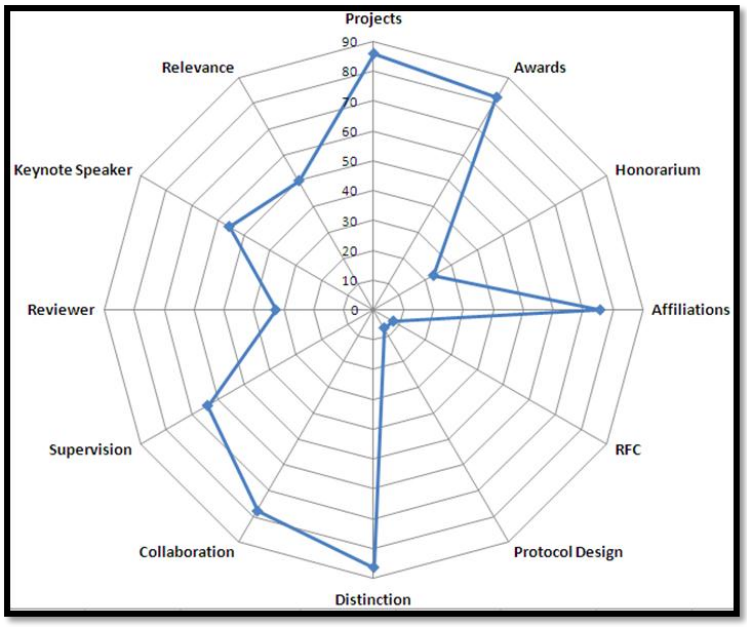

Figure 3. F-Measure for web weights used in the study

Lemma-2:

No equivalent relationship exists between both conventional scores and web scores such as: $\bigcup_{s_i \in S} s_i <-> \bigcup_{w_i \in W} w_i$.

Proof:

We must show that the relationships of both set S and W are tied into a relationship such as E^R. In order to prove it, we need to show that E^R is non-empty set. However, as R does not possess any reflexive, symmetric and transitive properties and it is already known that if a set holds these properties then members of each pair of this set exhibit equivalent classes in connection to their respective domain and range. In our case, the domain and range are non-web conventional scores (*S*) and web scores (*W*), respectively. This converges into the fact that both of these sets exhibit non-equivalent relationship.

7. Experimental Evaluation

In previous sections, we first argued sufficiently over the importance of identification of experts in any domain; secondly, we presented our results with their statistical analysis. However, identification and ranking of these experts is a debatable issue. We concluded that numbers of citations, size of co-author network or publication count alone are not sufficient for ranking experts. But other web factors, which we termed as multifaceted web parameters

or web weights, are also important. We shall cite an example of a notable professor at Nanyang Technological University Singapore. Dr. Sun Chengzheng is Professor at School of Computer Engineering. According to record set retrieved from *Growbag*, his publication count is 20 with citations count of 33 making a size of co-author network only 15 during period of 1996 to 2002. Apparently, these statistics show that the professor is not a high expert in the field. However, the actual facts are quite different. Professor Dr. Sun Chengzheng earned double PhD in two distant fields of computing. Since two decades, he has been vigorously active in projects related to computer networks and its allied technologies. He has been editor of many reputed journals as well as conference reviewers. He has collaboration with Australian and various Chinese universities. He worked in capacity of keynote speaker at various international industrial seminars. He runs half dozen research projects and the same number of research prototype systems. Moreover, he supervised 11 postgraduate students out of which seven hold PhD degree and are working in reputed organizations. This short example is enough to validate the fact that the conventional parameters of citations, publication counts, etc, are not enough but other more robust parameters should also be incorporated while ranking an expert. In support of this analogy, we shall cite a sentence from Academia Europaea Membership Nomination Form which states that *"mention Honours and Awards (Only mention major awards; max. 20; do not mention best paper awards or fellowships that one gets if one just pays a membership fee*"[11].

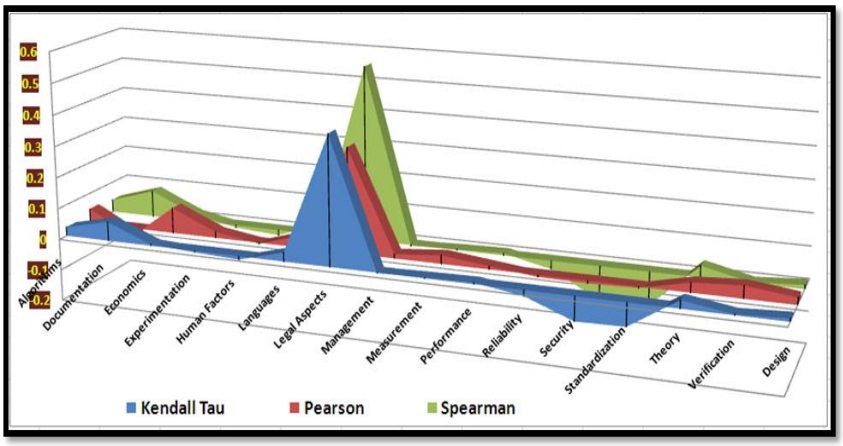

Figure 4. Correlation between non parametric values (web and non-web weights)

Another question that needs to be tweaked is: "What is the relationship between both sets of parameters". The answer lies in the statistical correlation measure for non-parametric features as shown in Figure 4. If we assume that there are two features, one is web faceted score and the other is non web faceted score. A careful examination of both of these features indicates that these features observe no probability distribution in general. There are a lot of correlation ranking measures for non-parametric features. However, considering the nature of scoring result set generated, we employed Kendall's tau-b, Pearson Correlation rank Spearman's rank correlation coefficient [51, 52]. Figure 4 indicates probabilistic existence of causation between the two kinds of parameters. While applying these correlation measures, we considered non-web-weights as criterion feature whereas the web weights were considered as predictor feature. We can conclude that a correlation was observed in case of

general term "legal aspect". However, in case of security and standardization, a weak or negligible correlation was found between both ranking weights.

8. Conclusions & Future Work

It has always been a desire of every organization to contact the most suitable and right person well in time. This study has addressed the issue of finding a better expert defined within several parameters. The study investigates the problem of topic's expert finding in *Growbag* dataset while using *dblpXML* to access author's homepages. An architecture MfWMA has been developed which was used in the context of identifying computer science topics experts and assigning reviewers. Prime contribution of this study is the introduction and implementation of novel idea of web mining with 12 web faceted parameters. For shrewd reader, complete result dataset can be asked from authors of this research. Our framework mined more than 2,500 Author's web pages on basis of 12 key parameters while parsing on each page for a large number of co-occurred keyword and all available general terms. Results presented evidence to validate our quantification measures of expertise in which we extracted most relevant experts in a growbag dataset. We delivered a credible and remarkable multi-facets mining technique which considerably enhance and helped research community to get their required domain expert. In future we have positive intention to tweak the peculiarities related to other domain converging into solution for building up a system in order to categorize the domain experts in the same way as we perceived and implemented in this study. Future work is aimed toward more robust, saleable and efficient optimization methodology in multi-objective direction focusing on complex expert judgments.

References

[1] T. J. Allen, "Managing the Flow of Technology", Cambridge, MA: MIT Press, **(1977)**.
[2] L. Foner and N. Yenta, "A Multi-Agent Referral-Based Matchmaking System", In Proceedings of the First International Conference on Autonomous Agents, Marina del Rey, CA, **(1997)**, pp. 301-307.
[3] M. S. Ackerman and C. Halverson, "Considering an Organization's Memory", Computer Supported Collaborative Work, **(1998)** Seattle, WA: ACM Press, pp. 39-48.
[4] B. Krulwich and B. Burkey, "The ContactFinder Agent: Answering Bulletin Board Questions with Referrals", Proc. 13th Nat. Conf. on AI, vol. 1, Portland, Oregon, **(1996)**, pp. 10-15.
[5] L. Olson and R. Shaffer, "Expertise Management – and Beyond", White paper in RGS Associates, **(2002)**
[6] Y. M. Li, C. H. Lin and C. Y. Lai, "Identifying influential reviewers for word-of-mouth marketing", Electronic Commerce Research and Applications, vol. 9, **(2010)**, pp. 294–304.
[7] M. T. Afzal and H. Maurer, "Expertise Recommender System for Scientific Community", Journal of Universal Computer Science, vol. 17, no. 11, **(2011)**, pp. 1529-1549.
[8] W. K. Balke and K. Mainzer, "Knowledge Representation and the Embodied Mind: Towards a Philosophy and Technology of Personalized Informatics", K. D. Altho, *et al.*, (Eds.): WM 2005, LNAI 3782, Springer-Verlag Berlin Heidelberg **(2005)**, pp. 586 – 597.
[9] Y. Kajikawa, K. Abe and S. Noda, "Filling the gap between researchers studying different materials and different methods: a proposal for structured keywords", Journal of Information Science, vol. 32, no. 6, **(2006)**, pp. 511–524.
[10] A. N. Zainab and S. M. D. Silva, "Expert systems in library and information services: publication trends, authorship patterns and expressiveness of published titles", Journal of Information Science, vol. 24, no. 5, **(1998)**, pp. 313–336.
[11] Academia Europea, http://www.ae-info.org/ae/Acad_Main/Sections/Informatics.
[12] PAS (Pakistan Academy of Science), http://www.paspk.org/indexa.htm.
[13] Y. L. Chen and X. H. Chen, "An evolutionary PageRank approach for journal ranking with expert judgements", Journal of Information Science, vol. 37, no. 3, **(2011)**, pp. 254–272.
[14] Z. Minjie, T. Xijin, B. Quan and G. U. Jifa, "Expert Discovery and Knowledge Mining In Complex Multi-Agent Systems", J Syst Sci Syst Eng., vol. 16, no. 2, **(2007)** June, pp. 222-234.

[15] D. Cameron, B. Aleman-Meza and I. B. Arpinar, "Collecting Expertise of Researchers for Finding for Relevant Experts in Peer-Review Setting", Proc. of 1st International Expert Finder Workshop Berlin, Germany, **(2007)** January 16.
[16] T. Mark and M. Mitre, Technical Report "Expert Finding Systems", **(2006)** September.
[17] P. Liu and P. Dew, "Using Semantic Web Technologies to Improve Expertise Matching within Academia", Proceedings of I-KNOW **(2004)** Graz, Austria, June 30 - July 2.
[18] B. G. Buchanan and E. H. Shortliffe, "Rule-Based Expert Systems", Addison-Wesley, **(1984).**
[19] J. R. Quinlan, "C4.5 Programs for Machine Learning", Morgan Kaufmann, **(1993)**, CA.
[20] L. Breiman, J. F'reidman, R. Olshen and C. Stone, "Classification and Regression trees", Belmont, CA: Wadsworth International Group, **(1984)**.
[21] R. S. Michalski, J. G. Carbonell and T. M. Mitchell, "A Theory and Methodology of Machine Learning – An Artificial Intelligence Approach", Morgan Kaufmann, Palo Alto, **(1983)**.
[22] R. S. Michalski, I. Mozetic, J. Hong and N. Lavrac, "The Multi-Purpose Incremental Learning System AQ15 and its Testing Application to Three Medical Domains", Proceedings of the fifth National Conference on Artificial Intelligence, **(1986)**, pp. 1041-1045, AAAI Press, Palo Alto.
[23] Z. Pawlak, "Rough Sets", Kluwer Academic Publishers, Dordrecht, **(1991)**.
[24] W. Ziarko, "Variable Precision Rough Set Model", Journal of Computer and System Sciences, vol. 46, **(1993)**, pp. 39-59.
[25] M. Maybury, R. D'amore and D. House, "Expert finding for collaborative virtual environments", Commun. ACM, vol. 44, no. 12, **(2001)**, pp. 55–56.
[26] A. Mockus and J. D. Herbsleb, "Expertise Browser: A Quantitative Approach to Identifying Expertise", In Proceedings on the International Conference on Software Engineering, Florida, USA, **(2002)** May 19-25, ICSE'02, pp. 503-312.
[27] L. Streeter and K. Lochbaum, "An Expert/Expert-Locating System Based on Automatic Representation of Semantic Structure", Proceedings of the Fourth Conference on Artificial Intelligence Applications, Computer Society of the IEEE, San Diego, CA, **(1988)** March, pp. 345-349.
[28] D. W. Mcdonald and M. S. Ackerman, "Expertise recommender: a flexible recommendation system and architecture", In Proc. of the 2000 ACM Conference on Computer Supported Cooperative Work (CSCW '00), **(2000)** December 2-5, Philadelphia, Pennsylvania, USA, pp. 231-240.
[29] T. J. Allen, "Managing the Flow of Technology", Cambridge, MA: MIT Press, **(1977)**.
[30] S. D. Neill, "The information analyst as a quality filter in the scientific communication process", Journal of Information Science, vol. 15, **(1989)**, pp. 3.
[31] A. N. Zainab and S. M. De Silva, "Expert systems in library and information services", Journal of Information Science, vol. 24, **(1998)**, pp. 313.
[32] P. R. Vecchio, "Journal Reviewer Ratings", Bulletin of Science Technology & Society, vol. 26, **(2006)**, pp. 228.
[33] A. S. Tsui and J. R. Hollenbeck, "Successful Authors and Effective Reviewers", Journal of Information Science, vol. 05, **(2008)**, pp. 6.
[34] G. M. Galal, D. J. Cook and B. L. Holder, "Exploiting Parallelism in a Structural Scientific Discovery System to Improve Scalability", Journal of the American Society for Information Science, vol. 50, no. 1, **(1999)**, pp. 65–73.
[35] H. A. Kautz, B. Selman and M. Shah, "Referral Web: Combining Social Networks and Collaborative Filtering", Communications of ACM, vol. 40, no. 3, **(1997)**, pp. 63-65.
[36] A. Vivacqua, "Agents for Expertise Location", In Proc. AAAI Spring Symposium on Intelligent Agents in Cyberspace, Stanford, CA, **(1999)**, pp. 9-13.
[37] I. Becerra-Fernandez, "Searching for experts on the Web: A review of contemporary expertise locator systems", ACM Trans. Internet Technol., vol. 6, **(2006)** November 4, pp. 333-355.
[38] R. Crowder, G. Hughes and W. Hall, "An agent based approach to finding expertise", In Proceedings of 4th International Conference on Practical Aspects of Knowledge Management, Berlin Heidelberg, **(2002)**, pp. 179-188.
[39] F. Sebastiani, "Machine Learning in Automated Text Categorization", ACM Computing Surveys, vol. 34, no. 1, **(2002)**, pp. 1-47.
[40] R. E. Kraut, C. Egido and J. Galegher, "Patterns of Contact and Communication in Scientific Research Collaboration", in Intellectual Teamwork: Social and Technological Foundations of Cooperative Work, J. Galegher, R.E. Kraut, and C. Egido, (Eds.), Lawrence Erlbaum Associates: Hillsdale, NJ, **(1990)**, pp. 149-171.
[41] Y. Sim, R. Crowder and G. Wills, "Expert Finding by Capturing Organizational Knowledge from Legacy Documents", In Proc. Int.l Conf. on Comp. & Comm. Eng., **(2006)** (ICCCE '06) KL, Malaysia.
[42] T. Bray, J. Paoli, C. M. Sperberg-McQueen and E. Maler, "Extensible Markup Language (XML) 1.0 (Second Edition)", W3C Recommendation, **(2000)** October.

[43] K. Razikin, D. H. Goh, A. Y. K. Chua and C. S. Lee, "Social tags for resource discovery: a comparison between machine learning and user-centric approaches", Journal of Information Science, vol. 37, no. 4, **(2011)**, pp. 391–404.
[44] C. Lu, J. Park and X. Hu, "User tags versus expert-assigned subject terms: A comparison of LibraryThing tags and Library of Congress Subject Headings", Journal of Information Science, vol. 36, no. 6, **(2010)**, pp. 763–779.
[45] S. Daniel, S. Florian and D. Schahram, "Expert Discovery and Interactions in Mixed Service-oriented Systems", IEEE Transactions on Services Computing, **(2012)**.
[46] M. Lopez, M. Vukovic and J. Laredo, "PeopleCloud Service for Enterprise Crowdsourcing", 2010 IEEE International Conference on Services Computing **(2010)** July 5-10, Miami, Florida.
[47] D. Yimam-Seid and A. Kobsa, "Expert Finding Systems for Organizations: Problems and Domain Analysis and the DEMOIR approach", Jrnl of Org.l Comp. & Electronic Commerce, vol. 13, **(2003)**, pp. 1-24.
[48] K. Ehrlich, C. Y. Lin and V. Griffiths-Fisher, "Searching for experts in the enterprise: combining text and social network analysis", In: GROUP '07: Proc. 2007 Int.l ACM conf. on Supporting group work. ACM, New York, NY, USA, **(2007)**, pp. 117-126.
[49] M. A. Rodriguez and J. Bollen, "An algorithm to determine peer-reviews (Technical Report)", Los Alamos National Laboratory, **(2006)**.
[50] M. Ley, "DBLP: Some Lessons Learned", Proc. VLDB Endow, vol. 2, no. 2, **(2009)**, pp. 1493–1500.
[51] V. Bagdonavicius, J. Kruopis and M. S. Nikulin, "Non-parametric tests for complete data", ISTE&WILEY: London & Hoboken, ISBN 9781848212695, **(2011)**.
[52] G. W. Corder and D. I. Foreman, "Nonparametric Statistics for Non-Statisticians: A Step-by-Step Approach", Wiley, ISBN 9780470454619, **(2009)**.