Original Article

Healthc Inform Res. 2013 December;19(4):243-249.
http://dx.doi.org/10.4258/hir.2013.19.4.243
pISSN 2093-3681 • eISSN 2093-369X

HIR
Healthcare Informatics Research

Developing a Biomedical Expert Finding System Using Medical Subject Headings

Harpreet Singh, PhD, Reema Singh, PhD, Arjun Malhotra, MSc, Manjit Kaur, MCA
Bioinformatics Centre, Indian Council of Medical Research, Ansari Nagar, New Delhi, India

Objectives: Efficient identification of subject experts or expert communities is vital for the growth of any organization. Most of the available expert finding systems are based on self-nomination, which can be biased, and are unable to rank experts. Thus, the objective of this work was to develop a robust and unbiased expert finding system which can quantitatively measure expertise. **Methods:** Medical Subject Headings (MeSH) is a controlled vocabulary developed by the National Library of Medicine (NLM) for indexing research publications, articles and books. Using the MeSH terms associated with peer-reviewed articles published from India and indexed in PubMed, we developed a Web-based program which can be used to identify subject experts and subjects associated with an expert. **Results:** We have extensively tested our system to identify experts from India in various subjects. The system provides a ranked list of experts where known experts rank at the top of the list. The system is general; since it uses information available with the PubMed, it can be implemented for any country. **Conclusions:** The expert finding system is able to successfully identify subject experts in India. Our system is unique because it allows the quantification of subject expertise, thus enabling the ranking of experts. Our system is based on peer-reviewed information. Use of MeSH terms as subjects has standardized the subject terminology. The system matches requirements of an ideal expert finding system.

Keywords: Medical Subject Headings, Data Mining, Online Systems, Expert Systems, Professional Competence

I. Introduction

The ability to rapidly identify subject experts is essential for the successful functioning of an organization. Some of the advantages of an expert finding system include 1) rapid formation of operational or proposal teams to accelerate research [1]; 2) identification of potential collaborators; 3) matching reviewers to submitted research proposals, manuscripts and other peer-reviewed documents [2]; 4) identification of expertise available within organizations [3]; 5) monitoring the research priorities of an organization; and 6) prediction of the effects of skill loss (attrition or retirement) or gain (merger or acquisition).

Though useful, locating and evaluating subject experts is a difficult task because experts are rare and unevenly distributed, while the requirements of expertise seekers are often poorly articulated. They often lack information about the past performance of experts, and it is difficult to classify and quantify expertise. Relocation of experts further complicates

Submitted: August 19, 2013
Revised: 1st, December 5, 2013; 2nd, December 12, 2013
Accepted: December 17, 2013

Corresponding Author
Harpreet Singh, PhD
Bioinformatics Centre, Indian Council of Medical Research, Ansari Nagar, New Delhi 110029, India. Tel: +91-11-26589556, Fax: +91-11-26588662, E-mail: hsingh@bmi.icmr.org.in

the task of identifying subject experts. Finally, complex proposals/problems/issues require the combined wisdom of experts from multiple subjects. Thus, there is a need to develop a computer-based system for finding experts. Developments in this field can reduce the time required for solving time-sensitive problems thus improving the overall efficiency of an organization.

An ideal expert finding system should have the following qualities:

- Should be robust—that is, it should be easily and cost effectively updated without much user intervention.
- Should be able to classify expertise into a standard subject classification schema.
- Should provide information about the past performance of experts.
- Should quantify expertise, enabling the ranking of experts.
- Should be based on authentic sources of information.
- Should be able to form expert communities.
- Should be able to identify locally available experts.

1. Available Expert Finding Systems

Efforts have been made during the past 20 years to develop an expert finding system. A number of expert finding systems have been developed at both national and international levels [3-6]. Broadly, the methods used for developing expert finding systems can be grouped into two categories.

2. Methods Based on Mining Unstructured Information

Unstructured information includes emails, corporate or personal Web pages, wiki, reports, etc. Text mining tools are used to index technical terms from unstructured documents, which can be queried to identify subject experts. Further, experts have been ranked using the number of occurrences of technical terms [7]. Some of the important tools using this information include email expertise extraction (e3) system [8], ContactFinder [9], MIT Expert Finder [10], etc.

Although this method is robust and useful, a major limitation is the authentication of information. Another issue is that, because of privacy, complete information may not be available for mining. Moreover, there is lack of implementation of a standard subject terminology.

3. Methods Based on Social Networking Sites or Contact Management Systems

Today there are many social networking sites, some specifically for the scientific community, such as ResearchGate (http://www.researchgate.net/), Nature Network (http://network.nature.com/), VIVOweb, etc. Experts have to feed in information about their subject expertise, domains, publications, credentials, etc. A major limitation of these methods is in the adding and updating of information. Many of the developed systems, particularly discussion forums and knowledge directories, have become obsolete due to decrease in interest of experts. Experts may not be interested in subscribing to social sites or responding to queries. Further, there is difficulty in ranking experts.

A common drawback to both approaches is that they are based on non-peer-reviewed information provided by the user; hence, they can be biased.

It is evident from the above discussion that identifying correct subject experts is extremely vital for the success of an organization. Most of the available expert finding systems are based on information provided by the individual or mining non-peer-reviewed information and can be biased as a result. For this reason, there is a strong need to develop an automated, unbiased expert finding system which is based on authentic information and can be easily updated.

In this article we present an expert finding system that is based on peer-reviewed information, can be updated regularly, and uses standardized subject vocabulary i.e., Medical Subject Headings (MeSH) associated with each article [11]. Using MeSH headings adds standardized subjects for querying. The latest release of the system can be used to search for experts in a particular subject and the subjects associated with a particular expert from India. The methodology is general and can be implemented to identify subject experts from any country.

II. Methods

1. Data Retrieval

PubMed is one of the largest repositories of peer-reviewed articles published worldwide. Publications originating from India (affiliation India) were downloaded in XML format using an in-house developed script. The developed script uses the Bio::Biblio module of Bioperl to interact with the PubMed database over the internet.

The 2013 MeSH subjects were downloaded from the MeSH browser (http://www.nlm.nih.gov/mesh/filelist.html) in tree format.

2. Data Pre-processing

Each XML record of articles downloaded from PubMed was parsed using the XML::Twig module, and relevant fields (Authors, Title, Journal, Abstract, Volume, Issue, Page, and MeSH) were extracted in an intermediate text file where each record begins with a 'START' tag and ends with an 'END' tag.

As we downloaded the MeSH as a text file, each MeSH

 http://dx.doi.org/10.4258/hir.2013.19.4.243

record was parsed on the basis of MeSH code using the in-house developed script, and the code for each left node (parent) and right node (child) were labeled.

3. Database Design: PubMed Data

A database consisting of two tables was designed to store information downloaded from PubMed, and their structure is shown in Table 1. The minimum fields required for the expert finding system were stored in the database. This reduces the size of the database and makes it suitable for developing cross platform standalone applications, such as Android or iOS apps.

For storing MeSH terms, we adopted a nested set model, which is suitable for storing and querying tree data structure. The MeSH data was pre-processed to identify the left (parent) and right (child) node for each given node. The structure of the MeSH data table is shown in Table 2.

The nested set is an efficient model for storing and searching through hierarchical data. The technique uses a method of storing metadata (left and right numbers) about the nodes (MeSH terms) contained in the tree in order to provide the SQL parser with information about how to "walk" the hierarchy of nodes. A critical aspect of the nested set model is that it alleviates the need for a recursive technique to find all children.

The following rules were used to calculate left and right numbers:

- For the root node in the hierarchy, the left side value will be 1, and the right side value will be 2*n, where n is the number of nodes in the tree.
- For all other nodes, the right side value will be left side + (2*n) + 1, where n is the total number of child nodes. For the leaf nodes (nodes without children), the right side value will always be equal to the left side value + 1.
- Left side value for any node is next free number, if we walk the tree counter-clockwise

4. Expertise Scoring Function

Our system uses a simple expertise scoring function, which is the number of publications from a given expert containing a selected MeSH term.

5. Statistical Significance of Subject Association

The statistical significance of identified experts for a given subject was estimated using a Z-score calculated from a contingency table (Figure 1):

$$z = \frac{|\hat{p}_1 - \hat{p}_2| - \left(\frac{1}{2n_1} + \frac{1}{2n_2}\right)}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

where

$$n_1 = (a+b) \quad and \quad n_2 = (c+d)$$

$$\hat{p} = \frac{(a+c)}{(a+b+c+d)} \quad and \quad \hat{q} = 1 - \hat{p}$$

$$\hat{p}_1 = \frac{a}{(a+b)} \quad and \quad \hat{p}_2 = \frac{c}{(c+d)}.$$

6. Web Interface

A Web interface is developed using PHP for finding experts using a browser (Figure 2).

III. Results

The objective of this study was to develop an unbiased and robust expert finding system using peer-reviewed information. We developed a Web-based system to find subject experts using MeSH associated with peer-reviewed articles indexed in PubMed. The system quantitatively ranks experts using MeSH terms associated with their peer-reviewed publications indexed in PubMed.

It is difficult to evaluate such a system as there are no benchmarks available; therefore, we evaluated the efficacy of the developed system using prior information about experts in subjects. The further statistical significance of the association between a subject and an expert was estimated using

Table 1. Structure of table containing PubMed records

Field	Type	Null	Key	Default	Extra
pmid	int(11)	No	PRI	Null	
pubyear	int(11)	Yes		Null	
mesh	longtext	Yes		Null	
author	longtext	Yes		Null	
affiliation	longtext	Yes		Null	
citation	longtext	Yes		Null	

Table 2. Structure of Medical Subject Headings table for storing hierarchical data

Field	Type	Null	Key	Default	Extra
category_id	int(11)	No	PRI	0	
name	text	Yes	MUL	Null	
left_side	int(11)	Yes		Null	
right_side	int(11)	Yes		Null	

	Given author	Excluding given author	
Given subject	a	b	**a+b** (number of articles in given subject)
Excluding given subject	c	d	**c+d** (number of articles in subjects other than given subject
	a+c (number of articles from given author)	**b+d** (number of articles from authors other than given author)	

Figure 1. Contingency table used to calculate statistical significance of association between given subject and expert.

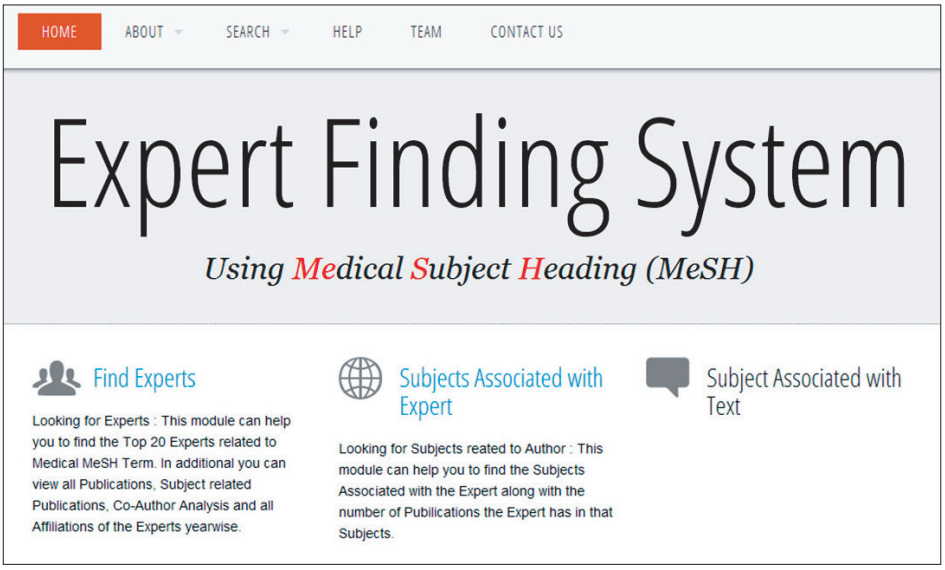

Figure 2. Home page of Expert Finding System available at (http://bmi.icmr.org.in/expert; http://202.141.106.122/expert).

Table 3. Experts from selected subjects along with their *p*-value for subject association calculated using Z-test

No.	Subject	Expert	*p*-value	Comment
1	Microbiology	T. Ramamurthy	<0.0001	Dr. Ramamurthy has nearly 70% publications in microbiology
2	Computational biology	Gajendra P. S. Raghava	<0.0001	Dr. Raghava is Bhatnagar awardee and a known international figure in bioinformatics
3	X-ray crystallography	M. Vijayan	<0.0001	With nearly 75% papers in the subject, Prof. M. Vijyan is known crystallographer
4	Database, proteins	Gajendra P. S. Raghava	<0.0001	Has developed nearly 70 Web services and databases
5	Genetics	Lalji Singh	<0.0001	Prof. Lalji Singh has worked extensively in genetics with more than 90% papers in the subject

Z-score (see Methods Section).

Table 3 shows the statistical significance of identified experts from known branches of sciences, such as 'Microbiology', 'X-Ray crystallography', 'Database', 'Genetics'. The identified experts are well known international experts on the subjects.

1. Description of the System

Below is the brief description of our system. Release 2.1 of our expert finding system allows users to find (1) experts from a particular subject and (2) subjects associated with a given expert.

1) Application: Finding experts in a given subject

To locate a subject expert the user has to enter either partial or complete name of the subject in the 'Subject' text box and click 'Submit' (Figure 3). On submission MeSH subjects similar to the query are displayed as a drop-down box (Fig-

 http://dx.doi.org/10.4258/hir.2013.19.4.243

ure 4). The user then selects the most relevant subject from the MeSH drop-down and clicks 'Submit'. A list of experts ranked on the basis of the number of articles published in a selected subject is displayed (Figure 5). The user can access a year-wise list of all affiliations of author, all co-authors, MeSH subjects and all publications. The publications are linked to PubMed using pmid.

To search the MeSH terms, we selected the entire subtree of the selected MeSH term. For example to find experts in microbiology, the entire subtrees, which include bacteriology, virology, etc., were also searched.

2) Application: Finding subjects associated with an expert

To locate subjects associated with an expert, the user enters either partial or complete name of the expert in the 'Last name, First name or initials' text box and clicks 'Submit'. On submission, the names of experts that are similar to the query are displayed as a drop-down box. The user then selects an expert from the drop-down box and clicks 'Submit'. A list of subjects ranked on the basis of the number of articles published in the subject is displayed (Figure 6). The user can access all publications of an expert or all of their publications in a given subject; they can also search for other experts in a given subject.

IV. Discussion

The expert finding system developed using MeSH terms has been tested using known subject experts from India and was found to be satisfactory. For example in the subject of crystallography Prof. M. Vijayan, Prof. T. P. Singh, Prof. A. Srinivasan, and Prof. M. R. Murthy are ranked as top experts. They are all well-known experts in crystallography in India. Similarly, the system was successfully tested for other subjects.

However, the system has some limitations, some of which will be addressed in future releases of the system. One major

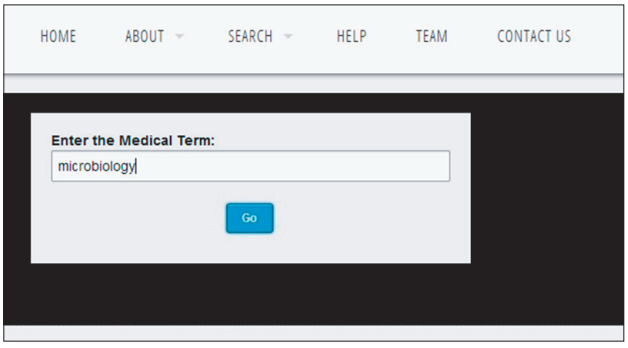

Figure 3. Finding subject experts: enter complete or partial subject.

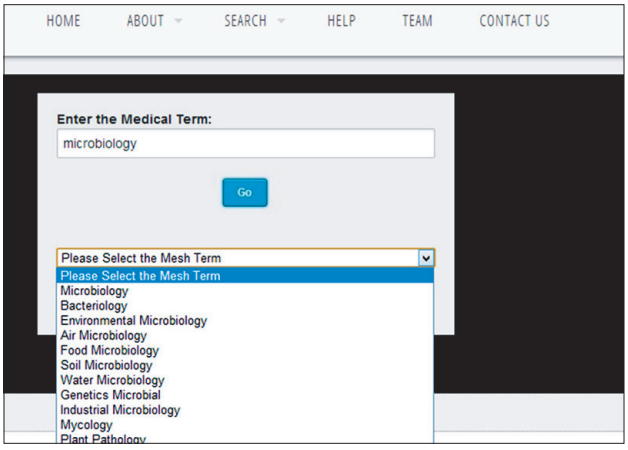

Figure 4. Selecting from list of Medical Subject Headings subjects related to entered text.

Figure 5. List of subject experts in 'microbiology'.

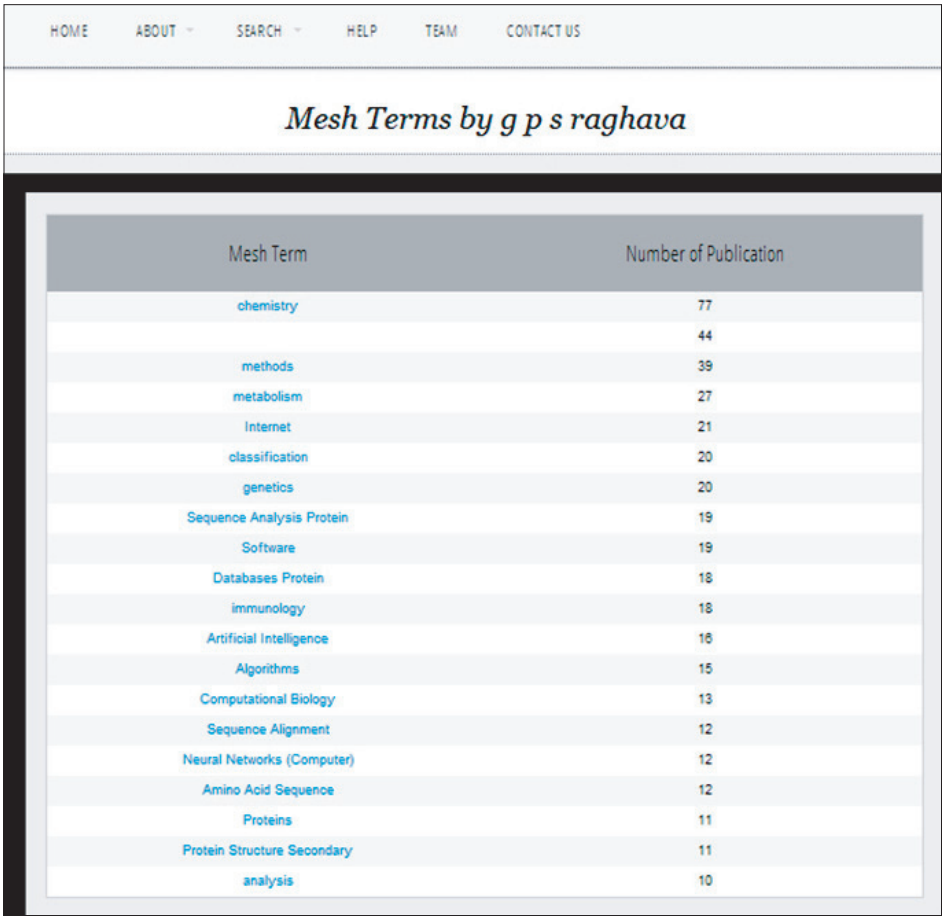

Figure 6. Searching subjects associated with an expert.

limitation of using MeSH terms associated with an article as expertise of all the authors of that paper is that some of the experts may have different expertise and may have been merely co-author in the paper. For example, in any biomedical work, statistical assistance is required, and any statistical expert may be associated with biomedical subject headings. However, given the large number of articles from which the data is collected, the probability of association of an expert with related subjects is higher than unrelated subjects. In our experience, any subject in which an author has more than 20 publications can be considered to be associated with author. To address this issue, we calculated two parameters:

(i) The percentage of contribution to the field which is calculated as

$$percent\ contribution\ to\ the\ field\ (pc) = \frac{n_s}{N},$$

where n_s is the number of publications of a given author in the selected subject, and N is the total number of publications by that author.

(ii) A statistical association test, based on the Z-score as described in the Methods Section.

We found correlation between the percentage of contribution to the field and the *p*-value calculated from the Z-score. However, both *pc* and *Z-score* were not correlated with the number of publications.

Another possible limitation is that many experts publish in journals that are not indexed in PubMed, and since we have used data extracted from PubMed, the ranking of subject experts may be incomplete. There are some Web services available for assigning MeSH terms to articles [12]. The integration of non-indexed journals will be done after verifying the credibility of sources.

The developed system is unique as it uses a purely objective measure to identify subject experts, the data used is peer-reviewed and reliable, and it allows the ranking of experts. For developing countries where resources for research are limited, developing such a system can improve the efficiency of research.

We are trying to improve the system by incorporating 1) graphical representation of co-author networks, 2) analysis tools for co-author networks, 3) clustering of affiliations to identify most probable affiliations, and 4) geographical mapping of expertise.

 http://dx.doi.org/10.4258/hir.2013.19.4.243

Conflict of Interest

No potential conflict of interest relevant to this article was reported.

Acknowledgments

The work was supported by funding under the project "Biomedical Informatics Centres of ICMR" funded by the Indian Council of Medical Research.

References

1. Maybury MT. Expert finding systems. Bedford (MA): MITRE Center for Integrated Intelligence Systems; 2006.
2. Mimno D, McCallum A. Expertise modeling for matching papers with reviewers. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2007 Aug 12-15; San Jose, CA. p. 500-9.
3. Yimam-Seid D, Kobsa A. Expert-finding systems for organizations: problem and domain analysis and the DEMOIR approach. J Organ Comput Electron Commer 2003;13(1):1-24.
4. Liu X, Croft WB, Koll M. Finding experts in community-based question-answering services. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management; 2005 Oct 31-Nov 5; Bremen, Germany. p. 315-6.
5. Serdyukov P, Hiemstra D. Modeling documents as mixtures of persons for expert finding. In: Advances in information retrieval. Heidelberg, Germany: Springer; 2008. p. 309-20.
6. Tang J, Zhang J, Yao L, Li J, Zhang L, Su Z. ArnetMiner: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2008 Aug 24-27; Las Vegas, NV. p. 990-8.
7. Wu CJ, Chung JM, Lu CY, Lee HM, Ho JM. Using Web-mining for academic measurement and scholar recommendation in expert finding system. In: Proceedings of the IEEE/WCI/ACM International Conference on Web Intelligence and Intelligent Agent Technology; 2011 Aug 22-27; Lyon, France. p. 288-91.
8. Krulwich B, Burkey C, Consulting A. The ContactFinder agent: answering bulletin board questions with referrals. In: Proceedings of the 13th National Conference on Artificial Intelligence; 1996 Aug 4-8; Portland, OR. p. 10-5.
9. Campbell CS, Maglio PP, Cozzi A, Dom B. Expertise identification using email communications. In: Proceedings of the 12th International Conference on Information and Knowledge Management; 2003 Nov 2-8; New Orleans, LA. p. 528-31.
10. Vivacqua AS. Agents for expertise location. In: Proceedings of the AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace; 1999 Mar 22-24; Palo Alto, CA. p. 9-13.
11. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc 2000;88(3):265-6.
12. Trieschnigg D, Pezik P, Lee V, de Jong F, Kraaij W, Rebholz-Schuhmann D. MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 2009;25(11):1412-8.