# Expert Finding by Capturing Organisational Knowledge from Legacy Documents

Yee-Wai Sim, Richard Crowder, Gary Wills
*School of Electronics and Computer Science*
*University of Southampton, UK*
*Email: {yws01, rmc, gbw}@ecs.soton.ac.uk*

## Abstract

*Organisations capitalise on their best knowledge through the improvement of shared expertise which leads to a higher level of productivity and competency. The recognition of the need to foster the sharing of expertise has led to the development of expert finder systems that hold pointers to experts who posses specific knowledge in organisations. This paper discusses an approach to locating an expert through the application of information retrieval and analysis processes to an organization's existing information resources, with specific reference to the engineering design domain. The approach taken was realised through an expert finder system framework. It enables the relationships of heterogeneous information sources with experts to be factored in modelling individuals' expertise. These valuable relationships are typically ignored by existing expert finder systems, which only focus on how documents relate to their content. The developed framework also provides an architecture that can be easily adapted to different organisational environments. In addition, it also allows users to access the expertise recognition logic, giving them greater trust in the systems implemented using this framework. The framework were applied to real world application and evaluated within a major engineering company.*

## 1. Introduction

Motivated by advances in information technology, organisations are giving more emphasis to the capitalisation of the increasing mass of knowledge accumulated in the course of their business [15]. More specifically, organisations aim to acquire knowledge from valued individuals and to analyse their activities, in order to learn from successes and failures. The recognition of the need to foster sharing of expertise has led to the development of systems that hold pointers to experts who possess specific knowledge in organisations. Still, there have remained problems of maintaining and retrieving expertise in these systems. These problems are related to the exploration of heterogeneous information sources, support for expertise analysis, and reusability and interoperability of these systems.

In addition, there is a need to derive a framework for providing up-to-date information used in expertise modelling. This is because organisational workers accumulate knowledge through task achievements and this output is a valuable source for capturing knowledge related to individuals' expertise. For instance, workers write project reports or document ongoing projects in order to meet their organisational functions. Exploiting metadata information from these documents can draw inferences to derive or update the knowledge about expertise associated with the workers.

The remainder of this paper is organised as follows: Section 2 discusses the problems identified in the surveyed expert finder systems. In response to the identified problems, a framework was proposed, as reported in Section 3; while Section 4 describes an expert finder system, implemented based on the proposed framework, in a real world application. An evaluation for the system is presented in Section 5 to justify its validity. Finally, the authors conclude this paper with some final remarks.

## 2. Problems in Reviewed Expert Finder Systems

Expert finder systems require a range of information as indicators of experts. To manage an expert finder system, there is a need for tools that gathers and consolidates this information in a form that is accessible by the system. The availability of large electronic repositories in organisations has led to the development of an autonomous approach to collect and analyse information in finding experts. The literature details a number of systems that undertake a fully automatic approach to locate experts, including, *Who Knows*[16], *Agent Amplified Communications*[7], *Contact Finder*[8], *Yenta*[5], *MEMOIR*[12], *Expertise Recommender*[11], *Expert Finder*[17], *SAGE*[1] and the *KCSR Expert Finder*[3]. In a review of these systems, problems related to heterogeneous information sources, expertise analysis support, and reusability and interoperability were identified.

### 2.1. Heterogeneous Information Sources

In order to effectively explore the organisational information space for expertise evidence, expert finder systems need the ability to handle the heterogeneity of the widely distributed information sources. This is reflected by the wide variety of expertise evidence, such as emails [7], electronic messages on bulletin boards [8], program codes [11][17], Web pages [5][12], and technical reports [3] used in expert finder systems. Hence, a framework that is adequately flexible in addressing this problem is required.

In this paper, the authors propose that the heterogeneity of information sources should be used as an indicator for reflecting experts' competencies. How well these expertise indicators (e.g. indexed terms) reflect expertise is mainly a factor of how the source in which these indicators occur relates to the expert. This idea is based on the assumption that terms found in different types of documents indicate expertise differently, irrespective of their statistical traits. For example, a specific term in a person's *Curriculum Vitae* or its occurrence in a journal publication may not have the same importance. Moreover, the occurrence of this term in the title of the document shows a different distance to his actual expertise, compared to its occurrence anywhere in the body of the document. Therefore, the relationship of expert-to-document needs to be determined before extracting indexed terms from the document.

### 2.2. Expertise Analysis Support

A user seeks individuals as sources of information and/or as individuals who can perform given organisational functions. Each of these purposes imposes its own requirements on the expert finder system's functionality. Hence it is the users who should select the appropriate filters depending on their needs, and the system should only support the expert finding process by providing analysis functionality. This means that including the ability to rank experts using different user-customisable criteria (rather than the mere provision of a linear listing based on pre-determined criteria) can considerably enhance the expert finding applications.

Another approach that can support users in expertise analysis is to increase the system's transparency. This can be done by providing interfaces to access the expertise evidence, as well as the expertise recognition logic. For example, the system can supply the scores for ranking the experts along with the expertise evidence (e.g. documents) associated with them. It allows users to evaluate for themselves the validity of the system's recognition logic. This in turn permits the incorporation of functionality, which can assist the users in evaluating and exploring the expertise evidence, i.e. spotting anomalies in the expert finding process, giving users greater trust in the system.

### 2.3. Reusability and Interoperability

Most of the expert finder systems surveyed only focus on solving a particular problem and coming up with a standalone solution. However, expert finder systems can be integrated into other organisational systems, such as information retrieval systems, recommender systems and *Computer Supported Cooperative Work* systems. Therefore the expert finder systems should be readily transferable from application to application and be interoperable with other systems.

Different applications and organisational environments may require tailored implementations of expert finder systems. This means the components in expert finder systems have to be flexible in order for them to be easily extended or replaced. Consider the case of the filters described in the previous section. They can be easily reused given the appropriate data. Hence, expert finder systems should be designed in a generic framework.

## 3. A Framework for an Expert Finder System

In view of the problems discussed above, the authors proposed a framework, shown in Figure 1, for an expert finder system. The framework is a collection of components for expertise evidences extraction, expertise modelling, querying, expertise matching and user interface. These components are flexible enough to address different organisational environments. By defining appropriate interfaces, the components can also be interoperable with other organisational systems. This can be done by using a generic low-level *Application Programming Interface* (API). The following sections describe the framework according to the system's general functions.

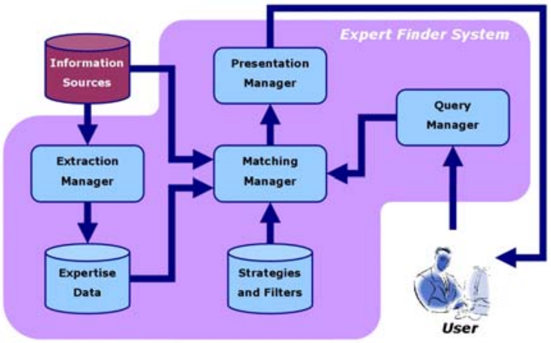

**Figure 1: The framework for an expert finder system described in this paper**

### 3.1. Expertise Extraction

The extraction manager is responsible for identifying potential information sources which contain expertise evidence. This will typically include shared or personal workspaces, document storage systems, or email archives. Therefore, it is essential for the extraction manager to be equipped with the capability to handle these workspaces.

### 3.2. Expertise Modelling

Regardless of the origin of the information sources, i.e. shared or personal, the extracted evidence is stored in a centralised server for analysis by the system. However, in order to address the heterogeneous information sources problem presented above, the process of building the expertise model needs to account for the type of relation of a given source to an expert. For instance, the occurrence of a term in the title of a document should be given a higher weighting, compared to its occurrence in the body.

### 3.3. Expertise Matching

The matching manager is initiated in response to a user request. Through the query manager, the users can indicate which strategies and/or filters they would like to apply in selecting expertise for their needs. In order to achieve the goal of selection, the matching manager provides access to the information space that maintains personal and organisationally related data, i.e. departmental affiliation, which can be used as criteria in finding experts. Based on these criteria, the set of candidate experts are reordered in different ranking positions and/or removed from the set to generate a refined list of experts.

### 3.4. User Interface

A ranked list of experts coupled with evidence (e.g. documents) retrieved for expertise modelling is presented to the users in response to their queries. The output result is routed to the users by the presentation manager. Increasing the system's transparency can provide aids to the users in expertise analysis. Such issues can be addressed by supplying expertise recognition logic and expertise evidence to the users via the presentation manager, so that they can evaluate the result for themselves. This in turn builds up users' trust towards the system.

## 4. A Demonstrator

A demonstrator is built based on the framework described above as part of an ongoing project at Southampton whose objectives is to develop tools to support the activities of the design engineering, particularly in areas of knowledge capture, sharing and reuse. The demonstrator is named *Relational Expert Finder System* (REFS) and its key processes are summarised in the following sections.

### 4.1. Extracting Expertise

To implement the demonstrator, data supplied by a major UK manufacture in the form of their internal publication database is used. This corresponds to a total of over 170,000 entries, covering a time period of fifty years. All the records in the database were entered manually by human operators, and ranged from technical reports to departmental memoranda. In practice, neither the data source could be guaranteed to be correctly maintained nor can be entries in any fields be guaranteed to be valid and/or consistent. Careful design is required as there are a large number of entries in the database. Otherwise, accessing such a resource can be very time consuming and the responses resulting from queries can easily overwhelm the system.

### 4.2. Modelling Expertise

Expertise models are created using text modelling algorithms based on the vector space model. TFIDF is a popular function employed by most vector space modelling applications [6] [14]. This function allows a term's importance in a given document to be reflected.

However, such a function ignores the document's structural elements, i.e. title or body, and document types, i.e. technical reports or memoranda, and treats all the text contained in that document as a bag of words, e.g. unstructured text. In view of this, the author modified the TFIDF to account for the structural elements,

$$w'\left(t_k, d_j\right) = w\left(t_k, d_j\right) + s_j$$

where $w'(t_k,d_j)$ denotes the weight of the term $t_k$ in a document $d_j$, $w(t_k,d_j)$ denotes the weight of $t_k$ in $d_j$ calculated by the TFIDF function, and $s_j$ denotes the weight of the structural element in which $t_k$ occurs at least once.

The value of $s_j$ was determined heuristically as part of the development process. As part of the process the raw documents were processed by a number of tools to extract the text under the various headings, i.e. project name, authors, report abstract, etc. Readers can refer to a paper published by the authors [4] for more details about the extraction tools. The development of the tool was significantly aided by the use of company standard document templates. The equation above reflects the fact that the more often a term occurs in a key part of a specific type of document, for example the title instead of the body of a technical report, the more it is representative of the content, and the more documents in which the term occurs, the less discriminating it is.

The resultant expertise models contained indexes resulting from the calculation of term space coverage and application of dimension reduction mechanisms. These models were then stored in a centralised database and used within the expertise matching and modelling processes.

### 4.3. Matching Expertise

The querying process is currently implemented using the Boolean AND operator. Using the query terms and the AND operator, expertise models containing the query terms are identified, and then combined by taking the intersection of the sets of retrieved expertise models.

The system implementation includes two expert finding strategies. The first strategy is based on the concept of organizational awareness [10], in which the system only considers an individual as an expert if he/she is linked to a large number of documents, which tends to promote one particular type of expert. However such an approach tends to reflect the interests of experts instead of their competency levels. For examples, an individual who has produced twenty journal publications will be treated as having the same expertise level as an individual who has written twenty memorandums. Hence, this strategy is only appropriate for users who seek individuals as sources of information.

The second strategy identifies experts by the importance of terms (supplied by the user at query time) in documents. The importance of such terms is calculated using a vector space algorithm, as defined by the above equation. This algorithm not only reflects the importance of terms in relation to their occurrences in a set of documents of a specific type, it also indicates the terms' importance in relation to the document structures and types containing them. Therefore, the computed terms' importance can then be used as indicators for an experts' competency. This strategy, coupled with or without the first strategy, can be used to find individuals who can perform given organizational functions.

### 4.4. Presenting Result

After the expertise matching process has been completed, the ranked experts' names with their associated scores are displayed, the scores for ranking the experts are supplied to assist the users in analyzing the recommendation. An interface is also provided to access the list of documents selected for expertise recognition, the documents are grouped by author's name for browsing purposes. This approach not only satisfies users' requirements in locating experts as information sources, it also allows users to evaluate the expertise recognition logic for themselves, hence, giving them greater trust as regards the recommendation.

## 5. Evaluation

Expertise retrieval effectiveness can be measured in terms of the information retrieval notions of *precision* and *recall* [2]. For obtaining estimates of *precision* and *recall* relative to multiple decisions when two or more queries are submitted, microaveraging was adopted as a global evaluation method [9].

The evaluation allowed us to compare the effectiveness of the proposed system against the system previously developed by Hughes and Crowder [3], the *KCSR Expert Finder*. As in the authors' approach, this expert finder system uses an organization's own resources to recommend experts. On entering a query, *KCSR Expert Finder* will return a list of experts ranked according to the number of documents associated with them. In addition, a list of documents, used as expertise evidences, is then displayed in the return result of query. However, the REFS and the *KCSR Expert Finder* approach the modelling of expertise differently. The system developed by the authors includes the type of relation of a given source to an expert, i.e. the documents' type and structural information, in constructing expertise models, while the *KCSR Expert Finder* ignores such information by representing its expertise models using full text indexes. In order for the experimental results on the two expert finder systems to be directly comparable, the experiments were performed using identical resource databases.

### 5.1. Test Data

The effectiveness of an expert finder system can be evaluated by test users relative to specific contexts. The most likely context is their experience accumulated from the workplace. In order to measure the effectiveness of *KCSR Expert Finder* and REFS in retrieving experts, a set of questions that can provide contextualised problem statements is needed. A total of 9 test users were interviewed to obtain the sample questions, for example "*How should I model a turbine blade and disc for analysis?*", and names of the individuals who have the expertise in answering these questions. The sample questions were chosen to be representative of the type of work problems that occur with reasonable frequency; so that they reflect the real problems arising in a work context.

### 5.2. Results

The experts recommended by the *KCSR Expert Finder* and REFS were compared with those identified by the test users. It was noted that:

- in the initial evaluation queries based on question 1, 2 and 6 failed to return any matches with both systems,
- queries based on questions 3, 4, 5 and 8 failed to return any matches with the *KCSR Expert Finder*,
- matches were obtained between both expert finder systems and the test users for questions 7 and 9.

In order to compare the systems' effectiveness, global effectiveness values for both of the expert finder systems in all sample questions were calculated using the microaveraging method. Calculated global effectiveness values for both systems are presented in Table 1.

**Table 1: Microaveraged effectiveness values for Expert Finder and REFS**

| | *Expert Finder* | *REFS* |
|---|---|---|
| *Global Effectiveness Value* | 0.12 | 0.28 |

Considering *KCSR Expert Finder*'s microaveraged effectiveness value as the baseline in assessing the performance of the REFS approach in identifying experts who have the relevant expertise for given sample questions, the improvement is 16%, which is statistically significant.

## 6. Conclusions

We have presented an approach to using capturing organisational knowledge for expert finding. A framework for expert finder systems is proposed by the authors that extends the relationships between information sources and expertise models. We have noted two significant activities for why experts need to be located, either users seek experts as sources of information or as collaborators in specific activities. Hence, we suggested that an expert finder system should provide analysis functionality, since it is the users who select the appropriate expert finding strategies depending on their needs. This was incorporated into the demonstrator presented in this paper by providing interfaces to access the expertise recognition logic and evidence.

In contrast to many expert finder systems that were designed to solve a particular problem within a specific organisational environment, the framework proposed by the authors is both flexible and modular; so that its components can be easily extended and replaced depending on requirements. The framework enables an implemented expert finder system to be interoperable with other organisational systems via appropriate API. This approach allows generated expertise data to be shared across an organisation.

The expert finder system developed by the authors has been compared with a similar system based on full-text indexing system which ignores structural information when analysing the source documentation. Although both of these systems shared the same data source, in initial testing the system developed by the authors outperformed the full-text indexing system in terms of expert retrieval effectiveness for a limited number of test cases. It is our view that the improvement in locating experts through the methodologies embodies in the demonstrator named

REFS will translate to a reduction in costs within an organisation as the correct expert is located more rapidly.

Although the approach demonstrates that the mechanism for extracting expertise data can be automated, it however trades the problems of increased workload and subjective self-assessments with the problem associated with `dirty' data. The authors invested a considerable amount of time and effort to formulate techniques for validating data and folding it to a format that can be processed by the systems. As such, the validity and consistency of the data plays an important role in determining the performance of expertise retrieval, and should be considered when deploying an expert finder system.

## References

[1] Becerra-Fernandez, I. The Role of Artificial Intelligent Technologies in the Implementation of People-Finder Knowledge Management Systems. *Knowledge-Based Systems*, 13, pp. 315-320, 2000.

[2] Cleverdon, C. W. On the inverse relationship of recall and precision, *Journal of Documentation*, 28, pp. 195-201, 1972.

[3] Crowder, R., Hughes, G. and Hall, W. An agent based approach to finding expertise. In *Proceedings of the Fourth International Conference on Practical Aspects of Knowledge Management*, Berlin Heidelberg, pp. 179-188, 2002.

[4] Crowder, R. and Sim, Y. W. An Approach to Extracting Knowledge from Legacy Documents. In *Proceedings of ASME International Design Engineering Technical Conferences and Computers and Information Engineering Conference*, Salt Lake City, 2004.

[5] Foner, l. N. Yenta: A Multi-Agent Referral-Based Matchmaking System. In *Proceedings of the First International Conference on Autonomous Agents*, Marina del Rey, CA, pp. 301-307, 1997.

[6] Joachims, T. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the European Conference on Machine Learning*, pp. 137-142, 1998.

[7] Kautz, H. A., Selman, B. and Shah, M. Referral Web: Combining Social Networks and Collaborative Filtering. *Communications of ACM*, 40(3), pp. 63-65, 1997.

[8] Krulwich, B. and Burkey, C. The ContactFinder Agent: Answering Bulletin Board Questions with Referrals. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence and Eighth Innovative Applications of Artificial Intelligence Conference*, Volume 1, Portland, Oregon, pp. 10-15, 1996.

[9] Lewis, D. D. Evaluating Text Categorization. In Proceedings of Speech and Natural Language Workshop, pp. 312-318, 1991.

[10] Maybury, M., Amore, D. and House, R. Awareness of organizational expertise. *MITRE*, Technical Papers, October, 2000.

[11] McDonald, D. W., and Ackerman, M. S. Expertise Recommender: A Flexible Recommendation System and Architecture. In *Proceedings of the ACM 2000 Conference on Computer-Supported Cooperative Work*, Philadelphia, PA, pp. 231-240, 2000.

[12] Pikrakis, A., Bitsikas, T., Sfakianakis, S., Hatzopoulos, M., DeRoure, D. C., Hall, W., Reich, S., Hill, G. J. and Stairmand, M. MEMOIR – Software Agents for Finding Similar Users by Trails. In *Proceedings of the Third International Conference and Exhibition on the Practical Application of Intelligent Agents and Multi-agents*, London, UK, pp. 453-466, 1998.

[13] Raghavan, V. V., Jung, G. S. and Bollman, P. A critical investigation of recall and precision as measures of retrieval system performance, *ACM Transaction on Information System*, 7(3), pp. 205-229, 1988.

[14] Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, 34(1), pp. 1-47, 2002.

[15] Shadbolt, N. R. and O'Hara, K. AKTuality: An Overview of the Aims, Ambitious and Assumptions of the Advanced Knowledge Technologies Interdisciplinary Research Collaboration. In Shadbolt, N. R. (Ed.) *Advanced Knowledge Technologies: Selected Papers 2003*, pp. 1-11, 2003.

[16] Streeter, L. A. and Lochbaum, K. E. An Expert/Expert Locating System Based on Automatic Representation of Semantic Structure. In *Proceedings of the Fourth IEEE Conference on Artificial Intelligence Applications, Computer Society of the IEEE*, San Diego, CA, pp. 345-349, 1988.

[17] Vivacqua, A. S. Agents for Expertise Location. In *Proceedings of the AAAI Spring Symposium on Intelligent Agents is Cyberspace*, Stanford, CA, pp. 9-13, 1999.