An Empirical Investigation of Expertise Matching within Academia

by

Ping Liu

Submitted in accordance with the requirements

for the degree of Doctor of Philosophy

The University of Leeds
School of Computing

September 2003

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgement

Firstly, I would like to thank my supervisor Professor Peter Dew. He has constantly provided excellent guidance and encourage on my research work. Dr. Jayne Curson provided most valuable guidance as my advisor. I particular benefited from her proof reading. The Informatics Research Institute is a great place to work. I want to thank both the present and former members. In particular Martin Thompson for installing RDFDB, Bill White, Vania Dimitrova, Khan Sharifullah, Zukri Ibrahim, Ahlian Kor, Daniela Romano, Seth Bullock, Jason Noble, Eric Atwell for discussion and support. Thanks also should be given to the PhD students who attended evaluations of the prototype systems.

A special thanks is reserved for Richard Drew at Symularity Ltd for his willingness to share advice and experience throughout the duration of this work. I also want to thank Professor Christine Leigh, Paul Micklethwaite, Colin Winnett for their work contributing to the REPIS database.

My parents and my brother receive my deepest gratitude and love for giving me incredible support throughout the years. Finally, but not least, I thank my husband, Hao Xie, who has been understanding and patient during my PhD study. I could not have come this far without his constant love and support.

Abstract

Many Organizations have realized that effective management of their knowledge assets is important to survival in today's competitive business environment. Consequently an Organizational Memory (OM) is used to store what has been learned from the past in order that it can be reused by current and future employees. Information retrieval techniques have been widely used to facilitate the retrieval of the right information in an OM at the right time. However, access to information alone is not sufficient since not all knowledge can be transferred into explicit documentation. Expertise, as one of the most important knowledge assets, is normally stored in people's heads and is difficult to codify. Expertise is shared when people communicate with each other. Therefore, finding the right person with the right expertise is recognized as being at least as important as retrieving documents. The typical approaches to find experts include knowledge brokers and expertise database. However, the former approach is impractical in large organizations and geographically disparate organizations whilst the latter approach relies heavily on individuals to specify their expertise and keep updated. This thesis focuses on two questions: (1) How to integrate multiple expertise indications existing in an organizational memory as complementary to the description by experts? (2) How to insure the relevant experts are not overlooked as well as irrelevant experts are minimized? To solve these problems, a conceptual model has been developed so that multiple expertise indications existing in the organizational memory can be semantically integrated. The heterogeneous data sources are integrated by using RDF(S) since RDF allows for a uniform representation of data and RDF Schema represents the conceptual model. In addition, the expertise profiles are extended to include both keyword form and concept form based on the domain ontology; this combined profile integrates the advantages of both keyword search and concept search. A prototype system, which aims to help PhD applicants locate their potential supervisors, has been designed and implemented to test the techniques and ideas. The results of the experiments using real data at the University of Leeds demonstrate the improved performance of expertise matching and also show the advantages of applying semantic web technologies (such as RDF, RDFS, ontologies) to the expertise matching problem.

Declarations

Some parts of the work presented in this thesis have been published or submitted in the following articles:

Liu, P., Curson, J., Dew, P. and Drew, R. (2001) "Expertise Matching using RDF," in *The First Semantic Web Working Symposium* Stanford University, California, USA, July 30 - August 1, 2001, page 53-54.
http://www.semanticweb.org/SWWS/program/position/SWWSpositionpapers.pdf

Liu, P., Curson, J. and Dew, P. (2002) "Exploring RDF for Expertise Matching within an Organizational Memory," in *Proceedings of the 14th International Conference on Advanced Information Systems Engineering*, CAiSE 2002, Toronto, Canada, May 27-31, *Lecture Notes in Computer Science 2348*, A. Banks Pidduck et al. (Eds.), Springer-Verlag Berlin Heidelberg 2002, pages 100-116.
http://www.mm.di.uoa.gr/~rouvas/ssi/caise2002/23480100.pdf

Liu, P., Curson, J. and Dew, P. (2003) "Use of RDF for Expertise Matching within Academia," submitted to the international journal of *Knowledge and Information Systems.*

Contents

LIST OF FIGURES

LIST OF TABLES

Chapter 1

Introduction

More and more organizations have realized that effective management of their knowledge asset is very important in order to win business in today's highly competitive environment [Abecker and Decker, 1999]. This has led to the idea of Organizational Memory (OM), which is used to store the knowledge and past experience of individuals or different groups. This knowledge and experience can then be reused by the current employees for effective decisions and actions.

There are two kinds of retrieval in an OM. One is "*information retrieval*" which aims to provide the information required for the task at hand. However, access to information alone is not sufficient since people may have problems in understanding the documented information. Furthermore, not all knowledge can be directly expressed in words. For example, expertise, as one of the highest-valued forms of knowledge, is stored in people's heads and cannot be easily codified. In order to find knowledge which is behind the explicit document, people need to communicate with each other. There is widespread agreement that employees learn more effectively by interacting with others and the real value of information systems is to connect people to people [Ackerman and Halverson, 1998; Bannon and Kuuti, 1996; Bennis and Biederman, 1997; Bishop, 2000; Choo, 2000; Cross and Baird, 2000; Gibson, 1996; Koskinen, 2001; Stewart, 1997; Wellins *et al.,* 1993; Yimam-Seid, 2003]. This brings a new problem – who should communicate with whom? The second kind of retrieval – "*people retrieval*" aims to solve this problem by facilitating people to locate others with similar interests in order that they can share their expertise and knowledge.

This thesis focuses on knowledge management in academia taking advantage of the understanding gained from the Leeds University Virtual Science Park project [Drew *et al.,* 1996; Lau *et al.,* 1999; Leigh *et al.,* 1999]. More specifically, the emphasis of this research is on "*people retrieval*". Instead of finding all the people with similar interests (such as Yenta [Foner, 1997]), this study concentrates on "**Expertise Matching**" - locating appropriate people with required expertise to help solve problems. The thesis analyses the problems with the current approaches to expertise matching in academia, and presents an empirical investigation aiming at improving the performance of expertise matching in terms of accuracy and efficiency.

In order to establish a common understanding of the key terms, explanations on the terminology are given below.

Data: raw numbers or facts. It has no meaning by itself.

Information: interpreted data. It is meaningful data.

Knowledge: useful information. It happens when people use information.

Skills: more intelligent, denoting familiar knowledge united with readiness and dexterity in execution or performance.

Expertise: possession of knowledge and skills, and the ability to deal with the unknown and the unexpected. In this thesis, it is defined as "a specialized, in-depth body of knowledge and skills in a particular academic area(s)/topic(s), and the ability to use them in creating new knowledge or apply it to new applications."

This chapter begins with a discussion of the motivation behind the investigation of expertise matching. Section 1.2 presents the typical approaches to expertise matching and examines their limitations. Section 1.3 describes the key research issues and objectives. Finally, Section 1.4 outlines the organization of the rest of this thesis.

1.1 Motivation

Knowledge is the most critical asset for a company [Grant, 1996]. Expertise, a major component of tacit knowledge, is the most important basis for the generation of new knowledge, therefore it is the most valuable knowledge [Wilson and Fredericksen, 2000]. Expertise is generally recognized as skills and experience[1] and is developed through individual's learning and

[1] The definitions of expertise are fully discussed in Chapter 3.

practice. Expertise defines the organization's unique capabilities and core competencies [Finley, 2001; Holloway, 2000; Olson and Shaffer, 2002]; organizations that make use of existing expertise are likely to have a distinct competitive edge over other players in terms of faster knowledge creation and innovation, more efficient and effective use of existing organizational know-how, and reduced risk of loss of valuable knowledge when people leave the organization and stronger team creation [Davenport and Prusak, 1998].

However, if the expertise of employees remains hidden in the individuals' heads and cannot be accessed by others when they need it, then the potential of expertise will be lost [O'Dell *et al.,* 1998]. The great value of expertise can only be exploited when an individual's expertise can be shared with others so that people can obtain the required knowledge and experience to accomplish their tasks at an optimum level in the shortest possible time to achieve maximum productivity [Finley, 2001].

There are two ways to share expertise. The first is to transfer expertise into explicit form such as documents or databases so that it can be disseminated easily. The major drawback is that the expertise may be difficult to express and codify. The other method is to enhance people to communicate with each other. Communication leads to expertise sharing through the free exchange of ideas and experience. For example, meetings, informal talks, and seminars are commonly used opportunities to discover what others are doing and learn from their "stories" or experiences. The value of this communication has long been recognised as an important mechanism for expertise sharing [Fagrell and Ljungberg, 1999]. If the "expertise seeker" is a novice, then the expertise sharing is single directional only; if the "expertise seeker" is also experienced, then the sharing is bi-directional, which means the expertise provider can also benefit from this sharing. The following scenario provides a useful explanation of the value of sharing knowledge.

> *"If I give you a dollar and you give me a dollar, then we each have a dollar. But if I give you an idea and you give me an idea, we each have two ideas… A dollar stays a dollar and doesn't increase in value even if I pass it on, but if I pass on an*

> *idea worth a dollar and discuss it with somebody else, I often receive a good tip. Then, all of a sudden, this idea is worth two dollars. The other person may also implement my idea and make a dollar fifty, or possible even three dollars, but we have both benefited!"*
>
> Cited from [D'Oosterlinck *et al.,* 2002 p.67]

The above scenario illustrates the importance of sharing ideas; it also works for sharing expertise. Although the expertise of each person is created from their own practice, it can be reapplied in different contexts and for specific purposes, and its value increases. In order to facilitate this sharing, it is necessary to be aware of others who have specific expertise. In a small organization, knowing who is an expert in what specific area is not a big problem as everybody knows each other or at least an expert (if one exists) may be located through asking a colleague. However, for large organizations, especially those which are geographically distributed, it is extremely difficult for employees to know each other's competencies and share their expertise. When these people have problems, they usually cannot quickly find experts in the organization with the required expertise. As a result, they have to spend time and effort on reinventing useful things, such as key business processes, systems, skills, relationships, and so on [Olson and Shaffer, 2002]. Expertise matching plays an important role in avoiding these duplicated efforts by identifying experts who have experience and knowledge. Furthermore, expertise matching is very useful when people seek a collaborator, team member, researcher, presenter and so on.

Nowadays, expertise matching is receiving more and more attention in universities. A university is a good example of a knowledge-based organization[2]. The knowledge and expertise of university staff who teach and research in different areas is the major asset that a university holds. In order to make use of this asset, there is a need to share expertise between staff as well as transfer expertise to industry. Normally it is difficult for university researchers to identify companies which could significantly benefit most from the application of their research. Hence, providing a facility to help industry locate experts with the specific expertise whenever they

[2] Knowledge-based organization: an organization whose functions revolve around knowledge of workers and knowledge embedded in artefacts and processes [cited from FAA knowledge sharing glossary http://km.faa.gov/ks.nsf/glossaryweb]

want is of growing importance for universities as their role expands to include knowledge transfer. Often the speed with which industry can find an expert or several experts from different disciplines can improve the chance of success for collaboration between university and industry. Further, there is an increasing requirement on multi-disciplinary research, which means that it is important for members within a university from different departments to be aware of each other when they are doing similar things. Thus expertise matching is also very important in this context.

The following section gives a brief description of basic approaches to expertise matching and the limitations associated with each approach.

1.2 Expertise Matching

1.2.1 Knowledge Broker

One of the classical approaches to expertise matching is to rely on specialized people with the capacity to span all areas of an organization and know what it is that everyone else knows. This kind of person can be thought of as a knowledge broker. Figure 1 shows that the knowledge broker is situated between knowledge seekers and the organization memory. Each component is further explained as follows.

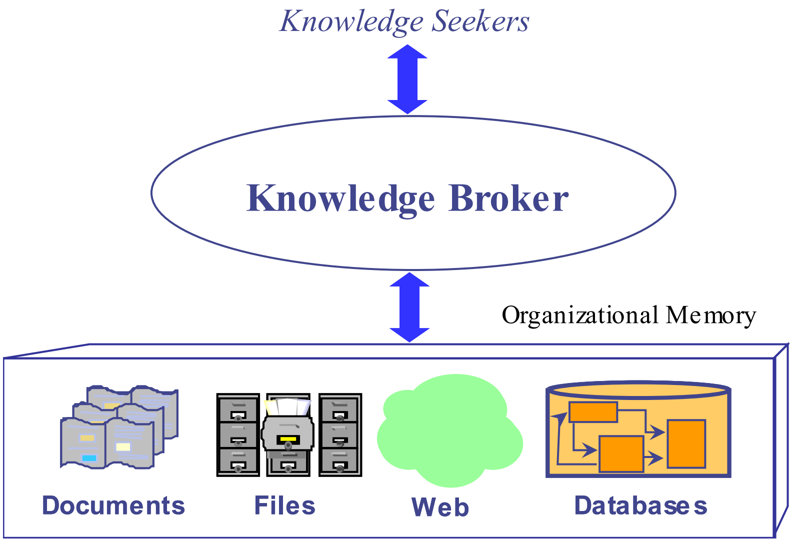


Figure 1-1 Knowledge broker searches OM to locate information or experts

Knowledge Seekers – members of an organization who require knowledge for particular purposes such as solving problems, collaboration and team formation.

Organizational Memory – an information repository used to store information created in the past intended for future use in a written record such as databases, documents and so on. Information which needs to be recorded includes corporate manuals, processes, procedures, project documents, expert directory, and so on.

Knowledge Broker – someone who brings together knowledge seekers and knowledge sources. The source of knowledge could reside in an explicit document, be the combination of several data sources, or it could be in the mind of experts. A knowledge broker should rapidly find and filter the relevant information [Eisenhart, 2002], and quickly locate the experts [Hellström, *et al.,* 2000]. Connecting individuals is considered as the dominant function of knowledge brokers because it facilitates learning from each other, converting tacit knowledge into real value for the company or the organization [Costello, 2000].

In order to serve these two roles, a knowledge broker needs to be an expert, who has experience with the company or organization [Hellström *et al.,* 2000]. The broker knows which data sources are relevant to the query of the knowledge seekers, what is the quality of each source, how to detect duplicated information if there is overlapping records, and how to sort the results back to the knowledge seekers. Furthermore, the broker should have a large contact network in order to identify specific people who are "extra knowledgeable" in some area.

Locating experts is more difficult for the knowledge broker than finding and filtering the relevant information. This is because: (1) The organisation is large and so it is impossible for the broker to know every expert; (2) People's expertise develop over time and it is difficult for the broker to capture this; (3) The members of the organisation are constantly changing; some may leave the organisation and other new members join. McDonald comments [2000, p.61],

> "*a single person may be able to keep track of many things, but in large organizations the number of people and the number of activities become too many and too varied for a single person to completely know and understand*."

1.2.2 Expertise Database

If users wish to search for information in web pages they can use search engines. However, if they want to locate somebody with the required expertise, there is no existing system which provides a satisfactory result[3]. Users have to manually check each "hit" to see if there is a link to the personal homepage of a suitable expert. In fact, not all experts have their own homepage and in these cases users have to search other data sources in order to find the information they need. Considering the huge amount of information that the organizational memory stores, it is no surprise that searching for people with specific expertise is a common problem in nearly every organization [Liao *et al.,* 1999]. Rather than relying on one or several knowledge broker(s), the alternative solution is to create an expertise database where individuals specify their expertise using several keywords or short sentences and users can then search these databases to find an expert. This solution is increasingly used by large organizations such as universities. This is because their staff normally work in small research groups or work alone, so it is unlikely that one person will know the expertise of everybody else throughout the department or organization.

A typical example of such an expertise database is Community of Science (COS)[4]. It is an Internet site for the global R&D community. COS brings together the world's most prominent scientists and researchers at more than 1,600 universities, corporations and government agencies worldwide. It is a knowledge management service for individuals and institutions. Currently, there are more than 480,000 personal profiles of researchers from over 1600 institutions worldwide stored in the COS expertise database. The fields in the COS Expertise Database include last name, first name, institution, past position(s), expertise, memberships, keywords[5], qualifications, patents and publication(s). Users can use keyword searching on one or several of these fields to locate experts.

[3] For example, when searching for experts in the area of "speech recognition" in the Leeds University domain 637 results are returned, which include presentation slides, thesis abstract, module introduction, training resources and so on, but no personal homepages were returned for the first 20 results.

[4] COS Expertise http://expertise.cos.com

[5] Professional editors at COS select the terms from controlled vocabulary and assign these terms to each profile added to the Expertise database.

A survey of experts finding systems among 27 universities has been conducted; the result is summarised in Appendix A. From the survey it can be seen that Experts Finding systems within a university are similar to COS because (i) most systems ask academics to create a profile themselves, it is up to them how much or little detail they supply. Take COS as an example, although there are 12 fields provided in the Expertise Database, only contact information and expertise information are compulsory, all the other fields are optional; (ii) the experts' information is stored in relational database or LDAP directory. Experts can be retrieved through browsing the simple subject tree or through keywords searching; most systems do not have the capability to rank experts which means that users have to check each expert's detail in order not to miss the most relevant expert; (iii) the task of maintaining the up-to-date profile is dependent on each expert although supporting team members can remind experts to do so periodically (for example, every 3 month or every year).

1.2.3 Expertise Matching at the University of Leeds

From the survey it can be found that the ULPD (University of Leeds Publications Database)[6] Expertise Matcher is representative since it includes common features (as well as common limitations) of most experts finding systems in the survey. Furthermore, it is one of the earliest expertise matching systems and the data is more accessible than other systems. Therefore it is selected as subject for the study.

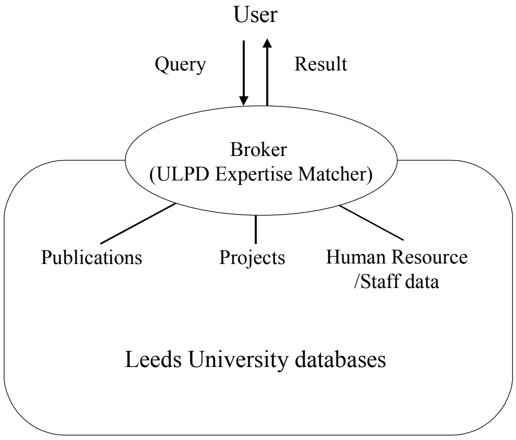


Figure 1-2 The ULPD Expertise Matcher as a knowledge broker linking users and experts

[6] http://ulpd.leeds.ac.uk/default2.asp; the previous version of the ULPD is REPIS (Research Expertise and Publication Information System).

The ULPD, a web-based information management system, is the attempt at the University of Leeds towards expertise management. ULPD stores information about the publications and research projects of academic staff from a variety of different sources including On-line System for the Computerised Administration of Research (OSCAR), Management Administration Information System (MAIS), Student Information Management System (SIMS) and the staff Phone/email directory. The principal objectives of the ULPD are to provide a central repository for information about all publications authored by University members of staff and research postgraduates, and to provide the opportunity for individual members of staff to create their own personal profile. Users can use the ULPD expertise matcher to locate experts with particular expertise in the University and obtain other associated information about each expert such as position, contact information, publications and completed research projects. The Expertise Matcher acts as a knowledge broker connecting expertise seekers and expertise providers, as shown in Figure 1-2.

One unique feature for the ULPD Expertise Matcher in the University of Leeds is that the outputs of the academic (publications and projects) are used as the complementary source to derive their expertise. This is because although an expert can express their expertise in their own words, this description may not be completely accurate and it can be difficult for them to indicate what is the difference between themselves and their peers. In many cases, the important evidence to show that they are experts in a particular area depends on the tangible outputs they have produced from applying their expertise [Stenmark, 1999]. In the ULPD system publications and projects are considered as these tangible outputs and are used to derive experts' expertise. Experts are retrieved if their publications or projects information match the keywords that users enter.

However, the ULPD system still suffers from several problems (the limitations of the ULPD expertise matcher are more fully analysed in Chapter 4). The first problem is the keyword searching problem; a single keyword may have multiple meanings in different contexts whilst the same meaning can be expressed using different keywords. This means some retrieved experts may not be relevant and other relevant experts may be missed. The second problem is that manually creating and maintaining a database to store all this information is very difficult and expensive. The third problem is that ULPD Expertise Matcher is unable to rank experts.

1.2.4 Expertise Lifecycle

"Knowledge Management takes the knowledge and expertise of people, plus the organizations' work processes and information repositories, and blends them into a comprehensive, collaborative environment" [Olson and Shaffer, 2002]. The Advanced Knowledge Technologies (AKT) project[7] is the state of the art knowledge management project which involves five universities throughout UK. To tackle the flow of knowledge around an organisation, the "knowledge lifecycle" has been studied [Shadbolt and O'Hara, 2003]. Expertise Management is the subset of knowledge management that focuses on the tacit knowledge stored in people's heads. Similarly, an expertise lifecycle is suggested in this thesis. It includes six activities as shown in Figure 1-3.

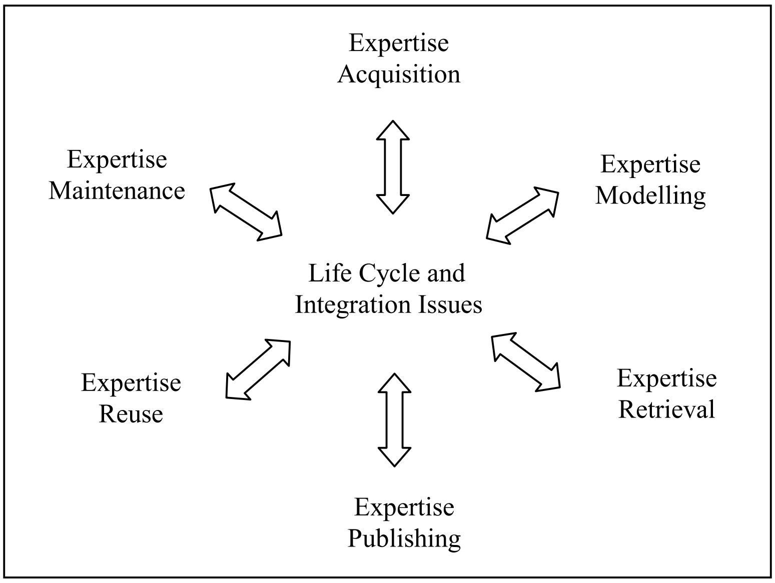


Figure 1-3 Expertise Lifecycle

A comparison of expertise lifecycle and knowledge lifecycle is shown in Table 1-1.

[7] AKT project http://www.aktor.org

Table 1-1 Comparison between expertise lifecycle and knowledge lifecycle

Expertise Lifecycle	Knowledge Lifecycle
Expertise Acquisition Capturing indications of expertise from diverse sources in the organizational memory	**Knowledge Acquisition** Capturing knowledge from diverse sources (e.g. experts, Web, electronic stores of data).
Expertise Modelling Expertise indicator extraction and expertise model representation	**Knowledge Modelling** Organising captured knowledge and describing it in formalised representation
Expertise Retrieval Identifying the experts with the required expertise	**Knowledge Retrieval** Finding the knowledge relevant to a particular problem from a repository
Expertise Publishing Presenting supported information for the retrieved experts so that users can select the appropriate experts easily	**Knowledge Publishing** Presenting modelled knowledge in different ways according to the users' requirements
Expertise Reuse Making expertise available for broader application rather than reinvention.	**Knowledge Reuse** Applying stored knowledge to new contexts instead of acquiring such knowledge afresh
Expertise Maintenance keeping the expertise information up to date (1) Updating expertise of current members. (2) Adding expertise of new members. (3) Removing the expertise of leaving people.	**Knowledge Maintenance** Keeping the knowledge up to date and discarding knowledge that is not useful any more

This thesis focuses on the expertise matching problem, therefore, it does not include expertise reuse activity[8]. A unique process for expertise matching consists of three steps as shown in Figure 1-4, where some activities of expertise lifecycle are regrouped.

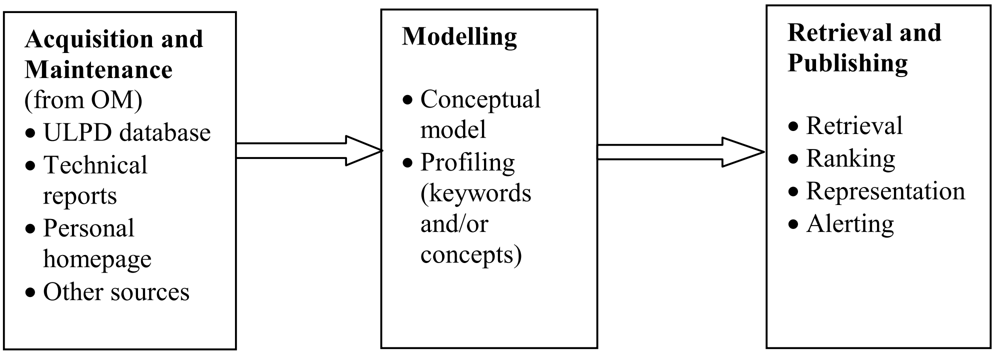


Figure 1-4 Expertise matching process

- In the *Acquisition and Maintenance* stage, the relevant data sources (expertise indications)[9] are collected, which include the direct statement by each expert (such as their personal homepage) or indirect evidence in the form of their outputs (such as technical reports). The acquisition activity is automatically repeated, through which some level of maintenance can be realised (except removing expertise of leaving people).

[8] The expertise reuse is realised through people interaction and communication.

[9] Expertise indication refers to evidence of expertise such as document authorship.

- In the *Modelling* stage, a conceptual model is built to integrate all these data sources and a domain ontology is used to store the main concepts in a domain of interest, the relationships between the concepts and the associated keywords to each concept. An expert's expertise is profiled in either keyword form and/or concept form.

- In the *Retrieval and Publishing* stage, the relevant experts with the required expertise are retrieved. Experts are ranked according to their expertise level and the detailed information of each expert integrated from the different data sources is presented to users in order to support them in selecting the appropriate experts. This retrieval process can be automatically repeated for those users who have comparable static requests so that the new experts will be identified more quickly.

1.3 Research Problems and Objectives

The specific problems addressed in this research have arisen from the analysis of the ULPD Expertise Matcher. Although it does not rely on each expert to specify their expertise, the aim of supporting users to locate experts quickly and accurately has not yet been realised. Furthermore, the burden for users in selecting the appropriate experts is still significant. In broad terms, the author's principal objectives for this research are to:

- Improve the performance of expertise matching in terms of precision and recall.
- Integrate and improve the quality of information provided for each expert in order to assist users to assess the experts' expertise.

A brief description of the research issues is presented below with details deferred to later chapters.

How to measure similarity between an expert's expertise and a user's request The significant drawback of the ULPD Expertise Matcher is that an exact match is required. This means that the experts will only be retrieved if their publications or projects information exactly match the keywords entered by a user. Therefore, there is no mechanism to rank the expertise of the retrieved experts. In order not to miss relevant experts, it is necessary to give users a flexible method of expressing their needs. In this research, the associated projects and publications information relevant to each expert are retrieved and processed using vector space model and an expertise profile is then generated. The similarity between an expert's expertise and a user's request is obtained by calculating the two vectors (user profile and expertise profile). Through this way, the retrieved experts can be ranked.

How to explore multiple expertise indications in order to build up a more accurate expertise profile The information stored in the ULPD database is very limited, hence experts may not be retrieved because the relevant information about them has not been recorded. Even for those retrieved experts, the associated information provided to users is restricted. In order to solve this problem, it is necessary to explore multiple expertise indications from data sources in the organizational memory. In this research, semantic web technologies have been used to integrate multiple expertise indications from diverse data sources to create a complete expertise profile. Therefore a more accurate match can be obtained, and high-quality information relevant to each expert can be provided to the users to facilitate them in selecting experts.

How to ensure the relevant experts are not overlooked as well as irrelevant experts are minimized The problem of keyword-based expertise matching is that some relevant experts are missed and irrelevant experts are retrieved. This problem is caused by the syntactic-oriented nature of the keyword search approach. In order to solve this problem, a concept matching approach has to be explored. The domain ontology plays an important role in concept matching since it includes all the major concepts in a domain as well as the relations between concepts. In this research, a concept based expertise profile is created as a complementary to the keyword based expertise profile.

How to extend single disciplinary expertise matching to multi disciplinary expertise matching Multi disciplinary expertise matching is a new area and no related work has been found so far. This research has conducted an initial investigation into this area. The domain of

"GeoComputing"[10] is selected as the start point. Building a multi-disciplinary expertise model is similar to mapping between ontologies. A two dimensional expertise domain model has been built and ranking mechanisms has been proposed. Some obstacles which hinder the multi-disciplinary expertise matching are discovered through the initial study and suggestions are given.

In order to test if the performance of expertise matching has been improved against the current ULPD Expertise Matcher, a prototype system called the Expertise Locator has been built to undertake an evaluation using real data in the University of Leeds. Participants are volunteered students who compare the Expertise Locator with the ULPD Expertise Matcher by identifying relevant experts from both search results. Data is collected through observation, conversations with participants, and also via questionnaire. Which system outperforms the other is largely decided by the two widely used evaluation metrics - precision and recall. In addition, time spent on retrieval and users' satisfaction on the detailed information of each expert provided by the system are also taken into account.

The major contribution of this thesis is the empirical investigation of how to improve the performance of expertise matching within the Leeds University and more broadly to academia. Both syntactic and semantic-oriented techniques are studied. The specific contributions can be summarised as follows:

- An academic expertise matching conceptual model which provides a uniform semantic view over the input sources.
- The application of semantic web technologies (RDF, RDFS, ontology) to the expertise matching problem; that leads to the effective integration of pieces of information relevant to each expert from heterogeneous data sources.
- A prototype system (Expertise Locator) has been implemented and evaluated; the superiority of the retrieval effectiveness of the prototype system over the traditional database approach has been demonstrated.

[10] Geocomputing is a new, innovative application area where information technology has been applied to the Geoscience environment.

- The first attempt, known to the author, to solve the multi-disciplinary expertise matching problems. The expertise domain model is proposed and some suggestions for future research are given.

1.4 Thesis Organization

The organization of the rest of this thesis is described below.

Chapter 2 describes the two kinds of knowledge ("tacit" and "explicit") and analyses the importance of tacit knowledge such as expertise. The conversation between these two kinds of knowledge is then examined followed by the reasons why it is difficult to codify expertise and why interaction between people is important to share expertise. The concept of a knowledge sharing environment, which facilitates people's awareness of each other and expertise sharing, is introduced using a number of examples.

Chapter 3 analyses in detail the nature of expertise and the different expertise indications as well as the criteria for evaluate expertise matching. It also describes the domain model of expert finding systems. The previous work on expertise matching is also discussed and compared against the criteria.

Chapter 4 examines the limitation of the ULPD Expertise Matcher. An extension of the current expertise matcher is proposed which employs the vector space model to build an expertise profile. An extended prototype expertise matcher is evaluated and the results are presented.

Chapter 5 starts with the limitations that have not solved in the extended Expertise Matcher and analyses the possible solutions to the remaining problems, especially how to apply the semantic web technologies to solve these problems. An expertise matching conceptual model and an RDF-based architecture are presented. A prototype system - Expertise Locator based on the conceptual model and architecture is described. Finally it compares the result of expertise matching performance between the Expertise Locator and extended Expertise Matcher.

Chapter 6 discussed how to extend the single discipline expertise matching to multi-disciplines. The differences between single and multi-disciplinary expertise matching have been analysed. The requirements of the multi-disciplinary brokering system are informed through a preliminary study. A modified architecture is described together with the expertise domain model for multi-disciplinary expertise. The initial studies are presented and suggestions are given.

Chapter 7 concludes with a short summary of the work in this thesis and discusses the broader application of the research. This chapter also gives a list of possible directions for research in the future.

Chapter 2

Context

Chapter 1 described the increasing need for organizations to more effectively manage their expertise. This chapter first provides an overview of the role of expertise for organizations, it then describes the two approaches to sharing expertise and explains why it is difficult to codify people's expertise. A number of knowledge sharing environments are discussed which facilitate sharing both explicit knowledge (such as documents) as well as tacit knowledge (such as expertise).

2.1 Expertise Management and the Learning Organization

2.1.1 Introduction

A learning organization is an organization "*skilled at creating, acquiring, and transferring knowledge, and at modifying its behaviour to reflect new knowledge and insights*" [Garvin, 1998]. Such organizations are adaptive to their external environment and continually enhance their capability to change [Skyrme and Farago, 1995]. To achieve this, learning organizations need to make use of "*the amazing mental capacity of all its members*" [Dixon, 1999] and facilitates collective learning. A crucial issue for organizational learning is how individuals' expertise, as a result of their long time learning, can be transferred to the organization [Huang, 1998]. This involves two activities - identify the expertise of employees and leverage the expertise to full potential by linking expertise provider and expertise seeker at the right time. Employees in the learning organizations should be able to quickly locate the "right experts" in order to reuse others' experience. Through expertise sharing, people at all levels, individually and collectively, are continually increasing their capacity.

2.1.2 Role of Tacit Knowledge

Knowledge is often considered as the most important strategic resource to enhance the organization's fundamental ability to compete [Zack, 1999]. Knowledge can be divided into two categories: explicit knowledge and tacit knowledge [Mahapatra and Chakrabarti, 2002]. Explicit knowledge refers to knowledge that can be articulated in written language and normally conveyed through manuals, documentation, files and other accessible sources [Nonaka and Takeuchi, 1995]. Tacit knowledge is the "*cognitive skills such as beliefs, images, intuition and mental models, as well as technical skills such as craft and know-how*" [Nonaka, 1994]. It is personal, subjective and experiential knowledge [Dyer, 2000]. Tacit knowledge is stored in people's heads and difficult to write down or collate in the form of documents. According to the Delphi Group's study on more than 700 U.S. companies, a large portion of corporate knowledge (42%) is tacit knowledge, which remains locked inside of employees' heads (as shown in Figure 2-1).

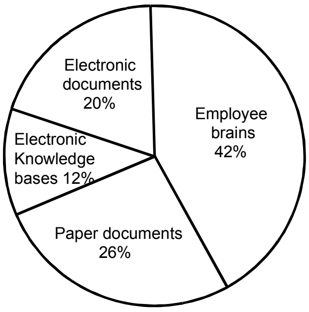


Figure 2-1 Distribution of corporate knowledge

(Source: The Delphi Group, cited from Hickins, 1999, p.100)

A key component of tacit knowledge is expertise, such as the skills and know-how. Expertise is acquired through a lifetime of experience. It provides the competitive advantage for an organization due to the following reasons.

- **Imitation** If an organization's advantage is based on explicit documents, it can be easily copied by its competitors [Teece, 1987; Reed and DeFilippi, 1990]. However, if the advantage of an organization is based on the expertise of its employees, it is difficult to imitate by competitors because expertise is deeply embodied in the person's personality, creativity, intelligence, perceptions, experiences, and so on [Fitzpatrick, 2003]. In order to create similar knowledge, competitors have to engage in similar learning experiences which takes time [Zack, 2002]. Hence, the competition of the organization will not be lost quickly.

- **Best Practice** Expertise represents the unique value added by the people who generate it when solving real problems. Compared with explicit documents, expertise reflects more closely the reality of how work actually gets done (in other words, work "practices" rather than business "processes"), which in turn can transfer best practices more effectively [Horvath, 2000].

- **Innovation** Expertise is strongly implicated in organization innovation. Innovation has two meanings: (i) to make changes, and (ii) to introduce new ideas, methods, and processes. Innovation requires insight and understanding of the current situation and continuous learning. Research shows a strong reciprocal relationship between prior knowledge and learning ability [Cohen and Leventhal, 1990], the more one already knows, the more one comprehends; the more one comprehends, the more one learns new knowledge. Since learning is a source of future innovation, the innovation is also largely dependent on people's expertise. Therefore, the more expertise employees have, the more capabilities they have to integrate new knowledge with their already knowing into new innovation.

2.1.3 Context Dependent and Reuse in Action

The content of information is important, however, it provides little value without associated contextual information. As Fitzpatrick [2003] states, putting content ('what') to work most effectively is critically dependent on knowing relevant contextual information ('how, why, where, who'). It is through knowledge of the context that content can be interpreted and communicated. Different people may have different interpretations based on the same information. Expertise is also context dependent. One person's expertise can be only benefited by others if they interconnect it with their own embodied knowledge and embed it in their own application. Expertise is integrated with people's existing knowledge to develop unique insights and create even more valuable knowledge. This is called "interconnectedness and complementarity" [Zack, 2002]. The value of expertise is only exploited when it is reused by different people at different times **in different ways** [Fitzpatrick, 2003].

2.1.4 Enterprise Requirements on Expertise

The roles of workers in the organizations have changed significantly from industrial age to information age [Nickols, 2000b]. The traditional knowledge management approach in the industrial age is that managers control the power while workers follow the procedures. As stated by Bekkedahl [1977], "*Knowledge held by a few, plus iron discipline over the many.*" Knowledge was narrowly concentrated by a few managers who made all the plans and decisions for their employees. It is believed that knowledge is embedded into procedures. The workers' task was to convert instructions and procedures into actions. What is converted is the materials only, from one form to another. In the new economy, the knowledge management approach places a higher value on people's intelligence and knowledge over rigid procedures. First, knowledge in organizations is widely distributed amongst the knowledge workers. Second, employees at every level have a significant amount of control over their work; the new task of workers is to convert knowledge into actions through which information is converted from one form to another. Workers continue learning new knowledge themselves in order to work effectively, their expertise and experience are augmented through daily work, which is very valuable for the organization.

The company that is able to make use of existing experiences and competencies quickest has a distinct competitive edge over other players [Gibbert *et al.,* 2002]. However, this knowledge and know-how cannot make great value for the organization if it is bound to an individual mind and cannot be accessed by others who need it. "*No amount of knowledge or insight can keep a company ahead if it is not properly distributed where it's needed"* [O'Dell *et al.,* 1998]. In today's knowledge economy, the pressures which most organizations are facing (such as distributed workforce, time to market, and fluid labour pool), require them to know what they know, to manage what they know, especially sharing expertise to accelerate innovation rates and retain core talent [CIO, 2002]. Bishop [2000] comments:

> *"It appears that many organisations today feel the only way to survive and prosper in a world characterised by speed, complexity, global competition, down-sizing and constant change, is to work smarter, not harder."*

Here "work smarter" means to encourage employees to collaborate with one another because: (1) employees depend on each other and it is very unlikely that individuals will undertake their tasks without the help of others; (2) the experience of one employee may be very useful for other employees. Only through collaboration and exchanging knowledge with each other regularly, can a group of people achieve greater than the sum of what can be achieved as individuals working alone [Bishop, 2000]. O'Dell *et al.,* [1998] points out:

> "*The major strategy for a company to achieve significantly higher levels of productivity is not by firing more people, not by buying more machines, not by forcing people to stay later and work harder, ... but by allowing people* ***to learn what works best in other areas and try it out in their own back yard****. And by ensuring* ***they have all the knowledge and experience they require to do their work at their best level****."*

The benefits of connecting people and encouraging them to share their expertise are summarised below.

- *Create organization-wide knowledge sharing*. This helps employees to capture and share undocumented knowledge.
- *Improve productivity*. Based on quick expertise location, employees can find otherwise unknown experts even when they are geographically separate and share their knowledge.

The time required for searching for knowledge from huge information repositories is saved which leads to productivity improvements and minimization of duplication.

- *Keep valuable expertise even when employees leave*. Through sharing, the expertise of individuals has been transferred to others before they leave. The loss to an organization is minimized.
- *Support collaboration.* If a team members have complementary expertise, then it normally leads more effective results.
- *Become more adaptive to changing conditions*. Quickly identify individuals who have accumulated many years of experience, expertise and insight is the key to unleashing the high levels of energy that enable organizations to become more effective, adaptive and responsive to changing industrial conditions.

2.1.5 Approaches to Sharing Expertise

Organizations may wish to capture their internal expertise and convert it into explicit knowledge so that it can be easily shared by large numbers of people. According to Nonaka and Takeuchi [1995], knowledge is not static; it may dynamically shift between tacit and explicit over time. Figure 2-2 shows four ways in which tacit knowledge (such as expertise) and explicit knowledge (such as documents) can be converted to each other.

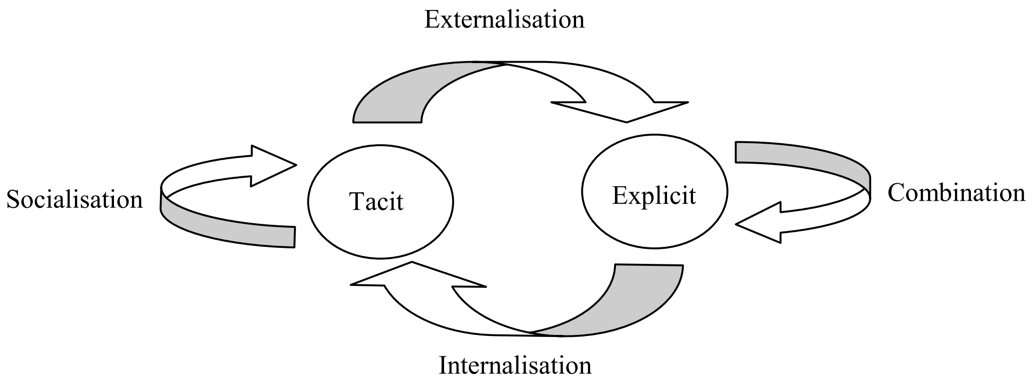


Figure 2-2 The processes whereby conversions between tacit knowledge and explicit knowledge occur (Nonaka [1994])

- *Tacit to Explicit (Externalisation)* This is the process of converting part of tacit knowledge to explicit through written language, for example, writing a paper.

- *Explicit to Explicit (Combination)* This is the process of merging diverse pieces of explicit knowledge into new explicit knowledge.
- *Explicit to Tacit (Internalisation)* This is the process of understanding and absorbing explicit knowledge into individual's own knowledge, such as "learn by doing".
- *Tacit to Tacit (Socialisation)* This is the process of creation of new tacit knowledge through discussion, observation and practice. In this process people expose their knowledge to others and test its validity.

From the above conversation process it can be seen that in addition to tacit-to-tacit process there is another way to share expertise, which includes two modes of interaction - externalisation (the process to codify expertise) and internalisation (the process to absorb knowledge from explicit form, adapt and adopt it in a new context). Expertise sharing is realized through tacit knowledge → explicit knowledge → tacit knowledge. This approach is criticized largely because of the difficulties in codifying expertise. The following section explains the reasons of the difficulties.

2.1.6 Difficulties in Codifying Expertise

The first process *Tacit to Explicit* transfer indicates expertise can be codified. However, there are three major barriers in this codification process [Stenmark, 2001]. These are described below.

- People are not fully aware of their tacit knowledge – "tacit knowledge incorporates so much accrued and embedded learning that its rules may be impossible to separate from how an individual acts" [Davenport and Prusak, 1997]. This is also called the "unknown knowledge" and to share this kind of knowledge needs skills observation, on-the-job experiences, and apprenticeships [Heimburger, 2001].
- On a personal level people do not need to make tacit knowledge explicit in order to use it since people are able to use their tacit knowledge naturally. In addition, people cannot directly benefit from codifying their tacit knowledge, which is normally a difficult and laborious task.

- There is a potential risk of losing power and competitive advantage by making it explicit – if the tacit knowledge provides an important competitive advantage, there is little reason to share it with others.

Even when people are aware of their tacit knowledge and are willing to share, only a small percentage of tacit knowledge can be codified due to its "embodied" characteristic [Horvath, 2000]. One study shows that 80% of the knowledge that needs to be transferred is in the non-codifiable area [Holloway, 2000]. Some attempts to codify tacit knowledge have yielded disappointing results such as the example below.

> *Xerox once attempted to embed the know-how of its service and repair technicians into an expert system that was installed in the copiers. They hoped that technicians responding to a call could be guided by the system and complete repairs from a distance. But it turned out that technicians could not solve problems using the system by itself. When the copier designers looked into the matter more closely, they discovered that technicians learned from one another by sharing stories about how they had fixed the machines. The expert system could not replicate the nuance and detail that were exchanged in face-to-face conversations.*
>
> cited from [Hansen, 1999 p.68]

The codification process is both difficult and costly, and the fact that the tacit knowledge must be externalised before it can be exploited limits its usefulness [Stenmark, 2001]. The primary disadvantage of documented knowledge is its lack of contextual richness [Lyons, 2000]. The writer lacks the insight or imagination to understand where the readers are coming from and the context in which they interpret his words.

The alternative expertise sharing approach, socialization, is recommended by many researchers, such as Horvath [2000] and Fitzpatrick [2003]. This process enables communication which allows people to capture the rich context and better adapt the content to their own situation [Davenport and Prusak, 1998; Lyons, 2000].

One limitation of socialization process is that physical proximity is typically required in order for the sharing of expertise to occur. This requirement is difficult to satisfy with large organizations, distributed organizations or virtual organizations. Sometimes, in order to enhance organizational flexibility, people are organized into cross-functional teams, these people spend all their time and energy dedicated to the projects and get less chances to find out the expertise of their peers because they disconnect with others. In order to realize expertise sharing, people need to find the right people with the required expertise. In most organizations, employees rely on personal information social networks to locate experts. However, this approach suffers from problems such as "potential unreliable, frequently limited in their effectiveness, cannot scale particularly well" [Bussell and Holter, 2002]. A knowledge sharing environment is then designed to help people locate experts no matter where they are.

2.2 Sharing Expertise through Knowledge Sharing Environments

A knowledge sharing environment (KSE) is an environment that supports the processes of sharing and transferring knowledge within network communities or project teams with the help of modern communication technologies such as the Internet. A KSE is characterized by virtual working across spatial boundaries and its ability to provide users access, sharing and management of various types of knowledge at different levels (individual, group and organizational).

2.2.1 Network Community

One important reason for developing a KSE is to facilitate a new mode of community – the network community. A network community is a group of people whose communication and collaboration over networks strengthens and facilitates their shared identity and goals (share expertise and solve problems together) [Carroll and Rosson, 1998]. The core idea is to extend the current organisational boundary, so that people can find and interact with others who have experience and expertise in a specific area, which brings down the barriers of physical localities that traditionally hindered knowledge sharing. People are organized together automatically if they have similar interests, and they can enhance the mutual understanding and trust through using public communication networks[1]. The network community is powerful and flexible with the following features [Ishida, 1998].

- People in distant places can join the same community.
- Each person can participate in multiple communities at the same time.
- There is no specific structure defined beforehand in the community, and its structure changes dynamically.
- A community itself is spontaneously created and modified, and possibly diminishes over time.

2.2.2 Examples of Knowledge Sharing Environments

A KSE can provide three major functions:

- *Facilitate finding experts or other people who have similar interests* Through locating experts, it provides the potential for users to locate experts with the required expertise and capture valuable tacit knowledge embodied in experts themselves.
- *Facilitate communication* Once experts or people who have similar interests are located, the KSE can help people maintain connections through collaborative tools such as chat rooms and videoconferences, which foster interactions that lead to increased trust and expertise sharing.
- *Facilitate access to community memory* It aims to share explicit knowledge. Network community members can quickly and easily access to community's information repository.

[1] A new type of scientific collaboration called e-science (http://www.lesc.ic.ac.uk/admin/escience.html) aims to provide support for large-scale scientific experiments by enabling distributed global collaborations through the formation of virtual co-laboratories. These will allow scientists to work together irrespective of location and permit universal access to scientific resources.

The following three examples of knowledge sharing environments have common features in terms of facilitating people to find others who have expertise or who have the similar interests in order to share tacit knowledge, and facilitating people to share explicitly documented knowledge.

2.2.2.1 Virtual Knowledge Park

The Virtual Knowledge Park (VKP) Project aims to support knowledge management and outreach activities within the University of Leeds. It facilitates knowledge transfer between the University and business by providing collaborative tools and access to the internal knowledge sources, such as university expertise; external knowledge sources from outside university can be extracted through collaborative and project based working.

Finding people An expertise matcher is built within the VKP to help users to search for people with specific skills and abilities, and to identify suitable individuals to form a project team. The search fields include: name, expertise, profile, skills, languages, geographical location, business sector, and previous employers. Users can browse a standardised classification list or use keyword search (based on publication and projects database) to find experts.

Support communication To encourage geographically separated team members to collaborate and to increase trust between people, the VKP uses a series of collaboration tools which support two kinds of communication -- synchronous (real-time communication, such as Netmeeting) and asynchronous (non-real-time communication, such as email).

A browser can be used alone to access the core set of collaborative work tools including document management, discussion groups, information resources, email account management, email notification and contact books. The VKP Assistant software can be used, in conjunction with a browser, to deliver synchronous communication tools alongside asynchronous collaborative working tools. This includes video and audio conferencing, application sharing (joint real-time document editing), instant messaging, chat rooms, file transfer and whiteboards.

Workspace – Shared Resource The VKP supports the Broadbent Knowledge map (shown in Figure 2-3) by creating three types of workspace to facilitate resource sharing. They are personal workspace, team workspace and public workspace. Access to any information within a workspace is possible only if the current user has explicit permission to access that information, either as an individual, or as a member of a team.

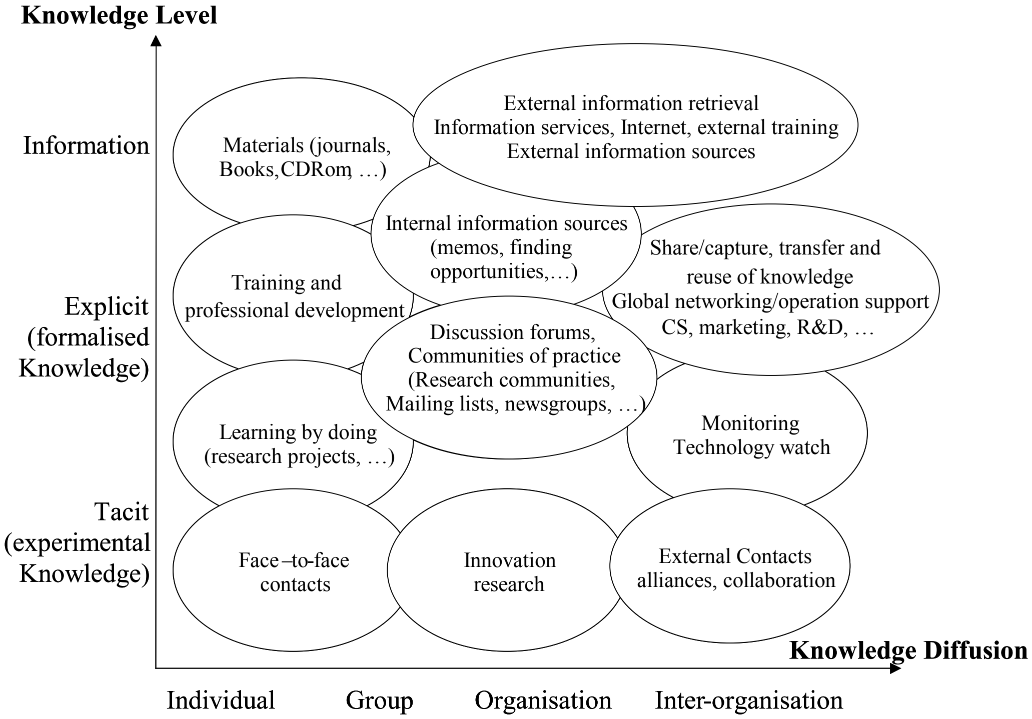

Knowledge Level
Information
Explicit (formalised Knowledge)
Tacit (experimental Knowledge)
Materials (journals, Books,CDRom, …)
External information retrieval
Information services, Internet, external training
External information sources
Internal information sources (memos, finding opportunities,…)
Training and professional development
Share/capture, transfer and reuse of knowledge
Global networking/operation support
CS, marketing, R&D, …
Discussion forums, Communities of practice (Research communities, Mailing lists, newsgroups, …)
Learning by doing (research projects, …)
Monitoring Technology watch
Face–to-face contacts
Innovation research
External Contacts alliances, collaboration
Knowledge Diffusion
Individual
Group
Organisation
Inter-organisation

Figure 2-3 Knowledge management map (adapted from Broadbent, 1998)

The knowledge map has three knowledge domains or levels (tacit knowledge, explicit knowledge, and information) and four knowledge locations. These locations represent the extent of knowledge diffusion: individuals, groups, the organization as a whole, and inter-organizational locations.

Individual Each registered user has a personal workspace where users can manage their personal documents, have access to an organization's resources (expertise), be part of formal and informal discussion groups, and manage personal contacts via a contact book.

Group People in a group have a team workspace where the team members can store, retrieve, and update internal documents belonging to the team; they can also attach personal notes to the documents. They can access the discussion forums. In order to help team members to share tacit knowledge with each other, the expertise matcher is used to locate team members with the required skills as members of the team might be spread across space or belong to different organizations and may not know each other. The Discussions component allows users to participate in topic centred forums. The Discussions component is accessed from the Navigator and exists in each user's Personal Workspace and in each project workspace.

Organization The organizational memory comprises information on the skills and expertise of Leeds University staff and explicit knowledge – documents produced or recommended by staff. Users can search the organizational memory to locate experts or documents based on the metadata such as title, abstract, author, filename, etc. structured, indexed document repository. Alerts enable users to be notified about events; that is, actions which take place to documents, folders, discussions, and users.

Inter-organization Collaborative work tools are provided to support the development of partnerships between the University and business. Information sources may be accessed by external users to support collaboration on projects.

2.2.2.2 BT-KSE

BT has developed the Knowledge Sharing Environment (KSE) [Davies *et al.,* 1998] to support virtual communities to interact in a virtual space, whose members may be geographically and temporally dispersed. It is a system of information agents which organizes users into small communities based on their common interests. Users coming from different organizations can join the same community and share knowledge from a number of sources. The possibility of the exchange of tacit knowledge is opened up by adding awareness of others with similar interests or concerns.

Finding people A user can search for others who share the similar interests by comparing the user’s profile with others’ using the vector space model [Baeze-Yates and Ribeiro-Neto, 1999]. The retrieved people are ranked according to the similarity level. Each user has a personal profile in a set of keywords initially provided by the user but can be adapted by the personal agent through observing user's behaviour. For example, if the profile of a user does not match the information being stored by him/her, the agent will suggest phrases which the user may elect to add to their profile.

Sharing explicit documented knowledge The information to be shared in the KSE comes from the Internet, from an organization's intranet or from other users. Only metadata is stored in KSE agents such as reference to the remote WWW document, a summary of the document, an annotation, date of storage, and the user who stored the information. Users are informed of the relevant information. This is realized through matching users' profiles with the content of the page using the vector space model.

In addition to being informed with the relevant information, a user can also ask his KSE agent for the most recently stored information. The agent then searches the KSE store and presents the user with a list of links to the most recently stored information, along with annotations where provided, date of storage, the storer's name, and an indication of how well the information matches the user's profile.

Communication However, the KSE does not support people communicating with each other. This drawback is overcome in another system called "Knowledge Garden" [Crossley, 1999]. Knowledge Garden provides an environment with a 3D information visualization tool which can help users to meet colleagues and share information. In this shared environment, users can see their fellow team members via their representational avatars within the people section of the garden. Avatars can meet and communicate via a number of media including text, audio, video, and electronic whiteboards.

In addition, knowledge garden assists users to select useful documents on their own rather than relying on the retrieved documents matching users' queries. This is because the similarities between documents are clearly presented in the knowledge garden, which can be seen as a complement of the retrieved documents. Information is seen as an organic resource that changes over time, and is represented as plants in a shared three-dimensional knowledge garden. Internet resources are clustered and the related information resources are grouped together as “plants”. When a user takes a cutting from information plants into their own personal environment, a set of key phrases is extracted from the document(s) represented by the stalk and this set of key phrases is sent to a search engine in order to retrieved relevant documents. The ten most relevant documents are then represented as stalks on a plant which grow in the 3D space.

2.2.2.3 GMD – Social Web

“Social Web” is an Internet-based infrastructure that facilitates social activities such as meeting people with similar interests, forming groups, and working together [GMD-FIT, 1998]. It aims to build a social space using computers and networks as a social medium to link people as well as documents. It is expected to offer “places”, or social spheres, where social activities take place, with more awareness about other members rather than in anonymity as featured in the present-day Internet.

Finding people GMD’s match-maker agent assists with finding experts or persons with similar interests in the community of users who might join a group or collaborate on a task according to their profiles. A user profile in the match-maker agent is expressed as a set of text vectors, which can be derived from a query, a task, or a set of bookmarks [Voss *et. al.*, 1998]. The matching process is similar to BT-KSE agent, however, the individual’s profile is static rather than adaptive as in the BT-KSE system.

Access to community knowledge Documents are seen as the main carrier of knowledge. When members of a community store documents to the information repository they also identify the significant concepts in each document. The cross-references between documents are then created by relations among concepts. This cross-referencing is called Concept Index. A Concept Index provides a shared vocabulary and enriches document relation rather than direct references and the physical location of documents [Nakata *et al.,* 1998]. It also facilitates users to quickly navigate documents and locate the most useful parts. Besides simply highlighting the keywords in the documents, users can add synonyms and related concepts in order to refine the index. Synonyms and related concepts can be selected from thesauri such as WordNet or discovered by text mining [Voss *et al.,* 1998].

Communication/collaboration GMD has developed the BSCW (Basic Support for Cooperative Work) Shared Workspace system with the goal of transforming the Internet from a primarily passive information repository to an active cooperation medium [Appelt, 1999]. The BSCW system extends the browsing and information download features of the Web with more sophisticated features, which are similar to that in the VKP workspace. For example, discussion forums, search facilities, document upload, document version management, access right control, synchronous communication, member and group administration, and event awareness (such as uploading a new document, downloading an existing document, renaming a document, and so on). This enable effective communication and collaboration among multiple people.

2.2.2.4 Summary

The above examples illustrate that knowledge sharing environments can help people share explicit documents as well as sharing expertise through identifying the similar people and providing communication tools. Compared with document management, connecting people is a new area and different approaches being used. The BT-KSE and GMD-Social Web built user profiles in a set of keywords and use vector space model to match people with similar interests. They did not distinguish experts with others who just have interests in a particular area. Therefore, it is possible that a returned "expert" or expertise provider is actually an expertise seeker. Although finding people with similar interests is useful, locating the experts with the required expertise is more important since the real experts can provide explanation, solutions to the questions. The VKP make an improvement by deriving people's expertise from their publication and project information (which are expertise indications). However, the VKP expertise matching cannot rank experts so all the retrieved experts seem equally important. All of three system use keyword search rather than concept search. It is found that in the GMD social web, a concept-index is created to link the concepts with documents, this can be further extended to include experts in order to support concept search. There are still some open issues which require further investigation, for example, how to identify people's expertise and how to improve the accuracy of the experts retrieval.

2.3 Conclusions

Expertise stored in employee's heads is important in retaining key competences in knowledge-based organizations. In order to make great value of individual's expertise, it is necessary for employees to share their expertise with each other. There are two approaches to sharing expertise – codification and socialization (personalization). Codifying expertise is expensive and sometimes less effective than sharing expertise through interaction between people (socialization). A basic requirement for this socialization process is to find the right people with the required expertise. Some knowledge sharing environments have been built to facilitate people sharing expertise and support collaborative learning. However, the emphasis has been put on explicit documents and the process of matching expertise is crude. This study investigates how to improve the performance of expertise matching in a Knowledge Sharing Environment; the focus is put on academic environment such as VKP Expertise matcher. In order to achieve this, it is necessary to have a deeper understanding of the nature of expertise and relevant approaches to expertise matching. These issues will be discussed in detail in the next chapter.

Chapter 3

Related Work on Expertise Matching

Chapter 2 analysed why sharing expertise is important for organizations and described a number of knowledge sharing environments where expertise sharing is supported. This chapter provides a comprehensive analysis of the nature of expertise. It then describes the domain model of expertise matching systems followed by a number of specific criteria of expertise matching. The previous work on expertise matching is also discussed and compared against the criteria. This chapter ends with a number of areas for further research towards an effective expertise matching.

3.1 Nature of Expertise

Researchers in cognitive psychology, cognitive science and computer science have conducted a significant amount of research on the nature of expertise over the last thirty years. Dozens of definitions have been given which indicate different understandings of the nature of expertise. These understandings are classified into four groups; each of these groups is discussed briefly below.

3.1.1 Expertise as "the possession of skills and knowledge"

One definition is that expertise is the possession of knowledge and procedural skill(s) [Bedard, 1991]. A similar definition can be found in Webster's dictionary[1], "*having, involving, or displaying special skill or knowledge derived from training or experience*"; Knowledge is defined as "*acquaintance with facts, truths, or principles, as from study or investigation*", whilst skill is defined as: "*the ability, coming from one's knowledge, practice, aptitude, etc., to do*

[1] http://www.m-w.com/home.htm

something well". These definitions consider skill or knowledge as some substance that may be possessed by the individual (in this case, it implies that knowledge is the substance underlying skill), and once a person has this substance, he/she has "*problem solving ability*" [Green and Gilhooly, 1992], and can "*perform qualitatively well in a particular domain*" [Frensch and Sternberg, 1989].

3.1.2 Expertise as "process"

Definitions that rely on knowledge or skills are often criticized because they assume expertise to be consistent and invariable. Gaines [1995] argues that expertise is not something that simply exists which can be captured and transferred to a computer, and the fact that people demonstrate "action centred"[2] skilled performance in a pre-defined task (such as typing) does not illustrate that they possess knowledge. Expertise is dynamically evolving. The real experts do not merely preserve their existing capabilities, but extend them continually in order to match dynamic situations, including unpredictable circumstances [Schön, 1983]. The definitions imply expertise as a property of individuals and focus on demonstratable skill, but ignore how a person becomes an expert, in other words, how they assimilate experience. Some researchers have noticed this problem and suggested a process component should be incorporated. For example, Marchant [1989] views expertise as "*a process by which individuals develop the ability to achieve task-specific superior performance*". In Dreyfus's model of expert skill acquisition, five stages of expertise are presented (Novice, Advanced Beginner, Competence, Proficient, and Expert) [Dreyfus, 2001]. This definition focuses on the learning process - the internalisation of given rules to deal with different situations where intellectual skills are needed.

This view of expertise is on the upper layer of skill understanding because know-how requires skills plus the ability to apply it to different contexts (such as judge which pattern is appropriate for the situation). These understandings are criticized because of their routinization. The assumption is through repetitions of routine tasks, people can perform better in terms of speed and accuracy [Bullard *et. al.*, 1995]. However, this performance is based on highly practiced,

[2] Zuboff distinguishes action centred skills and intellective skills in [Zuboff, 1988].

pre-programmed tasks. Although people can use pre-stored rules in a slightly changed situation, this situation is predictable. When facing an uncertain or unpredicted situation, "experts" do not always outperform "novices" [Engestrom and Engestrom 1986]. However, real experts can demonstrate *knowledge-based* performance in coping with very different or completely novel situations instead of *skill-based* or *rule-based* performance [Maurino *et. al.*, 1995]. What constitutes their expertise? The following section focuses on this question.

3.1.3 Expertise as "the creative capacity to deal with the unknown and unexpected"

> *"Knowledge is a capacity to behave adaptively within an environment; it cannot be reduced to (replaced by) representations of behavior or the environment*."
>
> [Clancey, 1995 p.230]

This adaptability is the key component for expertise. In order to obtain this adaptability, experts are keen to learn, not only what is there, but most importantly, to learn "*what is not yet there*" [Engestrom, 1992] through experience. Experts are involved in a progressive problem solving process, in which they continuously refine their knowledge and methods in order to solve bigger and bigger problems where no correct answer previously existed [Bereiter and Scardamalia, 1993]. Gadamer [1972] states that experts draw knowledge from many experiences, but they never stop and never feel satisfied with what they have learnt.

> "*The truth of experience always contains an orientation towards new experience. That is why a person who is called `expert' has become such not only through experiences, but is also open to new experiences. The perfection of his experience, the perfect form of what we call "expert", does not consist in the fact that someone already knows everything and knows better than anyone else. Rather, the expert person proves to be, on the contrary, someone who is radically undogmatic; who, because of the many experiences he has had and the knowledge he draws from them is particularly equipped to have new experiences and learn from them*"
>
> [Gadamer, 1972 p.412]

In complex and continuously changing situations where no rules can be followed experts have the ability to transfer prior knowledge and skills to new situations and create new solutions. They have the ability to "*influence the rules*" [Gray, 2000]. This is the core difference between experts and routine problem solvers.

The afore mentioned definitions consider expertise as skills, process, or creative ability in dealing with the unexpected, which corresponds with three performance levels - *skills-based*, *rule-based*, and *knowledge-based* respectively [Maurino *et al.,* 1995]. Capper [2000] points out that skills-based and rule-based performance can often be carried out by individuals, although the latter is more likely to produce optimal outcomes if it involves the discussion between two or more people. However,

> *"... knowledge based performance will generally be sub optimal if engaged in by an isolated individual, regardless of the level of formal expertise or experience of the individual. Knowledge based performance can only be optimised by the use of critical inquiry and collaborative discourse in groups."*
>
> [Capper, 2000 p.157]

The fact that knowledge-based performance can only be optimised during a collaborative activity leads to the fourth kind of understanding of expertise – expertise as collaborative activity which will be discussed in the next section.

3.1.4 Expertise as collaborative activity

Instead of considering expertise in isolation, some researchers argue that expertise should be considered as a collaborative activity. Vygotsky [1978] and Leont'ev [1981] model skills development and expertise as occurring within an 'activity system' consisting of the individual, co-workers, and the workplace community. Their ideas are supported by Engestrom [1992], who argues that in the continually changing environment, the lonely, unaided and narrowly task-oriented expert appears helpless. Accordingly, expertise derives from the capacity of individuals to work collaboratively to achieve continuous innovation, learning and improvement [Hill *et. al*., 1998].

"*Communities of practice are the basis for collaborative activity. Learning takes place as groups have a need to learn and as individuals within groups increase their ability, over time, to respond to authentic problems facing the group*" [Bull *et al.* 2000]. A community of practice is a group of people who are linked together by a common ability or a shared interest, and consequently possess common practical experience, specialist information and intuitive knowledge [Enkel *et al.,* 2002]. They share information, experience and insights and are supported by various tools. Informal COPs are important for the development and sharing of expertise within organization [Jim Eales, 2003]. The best practice, insights and lessons learned are spread and reused among the members, which results in "a sharper individual learning curve and generally a higher level of knowledge" [Franz *et al.,* 2002]. The combined and new knowledge is developed by means of various activities across hierarchical and group borders. The group expertise is accumulated through extensive communication. Collaborative activities increase the transformation of information into knowledge through questioning, discussing, and sharing of information.

Based on this distributed form of expertise, the quality of a group or team can be improved because the collective activity is far more important than the contribution of any one individual [Raeithel, 1993]. In addition, knowledge creation is accelerated through the process of collaborative learning [Argyris and Schon, 1996], in which the tacit knowledge of all team members is utilized [Nonaka and Takeuchi, 1995]. Hence, importance is attached to the communicational ability which is necessary to preserve expertise [Salas *et al*., 1997].

3.1.5 A Working Definition in the Academic Context

In the academic environment, research is always associated with innovation [Langford, 2002]. Innovation includes the new use of knowledge (creating new solutions to challenging tasks or solving tasks in new ways) or invention of new knowledge (generating new theories). The value of research is limited if the work is repetition of what has been done before. Researchers in academia are interested in the tasks when there is no definite or obvious answer. As stated by Michaelis [1997], "Uncertainty" is a constant companion in the life of most scientists or academicians who are fully immersed in the conduct of research. However, the most successful researchers transform these uncertainties into a significant investigation and experiments, through which they transfer information to knowledge and effectively exploit the knowledge. Thus, it enhances their expertise, which again motivates them to conduct more innovations.

The focus of this thesis is to improve the accuracy of expertise matching so that users can quickly locate experts. Accordingly, the working definition of expertise in this thesis is "***a specialized, in-depth body of knowledge and skills in a particular academic area(s)/topic(s), and the ability to use them in creating new knowledge or apply it to new applications***". The working definition combined the first and third understandings of expertise. In this thesis experts are ranked according to their expertise level, this corresponds to the second understanding. The fourth understanding of expertise is not reflected in the working definition because the initial goal of expertise matching is to locate individuals with the required expertise. However accurate expertise matching facilitates team formation and collaboration between individuals and further initiates the generation and development of the group expertise.

3.2 Expertise Matching

In this thesis, expertise matching can be defined as "*the process of finding experts* ***with the required expertise***". Experts can be retrieved in many ways (for example, name, location, position, and so on), the difference between expertise matching and other expert finding systems is that it focuses on the expertise of experts. This section first describes the domain model of expert finding systems, and then discusses the criteria for expertise matching systems. The related work on expert finding is also reviewed based on these criteria.

3.2.1 Domain Model of Experts Finding Systems

A domain model of expert finding systems is suggested by Yimam-Seid [2003], which includes seven domain factors.

- **Basis for expertise recognition**: this is the collection of various pieces of evidence which indicates the area(s) of expertise. These evidences can be grouped as *explicit* evidences such as self-declaration by experts and *implicit* evidences such as document authorship.
- **Expertise indicator[3] extraction**: the extraction techniques can be grouped as domain knowledge independent or domain knowledge driven.
- **Expertise models**: these can be dynamically generated at query time from expertise indicator sources, or extracted and stored either by personal agents or as aggregated models to which experts are associated.
- **Query mechanisms**: the system either requires users to explicitly specify their requirements or infers expertise need based on users' communications, activities, and so on.
- **Matching operations**: matching techniques include keywords matching (exact keyword matching or similarity matching such as vector space based methods) or concept matching. Inference mechanisms can be applied to concept matching.
- **Output presentations**: Experts need to be ranked uni-dimensionally or multi-dimensionally. A varying degree of personal detail may be presented as well as their social network.
- **Adaptation and learning operations**: The system should employ user models to compare the experts' competence level with that of the user's, make user-tailored rankings, and attempt to describe expertise at a level of granularity that matches queries.

3.2.2 Criteria of Expertise Matching

Some of domain factors described above do not directly influence the performance of expertise matching. For example, *query mechanism*, asking users to explicitly specify their requirements does not guarantee the better performance than inferring expertise needs based on users activities because not all the users are good at specifying their needs. Another example is *expertise indicator extraction*, the selection of the extraction techniques depends on how

[3] Expertise indicator means terms or phrases reflecting expertise; expertise indication means the evidence of expertise such as document authorship.

expertise is represented. In the other word, this factor heavily relies on *expertise model*. In order to achieve better performance five fields are summarized based on the five remaining factors, which act as the criteria to evaluate expertise matching systems (shown in Table 3-1). The criteria are described below.

- **Multiple expertise indications**: there are many indications of expertise, such as publications, projects, homepages and so on. These expertise indications are physically distributed across the organization and stored in various formats (databases, document repositories, web sites and the like).

- **Concept searching**: Users should be able to navigate the domain concepts (in hierarchical structure) to locate experts. In addition, for the users without domain knowledge (those prefer keywords input), the system can guide them to the appropriate concept(s) based on the keywords they specify.

- **Experts ranking**: the experts should be ranked according to their level of expertise in the particular area that a user is interested in. So users can limit the number of experts they will accept.

- **Clear output presentation**: Users should be supported in the selection of experts through the provision of integrated information of experts extracted from different data sources. This means that users do not have to manually search for relevant information on each expert on their own.

- **Adaptability**: the system should be able to use the feedback of users to learn users' expertise requirements in order to achieve the improved matching performance in the new retrieval.

3.2.3 Expertise Indications and Representations

The expertise indications determine what kind of data sources are to be collected while the expertise representations determine the form that expertise is stored and also how to match expertise.

3.2.3.1 Indications of Expertise

Expertise, as one kind of tacit knowledge, has the inherent characteristic of tacit knowledge - it is difficult for people to write down their expertise, they know but unable to express. Although expertise is embodied and embedded, it can often be observed through tangible results [Stenmark, 2002]. The following are diverse indications of expertise:

- **Answers to others questions**: people ask questions in discussion forums, newsgroups, bulletin boards. When a person always answers questions on a particular topic, he/she is very likely to be an expert in that field. The "quality" of the answers, rated by the questioners, can be seen as an indication of the expertise level.

- **Email**: people use email to communicate with their associates to share information, discuss problems and get answers on a daily basis. It has become an integral part of many workers' social interaction. By tailoring email contents it is possible to get vocabulary-based hints on the person's subjects of interest and knowledge level.

- **Browsing behaviour**: If one has expertise in a particular area, he/she may spend more time on searching/reading related documents on the web. So by tracking user behaviour, especially their preferences on the web it is possible to deduce their expertise.

- **Memberships/position/reputation**: association memberships can determine the areas of interest although it is not the same as expertise. Reputation is important in social network recommendations. A high reputation/position is always supported by a high level of expertise.

- **Publications**: such as journal articles, technical reports, seminar presentations, these are all good indications of a person's expertise. Through writing documents, part of the experts' tacit knowledge (expertise) can be converted into explicit knowledge.

- **Projects**: people usually acquire valuable knowledge and experience through undertaking projects. New knowledge is created via collaboration between team members. Individual's expertise is increased through communication and sharing.

- **Recommendations**: when a person recommends documents to a community, the quality of the documents can be evaluated by other users. If a person always recommends high quality documents on a topic, this person must be very familiar with this topic and should have expertise in the area. The assumption is that experts can find more quality information than ordinary people.

3.2.3.2 Expertise Representation

It is still an open issue how to best represent a person's expertise. Basically, there are two kinds of well-used expertise representation. One is keyword-based and the other is concept-based. In the former case, expertise is represented by a set of keywords based on which an exact match (using Boolean model) or a similarity match (using vector space model or probabilistic model). In the latter case, expertise is mapped to the pre-constructed concepts (such as concepts in a domain ontology); normally, users browse the concepts to locate experts. These keywords or concepts can be manually collected by experts or can be automatically extracted through analysing the relevant expertise indications using Information Retrieval techniques, Natural Language Processing techniques, and so on.

3.2.4 Existing Approaches to Expertise Matching

3.2.4.1 Expertise Database

The traditional approach is to create a database or directory of skills or expertise, also called "yellow pages". Individuals specify their expertise and the levels of their expertise in their own words or according to the pre-defined subjects; users are then able to search for experts based on this kind of database. This kind of skills databases is very popular in a wide range of organisations [Scarbrough, 1999]. However, it suffers from several limitations: (1) different people may describe similar expertise differently; (2) people have very different standards for judging the degree of their expertise; and (3) people's expertise is continually changing, and thus the skills database may be out of date quickly and be difficult to maintain. Compared with the traditional skills database, some intelligent systems have been developed where different expertise indications are used to identify experts, which will be described in the next section.

3.2.4.2 Information Repository

The problem of the static skills database is partially solved in a dynamic information repository such as Answer Garden [Ackerman and McDonald, 1996] which continually collects the answers to frequently asked questions. Expertise is implied in the answers and users might find the information they need in the database. If they cannot find the answer, the question is then sent to the appropriate expert, and the new answer together with the question is inserted into the answer garden.

An information repository approach does not focus on finding experts, but on reusing their codified expertise by storing and retrieving answers. The interaction between users and experts is not encouraged, so the expertise is not effectively reused and explored. In contrast, the ContactFinder system [Krulwich and Burkey, 1996] recommends the appropriate people by scanning and analysing messages in bulletin boards. It extracts topic areas from the messages based on heuristic keywords and associates the contact person to these topic areas. If a new message appears, the system can assist users by recommending an appropriate contact person who has answered questions in the same topic before. The limitation of this system is that the method of extracting a topic from each message (heuristic approach, such as finding words in upper case) is not accurate and needs to be improved.

Browsing Trail[4]:

Some systems consider people's web browsing patterns as indications of expertise. One example is the MEMOIR system [Pikrakis *et al.,* 1998], which searches not only people's homepages but also the URLs and associated keywords of webpages that people have visited in order to help users to find experts. Although the MEMOIR system can find people with similar interests, it cannot distinguish between those who have expertise and those who have interests only.

Expertise Browser [Cohen *et al.,* 1998] traces experts' browsing and searching behaviour in order to provide hints to users who are searching in similar areas. The assumption is that experts have abilities to find high quality information, and their expertise in the information filtering area can be reused by others through storing their information-browsing paths and patterns of content. However, the same information may have a different value for different people, and some good quality documents read by experts might be too difficult to understand by others. So users may not find what they need even when they follow experts' browsing paths. Another limitation of this system is that experts need to be pre-specified. This work has to be done manually and regularly in order to keep the experts database updated. Furthermore, if users do not know the experts themselves, then they have to scan through all the browse paths that match the query, which can be time consuming.

3.2.4.3 Keyword based Profile Searching

The above systems are based on keyword index, and do not have any expertise/interest profile about each expert. Expertise profile is necessary to conduct a more accurate match and to rank experts. The profile can be keyword-based (such as a set of keyword with different weights) or created through text analyses of different indications such as email, and work artefacts.

Emails

Yenta [Foner, 1997] is a multi-agent, referral-based matchmaking system. It functions in a decentralized fashion where every person has a personal agent which stores the interest profile

[4] Trail: A user's trail is the set of actions on documents that they have visited (such as opening the document) in pursuing a certain task.

of the person, and agents communicate with each other to find people with similar interests and introduce them to each other. By scanning all the users' emails, each user's interests profile (a set of weighted keywords) is created using a vector space model. Similar messages are clustered together by comparing these keyword vectors. Each cluster represents one interest of a user. Two users are considered to have a similar interest if they have at least one cluster similar to each other (the similarity between these two clusters has to be above a certain threshold).

Whilst Yenta only concentrates on grouping people based on their shared interests, a similar system called Expertise Locator [Kautz *et al.,* 1996] can further locate experts based on their emails. A user profile is a list of keywords that appear in any email message. Experts are ordered simply according to how frequently the keywords are mentioned in the email correspondence. Know-who email agent [Kanfer *et al.,* 1997] improved the Expertise Locator by adapting document retrieval methods in three ways: (i) people are represented in a vector space; (ii) relevance feedback is implemented to help user reformulate the queries; and (iii) the set of terms included in a query or person vectors are referred to an online dictionary in order to find the semantic relationships amongst the words. However, email is not a good indication to reflect people' expertise and scanning people's email involves the privacy problem.

Documents

Expert Finder (1) [Mattox *et. al.,* 1999] exploits organization's intranet documents to locate experts. The system ranks employees by the number of times a term or phrase is mentioned and its statistical association with the employee name either in corporate communications (such as newsletters or based on what they have published in their resume) or document folder. This system creates people's expertise profiles during the query time. Although it can capture the updated information and avoid some maintenance work, the system suffers from a high latency problem in query processing. The shortcoming of the query-time generated expertise model is also found by Yimam-Seid [2003] in a similar expert finding system developed for a research department.

An agent based Expertise Finder [Crowder *et al.,* 2002] is built for an academic research environment. It receives a user's query in keywords and retrieves publication repository to find all the publications which use the search terms. The associated authors are then listed as the relevant experts. They are ranked according to the number of occurrences each author appears in the returned publications. It suffers the same problem as in Expert Finder (1), furthermore, it only explores one type of expertise indication (publications).

Social network

McDonald and Ackerman (1998) distinguish two steps in finding expertise within organizations through the field study of expertise location in a software company. These two steps are expertise identification and expertise selection. Social networks are an important factor in the expertise selection process. There are systems which take into account social networks when recommending experts. One of them is ReferralWeb [Kautz *et al*., 1997a; 1997b], which aids users in finding "trusted" experts based on a "referral chain". The indicators of the social network between people include co-authorship on papers and team members in past projects. Furthermore, spiders were built to determine relatedness based on frequency of co-occurrence of names in the entire WWW. A social network is modelled by a graph, where nodes represent individuals, and the edges between nodes indicate that a direct relationship between the individuals has been detected. In addition to relationship, ReferralWeb also extracts evidence for expertise. The expertise database includes all the papers written by the individuals. The standard information retrieval vector space model is used to search for people with special expertise. Hence, users can find experts on a particular topic and those who have pre-existing social relationships with them.

Another example of using social network is the Expertise Recommender system [McDonald and Ackerman, 2000], which uses various heuristics to select an expert in a software company. Expertise identification is based on software change history and technical support database. A change history profile includes module name, version and date. The list order of experts is from those who touch the software most recently to who touch the software least recently. For technical support, the request text is parsed and three query vectors are created: one for symptoms; one for customers; and one for program modules. The profile database is then queried using the vector space index. Expertise selection is based on the organizational distance between the department of the person making the request and the department of each expert recommended, and how well the requester knows the expert (social network). Report on a system that uses various heuristics to select an expert, based on who has touched various files, who is 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 idea is to produce a very short list of recommended experts based on heuristics specified by the user. If no satisfactory expert is identified, the user can "escalate" the request, and the system will produce more potential experts, for example, by changing the threshold values in the heuristics.

Limitations of all these keywords-based profile systems are that (1) the search is based on syntax (if the keywords appear) rather than concept; and (2) experts in the result list are only sorted with respect to the given search terms. These limitations can be overcome in concept-based searching which will be described in the next section.

3.2.4.4 Concept-based search

Expert Recommendation [Yukawa and Kashara*,* 2001] is a system which locates engineers with a high level of expertise on a particular topic. The information source is a huge set of technical documents produced by experts. Again the vector space model technique is used to analyse these documents. Keywords and documents are mapped in the same multi-dimensional space through co-occurrence based thesaurus or dictionary-based concept base. Each personal profile is derived from associated documents and is represented as a weighted vector. In this system the keywords, the target documents, the authors of the documents and their organizations are all placed as vectors in the same multi-dimensional space, and the similarity between any two can be calculated as a cosine coefficient between vectors. The advantage of this technique is its flexibility, that is, it can accept not only a keyword but also sentences or even documents as a query and allows analysis and clustering of the results. However, there are issues remaining such as quantity and quality of the documents and multiple expertise.

The Expert Finder (2) [Vivacqua and Lieberman, 2000] agent assists novices in finding experts in the domain of Java programming. A user profile is automatically generated by a personal agent through scanning his or her Java programs. These files are parsed and analysed to find how many times the methods and libraries used. A profile contains a list of the user's areas of expertise and the associated levels (novice - beginner - intermediate - advanced - expert). Expertise level is determined by taking the number of times the user uses each class and divided by the overall class usage. Users can hide some areas of their expertise if they do not want others to know about them. The similarities between a user's profile and other experts' profiles are calculated using vector match by a matchmaking engine. The Java domain similarity model, which defines the features in the Java programming language and class libraries, is also exploited by a matchmaking engine to find a candidate expert whose knowledge lies in a more general or more specific category or related topic to the user's requirements.

Some recent commercial knowledge management systems such as Agentware Knowledge Server TM (from Autonomy http://www.autonomy.com) also provides features that support expertise matching in organizations. Agentware uses neural networks and advanced pattern-matching techniques to find the concept(s) of the documents that employees have accessed and then deduce their expertise. One system built on Autonomy's AgentWare platform to search expertise of others is the Volvo Information Portal (VIP) [Lindgren and Stenmark, 2002]. The Find Competence feature in the VIP was built to locate organizational members with a specific expertise through detecting their actions, such as searching for information related to a specified area. However, users' actions indicate more of one's interest rather than expertise, there can be a gap between users' interest and their competence.

Liao *et al.* [1999] propose a Competence Knowledge Base System (CKBS) which builds upon an ontology-based model of competence fields. In this approach the employees' competences are associated with the concepts in a domain ontology. Ontology-based retrieval heuristics are used to find experts who are indirectly linked to the search concept. These experts include people who have worked on a project applying the technology required or who have competence in the super- or sub-concept of the topic in question.

3.2.5 Summary

According to the criteria set in Section 3.2.2, the above expertise matching systems have been summarised in Table 3-1. Table 3-1 shows that challenges remain for dealing with expertise matching according to the criteria.

Multiple Expertise Indications: From Table 3-1 it can be seen that different systems use different data sources which include different expertise indications. There is no clear priority for these indicators; they complementary one another. Hence, an expertise matching system should include as many expertise indications as possible. A normal situation is that expertise indications are physically distributed across the organization and stored in various formats (databases, document repositories, web sites, and the like). In most systems only one type of indication (such as email) and/or only one data format is used to create an expertise model. In order to achieve a more accurate expertise model, there is a need to exploit the heterogeneity and the distributed nature of the information space as a source of expertise indications. This is called "source heterogeneity gap" [Yimam-Seid, 2003].

Concept search: Keyword search is still widely used in these systems. Few systems implement concept search with different approaches (dictionary-based, pattern matching, ontology-based). Among these approaches, ontology-based approach is widely accepted as the preferred method to deal with the problems of keyword searching. However, manually linking people or projects with the concepts in the ontology is still time consuming.

Experts Ranking: Nearly half of the systems return experts in a relevant order based on the vector model, although few of them use number of mentions of terms which a user specified (Expert Finder(1) [Mattox *et al.,* 1999]) or from the most recent touch to the least recent touch of the software (Expertise Recommender [McDonald and Ackerman, 2000]). If expertise is only represented by concepts such as CKBS, then it is difficult to rank experts. There is a need to combine keywords representation and concepts representation in order to address the problem of keyword searching and the ranking of experts.

Table 3-1 The comparison of the relevant works based on the criteria of expertise matching (see Section 3.2.2)

Systems	Criteria for Expertise Finding Systems				
	Expertise Indications used	**Expertise Representation**	**Expertise Model**	**Experts Ranked**	**Experts information provided**
Answer Garden	FAQ	Keyword /Databases	Query time generated	No	Name
Contact Finder	Messages in bulletin boards	Topic	Aggregate	No	Name, knowledge areas, contact information, previous emails
MEMOIR	Browsing trails	Keyword	Agent-based	Yes	Name, trails
Expertise Browser	Browsing trails	Keyword	Aggregate	No	Name, browse paths
Yenta	Email	Keyword	Agent-based	No	Name
Expertise locator	Email	Keyword	Agent-based	Yes	Name
Know-who	Email	Keyword	Agent-based	Yes	Name
Expert Finder(1)	Documents	Keyword	Query time generated	Yes	Name, contact information, documents
Referral Web	Papers and social network	Keyword	Aggregate	No	Name, social network
Expertise Recommender	Social network and technical reports	Keyword	Aggregate	Yes	Name, email, phone
Expertise Finder	Publications	Keyword	Query time generated	Yes	Name, email, phone, position, publications
VKP/ULPD	Publications and projects	Keyword, Research areas	Query time generated	No	Name, phone, email, projects, publications, classification terms
Expert Recommendation	Technical documents	Concept	Aggregate	Yes	Name, characterizing words, sentences, bibliography
Expert Finder(2)	Java Programming	Areas in domain ontology	Agent-based	Yes	Name, area of expertise, level of expertise
Find Competence	Documents	Concept	Aggregate	No	Name, email, company, phone, dept
CKBS	Self-described skills, projects	Concept	Aggregate	No	Name, phone, email, url, projects, competences
Expert Locator	Self-described skills	Technical thesaurus	Aggregate	No	Name

Output Presentation: Output presentation is very weak in these existing approaches. Normally, only personal contact information is provided. Ranking is not very useful for the users if there is no detailed information on each expert. It is difficult for users to agree or disagree on this order. The nature of expertise means that it is difficult for each person to accurately declare his expertise and level, and there is always a degree of "noise" in the ranked list of experts returned by the system based on the limited data sources available. After all it is users who select the appropriate experts. Thus there is a need to access detailed information about experts, which includes not only the personal contact information (such as organization, group, telephone number) but more importantly, their expertise indicator sources (such as homepages, publications, projects). Some systems provide the social network of each expert and the documents[5] they have produced, however, this is not sufficient. Yimam-Seid [2003] also noticed this problem and named it the "expertise analysis support gap".

Adaptability: This feature is not included in the Table 3-1 as very few systems (such as Know Who [Kanfer *et al.,* 1997]) possess this feature. Expertise matching seems similar to document matching in which user feedback can be collected and used to adapt users' profiles. However, expertise itself is more complex than single documents because of the intricacies of expertise, therefore the effectiveness of adaptability is less than with single documents.

3.3 Conclusions

The nature of expertise means that expertise itself cannot be simply expressed even by experts themselves and it cannot be quantified in the same way as data or documents. This makes expertise matching more difficult than searching documents. Through analysing existing approaches, it can be seen that the vector space model is a widely used technique for building keywords based expertise profile. Chapter 4 discusses how this technique could be adopted for the ULPD Expertise Matcher. Some serious shortcomings in the existing approaches are also discovered. Firstly, expertise indications are not well explored. Secondly, the output presentation is not sufficient. Thirdly, the combination of advantage of concept matching and

[5] In most time, only titles of the documents are displayed.

keywords matching is yet to be realised. Chapter 5 discusses how to employ semantic web technologies to solve these limitations.

Chapter 4

Extension to ULPD Expertise Matcher

The previous chapter described the nature and value of expertise, and existing matching approaches in general. This chapter focus on expertise matching in academia. Through the survey of expertise matching in academia (Appendix A), it was found that the ULPD expertise matcher in the University of Leeds is a representative approach. This chapter describes the ULPD Expertise Matcher, which explores the two major expertise indications (publications and projects[1]). The limitations of current approach to expertise matching is analysed followed by a brief introduction of three information retrieval models. An extended Expertise Matcher is then presented which adopts one of the information retrieval models, namely, the vector space model to build keyword-based expertise profile. A prototype system is then built to compare the retrieval performance between the current Expertise Matcher and the extended system. An evaluation experiment has been carried out with real user participation and the results are presented. Finally, areas which still need to be improved are discussed.

4.1 Current Approach of Expertise Matching in the ULPD

The ULPD system has been developed at the University of Leeds to better manage the expertise of staff in the University (The ULPD data model is shown in Appendix C). One of the aims is to facilitate collaboration between the University and industry through locating University experts with the required expertise to solve a particular set of industrial problems.

[1] Academic researchers acquire experience by working on projects, through which the researchers accumulate the abilities to solve problems using their knowledge. On the other hand, publications reflect researchers' insight, understanding of knowledge, and their contribution in terms of theory or applications. Hence publications and projects are two major sources from which to derive a person's expertise in the academic environment.

4.1.1 Data Collection in the ULPD

The ULPD contains a set of data which allow it to operate efficiently as a publications database. These data sets include:

Publication: an efficient procedure has been developed to facilitate the upload of any existing departmental publications databases into the ULPD given completion of the necessary data re-formatting. Validation and maintenance of ULPD publications data are undertaken by a Library Administrator using dedicated resource. The duplicate publication records can be checked. This is based on the "Publication Title" and "Publication Type" (e.g. chapter in book, journal article, etc.); authors (or administrators) are then free to continue with inputting the new publication record or can quit the data entry at that point.

Journals: the ULPD contains information such as journal title, ISSN, and publisher for approximately 65,000 academic journals. This data was originally downloaded from the ULRICH'S periodicals directory provided by Bowker and has been extended by the ULPD support team whenever users have requested another academic journal be added which did not already appear in the ULPD. This journals data is the responsibility of the ULPD Library Administrator.

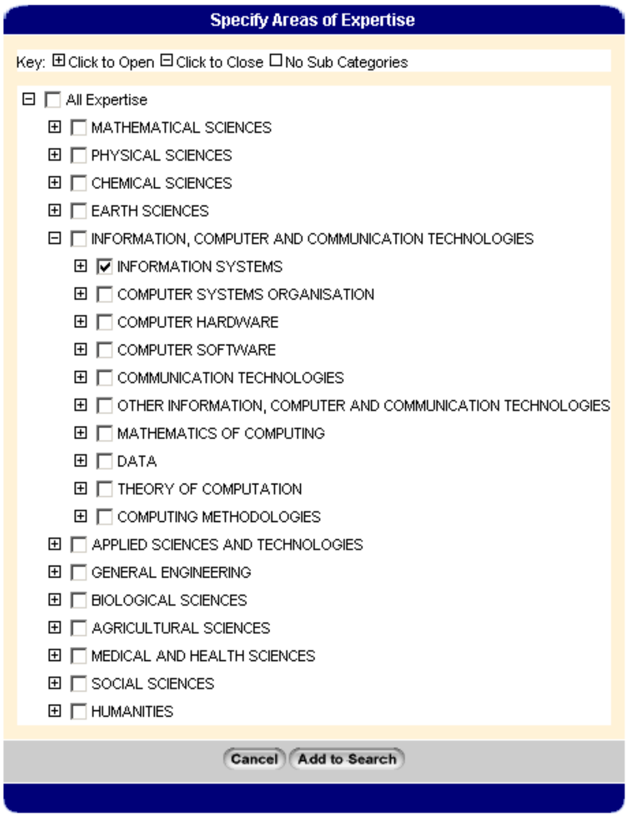
People: the lists of people come from a range of different sources. Information about academic staff, academic related staff, former staff members, and technical staff come from the University's central SAP system. Information about current research students and former research students come from the University's central **S**tudent **I**nformation **M**anagement **S**ystem (SIMS) system. The data which is taken from the University's SAP and SIMS systems is fed into the ULPD on a daily basis and this data cannot be edited by ULPD users.

Project: this data comes from the University's **O**n-line **S**ystem for the **C**omputerised **A**dministration of **R**esearch (OSCAR) and comprises details of research. Each research project in the ULPD includes details of project title, investigators, project start and end date, awarded value, and account number. This data which is taken from the University's OSCAR system is fed into the ULPD on a daily basis. Project details such as investigators and abstract can be edited by the departmental administrator whilst other fields such as project title cannot be edited.

A standard (academic) user has the ability to view and edit their own publications data only, that is, any publication records for which they are an author or co-author. They can also use the *Profiles* facility to add information to their own profile in the ULPD and export this as a report (for example as a Curriculum Vitae) or view it on-line as a web page. A Departmental Administrator user has the ability to view and edit publications data for members of any department to which they have been assigned administrator access rights.

4.1.2 ULPD Expertise Matcher

In ULPD there is an Expertise Matcher to help users search for experts with the required expertise. Each expert's expertise is derived from the associated publications and projects. All these publications and projects were classified according to the ULPD classification scheme. Experts can be found in two ways - by browsing fields of research classification terms (as shown in Figure 4-1) or by searching keyword. Firstly, users can navigate the ULPD classification scheme and select a particular classification term or topic, for example, *information systems*. The Expertise Matcher will then retrieve experts who have published papers or have worked on projects classified under the selected field of research. Secondly, users can also enter keywords, for example, "*multimedia and networking*" and the Expertise Matcher will retrieve people who have published papers or have worked on projects with titles and/or abstracts including these keywords. Boolean operations are supported, so that users can use Boolean operators (AND, OR, and NOT) to combine the search keywords. This search is implemented using Microsoft SQL Server 2000 through which the publication table, project table, and personal profile table are indexed.



Specify Areas of Expertise

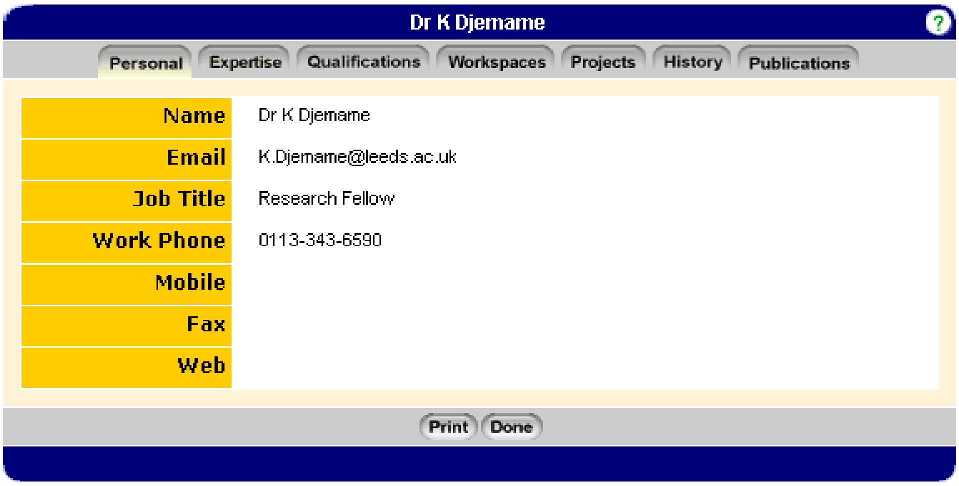
Key: ⊞ Click to Open ⊟ Click to Close ☐ No Sub Categories

- ⊟ ☐ All Expertise
  - ⊞ ☐ MATHEMATICAL SCIENCES
  - ⊞ ☐ PHYSICAL SCIENCES
  - ⊞ ☐ CHEMICAL SCIENCES
  - ⊞ ☐ EARTH SCIENCES
  - ⊟ ☐ INFORMATION, COMPUTER AND COMMUNICATION TECHNOLOGIES
    - ⊞ ☑ INFORMATION SYSTEMS
    - ⊞ ☐ COMPUTER SYSTEMS ORGANISATION
    - ⊞ ☐ COMPUTER HARDWARE
    - ⊞ ☐ COMPUTER SOFTWARE
    - ⊞ ☐ COMMUNICATION TECHNOLOGIES
    - ⊞ ☐ OTHER INFORMATION, COMPUTER AND COMMUNICATION TECHNOLOGIES
    - ⊞ ☐ MATHEMATICS OF COMPUTING
    - ⊞ ☐ DATA
    - ⊞ ☐ THEORY OF COMPUTATION
    - ⊞ ☐ COMPUTING METHODOLOGIES
  - ⊞ ☐ APPLIED SCIENCES AND TECHNOLOGIES
  - ⊞ ☐ GENERAL ENGINEERING
  - ⊞ ☐ BIOLOGICAL SCIENCES
  - ⊞ ☐ AGRICULTURAL SCIENCES
  - ⊞ ☐ MEDICAL AND HEALTH SCIENCES
  - ⊞ ☐ SOCIAL SCIENCES
  - ⊞ ☐ HUMANITIES

Figure 4-1 Selection of research field(s) from classification

The search results consists of a list of experts' names. In order to select the relevant expert, users can click on an individual's name to view a personal profile as shown in Figure 4-2, which includes the following information:

- *Personal details*: title, initials, surname, e-mail, extension, qualifications, research interests, membership of research groups, membership of committees and associations, current position and previous position(s), homepage, language skills.
- *Areas of expertise*: a collection of classification terms under which each expert's publications and projects are classified.
- *Project details*: project title, project abstract, start date, end date, other project investigators, sponsor(s), project value.
- *Publication details*: publication title, publication abstract, year of publication, other authors, etc (publication status, published in, pagination, confidentiality, editor, keywords).



Name	Dr K Djemame
Email	K.Djemame@leeds.ac.uk
Job Title	Research Fellow
Work Phone	0113-343-6590
Mobile	
Fax	
Web	

Figure 4-2 Example of a personal profile

4.1.3 Limitations of the Current Expertise Matcher

Both browsing and searching activity have their limitations. Browsing can help users find all the experts in a particular field of research without query formulation. However, it is only suitable for users who are familiar with the ULPD classification system. Furthermore, whether users can find the experts also depends on the administrators of each department, who associate the publications and projects with the classification terms. It is difficult to correctly classify every publication and project, and if there are some incorrect classifications, the relevant experts may not be retrieved and/or irrelevant experts may be returned.

In contrast, the search function is very helpful for those users who are not familiar with the classification. It is always quicker to get the results by entering several keywords than browsing. However, it suffers from the following drawbacks:

- Some irrelevant people are retrieved and some relevant experts are missed.
- Too many experts are retrieved, or too few. There may be dozens if a user inputs general terms in the query, or there may be none at all if a user uses several specific terms together. In the former case, it is difficult for users to check each person on the list to find the real experts. In the latter case, users have to reformulate the query in some way in order to find an expert. However, users do not know which term should be removed or

changed, and they may formulate the query several times before obtaining a more useful list of experts.

- Experts are listed according to alphabetical order rather than expertise level. There is no mechanism by which the results may be ranked in order of decreasing expertise level. Retrieved experts are equal to each other, and users have to check each of them in order to find the most appropriate expert.
- It is difficult for users to express their request using Boolean query. Users may have a rough idea of the expertise they are looking for, but cannot formulate a precise query.

The reasons for these problems can be summarised below.

- **Limitation of expertise data sources** The information stored in ULPD for expertise retrieval is limited and incomplete. Although publications and projects are important expertise indicators, not all the publications are stored, for example, technical reports are not included in the ULPD database. For the stored publications and projects information, only less then 10% have abstract information; the others include titles only. In addition, due to the fact that the manual collection of personal information is a tedious and time-consuming task, many of the fields in the person table are still left blank, such as *qualifications*, *research interest*, *URL of homepage* and so on. Some of them are important for expertise retrieval, such as *research interest*. The limited expertise data sources hinder the retrieval of relevant experts.

- **Lack of expertise profile** There is no pre-stored expertise profile. Expertise is derived at the time of the query. Although publication and project tables are indexed which reduces the searching time to some extent, some experts are overlooked if keywords given by the user appear in different publication titles. For example, if a user inputs "*A* AND *B*" in the query, and an expert uses keyword *A* in one publication and keyword *B* in another respectively, then this expert is not retrieved.

- **The Boolean search is conducted** Therefore experts will be retrieved if and only if their publications or projects information include the keywords that a user specifies in the query. All the experts are considered to be equally relevant to the query and have to be listed according to alphabetical order of their surname. In addition, users cannot have control over the size of the output produced by a particular query. The Expertise Matcher is unable to predict a priori how many experts will be returned to the user.

The following section will describe how information retrieval techniques are employed to reduce the problems.

4.2 Adaptation of Information Retrieval Techniques

Rather than querying data in a standard format, information retrieval can work with plain unformatted data. A fundamental idea within IR is that a document is relevant to a query if they are similar, which can be defined as string matching, similar vocabulary or same meaning of text [Monz and de Rijke, 2001]. There are many information retrieval models such as set theoretic models, algebraic models and probabilistic models [Baeza-Yates and Ribeiro-Neto, 1999]. Each of these models has its own advantages and disadvantages. The classical information retrieval models include the Boolean model, the vector space model, and the probabilistic model. They are classical models, not only because they were introduced in the early 70's, but also because they represent three classical problems of information retrieval respectively: structured queries; initial term weighting; and relevance feedback [Hiemstra, 2000]. This section introduces these three classic models and explains why the vector space model is chosen in building an expertise profile.

4.2.1 Information Retrieval Models

4.2.1.1 Boolean Model

The Boolean model is based on set theory and Boolean algebra. The Boolean model represents documents by a set of index terms, the value of an index term is "1" if this term appears in a document, otherwise, the value is “0”. A query is also specified as Boolean expressions, which is composed of index terms linked by the standard logical operators: AND, OR, and NOT. The query expression is then represented as a disjunction of conjunctive vectors. A document is predicted as relevant if it satisfies the query expression, that is, if it includes any of the conjunctive components.

The major drawback of the Boolean model is that a document is predicted to be either relevant or non-relevant without any notion of a partial match, which prevents good retrieval performance. The Boolean model is in reality much more a data retrieval model (the difference between data retrieval and information retrieval can be found in Appendix B). Thus, the data specified by the user is important. However, sometimes it is difficult to translate an information need into a Boolean expression. Another disadvantage of the Boolean model is that exact matching may lead to retrieval of too few or too many documents.

4.2.1.2 Vector Space Model

The vector space model [Salton *et al.,* 1975] realizes a partial match through associating weights with each index term appearing in the query and in each document. In the vector space model, non-binary weights are assigned to index terms. As shown in Table 4-1, documents are represented as n-dimensional vectors (n is the total number of index terms). User queries can be similarly mapped into the vector space. The similarity between a document d_j and a user q can be quantified by the cosine of the angle between these two vectors (see equation 4-1). The retrieved documents can be sorted in decreasing order of relevance which leads to more precise results than that of Boolean model.

Table 4-1 Documents are represented as an n-dimensional vector with different weights on each dimension

Document term	d_1	d_2	...	d_j	...	d_m
t_1	w_{11}	w_{12}	…	…	…	w_{1m}
t_2	w_{21}	w_{22}	…	…	…	w_{2m}
...	…	…	…	…	…	…
t_i	…	…	…	w_{ij}	…	…
...	…	…	…	…	…	…
t_n	w_{n1}	w_{n2}	…	…	…	w_{nm}

$$sim(d_j,q) = \frac{\vec{d_j} \bullet \vec{q}}{|\vec{d_j}| \times |\vec{q}|} = \frac{\sum_{i=1}^{n} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{n} w_{i,j}^2}\sqrt{\sum_{i=1}^{n} w_{i,q}^2}} \qquad (4\text{-}1)$$

where w_{ij} is the weight of the *i*th term in the document d_j, and
w_{iq} is the weight of the *i*th term in the query q.

Various methods for weighting index terms have been developed [Salton and Buckley, 1988]. Here two methods are introduced. The first simplistic method is based on Term Frequency (TF), which is the number of times a given term occurs in a document. Words that repeat multiple times in a document are considered salient. The second method is based on Term Frequency and Inverse Document Frequency (TF × IDF) based on the premise if a word appears in many documents, it is a common word and not very indicative representation of document content. IDF, proposed by Sparck-Jones [1972], is an appropriate indicator of how well a term distinguishes a relevant document from a non-relevant one. It measures the proportion of documents over the entire collection that contain a given term. TF and IDF represent intra-cluster similarity[2] and inter-cluster dissimilarity respectively.

In the BT KSE system [Davies *et al.,* 1998], the vector space model is used to retrieve documents relevant to a user's interest. First, each document is represented by an n-dimensional vector of terms. The weight of a term in a document matrix is calculated by its term frequency. A user profile is also a vector with term weight "1" if this user specifies this term to be his/her interest, otherwise the term weight is "0". The similarity between a document and a user profile is then calculated as the cosine product of the two associated vectors.

$$sim(d_j,p) = \frac{\vec{d_j} \bullet \vec{p}}{|\vec{d_j}| \times |\vec{p}|} = \frac{\sum_{i=1}^{n} w_{i,d} \times w_{i,p}}{\sqrt{\sum_{i=1}^{n} w_{i,d}^2}\sqrt{\sum_{i=1}^{n} w_{i,p}^2}} \qquad (4\text{-}2)$$

where $w_{i,p}$ is the weight of the *i*th term in the profile p,
and $w_{i,d}$ is the weight of the *i*th term in a document d.

The vector space model is a popular retrieval model nowadays. The main advantages of the vector space model are: (1) its term-weighting scheme improves retrieval performance; (2) its

[2] Here, the relevant documents are considered as a cluster, the non-relevant documents are considered as another cluster.

partial matching strategy allows retrieval of documents that approximate the query conditions; and, (3) its cosine ranking formula sorts the documents according to their degree of similarity to the query.

4.2.1.3 Probabilistic Model

The probabilistic model [Robertson, 1977] assumes that there is an ideal answer set which contains exactly the relevant documents to a given query and no other. The querying process is considered as a process of specifying the properties of an ideal answer set. Index terms are used to characterize these properties. The probabilistic model attempts to predict the probability that a given document will be relevant to a given query according to the terms included in this document, and the probability that these terms are present in a document randomly selected from the ideal set.

The probabilistic model improves on the Boolean model in that documents can be ranked in decreasing order of their probability of being relevant. However, it usually needs users assistance in the initial separation of documents into relevant and non-relevant sets. Furthermore, the term frequency in a document is not taken into account because all weights are binary.

Both the vector space model and the probabilistic model support natural language queries because they treat documents and queries in the same way. The results can be ranked using both models and relevance feedback can be supported. The major difference is that the vector space model assumes relevance and the probabilistic model relies on relevance judgements or estimates.

4.2.1.4 Selection of the Vector Space Model

The current ULPD Expertise Matcher uses an exact match, which suffers the same problem as in the Boolean model. It is likely that it could be improved by the vector space model or the probabilistic model. Salton and Buckley [1990] found that the vector space model is in general more effective than the probabilistic model. Despite its simplicity, the vector space model is a resilient ranking strategy with general collections. In general, the vector space model is either superior or almost as good as a large variety of alternative ranking methods [Baeze-Yates and Ribeiro-Neto, 1999]. The results are difficult to improve without query expansion or user relevance feedback. Furthermore, it is easy to compute and therefore fast. For these reasons, the vector space model is chosen to improve the current Expertise Matcher.

Some of the limitations of the current Expertise Matcher (described in Section 4.1.3) could be solved if an expertise profile is created using the vector space model.

- Pre-stored expertise profile can integrate publications or projects information so that there is more chance to find the experts even if the keywords that a user specifies appear in different titles of publications or projects.

- Based on the vector space model, the similarity between the expertise profile and a user query can be calculated. Therefore, the experts can be ranked according to the similarity degree and users can have control on the number of the experts returned.

- The measure of similarity between experts provides mechanisms to find otherwise missed experts. For example, two experts are doing similar research, and only one expert is retrieved because his publications include the keyword(s) specified by the user. Another one is missed. Using the vector space model, the experts with similar expertise can be found even when their publications or projects information do not include the specified keywords.

- It is possible to use “query refining” or expanding so that the new search will return more relevant experts. Initially, the user profile is a set of keywords specified by the user. It could perhaps be improved by adding the extracted keywords from experts which the user finds relevant and adjusting the associated weights. The new profile can be used to retrieve more experts and the user will then evaluate again. The process will be repeated until the user profile no longer changes drastically. For the long-term success, the construction of accurate user profiles is necessary.

- One additional benefit is that users can give emphasis on some keywords by giving different weights to different keywords if the vector space model is employed. Currently all terms in a query are considered as equally important.

4.2.2 Extending Expertise Matcher with the Vector Space Model

To find experts, the vector space model is used to map not only keywords and documents but also people who have written documents and worked on projects into the identical multi-dimensional space. The expertise profile of an expert *e* is represented as:

$$p_e = \frac{\sum_{j \in Pub(e) U \Pr oj(e)} D_j}{n} \tag{4-3}$$

where D_j refers to a vector for a document *j*, *Pub(e)* refers to a set of publications written by the expert *e*, *Proj(e)* refers to a set of projects that expert *e* has worked on, and *n* is the total number of publications and projects for an expert. Since the expertise profile, document and a query (combination of keywords) are treated in the same way, the system acquires an ability to discern the relevance between any combination of keywords, documents and people. The implementation consists of six processes as follows:

- **Lexical analysis** Lexical analysis of the text with the objective of treating digits, punctuation marks, and the case of letters.
- **Stopwords removal** Elimination of stopwords[3] with the objective of filtering out words with very low discrimination values for retrieval purposes.
- **Stemming** Stemming of the remaining words with the objective of removing affixes (i.e. prefixes and suffixes) and allowing the retrieval of documents containing syntactic

[3] Stopwords are listed in http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words

variations of query terms (e.g. connect, connecting, connected, etc) according to Porter's stemming algorithm [Porter, 1980].

- **Term indexing** Indexing the remaining terms according to alphabet order.
- **Term weighting** Each publication and project title is considered as a document. Each document is then represented as a vector consisting of a set of index terms. The weights of the terms are calculated using TF and TF-IDF respectively.
- **Expertise profile building** All the publications and projects associated with each expert are collected. The represented vectors are used to calculate the expertise profile according to formula 4-3.

When a query is submitted, it is also represented as an n-dimensional vector. The weights in the query vector are calculated using TF and TF-IDF respectively, and the similarity between an expert's expertise profile and a query is determined by the cosine of the angle between these two vectors according to formula 4-1.

4.3 Implementation

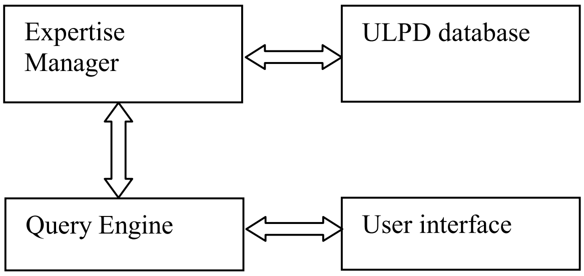


Figure 4-3 Architecture of extended Expertise Matcher

The architecture of the extended Expertise Matcher is shown in Figure 4-3, which consists of 4 components described below.

- User interface: receives the query from the user and sends the results of the ranked experts to the user.
- Query Engine: creates the user's profile using vector space model; retrieves and ranks the experts based on the similarity between user's profile and expertise profile.

- Expertise Manager: queries the ULPD database to obtain the relevant information of each expert; creates and maintains the expertise profiles.
- ULPD database: provides expertise information of each expert such as publications and projects.

The extended Expertise Matcher is implemented in Java and communicated with the ULPD system using JDBC. Since the dimensionality of keyword vectors is very large but most expertise profiles and queries do not contain most words, the vectors are sparse. The keyword-to-expert index is therefore implemented using a hash table (see below).

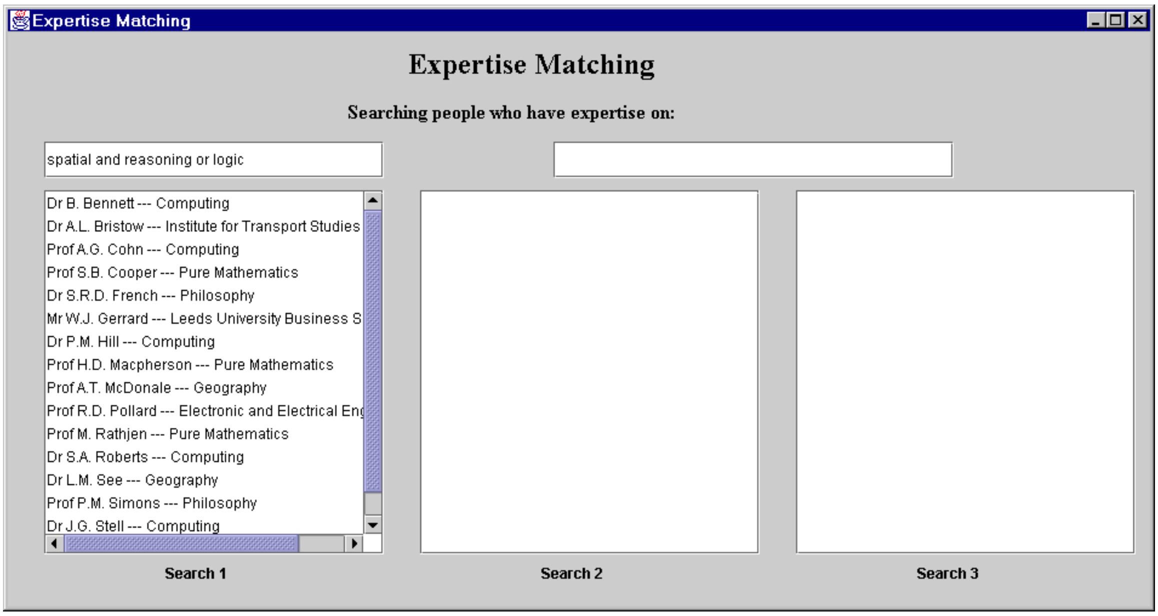
```
Create an empty HashMap, H;
For each expert, E, (i.e. retrieved relevant information from the ULPD database);
        Create a HashMap Vector, V, for E;
        For each (non-zero)token, T, in V;
                If T is not already in H, create an empty
                        TokenInfo for T and insert it into H;
                Create a TokenOccurrence for T in E and
                        add it to the occList in the TokenInfor for T;
Compute TF*IDF for all tokens in H;
Compute vector lengths for all experts in H.
```

4.4 Evaluation

4.4.1 System walk through

In order to test the retrieval performance after extending the current Expertise Matcher with the vector space model (both TF and TF-IDF strategies), a prototype system has been built and used to locate experts with the required expertise. In this prototype system, the current Expertise Matcher is called Search 1, the extended Expertise Matcher with TF strategy is called Search 2, and the extended Expertise Matcher with TF-IDF strategy is called Search 3. The testing process is as follows.

For testing the Search 1, the user inputs a few keywords (for example "*spatial and reasoning or logic*") to express his/her required expertise and links these keywords in Boolean operators (AND, OR and NOT). A list of names of experts is then displayed as shown in Figure 4-4. This is the result of Search 1. The experts are listed in alphabetical order.



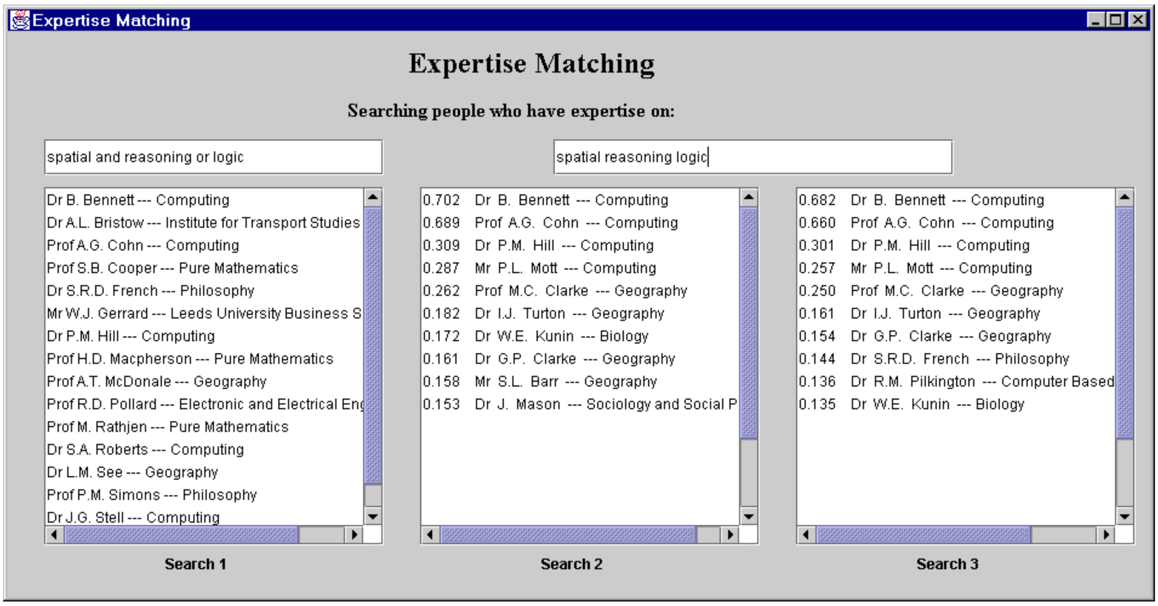
Expertise Matching

Searching people who have expertise on:

spatial and reasoning or logic

Dr B. Bennett --- Computing
Dr A.L. Bristow --- Institute for Transport Studies
Prof A.G. Cohn --- Computing
Prof S.B. Cooper --- Pure Mathematics
Dr S.R.D. French --- Philosophy
Mr W.J. Gerrard --- Leeds University Business S
Dr P.M. Hill --- Computing
Prof H.D. Macpherson --- Pure Mathematics
Prof A.T. McDonale --- Geography
Prof R.D. Pollard --- Electronic and Electrical En
Prof M. Rathjen --- Pure Mathematics
Dr S.A. Roberts --- Computing
Dr L.M. See --- Geography
Prof P.M. Simons --- Philosophy
Dr J.G. Stell --- Computing

Figure 4-4 Search 1 shows the searching result obtained from the current Expertise Matcher



Expertise Matching

Searching people who have expertise on:

spatial and reasoning or logic

spatial reasoning logic

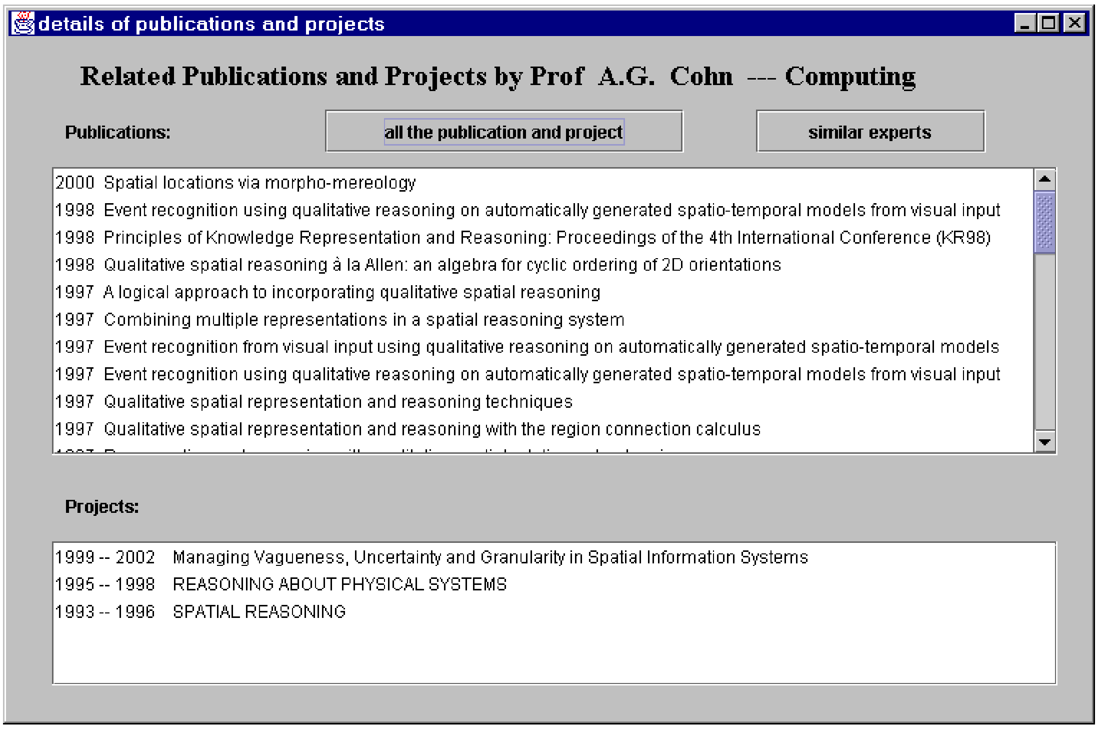
Dr B. Bennett --- Computing
Dr A.L. Bristow --- Institute for Transport Studies
Prof A.G. Cohn --- Computing
Prof S.B. Cooper --- Pure Mathematics
Dr S.R.D. French --- Philosophy
Mr W.J. Gerrard --- Leeds University Business S
Dr P.M. Hill --- Computing
Prof H.D. Macpherson --- Pure Mathematics
Prof A.T. McDonale --- Geography
Prof R.D. Pollard --- Electronic and Electrical En
Prof M. Rathjen --- Pure Mathematics
Dr S.A. Roberts --- Computing
Dr L.M. See --- Geography
Prof P.M. Simons --- Philosophy
Dr J.G. Stell --- Computing

0.702 Dr B. Bennett --- Computing
0.689 Prof A.G. Cohn --- Computing
0.309 Dr P.M. Hill --- Computing
0.287 Mr P.L. Mott --- Computing
0.262 Prof M.C. Clarke --- Geography
0.182 Dr I.J. Turton --- Geography
0.172 Dr W.E. Kunin --- Biology
0.161 Dr G.P. Clarke --- Geography
0.158 Mr S.L. Barr --- Geography
0.153 Dr J. Mason --- Sociology and Social P

0.682 Dr B. Bennett --- Computing
0.660 Prof A.G. Cohn --- Computing
0.301 Dr P.M. Hill --- Computing
0.257 Mr P.L. Mott --- Computing
0.250 Prof M.C. Clarke --- Geography
0.161 Dr I.J. Turton --- Geography
0.154 Dr G.P. Clarke --- Geography
0.144 Dr S.R.D. French --- Philosophy
0.136 Dr R.M. Pilkington --- Computer Based
0.135 Dr W.E. Kunin --- Biology

Figure 4-5 Extended Expertise Matcher (Search 2 and Search 3) ranking experts according to their relevance

For testing the Search 2 and 3, the same keywords (for example, "spatial reasoning logic") are input by the user but without the Boolean expression. Two lists of experts' names are then displayed (see Figure 4-5). In both Search 2 and 3, the experts are ranked. The value before the name of each person indicates how relevant that person is to the query. These two sets of results are very similar except that the score of each person is different which sometimes results in the different listing order.


details of publications and projects
all the publication and project
similar experts

Related Publications and Projects by Prof A.G. Cohn --- Computing

Publications:

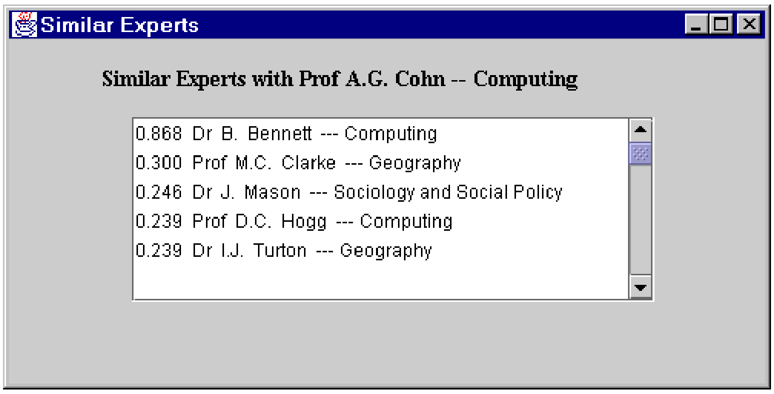
2000 Spatial locations via morpho-mereology
1998 Event recognition using qualitative reasoning on automatically generated spatio-temporal models from visual input
1998 Principles of Knowledge Representation and Reasoning: Proceedings of the 4th International Conference (KR98)
1998 Qualitative spatial reasoning à la Allen: an algebra for cyclic ordering of 2D orientations
1997 A logical approach to incorporating qualitative spatial reasoning
1997 Combining multiple representations in a spatial reasoning system
1997 Event recognition from visual input using qualitative reasoning on automatically generated spatio-temporal models
1997 Event recognition using qualitative reasoning on automatically generated spatio-temporal models from visual input
1997 Qualitative spatial representation and reasoning techniques
1997 Qualitative spatial representation and reasoning with the region connection calculus

Projects:

1999 -- 2002 Managing Vagueness, Uncertainty and Granularity in Spatial Information Systems
1995 -- 1998 REASONING ABOUT PHYSICAL SYSTEMS
1993 -- 1996 SPATIAL REASONING

Figure 4-6 An example of an expert's detail information

The publications and projects information associated to an expert can be displayed (as shown in Figure 4-6) if a user double clicks on the name of the expert. More specifically, each of these publication or project includes at least one of the keywords that the user has entered. A full list of publications and projects can be displayed by clicking "all the publications and projects" button. By clicking on "similar experts" button users can find other experts with the similar expertise (as shown in Figure 4-7).



Similar Experts with Prof A.G. Cohn -- Computing

0.868 Dr B. Bennett --- Computing
0.300 Prof M.C. Clarke --- Geography
0.246 Dr J. Mason --- Sociology and Social Policy
0.239 Prof D.C. Hogg --- Computing
0.239 Dr I.J. Turton --- Geography

Figure 4-7 Displaying similar experts

4.4.2 Experiment Results

In the initial experiment, 10 PhD students, who are ranged from 1st year to 4th year, are randomly selected from the School of Computing and invited to compare the three searches. Initially, they input a few keywords to express their research interests and obtained results from each Search. Then the participants selected the relevant experts from the returned lists of people and identified the position of their supervisor's name in each list. The participants added one more keyword in their queries and repeat the above process. For each search, the results before and after adding the new keyword were compared. In this experiment the test is based around four questions: (1) which search is more likely to locate the supervisors; (2) which search can help users find their supervisors in a shorter time; (3) which kind of query is easier to formulate; (4) which search is the most useful.

Table 4-2 Results obtained from 3 searches before and after adding a keyword
(Search 1 refers to the current ULPD Expertise Matcher; Search 2 refers to the extended Expertise Matcher with TF strategy; Search 3 refers to the extended Expertise Matcher with TF-IDF strategy)

Participant No.	**Position of the actual supervisor in the Search 1 list (before)**	**Position of the actual supervisor in the Search 2 list (before)**	**Position of the actual supervisor in the Search 3 list (before)**	**Position of the actual supervisor in the Search 1 list (after)**	**Position of the actual supervisor in the Search 2 list (after)**	**Position of the actual supervisor in the Search 3 list (after)**
1	-	8th	7th	-	2nd	1st
2	-	3rd	2nd	-	1st	1st
3	-	7th	4th	2nd	1st	1st
4	2nd	1st	1st	2nd	1st	1st
5	2nd	1st	1st	-	1st	1st
6[4]	(2nd)	(2nd)	(1st)	(2nd)	(2nd)	(1st)
7	9th	1st	1st	1st	1st	1st
8	-	4th	2nd	-	2nd	2nd
9	60th	4th	4th	152nd	3rd	3rd
10	1st	5th	5th	-	3rd	3rd

[4] No.6 participant did not find his supervisor in the three searches because his supervisor has retired and the information relating to his supervisor was removed from the ULPD database. A potential supervisor was identified in this case.

Table 4-2 shows how far down the actual supervisor of each participant appeared in the 3 sets of matching results (before and after adding a keyword). For example, No.1 participant did not find his supervisor's name in the list of Search 1, but did find it in the list of Search 2 and Search 3 positioned 8th and 7th respectively. After No. 1 participant added a keyword to the query, he still did not find his supervisor's name in the list of Search 1, but the positions of his supervisor's name in the list of Search 2 and Search 3 were changed to 2nd and 1st respectively.

From Table 4-2 it can be seen that 90% of actual supervisors were found using Search 2 and Search 3, whilst only 50% of them were found using Search 1. Among this 50%, 80% of actual supervisors were more easily found using Search 2 and Search 3 because they were listed in the top few. It was noticed that the supervisor of No. 10 participant was positioned 1st in the result list of Search 1. This is because in Search 1 the results are ordered alphabetically on surname and the first letter of the supervisor's surname is "B" (Dr. R.D. Boyle), hence it was displayed as the first result. Finding the supervisor in Search 1 at the top of the list only occurred with one student (10% of all participants). Compared with Search 2 and Search 3, in 40% of cases the supervisors were ranked higher in Search 3 than in Search 2. After adding a keyword, 60% participants found it easier to find their supervisors in Search 2 and Search 3 because their supervisors' names were ranked higher than before; 40% retained the same position because the supervisors were already listed 1st. In contrast, for Search 1, only 20% of participants found it easier to find their supervisors whilst 50% retained the same position; 30% found it more difficult as the positions of their supervisors were further down or it was not in the list at all. In summary, adding a keyword leads to more useful results in Search 2 and Search 3, but less useful in Search 1.

The precision and recall are two main criteria used to evaluate the performance of 3 searches. Precision means the proportion of relevant retrieved experts out of those retrieved experts whereas recall means the proportion of relevant retrieved experts out of all relevant experts. Table 4-3 shows the number of relevant experts found in Search 1 and Search 2 & 3, according to which, the precision of Search 1 and Search 2 & 3 can be calculated. This is shown in Figure 4-8. The average precision for Search 1 is 11.2%. The average precision for Search 2 is 18%.

The total number of relevant experts are unknown so that it is difficult to calculate the accurate value of recall. However, the average number of accepted potential supervisor in Search 2 and 3 is 1.8 whilst in Search 1, the average number is 0.7. This indicates that the recall in Search 2 and 3 is higher than Search 1.

Table 4-3 The number of retrieved relevant experts in Search 1 and Search 2 and 3

Participant No.	Number of relevant retrieved experts in Search 1	Number of retrieved experts in Search 1	Number of relevant retrieved experts in Search 2 and 3[5]	Number of retrieved experts in Search 2 and 3[6]
1	0	33	1	10
2	0	0	1	10
3	0	5	4	10
4	1	4	1	10
5	1	2	2	10
6	1	9	1	10
7	1	31	2	10
8	0	0	3	10
9	1	140	1	10
10	2	9	2	10

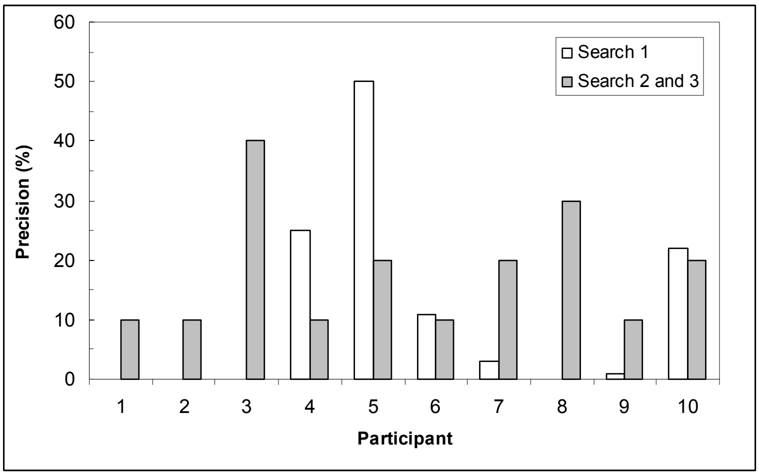


Figure 4-8 Comparison of precision of Search 1 and Search 2&3

[5] Number of relevant retrieved experts is the same for Search 2 and Search 3.
[6] Number of retrieved experts is the same for Search 2 and Search 3.

Despite the semantic indication associated with the Boolean model, the great majority of participants found it difficult to express their query requests in terms of a Boolean expression, especially for a long query (more than 3 keywords). The experiment showed that 80% of participants prefer to list a set of keywords without considering the logic behind them (as shown in Figure 4-9); only one student preferred to input a query with operators. He explained that "using an operator 'AND' could narrow the results". While it worked in his case, sometimes, it might be too narrow and no results were retrieved. Participants also found it difficult to expand a query. Two students still input "*A* AND *B* AND *C*" in their second round search after no results were obtained when they used "*A* AND *B*" in the first search. When they were advised to change the operators, they just simply changed "AND" to "OR", for example, "*A* OR *B* OR *C*". In Search 2 and Search 3, query expansion is much easier as participants can simply add as many keywords as they want.

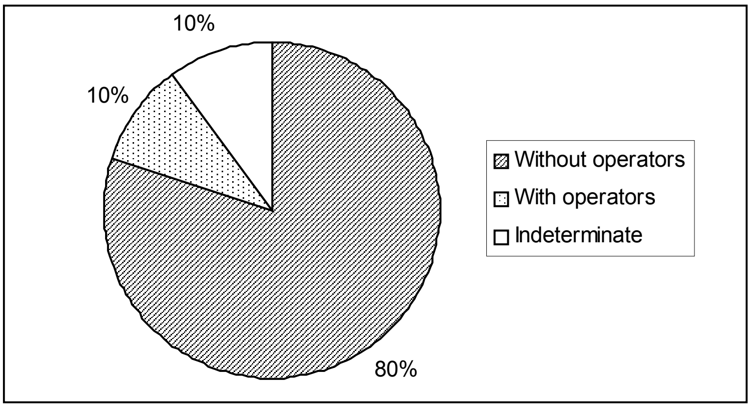


Figure 4-9 Participants' preference on operators in queries

The "Finding Similar Experts" function was also tested. The results are shown in Appendix D. A member was randomly selected from each research group in the School of Computing; the other experts who share similar interests were then retrieved. The results show that the most similar experts are always in the same research group with the selected expert. In the experiment, one participant found this function very useful. The initial search result did not return his supervisor's name but a colleague of his supervisor was returned instead. Through searching similar experts, the participant located his supervisor.

In summary, Search 2 and Search 3 are more effective than Search 1 for the following reasons:

- The precision and recall of the Search 2 & 3 are higher than Search 1.
- It is easier to formulate the query and expand the query.
- It gives more chances to find the experts (40% of participants found their supervisors in Search 2 and Search 3 when no result was obtained in Search 1).
- It has size control on the output so users do not have to check the detailed information of each expert if too many experts are retrieved.

Thus, it is no surprise that only 10% of participants (1 student) considered Search 1 to be the most useful search[7] (see Figure 4-10). When Search 2 is compared with Search 3, as they used the same model (vector space model), the actual experts retrieved were the same, only the orders of the lists were slightly different. 60% of participants considered Search 3 as the most useful search as they found the ranking result in Search 3 more appropriate than in Search 2 (the relevant experts were ranked higher in Search 3 than in Search 2). The other 30% participants found it difficult to decide which Search is better than the other as the results of Search 2 and Search 3 were very similar.

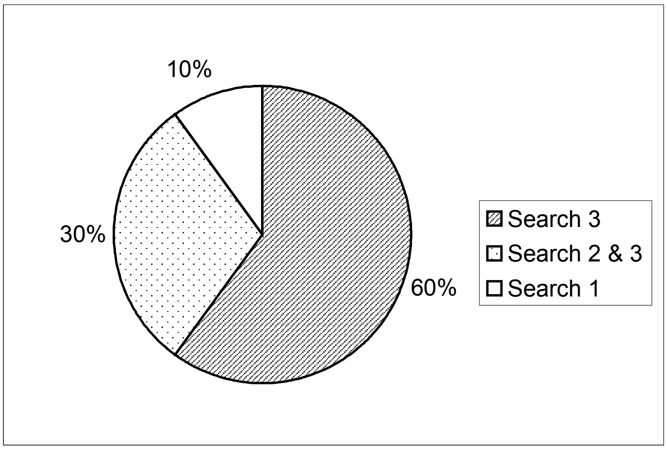


Figure 4-10 Comparisons of 3 searches on usefulness

[7] In this case, the participant was so lucky in choosing the keywords and retrieved his supervisor's name at the top of the Search1 results list and only another person was returned. This is a very rare case indeed.

4.4.3 Discussion

The extended Expertise Matcher creates the expertise profile using the vector space model. Some of the improved features of this as compared with the ULPD Expertise Matcher (see Section 4.1) are discussed below.

- It increases the possibility of finding experts. The expertise profile is obtained through combining all the publications and projects information of each expert. Even if the required keywords appeared in different publication/project titles then the system can still find the expert. This is impossible for the current ULPD Expertise Matcher. That is why sometimes there are no results when searching "*A* AND *B*" in Search 1, whilst Search 2 and Search 3 will retrieve some experts.

- The results can be ranked with a relevance rating. The expertise profile is expressed with a set of keywords with different weights, which is used to calculate the similarity between the expertise and the user query. Highly relevant experts can be displayed near the top of the list. The experiment results show that in 70% of cases the most relevant experts (supervisors of participants) were ranked in the top 3.

- The size of the results can be controlled. Due to the ranking ability, the number of results can be specified by the user. In the experiment, only the top 10 results found by Search 2 and Search 3 were displayed. Within the controlled number of results, Search 2 and Search 3 did not miss any experts that users found relevant in Search 1. Therefore, it normally saves users' time in locating the relevant experts.

- The keywords in the query are not treated equally. Search 3 automatically gives more weight to those keywords that appear less frequently in the collection of documents than the frequently used keywords. This avoids irrelevant people being ranked higher than more relevant ones due to more occurrences of frequently used keywords.

- It is easy to formulate a query. Users do not have to formulate precise queries with Boolean operators. They can simply list all the keywords or even give a document as their interest. This is because the system treats queries, documents, and experts in a uniform way.

- It is able to find similar experts. This alleviates the syntax match problem to some extent. The experts do not necessarily have to share the same keyword as specified by the user; they might be discovered through "similar experts" searching.

In theory, the vector space model supports adapting user profiles by gathering relevant information from user's feedback[8] [Rocchio, 1971; Ide, 1971; Salton and McGill, 1983]. Formally this is represented as:

$$\vec{f_m} = \alpha \times \vec{f} + \beta \times \frac{\sum_{\forall \vec{p_j} \in \mathrm{P_r}} \vec{p_j}}{|\mathrm{P_r}|} \qquad (4\text{-}4)$$

where $\vec{f_m}$ is the modified profile; $\vec{f}$ is the old profile; $\vec{p_j}$ refers to the expertise profile of the expert whom the user finds relevant; $\mathrm{P_r}$ refers to a set of relevant experts identified by the user among the retrieved experts; $|\mathrm{P_r}|$ refers to the number of experts in the set $\mathrm{P_r}$; α and β are tuning constants. However, the relevance feedback feature has not been tested in the evaluation process for two reasons. First, it takes a long time since users need to evaluate results a reasonable number of times before the adaptive user profile stops changing. Second, this relevance feedback feature is more useful in a large collection of experts than a small collection, which means through identifying the relevant experts in the initial search, more experts can be retrieved in the next search. This is only suitable if there are many experts relevant to each query. However in the experiment there are only 2 or 3 relevant experts in most cases so there is a possibility that no relevant expert will be returned in the initial retrieval, and then no relevant information can be gathered. In addition, the focus of this study is on improving the performance of the initial retrieval.

[8] Here only positive feedback is used since it is more important than negative feedback [Salton and McGill, 1983].

Despite these advantages, the extended Expertise Matcher also revealed some deficiencies. Firstly, the extended Expertise Matcher is still an isolated keyword base search (syntax-based search), terms are considered to be independent of each other. This means that an expert will be retrieved if his/her expertise profile includes the same keyword(s) given in the query. However, a single word can have two or more meanings (this is called polysemy) and the retrieved experts may not be relevant. On the other hand, relevant experts may not be retrieved even if their expertise profiles include the keywords which are semantically similar to the given keyword (synonymy) or highly relevant (such as hyponymy - subset/superset relations between two words). For example, a user inputs "information and integration" as the query, and did not find his supervisor although his supervisor uses the words "semantic sharing, information broker, mediator" in his publication titles. Studies show that the chances of two people choosing the same term to describe the same concept is less than 20% due to the diversity of the human language [Deerwester *et al.,* 1990].

One solution to this "*term mismatch*" problem is *query expansion* [Efthimiadis, 1996], which aims to retrieve a more relevant target by adding terms to the query. Collecting relevance feedback [Rocchio, 1971] is one kind of query expansion. Terms can also be selected from a thesaurus, such as finding the synonyms of the terms in the query. Manually building thesaurus is quite expensive and different techniques are used to automatically generate thesaurus, such as analysing word co-occurrence in the documents [Attar and Fraenkel, 1977]. However, this approach leads to rapid degradation of precision [Sparck-Jones, 1972]. Thesaurus-based query expansion causes a decline in retrieval performance generally [Hersh, *et al.,* 2000]. This is because synonyms are not equal to the original word, and if a synonym with multiple meanings is chosen, the situation is worse. Furnas *et al.* [1983] proposed the Latent Semantic Indexing model to map each document and query vector into a lower dimensional space which is associated with concepts. Thus it allows a match between queries and documents if they do not share the same word. Unfortunately, the high computational requirements of LSI and its difficulty in determining the number of dimensions limit its applicability[9] [O'Riordan, and

[9] On the one hand, the system will reduce to the vector space model if the number of dimensions is too large; on the other hand, significant semantic content of a particular domain will remain uncaptured if the number of dimensions is too small.

Sorensen, 1999; Karypis and Han, 2000]. In summary, these approaches are *recall-oriented* since they focus on synonymy rather than polysemy, and achieve very limited success in improving search effectiveness due to the lack of query context [Singhal, 2001].

Secondly, coverage of expertise data sources needs to be improved. In the ULPD system, the expertise of each expert is derived from the publications and projects database. This is not sufficient; some information stated in personal homepages is very valuable to derive their expertise. For example, the "research interests" section in experts' homepages clearly reflects their expertise. Hence, it should not only be included in the expertise information, but also be given higher weights than the publication and project titles. Another example of expertise data sources are technical reports.

Thirdly, output presentation needs to be enhanced. In the initial experiment, what the system provides about each expert's detailed information is only the titles of the experts' publications and projects. This is not sufficient for users to evaluate their expertise. It is not a serious problem in the initial experiment because the selected participants are PhD students in the School of Computing and they are supposed to know their supervisors and other relevant experts in the department. It is much more difficult for other users to evaluate the expertise just using the titles of the publications and projects. They need not only personal contact information, but also the research interests of each expert, the information of research groups they are members of, their work experience, and any online documents they have produced, and so on. Different users may have different requirements; not all the users seek the expert with the most experience and expertise. The system should support their selection process by providing general relevant information about each expert.

4.5 Conclusions

This chapter has described the use of the vector space model to extend the current ULPD Expertise Matcher. This approach treats user query, publication, project and expert in the same way (weighted keyword vector), and relevance is measured by the cosine between two vectors.

Therefore, the extended system can rank experts according to their relevance to the query and implement a partial match. Furthermore, users can easily form a natural language query. The initial experiment results are promising in that most of the drawback of the current Expertise Matcher have been solved. This experiment illustrated that the traditional IR method (vector space model) remains effective when applied to finding experts.

The extended system still leaves a number of issues unresolved which serve as the basis for continuing research. This includes syntactic search limitation, limited expertise information, and poor presentation. How to solve these issues whilst retaining the advantages of the vector space model is the focus of future work.

Chapter 5

Use of RDF in Expertise Matching

The previous chapter analysed the limitations of the Expertise Matcher in the ULPD system and presented an extended Expertise Matcher which uses the vector space model to build an expertise profile. The results of the experiment show that some of the limitations have been solved, however a number of issues still remain. This chapter analyses the possible solutions for the remaining problems and examines how the semantic web technologies such as RDF/RDFS, XSLT, and ontologies can be used to address these issues. To test the applicability of these technologies in expertise matching, a prototype system called Expertise Locator has been built which aims to help PhD applicants (expertise seekers) locate potential supervisors (expertise providers). The evaluation of the Expertise Locator has been conducted through an experiment with real users and the key results are presented. Finally, a comparison between the Expertise Locator and other related work is undertaken.

5.1 The Remaining Issues and Possible Solutions

The previous experiment described in Chapter 4 has identified the critical success factors and factors to be improved in the extended Expertise Matcher. They are highlighted in Table 5-1 below.

Table 5-1 Comparison between success factors and limitations in the extended Expertise Matcher

Success Factors	**Limitations**
Ability to build expertise profile	This profile is built based on the ULPD database only
Ability to rank experts	Ranking results depends heavily on the keywords that user specified
Ability to retrieve similar experts	Similarity is calculated based on keywords rather than concepts. If an expert has many research interests, then the retrieved similar experts may have expertise in different areas.
Ability to display the publication and project information relevant to each expert	The provided available about each expert is limited, some other information from different data sources useful to expertise assessment and expert selection is overlooked.

The limitations presented in the table 5-1 are due to (i) lack of integration of expertise indications from heterogeneous data sources (ULPD database is the only data source), (ii) syntactic search (retrieval and ranking heavily depend on the keywords). In order to overcome these limitations, multiple expertise indications from heterogeneous data sources should be integrated and concept search should be designed, hence the integrated relevant information associated to each expert can be provided to users in helping them assess the expertise of an individual. The rest of this section analyses the approaches to alleviate these limitations.

5.1.1 Heterogeneous Data Sources in Reflecting Expertise

5.1.1.1 Heterogeneity Problem

The expertise indicators extracted from different data sources are the foundation for the intelligent expertise matching systems. These indicators are physically distributed in different sources with different formats across the organization. For example, some departments such as School of Computing have its own database which stores publication information about its staff. Some experts have their own homepages from which personal updated information can be obtained. Manually creating a database such as ULPD to store all this information is very difficult and expensive. Furthermore, there is a critical problem of maintaining up-to-date information. A person’s expertise changes over time and it is not feasible to rely on the individual to report developments to their expertise profile and even so, the database maintenance task would be significant if hundreds or even thousands of individuals were involved.

The above analysis leads to the question: “*Is it possible to automatically extract the relevant information from disparate data sources and integrate them?*” To answer this question, it is necessary to examine closely what type of information is available which includes expertise indicators. There are a number of different data sources varying from structured data (such as databases), semi-structured data (such as web pages), to unstructured data (such as text files). This heterogeneity brings many difficulties to the task of information integration. Busse et al.

[1999], Seligman and Rosenthal [2001], and Sheth [1998] present different classifications of heterogeneity which can be summarized below:

- *Heterogeneous systems* This includes platform heterogeneity (such as operating system and hardware) as well as information system heterogeneity. For example, different types of DBMS support different data models (such as relational, hierarchical, object-oriented models) and different query languages (such as SQL and OQL).

- *Heterogeneous attribute representations* This is also called syntax heterogeneity, which includes data type and format differences. For example, in one source date is measured by year only (1993), whilst in another source it is measured by day, month and year (10/12/93).

- *Heterogeneous schemas* This means that the same information elements can be assembled into many different structures. For example, one system might store all publications information in one renormalized table, while others might split it among several tables.

- *Heterogeneous semantics* This refers to the meaning of the terms. The relations and attributes in a schema have names only, the implicit semantic (concept they stand for) are interpreted by people. The understanding by different people may be different. Semantic conflicts can occur when different names stand for the same concept or the same names denote different concepts. For example, one system might use "author" while another system might use "creator" to express the same meaning. Differences in semantics are more challenging than representation heterogeneity.

- *Object identification* When the relevant information for the same object is stored in separate sources then how does the system recognise that they are referring to the same thing but with different attributes? For example, the central administration office of a university may have a record of a person ("Smith Black, 1970"), and the individual school or department may also have a record of this person ("Smith Black, 1970") but the problem is how to identify that they are the same object.

5.1.1.2 Approaches to Solving the Heterogeneity Problem

The problem of heterogeneity has been studied in the database research community for well over two decades. The representatives of traditional approaches are multidatabases [Litwin *et al.,* 1982; Dayal and Hwang, 1984; Rahimi *et al.,* 1982] or federated database systems [Heimbigner and McLeod, 1985; Sheth and Larson, 1990]. The latter is a special type of multidatabase systems (tightly coupled) because an integrated schema is provided. These approaches put more emphasis on system heterogeneity (such as database heterogeneity) than syntax and structure heterogeneity. There are two classes: (i) multidatabases and federated database systems use the virtual approach[1] (the actual data is still stored in the original data sources), (ii) the warehousing approach [Hammer *et al.,* 1995] uses the materialized approach[2] where relevant information is extracted, filtered and integrated in a repository. When a query is posed, the query is evaluated directly at this repository, without assessing the original information sources. However, it suffers from problems of data becoming out of date and consistency maintaining [Widom, 1995].

In order to deal with a variety of data sources (structured, semi-structured and unstructured data sources), mediator-based systems [Wiederhold, 1992] have been developed. These systems use the virtual approach to provide up-to-date information. Mediator-based systems are usually developed using a top-down approach, that is, starting with a global information need and sources that can contribute to this need can be plugged in later[3]. In mediator-based systems, a mediator provides a unified schema as an interface to a dynamically changing collection of heterogeneous information sources. A main component of a mediator-based system is the wrapper, which encapsulate data sources and translate the local data model and language into a common data model and common language. There are two techniques to map the source

[1] Also called lazy approach or on-demand approach

[2] Also called eager approach or in advance approach

[3] If the data sources are known before the integration, a bottom-up strategy is used such as federated databases systems and data warehousing.

schemas to the mediator schema. One is Global-as-View (GaV) where the mediator schema is defined as views over the sources schemas for each class, as in Information Manifold [Levy *et al.,* 1996], and SIMS [Arens *et al.,* 1996]. Although the query decomposing is fast, when information needs or sources change, a new mediator schema should be generated. The other is Local-as-View (LaV) where the source schemas are described by giving equivalent views on the global schema, such as in Garlic [Carey *et al.,* 1995], and TSIMMIS [Garcia-Molina *et al.,* 1995]. Rules are used to construct these views. Mediators contain mechanisms to rewrite queries according to the rules. The emphasis for mediator-based systems is on heterogeneous syntax (attribute) and structure (schema) rather than heterogeneous systems (such as heterogeneous DBMS).

In order to support interoperability and integration of a variety of data sources, a broad variety of metadata is exploited. The role of metadata for semi-structured and unstructured data sources is like schema for a database. Kashyap *et al,* [1995] classified metadata into content-independent metadata (such as modification data of a document) and content-dependent metadata (such as size of document). Content-dependent metadata can be further subdivided into direct content-based metadata (such as full-text indexes); content-descriptive metadata (such as textual annotations of a page); domain-independent metadata (such as structure metadata); and domain-specific metadata (such as terms chosen from domain-specific ontologies). Mediator-based information systems require the software developers to have a clear understanding of a variety of metadata, as well as a comprehensive understanding of schematic heterogeneity [Sheth, 1998]. In rule-based mediators, rules are mainly designed in order to reconcile structural heterogeneity [Garcia-Molina *et al.,* 1995], whilst for the reconciliation of the semantic heterogeneity problems, the semantic level also has to be considered [Stuckenschmidt, 2000]. The literature on integration is concentrated on syntax and structure with few people focusing on semantic interoperability (see for example [Fensel *et al.,* 1999], [Stuckenschmidt, 2000]).

5.1.1.3 The Roles of Semantic Web Technologies

Extensible Markup Language (XML) [Bray *et al.,* 2000] is accepted as the standard for data interchange on the web. XML is a neutral syntax that can transform diverse data structures into graph-structured data as nested tagged elements [Seligman and Rosenthal, 2001]. In this way, heterogeneous data structures can be represented in a uniform syntax – XML. XML also helps by providing a convenient mechanism for attaching descriptive metadata to attributes of both the source and target schemas. XSLT [Clark, 1999] can define the mapping between the heterogeneous schemas. Using XML, three problems listed in the section 5.1.1.1 can be alleviated, they are heterogeneous DBMSs, heterogeneous attribute representations, and heterogeneous schemas. However, XML cannot support integration at the semantic level. For example, suppose there are two expressions: <Surname> Black </Surname> and <Lastname> Black </Lastname>, which seem to carry some semantics. However, from a computational perspective, a tag such as <Surname> carries as much semantics as a tag such as <H1>. Hence the system does not understand that Surname and Lastname mean the same thing and that they are related to another concept - "Person". An XML Schema provides support for explicit structural cardinality and data typing constraints, but does not provide much support for the semantic knowledge necessary to integrate information [Hunter and Lagoze, 2001]. Further, XML does not play a very significant role in object identification.

RDF (Resource Description Framework) [Lassila and Swick, 1999] and RDFS (the Schema Language for RDF) [Brickley and Guha, 2000] are W3C standards for describing metadata on the web. They can be used to solve the semantic heterogeneity problem. It is useful for "semi-structured" or schema-less data [Brickley, 2001]. RDF provides a standard representation language for web metadata based on directed labelled graphs [Karvounarakis *et al.,* 2000]. It consists of three object types: resource, property and statement[4]. Every resource has a Uniform Resource Identifier (URI). The use of URIs to unambiguously denote objects, and the use of properties to describe relationships between objects, distinguish it fundamentally from XML's tree-based data model [Decker *et al.,* 1999]. The RDF data model is just a triple of {subject, predicate, object} and the order of information is not significant. The same RDF tree can be

[4] The more detailed information of the RDF data model can be found in Appendix E.

expressed differently in many XML trees because the order of elements in an XML document is very meaningful. Therefore RDF successfully avoids the problem of querying XML trees which attempt to convert the set of all possible representations of a fact into one statement [Berners-Lee, 1998].

In addition, RDF vocabularies can be described using an RDF Schema which is also written in RDF. An RDF Schema further allows simple semantics to be associated with terms; classes may have multiple subclasses or super-classes, properties may have sub properties, domain and range [Heflin and Dale*,* 2002]. RDF adds value in comparison to traditional DTD or XML schema approaches in the XML world. A DTD focuses on the structure of an XML document. It gives the name of elements, the associated attributes each element has, and the order of elements in an XML file. An XML Schema provides a means of specifying element content in terms of data types, so that document type designers can provide criteria for validating the content of elements. Either XML schema or DTD provide poor support for semantics [Hunter and Lagoze, 2001], in contrast, RDF schema (RDFS) defines the types of resources that a document might describe, the types of properties (attributes and relationships) that can be possessed by the resources and restricts the ranges of the properties.

In order to solve the heterogeneous semantics problem, there is a need to agree on the meaning of the terms used in the different data sources. The description of a shared set of terms in an application domain is called an ontology or a conceptual model instance[5], which includes not only the definition of the terms, but also the relationships between these terms. RDFS can be seen as the first language to describe ontology [Hunter and Lagoze, 2001]. Through using ontologies to make the implicit meaning of their different terminologies explicit, it is then possible to dynamically locate relevant data sources based on their content and to integrate them as the need arises [Cui *et al.,* 1999]. Global specific ontologies act as "semantic conceptual views" over the heterogeneity of data sources. The problem of mapping structure and semantics

[5] The difference between Ontology and Conceptual Model is that "Ontology is external to information systems and is a specification of possible worlds in some particular domain that covers multiple and often a priori unknown information systems while a conceptual model is internal to information systems and is a specification of one possible world of that domain" [Bishr and Kuhn, 2000].

of data between data sources is then changed to mapping the metadata of individual data source against the global ontology.

5.1.2 Output Presentation

In order to help expertise seekers assess the expertise of each expert, the final description of each expert returned to users should provide an integrated view of each expert in the same way that users might manually select and integrate pieces of relevant information from diverse data sources. The duplicated information from diverse data sources should be removed.

As analysed in Section 3.2.4, most experts finding systems did not provide sufficient information on each expert. Normally, the result of searching for experts is a set of experts and their contact information. Some systems display the publication titles and/or a few keywords to describe the expertise. This is usually not sufficient for users to assess the expertise of each expert. The extended Expertise Matcher did not solve this problem where the output presentation is just the information of the publications and projects. Although different users may be interested in different aspects of the experts, some common interesting facts can be pre-specified in the conceptual model (application ontology), and the output presentation can be created based on the conceptual model.

XML and XSLT are very useful for the presentation of the output of a search. XML separates the structure of a document from its presentation, and XSLT can be used to provide different presentations to different users based on the same content. The output presentation is similar to a personal homepage, but is dynamically and automatically created by the expertise matching system through integrating heterogeneous data sources. If any new information is found in any of these data sources then the output will reflect this change.

5.1.3 Concept Search

Even if the conceptual model (application ontology) is created as a global schema and relevant information about each expert is extracted from different data sources and integrated semantically, there is still a key problem, that is, syntax search. Although users can conduct a restricted field search, for example, "show me all the people who have published papers on 'information search and retrieval'", the searching process is still based on the syntax match, hence it might not find a paper with the title "document clustering" as there is no common keywords between the user query and the paper title. Even when there are common words the meaning may be different. For example, vision could mean "act or power of imagination", but it could also mean "frames enable the division of a browser window into independent areas". The integrated publications information of each expert from the different data sources increases the chance of finding experts. It is, however, for recall only. Precision suffers because information retrieval systems are unable to distinguish which meaning was used in queries or in documents [Egnor and Lord, 2000].

Another kind of ontology – domain ontology - has been viewed as a promising means to tackle this problem. Ontologies help to de-couple description and query vocabulary and increase precision as well as recall [Guarino *et al.,* 1999]. Domain ontology characterises the body of knowledge associated with the particular domain of a task, such as, the definition of the concepts, the attributes of the concepts (for example, synonyms, abbreviations), and the relations between concepts (for example, is-a and part-of). If both the users queries and the experts profiles can be linked to the concepts in the ontology, then the searching precision is probably higher than simply keyword searching.

This linking is difficult to implement automatically due to the nature of the English language. It is very difficult for a machine to understand the meaning of a question posed by a user. Although Natural Language Processing researchers have conducted research on extracting meaning/concept from documents, the technology is not mature enough to be satisfied [Li *et al.,* 2001]. The same problem is found when processing the integrated information of each expert and extracting the concepts of their expertise. Due to the difficulties in automatically linking expertise profile and user queries with concepts, a semi-automatic approach is proposed. As shown in Figure 5-1, for each concept, a set of keywords is extracted as "relevant keywords". Based on the expertise profile (a set of keyword with weights), the relevant concepts are retrieved if the description of the concept contains some keywords in the expertise profile. The concept whose description contains the most keywords is listed on the top. Each expert can then confirm if these concept(s) reflect their expertise. Thus, the expertise profile of each expert is built up which includes a set of keywords (with weights) and a set of concepts. A similar process will be applied in confirming the context of user queries. Once the concepts are selected by the user, the user query is replaced by the short explanation of a concept (in a set of keywords), which will then be used to search for an expert. Only those experts whose expertise profiles include the specified concept are retrieved. The experts are ranked according to the similarity between the keyword profile of experts and the new user query.

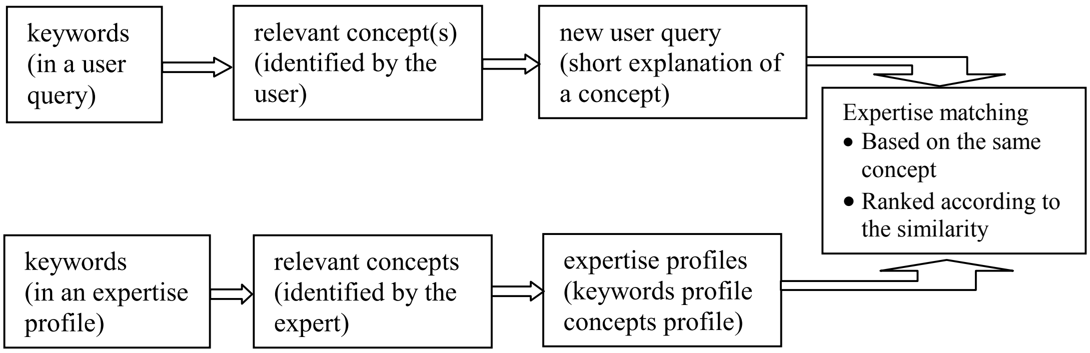

keywords (in a user query)
relevant concept(s) (identified by the user)
new user query (short explanation of a concept)
Expertise matching
• Based on the same concept
• Ranked according to the similarity
keywords (in an expertise profile)
relevant concepts (identified by the expert)
expertise profiles (keywords profile concepts profile)

Figure 5-1 Matching between user query and expertise profile

Having analysed the limitations in the extended Expertise Matcher and also justified the role of RDF and ontologies in solving these problems, the use of RDF and ontologies in expertise matching will now be explored. A prototype system, namely Expertise Locator, is then designed and built. The Expertise Locator is designed to help PhD applicants select their potential supervisors in the School of Computing at the University of Leeds prior to them making a formal application. The aims of Expertise Locator are to improve the accuracy of searching for experts and provide a coherent and meaningful view of the integrated heterogeneous information sources associated with each particular expert.

5.2 Experiment and Rationale

The School of Computing in the University of Leeds is a large department and each year there are approximately 50 applicants who formally apply to the School as research students, and there are also many more general enquiries from individuals considering making an application. The case study detailed below has been designed to deal with both the actual applications as well as the general enquiries. The first step for a potential student is to discern whether anybody in the School has the expertise in the research area they are interested in and whether they could possibly be their supervisor. Normally, the students can get access to most web-based information sources, but they may well have to spend a long time checking each web page and searching each database (such as publications) to find pieces of information and to integrate them manually. It is a significant burden on the user to select, search, filter, and integrate the information they want. As a result of this, many students simply ignore this process - what they do is simply write down their research interests and leave the School's PhD Admissions Tutor to try and select a suitable supervisor for them based on their proposed research topic. The better the PhD Admissions Tutor understands the expertise of each academic in the School then the better the match between supervisor and student will be. Sometimes even when a PhD Admissions Tutor has worked for many years in the school, it is still very difficult for him/her to recall up-to-date details of all the expertise and research interests for each individual. This is because the number of researchers in the School of Computing who could be supervisors is large and their expertise and research interests may continually change and develop. Furthermore, the PhD Admissions Tutor may not fully understand the applicants' intents because some applicants use quite specific and often inaccurate technical terminology. As a result, the supervisor that the PhD Admissions Tutor recommends may not be the most suitable person, and there exists a real possibility that some appropriate applicants are rejected because their needs cannot be appropriately matched in this way.

5.2.1 Business Objectives of the Expertise Locator

The above problems are addressed in the design of the Expertise Locator System, which aims to improve the process of matching supervisors and potential research students by enabling the potential applicant to make more informed choices about their supervisors before they formally apply to the University. Both applicants and the School could be benefited in the following areas:

- The applicant could search Expertise Locator for potential supervisors themselves and retrieve integrated information on each supervisor without having to browse many webpages, thus they can make a more accurate selection of their preferred supervisor(s).

- The burden on the broker (PhD Admissions Tutor) for matching between applicants and supervisors could be reduced if the preferred supervisors were stated by the applicants in the application forms, or if the broker could use Expertise Locator to locate potential supervisors.

- On some occasions there is no directly related expertise available and the broker may recommend the applicant to other research areas. It will take time for the applicants to make a decision to accept or reject the offer and for the broker to get feedback from them. This problem will be solved by the Expertise Locator System as the applicants could make the alternative selections themselves immediately.

- The applicants may change their mind (for example, apply to another university) if there is no expert in their preferred research area. This also saves time for both the PhD Admissions Tutor and the applicants.

5.2.2 User Study

To identify the support tasks needed in the Expertise Locator System, consider the following scenario, which represents a typical case for the problem described above:

> *Mary is a Masters student in the University of Manchester and is graduating soon. Since her plan is to continue studying as a PhD student, she is searching the web pages of several universities, including the University of Leeds, in order to decide which university is the best one for her. Her preferred research interest is "heterogeneous database systems". Mary first navigates the School of Computing website at the University of Leeds and browses the homepage of each member of staff. She quickly finds that there are a large number of staff in the School and*

> *many of them are not active researchers. Then she changes her mind and decides to browse the research groups in order to quickly locate a potential supervisor. She finds these websites are not well organized. Although she searches very carefully, she still does not find a researcher who exactly matches her requirements. She is thinking that maybe there are no academics conducting research in this area and she should give up applying to Leeds University.*

This is not the desired outcome as there are people who can supervise her in her preferred research area at Leeds University. The scenario draws attention to the following problems involved in identifying the potential supervisor(s):

- **Low recall** This means that some relevant people are missed. This is mainly due to: (1) There is a large number of staff in the School and it is a very time consuming task for the user to access each person's homepage; users may stop after they have browsed a dozen of the staff's homepages. (2) The web page of each research group does not give detailed information on the individuals in the group. As a consequence, the user may not find the relevant person even when searching carefully.

- **Low precision** This means that some of the people found are not experts in the preferred research areas. It is not always the case that researchers working in the same research group have very similar research interests or expertise. Users still need to conduct further assessment by looking carefully at the detail of each researcher in order to determine if that individual is a suitable supervisor. Therefore, the number of real experts is very small compared to the total number of people retrieved.

The following is the ideal situation that Mary wants the system to provide:

> *When Mary conducts a search by entering her research interests – "heterogeneous database systems", several relevant research areas available are returned. Mary chooses "Information Integration and Databases" as her preferred research area, and two related researchers are displayed with the relevant score. Each researcher*

> *has his/her own detailed information page including research interests, the projects they are working on or have worked on in the past, the papers they have published, technical reports which can be downloaded, and so on. Mary compares these two researchers and reads abstracts of 2 papers, she then chooses one of the two to be her potential supervisor and starts completing the application form and indicating the name of the potential supervisor on the form.*

From the ideal situation in the above scenario, the most significant support tasks required of the Expertise Locator System can be identified. These are summarized as follows:

- Identification of expertise requirements;
- Conducting concept search by prelinking experts' expertise with the domain concepts;
- Ranking experts according to their expertise level so that the chance of missing most relevant experts is reduced;
- Capturing the relevant information of each expert from diverse information sources in the organizational memory and providing an integrated view of each expert to the user.

5.3 The Conceptual Model and Architecture of the Expertise Locator System

5.3.1 Conceptual Model

A common conceptual model is necessary in order to integrate different expertise indications. Figure 5-2 shows a simplified conceptual model for expertise matching within academia, and hierarchical relationships have not been included due to space constraints. An example of the underlying hierarchical structure associated to the concept "Person" is given in Figure 5-3. The major concept in Figure 5-2 is "Person"; the others are "Publication", "Expertise", "Project", "Research_Group" and "Classification". The relationships between the concepts and the attributes related to each concept are also specified in the conceptual model. For example, a resource of type "Person" may have a property "author_of" whose value is a resource of type "Publication". In the meantime, it can have another property "email" with value "Literal". "author_of" represents the relation between concepts "Person" and "Publication" while "email" represents the attribute related to the concept "Person". The full model is listed in Appendix F which is represented in RDF. This conceptual model is created by the application designer using ontology editor such as Protégé-2000[6].

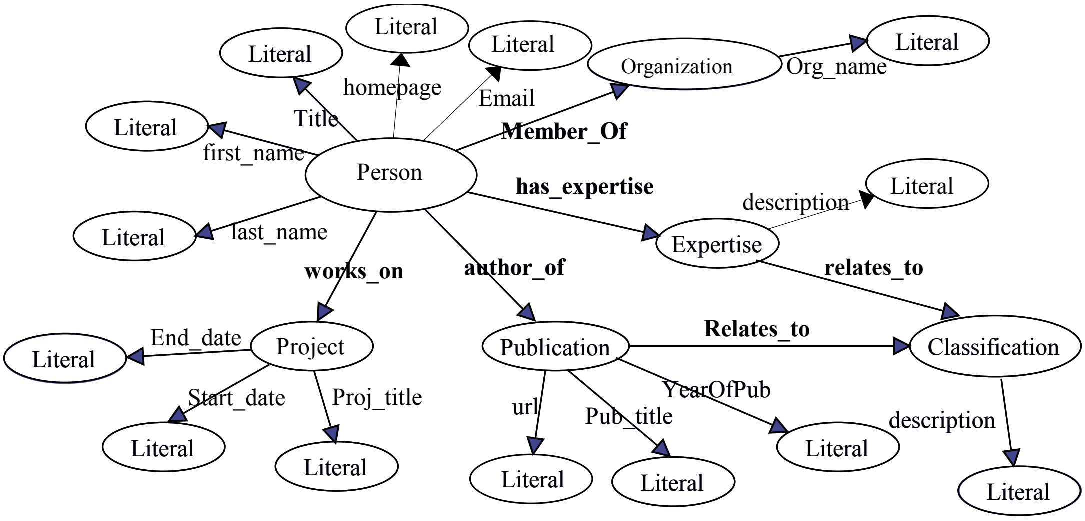


Figure 5-2 Sample conceptual model used in the Brokering System

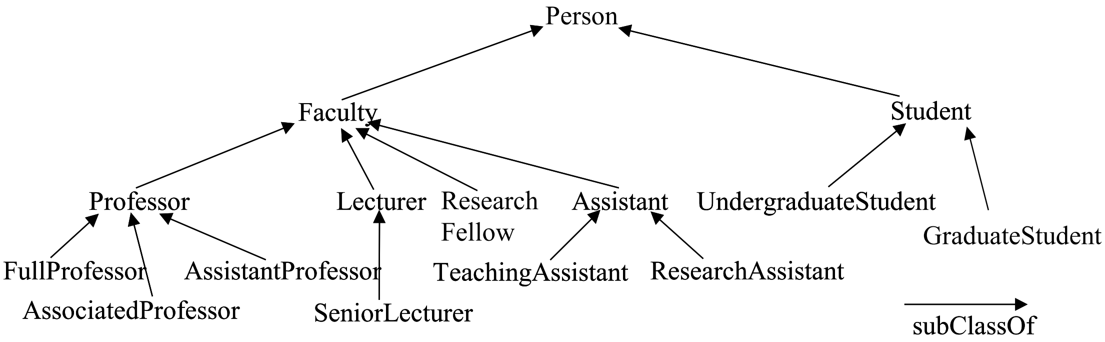


Figure 5-3 An example of the underlying hierarchical structure associated with the concept "Person"

5.3.2 Architecture of the Expertise Locator System

The architecture of the Expertise Locator System is shown in Figure 5-4 (Figure 5-4 also illustrates the different data sources used in the case study). The architecture can be divided into two layers, namely, i) semantic information integration; ii) expertise management. The first

[6] http://protege.stanford.edu/

layer was developed based on [Vdovjak and Houben, 2001]. Each component in the architecture is described in detail below:

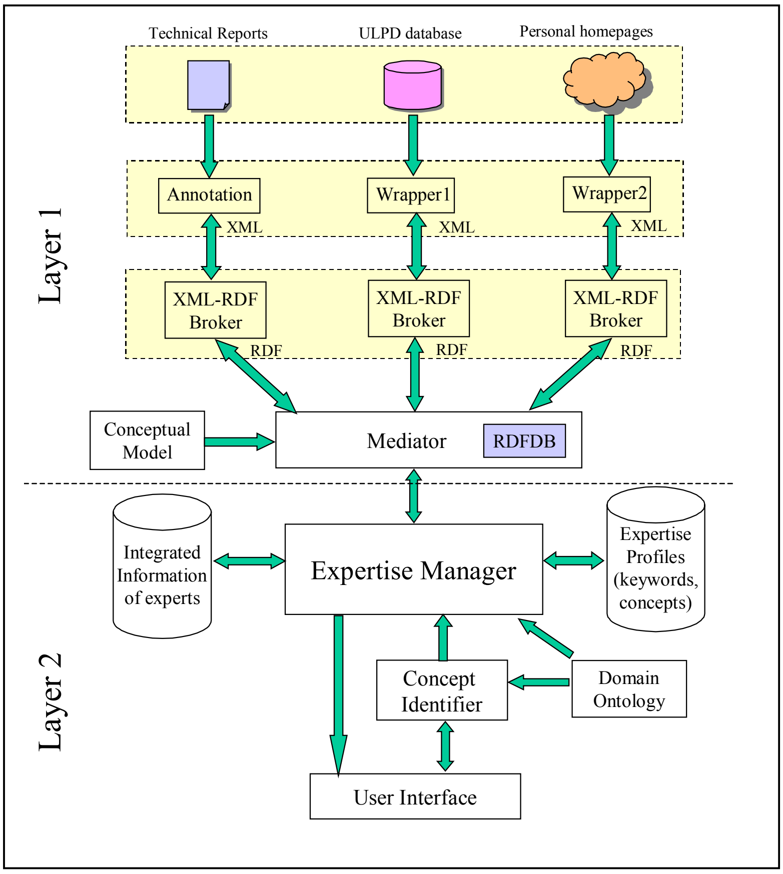


Figure 5-4 Architecture of expertise matching based on integration of heterogeneous information sources

- **Source** Contains data sources that are relevant to identifying the expertise of each potential supervisor such as personal homepages which include personal contact information, research interests, associated research group(s), and recent publications; the ULPD database which stores information about publications and projects by members of staff across the University of Leeds; and technical reports which are online documents stored in the School of Computing database. These data sources are built by different people for different objectives or different users, some of the data across these three data sources is duplicated. For example, information on a particular publication authored by a member of staff may be stored in all three data sources.

- **Wrapper** Different wrappers such as DB-XML wrappers or HTML-XML wrappers are used to extract relevant information from the original data source and present it to the serialized XML data. For these unstructured data, some manual processes are needed such as adding metadata in XML according to the vocabularies stored in the Conceptual Model.

- **XML-RDF Broker** Identifies the relevant concepts in the XML sources and replaces them with the concepts in the Conceptual Model; the mapping rules are specified in XSLT. These mapping rules are defined by the application designer and can be modified if the concepts of the source change. However, the underlying Conceptual Model should be stable as it is the basis for the semantic integration; if it has to be changed, then the RDF model and the mapping rules should be modified accordingly. The XML-RDF broker also receives the queries from the mediator and response with a set of RDF statements by searching the XML source.

- **Mediator** Maintains the Conceptual Model (shown in Figure 5-2). This layer identifies which data sources are relevant to the query, transfers the query to subqueries, and retrieves subresults from brokers. These subresults are input into RDFDB , and through searching RDFDB, the final results (the semantically integrated information of each expert) is delivered to the expertise manager.

- **Expertise Manager** In addition to maintaining experts' information (experts profiles), the expertise manager also creates, stores and retrieves expertise profiles which consist of two forms – keywords and concepts. It receives the extended query and specified concept from the concept identifier and retrieves the experts whose expertise includes the required concept. The Expertise Manager ranks experts according to the similarity between their keywords profiles and the user query. The ranked experts with their integrated information are then sent directly to the user interface.

- **Concept Identifier** Receives a user's query and provides the relevant concepts according to the domain ontology. After a user confirms the relevant concept(s), the concept identifier extends a user query with the description of the relevant concept and sends the specified concept and the extended query to the expertise manager.

- **User Interface** Receives the query from the user and sends the results of the ranked experts together with the detailed information of the experts to the user.

5.4 Implementation and Quality Control

The implementation of the architecture described in the previous section includes several crucial aspects which are briefly described below:

- **Indexing and retrieval of concepts** The concepts and associated keywords are chosen from the ACM Computing Classification[7] and an online computing dictionary - FOLDOC[8]. The ACM Computing Classification System has roughly 100 third-level headings and provides a relatively stable scheme that covers all research in computing [Halpern, 1998]. FOLDOC is a searchable dictionary of computing contributed by 1500 people. The dictionary has been growing since 1985 and now contains over 13500 definitions totalling nearly five megabytes of text. Entries are cross-referenced to each other and to related resources elsewhere on the Internet. The concepts and their associated keywords and supervisors are stored in a relational database. This database is connected to the Java system code via JDBC. The possible relevant concepts are retrieved based upon the research interests that the user inputs.

- **Constructing the detailed information for supervisors** Firstly, relevant information from the diverse data sources should be collected. The information is stored in the web pages and the ULPD database is transformed into XML form using wrappers. Some

[7] ACM Computing Classification http://www.acm.org/class/1998

[8] The Free On-line Dictionary of Computing http://foldoc.doc.ic.ac.uk/foldoc/index.html

manual annotations are needed for interpreting the information stored in the unstructured data sources. Manual annotation is time consuming and it is noticed that some annotation tools such as, MnM[9] and Ontomat[10] are becoming available to provide a degree of automatic annotation. Secondly, these XML files are then transformed into RDF using XSLT (an example is given in Figure 5-5). Thirdly, the separate RDF data is input into an RDF database -- RDFDB. Fourthly, a search is conducted on RDFDB to produce the complete detailed information for each supervisor. Duplicate information is removed at this step automatically. The third and fourth steps are implemented through a Java interface for RDFDB.

```
<Researchers>
…
<Researcher>
        <id>id01</id>
        <position>Research Fellow</position>
        <name>Jason Noble</name>
        <homepage>http://www.comp.leeds.ac.uk/jasonn/</homepage>
        <email>jasonn@comp.leeds.ac.uk</email>
        <publication>
                <id>id233</id>
                <title>Conditions for the evolution of mimicry</title>
                <year>2002</year>
        </publication>
      </Researcher>
…
      </Researchers>
```

Figure 5-5A The original XML files

[9] http://kmi.open.ac.uk/projects/akt/MnM

[10] http://annotation.semanticweb.org/tools/ontomat

```
 <?xml version="1.0"?>
<xsl:stylesheet xmlns:xsl="http://www.w3.org/1999/XSL/Transform" version="1.0">
<xsl:output method="xml" indent="yes"/>
<xsl:template match="Researchers">
<rdf:RDF xml:lang="en" xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
        <xsl:apply-templates select="Researcher"/>
</rdf:RDF>
</xsl:template>
<xsl:template match="Researcher">
        <xsl:variable name="Rid" select="Researcher/id"/>
        <Person rdf:about="{$Rid}">
                <name>
                        <xsl:apply-templates select="name"/>
                </name>
                <position>
                        <xsl:apply-templates select="Position"/>
                </position>
                <xsl:if test="author_of">
                        <author_of>
                            <xsl:apply-templates select="author_of"/>
                        </author_of>
                </xsl:if>
                <Email>
                        <xsl:apply-templates select="email"/>
                </Email>
                <xsl:if test="homepage">
                        <homepage>
                            <xsl:apply-templates select="homepage"/>
                        </homepage>
                </xsl:if>
        </Person>
 </xsl:template>
 <xsl:template match="author_of">
        <xsl:for-each select="Publication">
                <xsl:variable name="Pid" select="id"/>
                <Publication rdf:about="{$Pid}">
                        <Pub_title><xsl:value-of select="title"/></ Pub_title>
                        <YearOfPub><xsl:value-of select="year"/></YearOfPub>
                </Publication>
         </xsl:for-each>
 </xsl:template>
</xsl:stylesheet>
```

Figure 5-5B XSLT template file, which is used to transform data from XML to RDF

<id01,	*rdf:type,*	*'Person'>*
<id01,	*position,*	*'Research Fellow'>*
<id01,	*name*	*'Jason Noble'>*
<id01,	*homepage*	*'http://www.comp.leeds.ac.uk/jasonn/'>*
<id01,	*email*	*'jasonn@comp.leeds.ac.uk'>*
<id01,	*author_of*	*id233>*
<id233,	*Pub_title*	*'Conditions for the evolution of mimicry'>*
<id233,	*YearOfPub*	*'2002'>*

Figure 5-5C A set of RDF triplets after translation:

Figure 5-5 An example of using template rules to transfer XML files into RDF triples

One problem of integrating pieces information from diverse data sources is how to identify that the two descriptions refer to the same object. RDF is good in that URIs can be used to unambiguously denote objects so that multiple data sources can be joined through the same URIs. This is one important reason why RDF data model is chosen rather than XML model. A "*semantically meaningful object identifier*" [Papakonstantinou *et al.,* 1996] is important in this context. For example, the email address of a person can be considered as an identifier because if two people have the same email address, then these two people normally be the same person. However, in many cases, not all resources can be easily given a URI and any given piece of RDF might mention a resource "in passing" without bothering to mention the URI name for that resource. This is called "anonymous node" or "anonymous resource" [Brickley, 2001]. In the above example, publication id233 is an example of temporary id for a resource. The same publication can be given a different id (for example, id785) in another resource. In order to avoid misleading the user (the same object was considered as two different objects), a "rename" operation is added, that is, id number is rewritten according to the identifiers for these resources. For each publication, the identifier is the combination of the title and the year of publication. Accordingly, *id233* is changed to '*Conditions for the evolution of mimicry2002*'.

- **Creating expertise profiles and ranking the expertise of potential supervisors** The integrated information of each expert is considered as one document stored in a repository. Through scanning all the documents in the repository the keyword profile of each expert (represented as vectors of keywords) is created by the expertise manager using the vector space model technique (TF-IDF) [Baeze-Yates and Ribeiro-Neto, 1999]. The relevance of each potential supervisor is calculated through the similarity between the profile of each potential supervisor and the extended user query (adding the description of a concept to the original query). The weight attributed to each potential supervisor is then converted into a percentage value by dividing the weight attributed to the individual by the sum of the weights of all the potential supervisors. In addition to keyword profiles, the expertise manager retrieves the relevant concepts according to the domain ontology, these concepts are then confirmed by experts. The confirmed concept(s) of each expert is stored by the expertise manager in a repository.

- **Displaying the semantically integrated information of potential supervisors** This is also implemented in Java. The search results from RDFDB are firstly constructed into a XML file, and then through XSLT into an HTML file which is presented to users.

In order to provide high quality information about each expert, it is essential that accurate information is available. For example, personal homepages and technical reports should be updated annually. In particular, the ULPD database, as a core data source, is heavily relied upon. This is because the data stored in ULPD on individual academics has been validated by the administrator of each department. There are also a number of automated validation processes built into ULPD. For example, one data source held in ULPD is ULRICHs[11], the authorative serials bibliographic database providing details of title and the International Standard Serial Number (ISSN) for journals published throughout the world. If an administrator attempts to input details for a publication type of 'academic journal paper' and indicates an incorrect journal title and/or ISSN then they will be automatically informed of this and provided with the correct details. The other data sources (such as personal homepage and technical reports) are complementary to ULPD in order to provide a richer description of each expert. If there is conflict between the ULPD database and other departmental source, then the ULPD database takes precedence. The duplicate information (for example, the same information about a paper stored in different places) will be deleted according to the predefined rules. For example, if the two papers have been published in the same year with the same title, then it is assumed that these two papers are the same and only one paper will be displayed in the final presentation.

5.5 System Walk Through

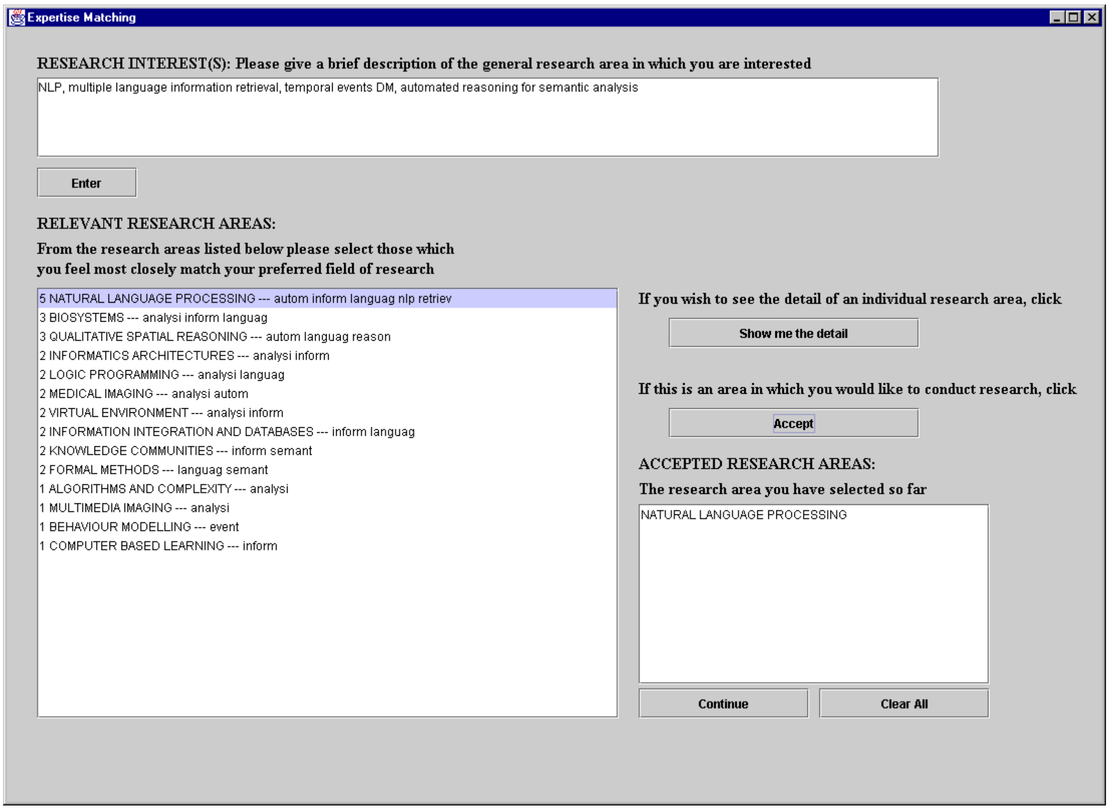
A prototype brokering system is implemented on the architecture. It aims to help PhD applicants find potential supervisors. The search for potential supervisor(s) follows 3 steps which are summarised here and described in more detail below:

[11] ULRICHs http://www.ulrichsweb.com/

1. The user inputs a description of their preferred research interest(s) and selects those individual research areas which are the most relevant.
2. The user views a list of the names of academics working in the relevant research area.
3. The user views the detail of each academic and selects one as their preferred supervisor.

These steps are described below.

Step1 Initially the user inputs a brief description of their general research interests. This description is formulated in natural language. A list of relevant research areas will then be displayed (Figure 5-6). The relevant research areas are ranked according to the number of keywords contained in the research interest field that was entered by the user and which are relevant to each research area. Each result consists of three parts: (i) a value indicating the number of keywords that the user inputs which are relevant to the research area; (ii) the research area which is displayed in upper case; and (iii) a list of relevant keyword stems which are used to search all variants of the same keyword. The user can view the detailed information of each research area by clicking on "Show me the detail" or they can "Accept" the research area if they feel this is an area in which they would like to conduct research. They may accept as many research areas as they wish.



RESEARCH INTEREST(S): Please give a brief description of the general research area in which you are interested

NLP, multiple language information retrieval, temporal events DM, automated reasoning for semantic analysis

RELEVANT RESEARCH AREAS:
From the research areas listed below please select those which you feel most closely match your preferred field of research

5 NATURAL LANGUAGE PROCESSING --- autom inform languag nlp retriev
3 BIOSYSTEMS --- analysi inform languag
3 QUALITATIVE SPATIAL REASONING --- autom languag reason
2 INFORMATICS ARCHITECTURES --- analysi inform
2 LOGIC PROGRAMMING --- analysi languag
2 MEDICAL IMAGING --- analysi autom
2 VIRTUAL ENVIRONMENT --- analysi inform
2 INFORMATION INTEGRATION AND DATABASES --- inform languag
2 KNOWLEDGE COMMUNITIES --- inform semant
2 FORMAL METHODS --- languag semant
1 ALGORITHMS AND COMPLEXITY --- analysi
1 MULTIMEDIA IMAGING --- analysi
1 BEHAVIOUR MODELLING --- event
1 COMPUTER BASED LEARNING --- inform

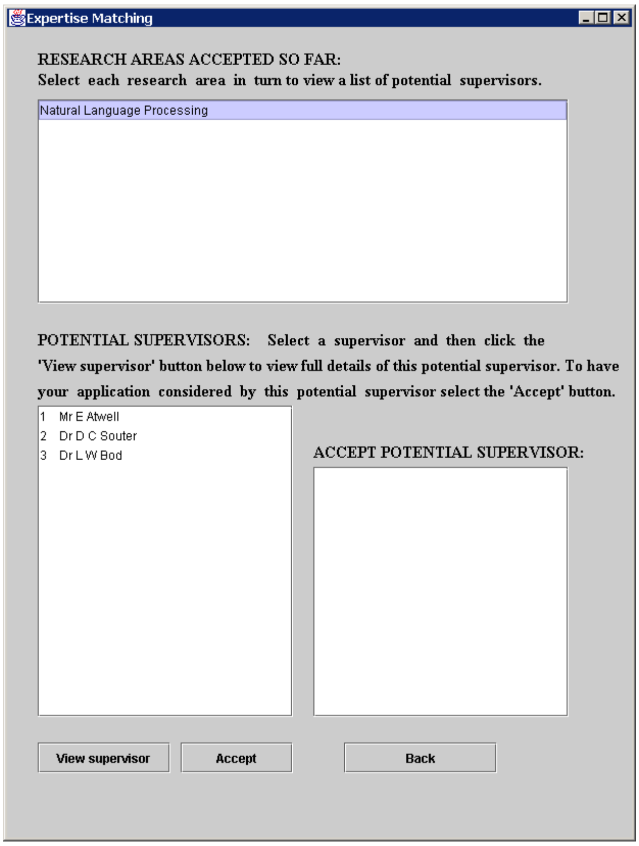
If you wish to see the detail of an individual research area, click

If this is an area in which you would like to conduct research, click

ACCEPTED RESEARCH AREAS:
The research area you have selected so far

NATURAL LANGUAGE PROCESSING

Figure 5-6 Step 1: User interface for inputting research interests



RESEARCH AREAS ACCEPTED SO FAR:
Select each research area in turn to view a list of potential supervisors.

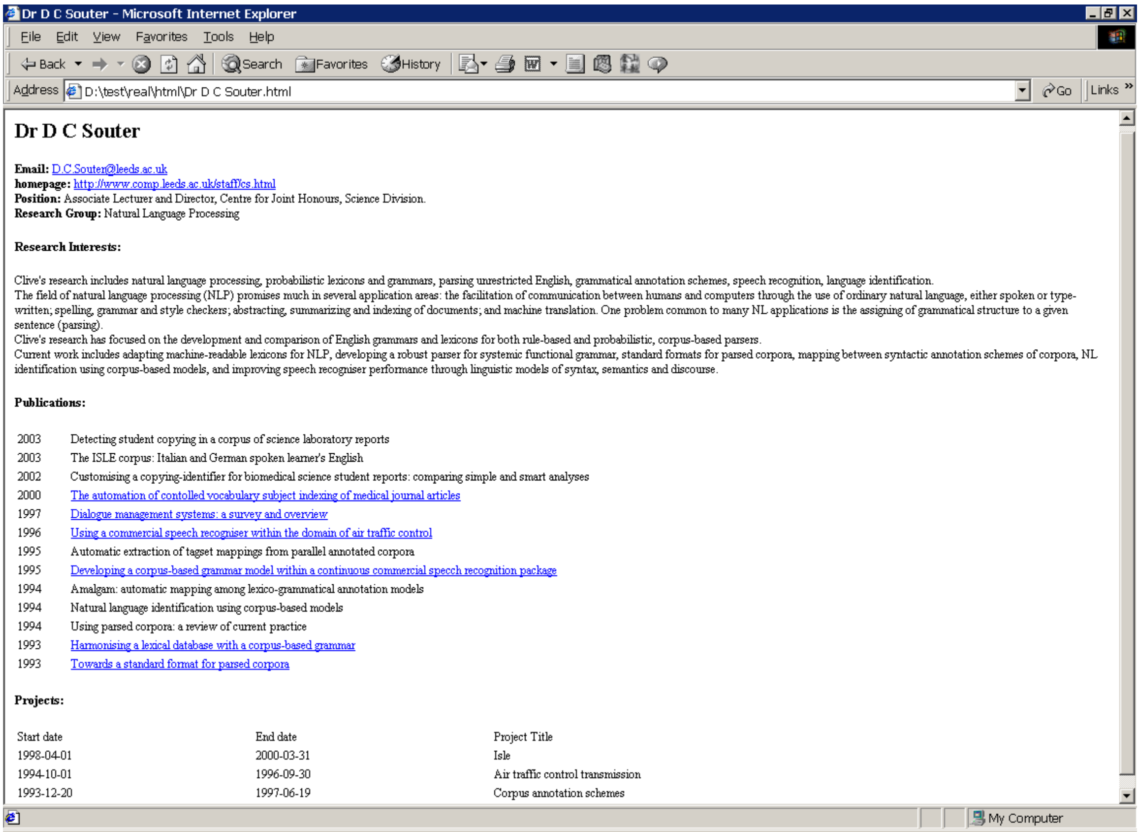
Natural Language Processing

POTENTIAL SUPERVISORS: Select a supervisor and then click the 'View supervisor' button below to view full details of this potential supervisor. To have your application considered by this potential supervisor select the 'Accept' button.

1 Mr E Atwell
2 Dr D C Souter
3 Dr L W Bod

ACCEPT POTENTIAL SUPERVISOR:

Figure 5-7 Step 2: Display the potential supervisor(s) for each preferred research area selected


Dr D C Souter - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address D:\test\real\html\Dr D C Souter.html
Go Links
My Computer

Dr D C Souter

Email: D.C.Souter@leeds.ac.uk
homepage: http://www.comp.leeds.ac.uk/staff/cs.html
Position: Associate Lecturer and Director, Centre for Joint Honours, Science Division.
Research Group: Natural Language Processing

Research Interests:

Clive's research includes natural language processing, probabilistic lexicons and grammars, parsing unrestricted English, grammatical annotation schemes, speech recognition, language identification.
The field of natural language processing (NLP) promises much in several application areas: the facilitation of communication between humans and computers through the use of ordinary natural language, either spoken or type-written; spelling, grammar and style checkers; abstracting, summarizing and indexing of documents; and machine translation. One problem common to many NL applications is the assigning of grammatical structure to a given sentence (parsing).
Clive's research has focused on the development and comparison of English grammars and lexicons for both rule-based and probabilistic, corpus-based parsers.
Current work includes adapting machine-readable lexicons for NLP, developing a robust parser for systemic functional grammar, standard formats for parsed corpora, mapping between syntactic annotation schemes of corpora, NL identification using corpus-based models, and improving speech recogniser performance through linguistic models of syntax, semantics and discourse.

Publications:

2003	Detecting student copying in a corpus of science laboratory reports
2003	The ISLE corpus: Italian and German spoken learner's English
2002	Customising a copying-identifier for biomedical science student reports: comparing simple and smart analyses
2000	The automation of contolled vocabulary subject indexing of medical journal articles
1997	Dialogue management systems: a survey and overview
1996	Using a commercial speech recogniser within the domain of air traffic control
1995	Automatic extraction of tagset mappings from parallel annotated corpora
1995	Developing a corpus-based grammar model within a continuous commercial speech recognition package
1994	Amalgam: automatic mapping among lexico-grammatical annotation models
1994	Natural language identification using corpus-based models
1994	Using parsed corpora: a review of current practice
1993	Harmonising a lexical database with a corpus-based grammar
1993	Towards a standard format for parsed corpora

Projects:

Start date	End date	Project Title
1998-04-01	2000-03-31	Isle
1994-10-01	1996-09-30	Air traffic control transmission
1993-12-20	1997-06-19	Corpus annotation schemes

Figure 5-8 Step 3: Display detailed information on the selected potential supervisor

Step2 The user can select any relevant research area in order to view a list of potential supervisors working in that research area (as shown in Figure 5-7). The potential supervisors are ranked according to how likely it is that this person will be selected as the potential supervisor. The example shown in Figure 5-7 indicates that Mr E. Atwell is the most relevant expert and very likely to be selected as the potential supervisor, Dr D.C. Souter has less expertise than Mr E. Atwell but more expertise than Dr L.W. Bod.

Step 3 The detailed information of the potential supervisor (as shown in Figure 5-8) will be displayed if the user clicks on "View supervisor". The full detail page of Dr D.C. Souter appears just like a standard personal homepage which might currently exist in the School of Computing, but in fact the information is taken from different data sources. As shown in Figure 5-8, the data is retrieved as follows: (1) The personal contact information and research interests are retrieved from the personal homepage; (2) The publication section is a combination of information from the personal homepage and from a series of technical reports which can be downloaded from the ULPD database. The duplicate information is deleted and the final results are reorganized into a consistent format so that the user is not aware that this data has come from disparate sources; (3) The project information is also retrieved from the ULPD database.

5.6 Evaluation

In the evaluation process, the extended Expertise Matcher (Search A, which is named Search 3 in the Chapter 4) was used as the baseline against which to judge the Expertise Locator system (Search B). In Search A, the algorithm used to calculate the similarity between the expert's profile and the user's query is as follows:

$$sim(p, q) = \sum_{i=1,n}(t_{ip} * t_{iq}) \, / \, \sqrt{\sum_{i=1,n} t_{ip}^2 * \sum_{i=1,n} t_{iq}^2}$$

where t_{ip} is the weight of the i^{th} term in an expert's profile p, and t_{iq} is the weight of the i^{th} term in the query q. t_{iq}=1 if the i^{th} term appears in the user's query, otherwise t_{iq}=0. In Search A, experts' profiles are calculated through their publications and projects.

Search B uses the same algorithm except that t_{ic} (the weight of the i^{th} term in a concept c) is used instead of t_{iq} in order to calculate the similarity between the expert's profile and the concept that the user specifies. t_{ic}=1 if the i^{th} term appears in the concept description, otherwise t_{ic}=0. In Search B, experts' profiles are calculated based on their integrated personal detailed information, this includes their research interests, their publications and projects, and technical reports.

The success of the Expertise Locator system (Search B) is measured in terms of whether the Expertise Locator system achieves the following benefits when compared with Search A: (i) Saves time in locating experts; (ii) Improves the accuracy of the search results; (iii) Provides richer descriptions of individual experts for selection purposes. To measure this, the following questions need to be answered, including:

- How long does it take to find potential supervisors in each search?
- How many people in the returned list can be potential supervisors?
- How useful is the content of each potential supervisor's detail page in terms of expertise assessment?
- How useful is the ranking in each search?
- Which search is preferred by the participants (keyword or concept)?

The experiment was conducted in the School of Computing, University of Leeds. Participants of the experiment were asked to volunteer from the current PhD students[12] in the School. 50% of all the current PhD students attended the experiment. They ranged from 1st year to 3rd year and their research interests were very varied (in fact, their research areas covered all the possible research groups in the School). Participants were asked to compare between two searches and they were given full instructions as well as demonstration. Participants started with their

[12] The accuracy of the expertise matching relies on: (1) Whether the retrieved people are relevant experts in the specific area; and (2) Whether the ranking order of the retrieved experts is appropriate, in other words, the expert with more expertise is ranked higher than those with less expertise. Users need to have a certain background knowledge in order to answer these two questions. Although PhD applicants are the real users of the brokering system, it is found that they are less suitable to test the system than the current PhD students. This is because the current PhD students have more knowledge in their specific area and they know the relevant experts in the School and have more ability in judging experts' expertise. This is confirmed through interviews with individual PhD students. Therefore, in the evaluation process, participants are current PhD students rather than PhD applicants.

research interests that they input in their application forms and then conducted a search for potential supervisor(s) from two sets of results returned by Search A and Search B. After that the participants were encouraged to give their thoughts on the brokering system. After the feedback sessions each participant was asked to complete an evaluation form assessing the utility and perceived usability of the system, and the ways in which the brokering system performs better or worse than the extended ULPD expertise matcher.

Table 5-2 Results obtained using Search A (keyword searching)

Participant	**Number of potential supervisors found**	**Number of final accepted potential supervisors**	**Position of accepted potential supervisors in the list**	**Position of the actual supervisor in the list**
1	22	2	1st, 3rd	3rd
2	24	4	1st, 2nd, 8th, 9th	1st
3	27	5	4th, 6th, 11th, 13th,16th	13th
4	21	3	5th, 10th, 11th	11th
5	16	None	none	none
6	15	3	2nd, 5th, 7th	7th
7	19	2	2nd, 13th	2nd
8	25	4	1st, 3rd, 4th, 17th	1st, 17th
9	26	3	1st, 5th, 19th	19th
10	23	3	3rd, 13th, 16th	13th
11	23	1	2nd	2nd
12	25	4	3rd, 8th, 9th, 14th	8th (1/2)[13]
13	23	2	1st,15th	1st
14	25	3	1st, 2nd, 4th	2nd (1/2)
15	27	5	1st, 6th, 7th, 10th, 12th	1st
16	12	2	2nd, 10th	Not found
17	28	2	2nd, 4th	2nd, 4th
18	20	2	3rd, 4th	4th (1/2)
19	25	5	1st, 2nd, 3rd, 5th, 16th	1st, 2nd
20	7	3	1st, 4th, 6th	1st

[13] (1/2) means only one supervisor is found, the joint supervisor is not retrieved.

Table 5-2 shows the results of selecting relevant experts from results returned by Search A. For example, No. 1 participant found 22 potential supervisors in the list after he input his research interests. Among these 22 potential supervisors, 2 were selected as relevant potential supervisors, and they were positioned 1st and 3rd on the list. The actual supervisor of the student was found and was positioned 3rd on the list.

From Table 5-2 it can be seen that there were a large number of potential supervisors returned by the system in most cases. The only way for the participants to evaluate the potential supervisors on the list was to check each person’s publication and project titles as extracted from the ULPD database. Participants started to lose patience after they had checked about 7 or 8 potential supervisors. Under this situation, ranking was very important in order to list the most relevant potential supervisors on the top of the list. Unfortunately, the testing results showed that the ranking was not correct and not useful in helping participants locating the potential supervisors. From Table 5-2 it can be seen that 45% of actual supervisors were positioned below 10th position on the list or not found at all. As a consequence, it is no surprise that 55% of participants believed that the ranking was incorrect and not useful; whilst 40% of participants thought that the ranking was partially useful (see Figure 5-9). The precision of Search A was calculated by dividing the number of accepted potential supervisors by the total number of potential supervisors (see Figure 5-10). The average precision of Search A was 14.6%. If the number of returned potential supervisors was limited to 10, then the precision was increased to 22.1%.

Table 5-3 shows the results of selecting relevant experts from the results returned by Search B[14]. For example, No. 1 participant found 2 research areas relevant to his research interests. There were 4 potential supervisors associated with the first research area and another 2 potential

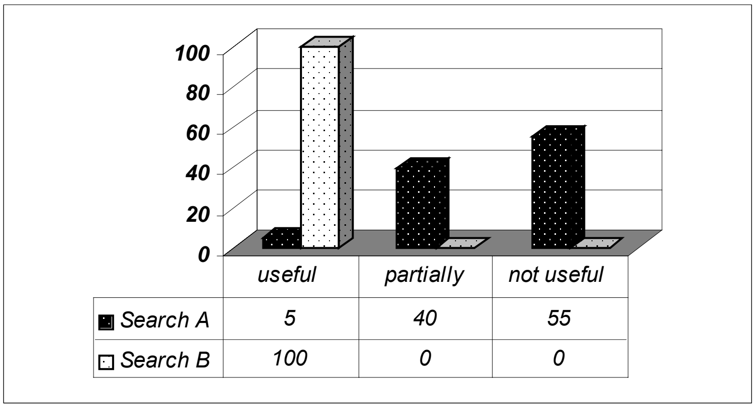
[14] In the column “No. of potential supervisors accepted and their positions in each list”, the actual supervisor was highlighted in Bold and Italic, where the same supervisor appeared more than once in the list, they are marked by underlining in a particular style, e.g., “_” or “﹏”. For example, No.5 participant chose 2 experts as potential supervisors in each research area. The first expert in the first research area is also listed in the second place for the second research area (marked with “_”); and the second expert in the first research area is also listed on the top of the second research area (marked with “﹏”). Both of the experts are the actual joint supervisors for the participant (highlighted in Bold and Italic).

Table 5-3 Results obtained using Search B (concept searching)

Participant No.	No. of relevant research areas	No. of potential supervisors for each research area	No. of potential supervisors in total	No. of potential supervisors accepted and their positions in each list	No. of final accepted potential supervisors
1	2	4	5	2---{1st, ***2nd***}	2
		2		1---{1st}	
2	3	2	4	2---{***1st***, 2nd}	4
		1		1---{1st}	
		1		1---{1st}	
3	1	2	2	1---{***1st***, 2nd}	2
4	1	2	2	2---{***1st***, 2nd}	2
5	2	2	2	2---{***1st***, ***2nd***}	2
		2		2---{***1st***, ***2nd***}	
6	3	2	8	1---{***1st***}	4
		4		1---{***2nd***}	
		2		2---{1st, 2nd}	
7	1	2	2	2---{***1st***, 2nd}	2
8	5	2	6	2---{***1st***, ***2nd***}	2
		2		2---{***1st***, ***2nd***}	
		4		1---{***3rd***}	
		2		1---{***2nd***}	
9	3	1	7	1---{***1st***}	3
		2		0	
		4		2---{***1st***,2nd}	
10	2	1	2	1---{***1st***}	2
		1		1---{***1st***}	
11	1	3	3	3---{***1st***,2nd,3rd}	3
12	3	1	5	1---{***1st***}	3
		1		1---{***1st***}	
		4		2---{1st, ***4th***}	
13	2	3	3	2---{***1st***, 2nd}	2
14	5	2	8	2---{***1st***, ***2nd***}	4
		4		2---{2nd, ***3rd***}	
		2		1---{***1st***, ***2nd*** }	
		2		0	
		2		2---{1st, ***2nd***}	
15	1	2	2	2---{***1st***, 2nd}	2
16	3	2	6	1---{***1st***}	4
		2		2---{1st, ***2nd***}	
		4		2---{2nd, ***3rd***}	
17	2	2	3	2---{***1st***, ***2nd***}	2
		2		1---{***1st***}	
18	3	2	3	2---{***1st***, ***2nd***}	2
		2		2---{***1st***, ***2nd***}	
		2		1---{***2nd***}	
19	3	4	6	2---{***1st***, ***2nd***}	4
		2		2---{***1st***, 2nd}	
		1		1---{1st}	
20	3	3	5	1---{1st}	3
		1		1---{***1st***}	
		1		1---{1st}	

supervisors associated with the second research area (a total of 6 potential supervisors). After checking the detailed information of each potential supervisor, the user selected two potential supervisors who were positioned 1st and 2nd on the list for the first research area. One of the two selected potential supervisors turned out to be his actual supervisor.

From Table 5-3 it can be seen that the number of possible supervisors returned for each participant by the system was reduced. This is because the system searched relevant research areas first which quickly narrowed down the possible relevant supervisors. The accepted potential supervisors (relevant experts) were positioned 1st or 2nd in the list for each accepted research area in most cases. It is noticed that all the actual supervisors of the PhD students were listed (in most cases, they were positioned at the top of the list). It should be noticed that the actual supervisor of each student was selected manually and methodically by the students themselves and the PhD Admissions Tutor together. This means that if the names of the actual supervisors are placed at the top of the results list most of the time then the system is considered to be successful. The precision of Search B was improved with an average precision of 68.7% (see Figure 5-10). The ranking was more appropriate than Search A as 100% of participants believed that the ranking was correct and useful. The differences between the results obtained from Search A and Search B are significant as shown in Figure 5-11, with 95% of participants indicating the results of Search B as more appropriate than those of Search A.



	useful	partially	not useful
■ Search A	5	40	55
□ Search B	100	0	0

Figure 5-9 Usefulness of the rankings in Search A and Search B

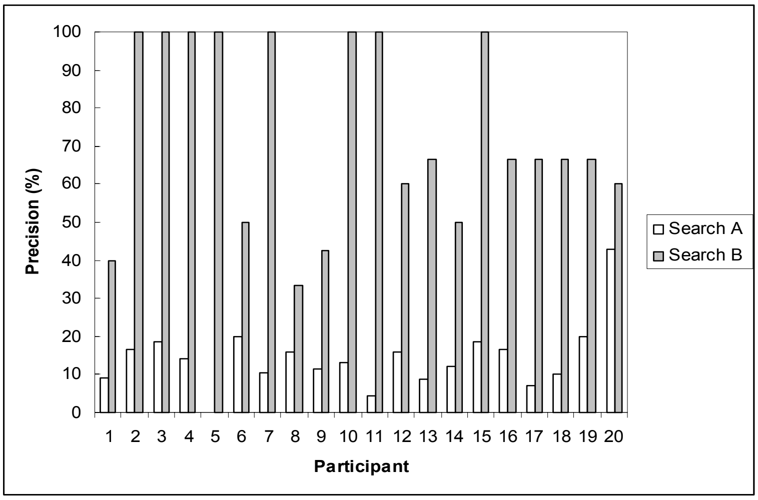


Figure 5-10 Precisions in Search A and Search B

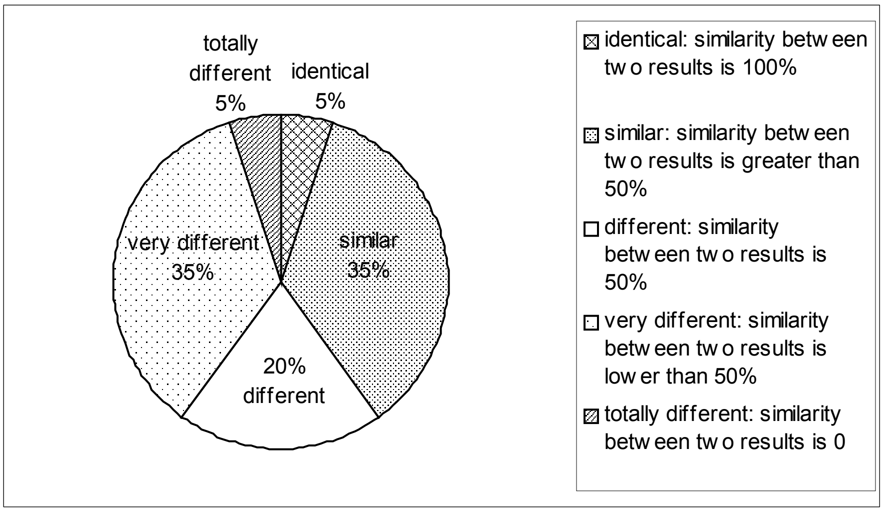


Figure 5-11 The difference between the results of Search A and Search B

In conclusion, Search B provides better performance than Search A in six fields as listed in Table 5-4. A brief discussion of this then follows.

Table 5-4 Comparison of Search A and Search B

Fields	**Search A**	**Search B**
Number of experts retrieved (average)	21.7	4.2
Average time spent on searching (minutes)	8.9	4.6
Precision (average)	14% (22%)	73%
Content information	Limited	Detailed and participants satisfied
Ranking	55% not useful; 40% partially useful	100% useful
Recall	Lower	Higher

- **Number of potential supervisors returned** The number of experts retrieved by Search B is much less than Search A since Search B looked for the relevant research area first. Search B narrowed down the number of potential supervisors.

- **Average time spent on searching** Users spent much less time in Search B than Search A not only because of the fewer experts retrieved, but also because of the more detailed personal information available.

- **Precision** Search B provides higher precision than Search A which means that users have more chance of finding the relevant potential supervisor in Search B rather than in Search A.

- **Content of detailed personal page** It is easier to evaluate the expertise of the potential supervisor in Search B than Search A. In Search A participants can only find the titles of publications and projects, which makes it difficult to assess the expertise of the potential supervisor. In contrast, richer information for each potential supervisor is provided in Search B. Besides the information of personal publications and projects provided in Search A, more detailed information such as personal position, research group membership, research interests, and online downloadable documents are given in Search B. All the participants were satisfied with the detailed personal information provided in Search B.

- **Ranking** The ranking in Search B is more appropriate than in Search A. The reason for this is that the ranking in Search A is based on the keywords input by the user, so some irrelevant researchers may be ranked much higher than an appropriate supervisor only because they have published papers including the particular keyword. In contrast, ranking in Search B is based on the research area (concept), and the profile of each research area is a short document which includes more relevant keywords in this research area. This profile can better present the meaning of the preferred research area than a short list of keywords, so the ranking results are improved.

- **Recall** Search B provides higher recall than Search A. Recall means the ratio of the total number of relevant people retrieved by the total number of relevant people available. Although the number of relevant people retrieved is known, it is difficult to find all the relevant people. However, the total number of relevant people should be the same for both searches, so what is important is which search provides a larger number of relevant people. In Search A, the average number of the accepted potential supervisors is 2.9 (in average) whilst in Search B, the average number is 4.2.

5.7 Discussion

The strengths and weaknesses of the work reported here are compared with related expert finding systems which were described in the Section 3.2.3. The major differences between the Expertise Locator system and other related systems are: (1) Expertise matching is based on the semantic integration of heterogeneous information stored in an organizational memory rather than a single data source such as publications or projects; (2) The hybrid approach combines the advantage of flexibility of keyword search and accuracy of ontology-based search. Although ontology-based search can quickly narrow down the relevant experts, it cannot distinguish one expert from another. In contrast, the vector space model is good at ranking the expertise of experts, but a syntax search may bring some irrelevant experts into the results; (3) The output presentation of experts in most experts finder systems is quite simple, only “expertise identification” [McDonald and Ackerman, 1998] is targeted. In the Expertise Locator system “expertise selection” is supported by providing high quality information relevant to each expert.

5.8 Conclusions

This chapter discusses how to apply semantic web technology - RDF/RDFS, XSLT, ontologies - to solving the three remaining problems of the extended Expertise Matcher described in Chapter

4. In summary, it provides semantically integrated information from heterogeneous data sources by using RDF/RFDS and an application ontology, a flexible output presentation by using XSLT, and a concept matching by using domain ontology. The evaluation of the prototype system indicates the benefits of using RDF/S in Expertise Matching against the extended Expertise Matcher in the following areas: (1) the accuracy of expertise matching has been improved in terms of precision and recall; (2) more detailed information of each expert can be obtained which facilitates uses in assessing expertise; (3) the burden of maintenance is alleviated since up-to-date information can be automatically extracted from heterogeneous data sources and presented in the final result.

In more detail, the brokering system offers superior expertise matching as a result of the following features:

- Keywords are associated with concepts. This not only increases the accuracy of searching, but also helps users to select the relevant concept(s) even when they are not familiar with the domain structure;

- Experts are ranked based on the combination of concept description and keywords that the user specifies. This combined information includes more relevant keywords which increases the possibility of matching with an expertise profile. This alleviates the problem that arises from users and experts using different words to express similar meaning;

- Clusters experts based on the concept rather than the similarity of experts' keywords profiles. Thus users do not have to find "similar experts" since all the experts relevant to one concept are automatically retrieved;

- Extracts the relevant information of each expert from different data sources and provides the combined results to the users. This helps users to compare the expertise of each expert and make an informed decision.

Due to the limited expressive character of RDFS, the Expertise Locator system did not implement the guiding function (i.e. find the adjacent research area and relevant experts if there are any) in case no expert in the area was specified by the user. This can be improved by using DAML+OIL[15] [Horrocks, 2001] which extends RDF/RDFS with richer modelling primitives to support more reasoning function.

The Expertise Locator system is only designed and tested in a single discipline. To widen the application area, multi-disciplinary expertise matching should be considered due to the increasing requirements of sharing knowledge across disciplines. The next chapter describes the initial attempt in solving this problem.

[15] The query language of DAML+OIL was still in development and was not available when the system was developed.

Chapter 6

Matching Experts with Multi-disciplinary Research Interests

6.1 Introduction

Expertise Matching presented in Chapter 5 is used to help people locate experts and share knowledge within a single discipline. However, there are an increasing number of teams whose members are from different disciplines. They are working together to create new knowledge and this leads to new multidisciplinary subjects such as bioinformatics. These experts whose expertise and research interests span across more than one discipline are called multidisciplinary experts. Expertise Matching should not only support locating single disciplinary experts as in the previous brokering system, but also multi-disciplinary experts. Collaboration between researchers from different disciplines will be facilitated by locating multi-disciplinary experts and this is the first step towards successful knowledge sharing. However, there is very little research on multi-disciplinary expertise matching.

In order to help people find experts with multi-disciplinary research interests, a multi-disciplinary brokering system will be proposed as the extension of the original single disciplinary brokering system which was presented in Chapter 5. This Chapter begins with a brief description of multi-disciplinary research, followed with an analysis of the need for multi-disciplinary expertise matching. The issues that have to be solved for the matching to take place are presented. To better understand the problem, a comparison between single- and multi-disciplinary expertise matching is given in Section 6.3. In Section 6.4, the multi-disciplinary brokering systems requirements are informed through a preliminary study. The expertise domain model is proposed in Section 6.5, together with the initial study. Finally, the suggestions which have emerged from the initial study are detailed.

6.2 Analysis of the Problem

It is difficult to distinguish between multi- and inter- disciplinary research and therefore in this chapter the term "multi-disciplinary research" is used to refer to both multi-disciplinary research and interdisciplinary research. Multi-disciplinary is an adjective describing the interaction among two or more different disciplines. This interaction may range from simple communication of fields to the mutual integration of organising concepts, methodology, procedures, epistemology, terminology, data and the organisation of research and education in a fairly large field[1]. Multi-disciplinary research implies that the research involves knowledge from different disciplines in undertaking tasks of increasing scale, depth and complexity which cannot be solved within a single discipline. Multi-disciplinary experts work in teams to solve specific problems across traditional academic boundaries.

6.2.1 What Prompts Multi-Disciplinary Expertise Matching

The reasons for locating multi-disciplinary experts can be summarised as follows:

- Research is undertaken at the intersection where a number of disciplines come together. An example of multidisciplinary research is geoinformatics[2], which is a collaborative research undertaken by geography and computer scientists. It aims to establish a system of seamlessly operating geoscience data and information network. For this purpose, a robust set of software tools for access, analysis, visualization, and modelling has to be fully integrated. This geoinformatics research overcomes the growing and pressing need for utilizing multi-disciplinary geoscience data sets and tools to fully understand the complex dynamics of geographic systems. Researchers in the Geoinformatics research group at the University of Leeds use computer techniques to study natural systems where there is often a more complex mix of factors acting than in the pure sciences[3]. These

[1] Guidelines for the Preparation and Review of Applications in Interdisciplinary Research http://www.nserc.ca/professors_e.asp?nav=profnav&lbi=intre

[2] Source: http://www.geoinformaticsnetwork.org/

[3] Geoinformatics Research Group http://www.geog.leeds.ac.uk/research/geoinfo/

computing techniques include neural network, spatial data analysis, GIS, simulation and modelling, visualization and so on.

- Nowadays when government or industry propose policy, multidisciplinary projects, and so on, frequently multidisciplinary experts are needed for consulting. For example, the World Bank has received a trust fund for a regional project to promote landfill gas (LFG) recovery and utilization for energy in the Latin America and Caribbean (LAC) Region. The project includes the production of a handbook for the preparation of LFG-to-energy projects. The handbook will give equal emphasis to technical issues, business planning, and financing. The World Bank is therefore requesting multidisciplinary consultants in the following areas: engineering design, construction and operation, energy policy, legislation and regulation, environmental and waste management policy, economic and financial analysis, energy markets and carbon finance[4].

- Frequently major projects have a very broad topic that makes multi-disciplinary experts more appropriate than single disciplinary experts. For example, one project proposed under European Sixth Framework Programme[5] titled “European Research Community Network” intends to build an information technology social network, this network brings together many researchers from different university research groups, research and technology organisations and enterprises (7 countries involved) in order to exploit the significant breadth of competencies, knowledge and resources. The research areas include Knowledge Representation and Engineering Design, Digital Content and Industrial Design, and Intelligent Interfaces and Human Factors. The integration of different research enables rapid and flexible design and introduction of new products that effectively meet the needs of individual citizens while creating wealth and maintaining market share for European businesses. The cooperation partner in this project normally is competitive in a specific research area, and have an ability to understand the fundament of the project. For example, as one of the project cooperation partners, the Keyworth

[4] Source: http://www.worldbank.org/html/opr/busop/December%2030/LFG-to-energy.doc

[5] The Sixth Framework Programme (2002-2006) http://europa.eu.int/comm/research/fpb/index_en.html

Institute at the University of Leeds has experts whose expertise across several areas such as human-artefact integrated affective design, virtual and physical prototyping, and virtual reality environments.

- Multi-disciplinary experts fulfil key roles which break down the boundaries and merge knowledge between subjects. They have the ability to link different disciplines. Through searching for multi-disciplinary experts, single disciplinary experts may discover the associated disciplines where their expertise can be applied. Firstly, researchers who are conducting theory research may find applications to test their hypotheses through locating multi-disciplinary experts. For example, neural network researchers are able to locate an application in flood prediction. Secondly, applied researchers are looking for new techniques from other areas to solve a sophisticated problem which cannot be solved by traditional methods or techniques. For example, geography researchers have made significant advances by employing modelling techniques from computing and applying them to population and migration problems. Furthermore, the single disciplinary experts and multi-disciplinary experts can work in a team so that they learn from each other and create new knowledge and emerge a new multidisciplinary subject.

Unfortunately, there is no system providing such a multi-dimensional searching function. People rely on traditional informal social networks to find multi-disciplinary experts. This kind of social network is based on personal contact between individuals and can have some drawbacks. Firstly, the chance of finding multi-disciplinary experts is very low due to the limited links associated with each person. Secondly, it is inflexible because the tie will be broken if one person leaves.

The Informatics Network at the University of Leeds is one example which was set up because of the limitations of traditional informal social networks. The Informatics Research Institute (IRI) is the hub of a growing Informatics Network which range across computational geography, complex systems, ecology and evolutionary biology, medical physics, health informatics, and bioinformatics. The Informatics Network offers a unique approach to the development of sophisticated computational skills and their application to challenging real world problems from a wide range of domains. Through co-ordinating cross-disciplinary collaboration (e.g. bringing together ecologists and economists, bioinformaticians and artificial intelligence researchers, etc.), and thereby connecting the various informatics communities, the Informatics Network will allow ideas and techniques currently specific to individual domains to percolate through the various informatics research enterprises[6]. In this situation, how to attract experts from other domains to join the Informatics Network is a critical issue. Multi-disciplinary expertise matching will play an important role in locating scientists who are from other domains and can contribute relevant expertise to form new communities.

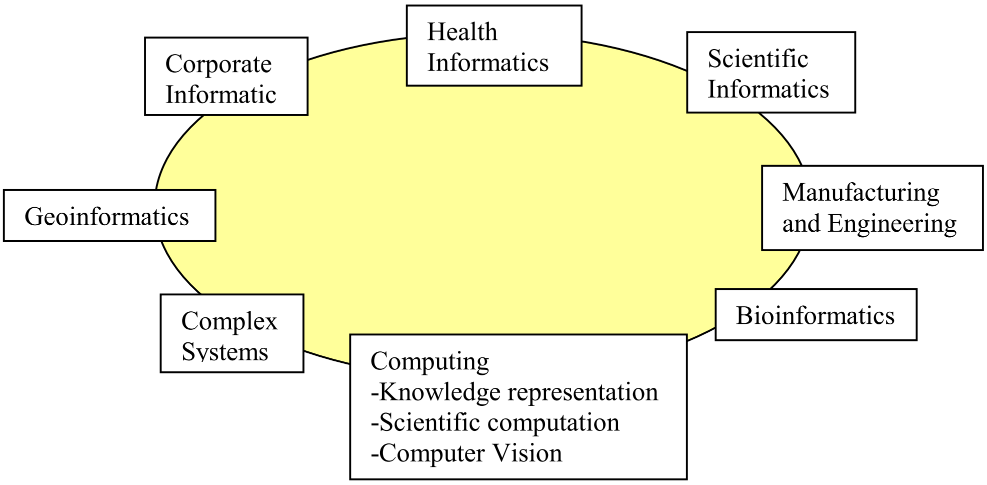


Figure 6-1 Informatics network

6.2.2 Comparison of Single Disciplinary Expertise Matching and Multi-Disciplinary Expertise Matching

As stated in Section 3.2.1, there are 7 domain factors in the Experts Finding Systems domain model, namely: (1) Basis for expertise recognition; (2) Expertise indicator extraction; (3) Expertise models; (4) Query mechanisms; (5) Matching operations; (6) Output presentations; and, (7) Adaptation and learning operations. Among these 7 factors, items (2), (3), (4), (5), (6)

[6] Source: http://www.iri.leeds.ac.uk/overview/network.html

will be different when matching multi-disciplinary experts rather than single disciplinary experts.

- **Expertise Indicator Extraction** Ideally, this should be domain-knowledge driven. For single disciplinary experts, only knowledge of one domain is required. However, multi-disciplinary research areas are very new and continuously changing; there may be no mature domain knowledge available. It has to be a combination of the knowledge from two domains.

- **Expertise Models** The major difference between a single-disciplinary expertise model and a multi-disciplinary expertise model is that the keywords and concepts have to be clustered into groups according to how many disciplines are involved. In addition, the mappings between the concepts in different disciplines have to be built first.

- **Query Mechanisms** When seeking single disciplinary experts, users are required to input keywords from the same discipline. When seeking multi-disciplinary, ideally, users are able to input keywords associated with each discipline. The situation that users may be familiar with only one discipline should be taken into account.

- **Matching Operations** Exact keyword matching or statistical/similar based matching can be used in seeking single disciplinary experts. When seeking multi-disciplinary experts, both experts' profiles and users' profiles should be grouped according to how many disciplines are involved. The matching should then be conducted separately and the separated matching results should be combined in an appropriate way.

- **Output Presentation** In single disciplinary experts matching the experts will be ranked according to their expertise level on a particular concept whilst multi-disciplinary experts matching will have more than one criteria due to the variety of user requirements.

6.3 User Requirements

To establish the multi-disciplinary brokering system requirements, the following preliminary study at the University of Leeds was performed.

The School of Geography at the University of Leeds is one of the largest geography departments in the UK. It consists of 4 research groups in which wide-ranging research is being conducted. The collection, management, analysis, modelling and visualization of spatial data (geodata) with the help of database systems, GIS, image processing systems and so on, has become a very important field of study and practical activity during the last few years. One new research group that is emerging is known as geoinformatics. Although computing techniques play a very important role in these application problems, in practice, few researchers in the geoinformatics research group contact experts from the School of Computing to request their expertise. Some researchers learn the required computing techniques themselves and use what they have learned in the projects they are working on, but this can be very time-consuming and, as geoinformatics researchers are not experts in computing techniques, although they may partly solve the problem using one technique, their implementation may not be the optimum one. Furthermore, computing experts have the expertise but may miss opportunities to use it in real applications.

This kind of separation also brings problems for potential PhD students when they want to apply to this multi-disciplinary research area. As there is no multi-disciplinary department and the potential PhD students are not permitted to indicate “geography and computing” in one application form, they have to choose either the School of Computing or the School of Geography as their target. However, it is not an easy decision for them. Some students may apply to both departments; the problem is that when both departments apply for funding for the same student, they will be told that only one department can proceed. Finally, the students still have to face the problem of choosing only one department. For those potential PhD students who only apply to the School of Geography or the School of Computing, they may miss the more appropriate potential supervisor who may reside in the other department. They may succeed in applying but have to change their research interest slightly according to the research interests of their potential supervisor, but they may not know there are experts in another department with closer matching expertise.

The following two scenarios illustrate the problems faced by computing researchers and PhD applicants.

> *Scenario 1:*
> *Dr. Henderson became a new research fellow in the School of Computing after he finished his PhD one month ago. He did very well in his PhD studies and has proposed a new method in neural networking. He now requires some real data to allow him to evaluate his new method. He spent a long time seeking a suitable application before he met a professor in the School of Geography who is starting to explore neural networking techniques in a problem which has not been totally solved using only Geography techniques for many years. Since the professor is not an expert in neural networking, he is very happy to work with Dr. Henderson.*
>
> *Scenario 2:*
> *Mary is a Masters student at the University of Edinburgh and plans to study for a PhD. Although her background is in geography, she finds that she is increasingly interested in computing and hopes to conduct PhD research in a combined area such as the application of AI-based technologies to hydrological modelling. When she is completing the application form, she does not know which department she should apply to, School of Geography or School of Computing? She searches the webpages of the two schools, but unfortunately she does not find anybody who has the required expertise in both areas. Finally, she considers it may be better for her to apply to both schools and leave the PhD admission tutors to help her select a suitable supervisor.*

Both these scenarios will be addressed in the design of the multi-disciplinary brokering system which aims to improve the process of matching multi-disciplinary experts with potential research students and computing researchers. Both PhD applicants and computing researchers could be benefited in the following ways:

- The applicant could search for multi-disciplinary experts across departments rather than having to browse the web pages of each department individually; the problem of choosing which department to apply to is alleviated.
- It is more likely that the applicants will find the appropriate supervisor themselves; the chances of missing relevant potential supervisors is reduced.
- The conflict arising from two individual departments applying for funding for the same student will be eliminated.
- Researchers who are conducting technique-based research may have more chance of being aware of others who have applied these techniques in solving problems. Based on this awareness, they may build teams and share expertise in future projects.

From the above scenarios, the most significant system requirements of the Brokering System can be identified and summarized as follows:

- Providing multi-disciplinary concept matching rather than simple keyword searching.
- Providing an integrated view of each expert from the diverse information sources in order to help users assess experts' expertise.
- Capturing changes to the expertise profile of experts in order to provide updated information on each expert.
- Ranking each expert's expertise based on several disciplines rather than a single discipline.

It can be seen that the second and third requirements are the same as in the previous experiment which focused on expertise matching within one discipline. However, requirements 1 and 4 are now more complex as several disciplines are involved.

6.4 Proposed Approach

In this section, the architecture of a multi-disciplinary brokering system is proposed to satisfy the user requirements as described in Section 4. In addition, the expertise domain model is recommended which is the core part of the system. The initial studies are also described.

6.4.1 Proposed Architecture

The architecture for multi-disciplinary expertise matching is the same architecture for single-disciplinary expertise matching (presented in the Section 5.3) except the domain ontologies consist of more than one discipline. Figure 6-2 shows a simplified architecture for matching expertise in both Computing and Geography areas. The components which are different from single disciplinary expertise matching are described below.

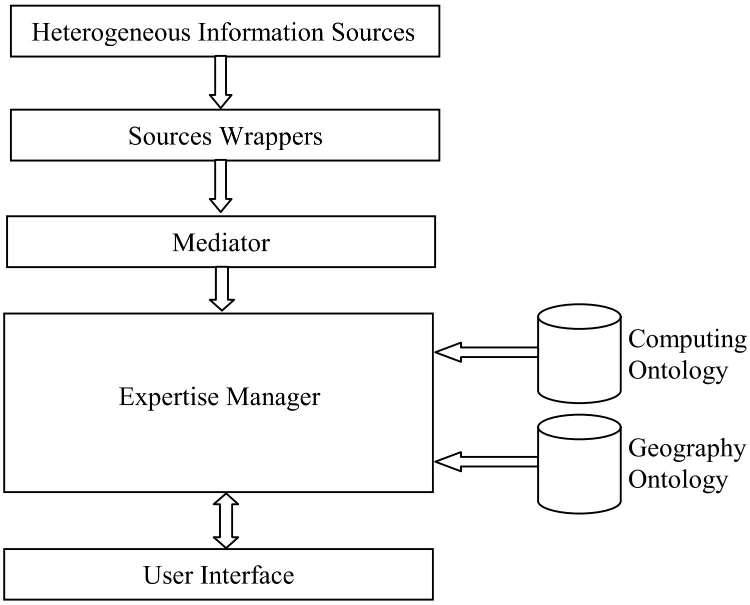


Figure 6-2 Multi-disciplinary expertise matching architecture
(It shows an example of matching expertise with both computing and geography areas)

- **Heterogeneous information sources** These sources are all relevant to identifying the expertise of each expert such as personal homepages, the ULPD database which stores information about publications and projects by members of staff at the University of Leeds, online documents published by either the School of Computing or the School of

Geography on their respective websites, and personal homepage for each expert in both departments.

- **Expertise Manager** Responsible for the overall management of the expertise domain model (see Section 6.5.2 for detailed information) based on the two ontologies supplied. The Expertise Model Manager clusters the experts' expertise profiles (combined keyword profile and concept profile) in groups and executes requests from the user interface.

6.4.2 Proposed Multi-Disciplinary Expertise Domain Model

The Expertise Domain Model described below is the core of the multi-disciplinary expertise matching approach. While the single disciplinary expertise domain model is one dimensional, the multi-disciplinary expertise domain model is two or more dimensional depending on how many disciplines are involved. Table 6-1 shows the expertise domain model for Geoinformatics. This can be obtained through the co-occurrence analysis of linking concepts in Computing and Geographic concepts. In the last Chapter it has been demonstrated that concept matching results in better performance than keyword matching, the same conclusion is also obtained in other research (such as [Brasethvik and Gulla, 2002]). Hence, concepts have been used rather than keywords in the expertise domain model proposed here. One dimension represents computing concepts such as C_1, C_2; the other dimension represents geographic concepts such as G_1, G_2. C_i-G_j means that there is a link between the *i*th computing concept C_i and the *j*th geographic concept G_j, otherwise 0 is displayed. For example, suppose C_2 represents "*neural network*", G_1 represents "*water policy and development*", G_2 represents "*historical geography*", C_2-G_1 means "**neural network technique** for water policy and development" and G_1-C_2 means "**water policy and development** by neural network" technique. The former focuses on a computing technique and latter focuses on geographic application. Not all computing concepts and geography concepts can be combined. For example, there is no connection between "*neural network*" and "*historical geography*". This table can be seen as a central representation of Geoinformatics expertise. The expertise model of each expert can be expressed as a collection of the selected items from the table, for example, {C_2-G_1, C_2-G_3, C_3-G_1}. The domain expertise model can be used to support users in searching and visualizing the expertise information.

Table 6-1 Two dimensional expertise domain model for geoinformatics

		Geographic Concepts				
		G_1	**G_2**	**G_3**	**G_4**	…
Computing Concepts	**C_1**	0	C_1-G_2 G_2-C_1	0	C_1-G_4 G_4-C_1	…
	C_2	C_2-G_1 G_1-C_2	0	C_2-G_3 G_3-C_2	0	…
	C_3	C_3-G_1 G_1-C_3	0	0	0	…
	C_4	0	0	C_4-G_3 G_3-C_4	C_4-G_4 G_4-C_4	…
	…	…	…	…	…	…

6.4.3 Initial Studies

Initial studies include building the expertise domain model, building the expertise profile, locating experts, and ranking their expertise.

6.4.3.1 Building the Expertise Domain Model

In order to produce the 2-D Expertise Domain Model, the concepts in each discipline, and the relations between these concepts need to be extracted and calculated. In theory, an ontology should be used in order to discover the key concepts in the domain, the associated keywords linked to each concept, as well as the relationship between the concepts. Whilst there is no existing Computing ontology or Geography ontology which can be used, the classifications are used instead. On the one hand, the ACM Computing Classification is used; on the other hand, as there is no common geography classification available, a simple classification (as shown in Figure 6-3) for the geographic-applied research areas (relevant to Leeds University) has been created which comprises two levels and seventeen items in total.

1 POPULATION AND MIGRATION
2 SOCIAL GEOGRAPHY
- 2.1 Urban or regional geography
- 2.2 Historical geography
- 2.3 Economic geography

3 ENVIRONMENT MANAGEMENT
- 3.1 Water policy and development
- 3.2 Sustainable development and resource geography
- 3.3 Global environmental history and change
- 3.4 Geomorphology
- 3.5 Hydrology

4 PHYSICAL GEOGRAPHY
- 4.1 Rivers and groundwater
- 4.2 Glaciers and snow
- 4.3 Soil and soil erosion
- 4.4 Tropical Environment
- 4.5 European/Mediterranean Quaternary Environments

Figure 6-3 Classification of geographic research at University of Leeds

In order to determine the relevant computing concepts from the ACM computing classification which are linked to the geography concepts, the topics appearing in recent geocomputation conferences were collected from Internet searches. From these the relevant computing techniques that have been used in solving geographic problems were extracted (These computing techniques are listed in Appendix I). Through analysing the information (such as title, abstract, summary) of previous Geoinformatics projects (22 in total) the co-occurrence between computing concepts and geography concepts has been calculated. In theory, information retrieval techniques can be applied to extract the keywords from the summary of each project, after that, these keywords will be processed using inference mechanisms to induce the key concepts. However, due to the lack of ontology and the limited information sources available, it is difficult to implement this automatically. Some human involvement is inevitable. For example, domain experts are needed to specify the computing techniques and the applied geographic areas associated with each multi-disciplinary project. Figure 6-4 shows one example of identifying the co-occurrence between computing concepts and geography concepts through the summary information for a single multi-disciplinary project.

Project title: Modelling urban residential development using geographic information system

Summary: The study attempts to allocate residential land use development using GIS. Cellular automata model will be integrated with GIS in simulating urban growth and predicting new residential development. The model will then determine the number of houses to be built on developable sites. The model will be tested in selected study areas to examine the impact of certain development to communities.

Applied area(s): POPULATION AND MIGRATION/ Urban or regional geography (from Figure 3)
Computing techniques: Pattern Recognition, Spatial Analysis, Simulating, Modelling, Distributed GIS Environment, Spatial Decision Making (from Appendix I)

Figure 6-4 An example of analysing a multi-disciplinary project

Of the 17 items in the geography classification (Figure 3), only 6 items (1, 2.1, 2.3, 3.1, 3.2, 3.3) are related to computing techniques due to the abstract nature of the classification. After mapping between classification items, the next step is to extract the associated keywords for these items. The keywords associated with the 6 items in the geography classification are collected manually (listed in Appendix J). These keywords will be used in creating the Expertise Model for each expert. For the computing techniques, it is difficult to find the associated keywords relevant to each computing technique, such as simulation and modelling. This is because these computing techniques are already in the lowest level of the computing classification.

6.4.3.2 Building the Expertise Profile

The expertise model should be expressed in two ways – concept representation and keyword representation. Basically, it is the same as for single disciplinary experts matching, however, concepts are selected from the Expertise Domain Model, for example, $\{C_2\text{-}G_1, C_2\text{-}G_3, C_3\text{-}G_1\}$ and keywords are selected from both disciplines. The concept representation is very difficult to implement automatically. The common solution is to ask the multi-disciplinary experts themselves to indicate their relevant expertise from the expertise domain model. The second representation, a set of keywords with weights $\{K_1(w_1), K_2(w_2), \ldots, K_m(w_m)\}$, will be used for ranking experts. The weight of each keyword can be calculated through traditional IR (Vector Space Model) techniques after source wrappers extract the expertise indicators from the heterogeneous information sources. One problem which arises when extracting multi-disciplinary research interests is that sometimes the identified keywords are relevant to only one discipline. For example, computing techniques cannot be easily extracted from the titles of geography publications such as “Release and dispersal of Pb and Zn contaminated mining sediments in an Arctic braided river system”.

6.4.3.3 Locating Experts

There are two ways of locating experts: navigate and search. To navigate, the expertise domain model is displayed to the users so that they are able to browse the system. Users can click on the link between two disciplines, for example, C_i-G_j, in order to obtain a set of experts with multi-disciplinary research interests. To search, users can input keywords expressing their research interests. If the keywords entered by the user are associated with both disciplines, the system then identifies the relevant concepts which are linked to these keywords. These concepts should be confirmed by the user. After that, the system is able to retrieve the link C_i-G_j from the expertise domain model and search for experts with expertise in C_i-G_j. If the keywords entered by the user are only associated with one discipline, the system will highlight the possible concept(s) in this discipline which are relevant to these keywords. The user needs to select one concept which best reflects his/her interests, the system then searches the expertise domain model and returns all the concepts in the other discipline linked to the concept specified by the user.

6.4.3.4 Ranking Expertise

Regardless of whether the user navigates or searches the system, the list of experts returned to the user should be ranked according to their expertise level. However, ranking experts with expertise in more than one discipline is more problematic than within a single discipline. Ranking computing expertise for geography experts and ranking geography expertise for computing experts is a very difficult task since each multi-disciplinary expert each has their own emphasis. Ranking consists of two parts: ranking of the geography applied areas and ranking of the computing techniques. The first is based on concept ranking; the second is based on keyword ranking (as computing concepts are already in the lowest level of the classification and thus operate in the same way as keywords). As described in Section 6.5.3.2, the expertise of each multi-disciplinary expert is represented in two forms. One is a set of concepts; the other is a set of keywords with weights. These keywords are divided into two groups: one set is from the geography domain, the other set is from the computing domain. These two sets of keywords can be seen as two sets of vectors. The ranking in each domain is based on the vector space model which is the same as in single disciplinary expert matching. The combined expertise can be calculated through the vector space model again as shown in Figure 6-5. For example, a user's requirement is "neural network" (0.3) and "water policy and development" (0.7), then expert B is more relevant to the user query than expert A.

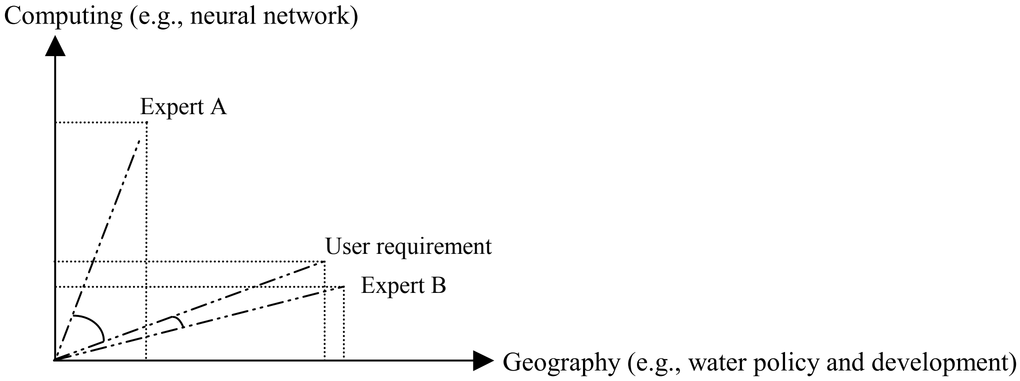

Computing (e.g., neural network)
Expert A
User requirement
Expert B
Geography (e.g., water policy and development)

Figure 6-5 Multi-disciplinary expertise matching using the vector space model

6.4.3.5 Evaluation

Evaluating the multi-disciplinary brokering system is more difficult than in the case of single disciplinary brokering system. Some major reasons are listed below.

- **Precision** This is the critical factor for testing the usability of the system. Precision refers to the percentage of experts returned by the system who are real experts in the multi-disciplinary areas. In order to assess each expert, users should be provided with a complete profile of the individual including their research interests, their publications, and the projects they are working on or have worked on in the past. However, normally, there is more than one expert involved in each project so it is very difficult to identify who plays which role in the project. For example, if four people are involved in a project in which AI techniques are used, can we say that all four people have the same expertise in AI techniques? The answer is most likely to be 'no'. Hence it is difficult for the user to assess the experts returned by the system.

- **Ranking order** Correct ranking order is important especially when the number of multi-disciplinary experts is large. If the single disciplinary brokering system ranks experts in one particular research area, then the multi-disciplinary brokering system can rank experts in each particular research area or in both areas. Suppose there are two multi-disciplinary experts in economic geography and they also have expertise in visualization techniques (computing). It is not too difficult for users to assess who has more expertise in economic geography, but to compare their expertise level in visualization is very difficult as these two experts may put more emphasis on solving the problem in economic geography rather than exploring visualization techniques.

- **Adaptability** This means that if there is nobody in the specified multi-disciplinary areas, the system should be able to provide users with alternative choices. This ability depends heavily on a well-defined ontology. If a classification is used instead of an ontology, the relationships between the classification terms are very limited (only super-class and sub-class). It is not always the case that two classification terms are similar to each other when they share the same super-class. For example, both historical geography and economic geography occur under the classification social geography, however, they are not related directly to each other. On the other hand, computing techniques are more likely to be flat structures. The lack of rich relationships between the concepts results in difficulties in adaptability.

6.4.4 Suggestions

From the initial study it was found that due to the lack of ontology and limited multi-disciplinary projects available, the prototype system was difficult to build and evaluate. However, some suggestions can be given for future research.

Since it may be difficult for domain experts to analyse each multidisciplinary project and publication information, it is recommended that this annotation work can be done by the authors of the publications or the participants of the projects. That is, whenever a new multi-disciplinary project or publication emerges, the author of the publication or the leader of the project provides information about the associated research areas. It seems that this is a tedious job, however, only the authors understand the link between the multi-disciplinary research areas and this annotation process can be supported by tools such as MnM[7] and Ontomat[8]. Mapping between the ontologies is normally considered as identifying the similar concepts in the different ontologies. In the context of this study, the mapping is the combined concepts between different domains. At the beginning, the work may be time consuming since there are very few links acknowledged by the system. So the authors can create links through highlighting the relevant texts as the two research areas. However each time the author identifies a link between the concepts, the system will record this link. Therefore with the increasing number of the multi-disciplinary projects/publications being annotated, the most commonly used concepts and associated links can be identified, the easier to annotate the new projects/publications. This process is similar to the **concept-index** creation process [Nakata *et al.,* 1998] where members of a community highlight the key concept(s) used to describe a document, and the documents in the community memory can be navigated by means of the concept relations.

In the process of building an expertise profile, it is recommended that a concept-based expertise profile is built rather than a keyword-based profile. The reason is that the concept-based profile can be easily built based on the annotation provided by the key authors or the key managers. The sequence of the research areas can be decided by the department that an expert belongs to. For example, if two experts collaborate in the same multi-disciplinary project and/or are the co-authors of multi-disciplinary publications then expertise profile of the expert who is working in the Computing would be $\{C_i\text{-}G_j\}$ whilst the expertise profile for the expert who is working in the Geography would be $\{G_j\text{-}C_i\}$. Building keyword based expertise profile is a long-term goal and cannot be realized in a short time. This is because of the difficulties in identifying the relevant keywords for each concept and also the combined concepts make the keywords ranking less accurate than in a single discipline.

[7] http://kmi.open.ac.uk/projects/akt/MnM

[8] http://annotation.semanticweb.org/tools/ontomat

Users can browse the existing multi-disciplinary concept base in order to specify the relevant concept(s), and the experts are retrieved if their expertise profiles include the specified concepts. The ranking can be based on the number of multi-disciplinary projects each expert has worked on or the multi-disciplinary publications he/she has published. Through this way, the limitations due to the lack of ontology for the immature subject can be overcome.

6.5 Conclusions

This chapter analysed the need for multi-disciplinary expertise matching and discussed the issues that have to be solved for the matching to occur successfully. The modified architecture based on the single disciplinary brokering system was presented. Furthermore, the expertise domain model was detailed and the initial studies also described.

Through investigating the multi-disciplinary brokering system, ontology is found as the most important factor since it influences other operations such as expertise indicator extraction, building of the expertise model, ranking experts, and providing adaptability. For example, if for each concept the sufficient or necessary keywords are defined in the ontology, then the concepts expertise model would be automatically obtained without involvement of each expert. The better the ontology, then the better the results which can be obtained. Consequently, in order to build an effective multi-disciplinary brokering system it is critical to build ontologies first. However, building a formal ontology is difficult, especially when building an ontology for an immature or emerging subject. Based on this fact, an alternative suggestion is given where every author of a multi-disciplinary publication or every member of a multi-disciplinary project contributes to the experts finding system by adding annotation on the associated research areas to each publication and project. It is expected that through this accumulated process, the correspondence mapping between disciplines can be built up.

Chapter 7

Conclusions and Future Work

This thesis started with the view that sharing expertise within and across organizations is very important and expertise matching is the foundation for expertise sharing. The research presented in this thesis focused on an investigation into how expertise matching can be improved within academia. More specifically, it has analysed the limitations of the current ULPD expertise matcher at the Leeds University (which is representative for expertise matching systems within academia) and investigated ways to improve the accuracy of expertise identification and provide support in finding appropriate experts. This final chapter presents the key findings from the investigation, suggests directions for future research, and also discusses the implications of this research.

7.1 Results and Major Findings

From the empirical study of expertise matching undertaken using the real data at the University of Leeds, the following conclusion can be drown.

- Traditional Information Retrieval model (in particular, the vectors space model) is still useful in ranking expertise. Through the first experiment (comparison of the extended Expertise Matcher with the current ULPD Expertise Matcher, presented in Chapter 4) it was found that if the retrieved experts were not ranked according to their expertise level then the number of the returned experts could not be controlled and users had to check each returned expert. This places a significant burden on users. It was also found that most users are not usually able to express their query requirements in the form of a Boolean query. To solve these two problems, the vector space model was employed to build both a user's profile and an expert's profile and to calculate the similarity between these two profiles (as presented in Chapter 4). In this way, a user can easily form a query (in a few keywords) and experts with more expertise are more likely to be displayed at the top of the list.

- Semantic web technologies (RDF, RDFS, ontologies) are good candidates for the integration the multiple expertise indications. As analysed in Chapter 3, expertise is different to explicit information such as documents; whilst documents are static, independent[1], and explicit; expertise is dynamic, hidden in the "heads" of experts, and reflected in many things. In order to obtain an updated and high quality expertise model, multiple expertise indications have to be explored. An expertise conceptual model (application ontology) was created to integrate the expertise indications (as presented in Chapter 5). RDFS is used to specify the classes and properties in the expertise model. RDF provides a uniform representation so that different data sources can be integrated. This integration has two roles: (1) It improves the quality of expertise profile, and (2) It helps users in assessing experts' expertise. These two features are special when compared with most approaches where only one expertise indication is used to determine experts' expertise and the output presentation of the expert's detailed information is very simple.

- The combination of keyword based expertise model and concept based expertise model is an important contribution towards expertise identification. Concept search (ontology-based, thesaurus-based) is normally more accurate than keyword searching in both precision and recall [Khan, 2000]. However, experts who are associated with a concept are considered to be equal in their expertise level which makes the selection difficult, especially when the number of experts is large. In this study, expertise model has been extended to include both keyword-based representation and concept-based representation. A domain ontology is built to link the concepts with the relevant keywords and help experts and users in selecting the relevant concepts. The extended expertise model combines the ranking ability of keyword search and accuracy of concept search, and

[1] Independent does not mean there is no link with other documents; here it means a document can exist on its own.

therefore leads to the improved performance of expertise matching (not only more relevant experts are retrieved but also they are listed in a relevance order).

An architecture for supporting both the application ontology and the domain ontology has been proposed and a prototype system (Expertise Locator) has been developed based on this architecture (as presented in Chapter 5). An experimental study with the system has been conducted in the Computing domain in order to discover the advantages and problems of this approach (see Section 7.2). To a reasonable degree the objectives (i.e. to support expertise identification and expertise selection) have been achieved. The precision and recall of expertise matching have been improved significantly and users were satisfied about the output presentation of the details of retrieved experts.

In the process of extending the single disciplinary expertise matching to multi-disciplinary expertise matching, it is found that not only matching itself is more complicated, personal desire and political reasons may even hinder the collaboration between multi-disciplines. Social navigation is still preferred. The results of the investigation also gave valuable insight to the problems of matching people with multi-disciplinary expertise; it is argued that some problems can only be solved as the need for multi-disciplinary research grows and the understanding of how to classify multi-disciplinary research grow. This thesis makes a good start.

7.2 Future Directions

There are several directions in which this research might be extended. These directions can be divided into six areas: improved expertise model, visualization support, reasoning support, improved user control, communication support, and information extraction support. Each of these areas will be discussed below.

7.2.1 Improved Expertise Model

The expertise model is created based on the collected implicit expertise evidence from diverse data sources. In this study, three different types of evidence were collected. research interests;

publications; and projects. Currently, they are considered to be of equal importance by the expertise manager. However, there are differences between these three. Research interests in personal homepages may be the most important form of expertise evidence since they are declared by the experts themselves; the publications of each expert are also important since they are externally validated by others; and the projects that the experts participated in may be of less importance than their publications. Therefore, different weights could be given to each expertise indication. For example, research interests (1.0), publications (0.8), and projects (0.6). Although the optimal values of the weights are difficult to determine, machine-learning techniques can be used to adjust these weights automatically based on a significant amount of user feedback.

Another way to improve the expertise model is to divide publications into several categories according to their quality (this can be roughly derived from where they are published). There is a clear difference between one expert who has published two papers in a world-leading journal and another with two papers in national conferences. The expertise manager may assign different weights to the different types of publication before building the expertise model.

Compared with the keywords profile, the creation of a concept profile still needs the involvement of experts. This process can be simplified by exploiting natural language processing techniques. Basically if the associated concepts of each document can be identified, then the author(s) will be automatically linked to these concepts. Duan [2002] used a lexical knowledge based method for meaning trend representation or theme representation. It is worthwhile to examine the effectiveness of the lexical knowledge based method in identifying the relevant concepts for a publication.

7.2.2 Visualization Tool

Instead of forming a query, users can browse a hierarchical classification or ontology to find the areas of interest. Although it seems a good way to begin searching for experts, it can be time consuming if users take the wrong paths through the ontology. A field study undertaken by Reimer [Reimer *et al.,* 2003] indicates that users are not willing to browse the ontology especially if it is large; a similar result is obtained from interviews with the participants of the experiment. Most participants felt that it was more convenient for them to input keywords first and then to select the query context – a concept from a list. However, some participants made comments that they would like to see a "small ontology space" returned relevant to each concept, in other words, display where the concept is positioned in the classification or indicate the related concept (for example, broader concept, narrow concept, similar concept) in the ontology. A visualization tool which provides access to a local map may help in this respect.

7.2.3 Reasoning Support

RDF provides a data model for describing machine-processable semantics of data. The basic semantics is specified by RDFS, which can be regarded as a very simple ontology language since it introduces basic ontological modelling primitives (class, subclass, subproperties, domain and range restrictions of properties). However, RDFS cannot provide enough semantic support due to limited expressivity (many types of knowledge cannot be expressed in this simple language, such as min, max, string, number, constraints). This results in limited reasoning opportunities. Knowledge representation languages such as DAML+OIL extend RDF and add more primitives to define precise semantics and support reasoning. Tools which support DAML+OIL are becoming available, such as Sesame[2]. Consequently one of the future directions of this research is to use DAML+OIL to support adaptive matching. For example, when there is no expert in the specified area, the related areas (such as broader or narrower areas) are automatically searched to find experts. In addition, more knowledge can be obtained using inference rules. For example, finding the collaborators of an expert is possible when a rule is added such as "if two different people work in the same project, then these two people are considered to be collaborators".

[2] http://sesame.aidministrator.nl/

7.2.4 Improved User Control

The current expertise matching focuses on locating experts in a single area. However, not all the users can easily find the single concept from the ontology which exactly matches their requirements. Sometimes, what they need is a joint concept, for example, the combination of concept *a* and concept *b*. In these cases, users have to conduct a search for experts in each concept respectively and perform analyses to obtain the answer. One future direction is to allow users to perform advanced searches using logical operators (AND, OR, NOT). For example, users can search for experts with expertise in “visualization” and “virtual environment”, and the experts with expertise in a single research area will not be displayed so that it is quicker for users to locate the experts they need. Furthermore, users can be supported in expressing their preferences by assigning different weights to each area using the vector space model again.

The theme of this thesis is expertise matching, and the assumptive question is “who are the experts in area X”. The conceptual model which is used as the semantic backbone to integrate information is hidden to users. One future direction is to make this conceptual model visible to users (as semantic interface) so that they know what kind of information is stored and they can conduct a more complicated query based on this conceptual model. For example, “show me the experts in ‘natural language processing’ and their current projects.”

7.2.5 Communication Support

The major aim of expertise matching is to support expertise identification and expertise selection, in other words, support users in identifying the most appropriate expert to contact. It could be extended by adding facilities to support people connection such as email or Netmeeting so that users can send their questions to the selected experts or even talk to them directly via the Internet. However, experts are normally quite busy and it is not feasible for them to accept all communication requests. Therefore, access might be controlled by the experts themselves. An alternative solution is to integrate the expertise matcher with other knowledge management systems which provide collaborative tools such as Virtual Knowledge Park.

7.2.6 Information Extraction Support

Many expertise indications are available from the web such as personal homepages, publications and projects descriptions, and so on. In order to integrate the multiple expertise indications, wrappers were used to extract relevant information from web documents. However, hand-coded wrappers are difficult to build and costly to maintain [Temelkuran, 2003]. Alani and his colleague have presented a new tool to automatically extract information from web documents [Alani *et al.,* 2003a]. This extraction tool is guided by an ontology so that it understands which type of information needs to be extracted even if the web page is changed. It would be sensible to adopt this flexible approach to extract the relevant information.

7.3 Implication

In practice, this thesis contributes to the expertise matching problem within academia. From the survey it can be found that most expert finding systems in academia rely on experts to specify their expertise in keywords or link to a simple classification term. This kind of expertise database always suffers from the keyword search problems in identifying experts and difficulties in maintaining the data. To solve these problems, this thesis has demonstrated an original approach which utilises multiple expertise indications to build expertise profiles. In addition, this thesis provided a conceptual model and an architecture which can be reused by other universities in building experts finding systems.

Recently there is an increasing requirement for expertise matching in industry, especially for identifying and forming communities of practice (as described in Chapter 2). Since most of the evidence discussed pertains to the academic environment, it would be inappropriate to generalize the findings or conclusions directly to the industrial environment. However, expertise matching within academia is expected to be similar to that within those knowledge-based organizations[3]. After examining a number of experts finding systems in industry (such as Expertise Recommender [McDonald, 2000], Referral Web [Kautz *et al.,* 1997a], Expert Finder

[3] Maurino [1995] presents three performance levels of expertise, skills-based, rule-based, and knowledge-based. This thesis focuses on knowledge-based expertise matching only.

[Mattox *et al.,* 1999], see Chapter 3), it can be found that there are some similarities. For example, most of these systems rely on the indications of expertise (such as publications and projects) to retrieve experts. The method of exploring semantic web technologies (RDF, RDFS, ontologies) to integrate multiple expertise indications can be used in knowledge-based organizations to improve the accuracy of expertise matching and output presentation.

There are two types of knowledge management, externalising knowledge and sharing expertise [Ackerman *et al.,* 2003]. Most current knowledge management programmes tend to focus on gathering, organising, and retrieving information. It is noticed that in the AKT project knowledge technologies are developed to interpret information into actionable knowledge, which aims to provide "*the right content to the right place at right time and in the right form*" [Shadbolt and O'Hara, 2003]. However "*not all the knowledge needed in a problem situation can be made explicit or stored in a knowledge base*" and "*there are many occasions where the best answer comes from finding the right person rather than the right information*" [Ehrlich, 2003]. Therefore a true knowledge management solution must address the organization and transfer of both tangible and tacit knowledge [Oakes and Rengarajan, 2002]. This work contributes to the second kind of knowledge management – expertise sharing by retrieving experts with the required expertise. It can be viewed as complementary to many knowledge management projects (such as AKT).

The work on Community Of Practice (COP) and ontology underpins the future development and application of expertise matching. A COP consists of people with common interests who interact with each other to share information and to solve problems in their areas of expertise. Informal COPs are important for the development and sharing of expertise within an organisation. In academia, members of COP can come from different disciplines. O'Hara *et al.* [2002] attempt to identify potential COPs through ontology network analysis. Connections or relations between entities in an ontology can be measured to provide metrics of connectedness. When a person instance has been selected, the close instances in the knowledge base can be identified as the potential COP. However, the connections between entities can be quite arbitrary and entities retrieved may not be in the same COP and the common interest of the identified COP is not discovered. The approach presented in this thesis has the advantage of recommending people who are interested in the same topic and could be potential members of COPs. It facilitates the development of COPs by introducing those people who are not aware of each other before, it also opens a large potential for expertise sharing between the members of COPs.

Ontology is a key technology that allows knowledge sharing and concept-based information retrieval. In this work, the performance of expertise matching is improved largely because of building an application ontology to integrate diverse data sources and a domain ontology to conduct the concept searching. This work contributes to the wider knowledge management agenda (knowledge acquisition, knowledge modelling, knowledge retrieval), in particularly, to the understanding of knowledge management technologies around ontology such as ontology-based information extraction to support knowledge acquisition, construction of ontologies and ontology mapping to support knowledge modelling, ontology-based answering to support knowledge retrieval.

Reference:

[Abecker and Decker, 1999] Abecker, A. and Decker, S. Organizational Memory: Knowledge Acquisition, Integration, and Retrieval Issues in Knowledge-Based Systems, Lecture Notes in Artificial Intelligence, Vol. 1570, Springer-Verlag, Verlin, Heidelberg, pages 113-124, 1999.

[Ackerman *et al.,* 2003] Ackerman, Mark, Pipek, Volkmar, and Wulf, Volker. eds. *Sharing expertise: beyond knowledge management.* Cambridge, MA: MIT Press, 2003.

[Ackerman and Halverson, 1998] Ackerman, M. S. and Halverson, C. (1998) "Considering an Organization's Memory," in *Conference on CSCW'98*, 1998, Seattle, WA, ACM Press, pages 39-48.

[Ackerman and McDonald, 1996] Ackerman, M.S. and McDonald, D.W. (1996) "Answer Garden 2: Merging Organizational Memory with Collaborative Help," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW'96)*, 1996, Boston, MA, pages. 97–105.

[Alani *et al.,* 2003a] Alani, H.; Sanghee Kim; Millard, D.E.; Weal, M.J.; Hall, W.; Lewis, P.H.; Shadbolt, N.R. (2003) "Automatic ontology-based knowledge extraction from Web documents," *Intelligent Systems, IEEE*, Vol.18 No.1, 2003, pages.14-21.

[Alani *et al.,* 2003b] Alani, H.; Dasmahapatra, S.; O'Hara, K.; Shadbolt, N.(2003) "Identifying communities of practice through ontology network analysis," *Intelligent Systems, IEEE*, Vol.18 No.2, 2003, pages. 18-25.

[Appelt, 1999] Appelt, W. (1999) "WWW based collaboration with the BSCW System," in *Proceedings of the 26th Conference on Current Trends in Theory and Practice of Informatics,* Springer-Verlag LNCS 1725, 1999, pages.66-78.

[Arens *et al.,* 1996] Arens, Y. Hsu, C. and Knoblock, C.A. (1996) "Query processing in the SIMS information mediator," in *Advanced Planning Technology*, Austin Tate, Ed, AAAI Press, Menlo Park, California, 1996, pages. 61-69.

[Argyris and Schon, 1996] Argyris, C. and Schon, D.A. (1996) *Organizational Learning II: Theory, Method, and Practice*, Addison-Wesley Publishing Co, Reading, Mass

[Attar and Fraendel, 1977] Attar, R. and Fraenkel, A.S. (1977). "Local feedback in full-text retrieval systems," *Journal of the Association for Computing Machinery*, Vol.24 No.3, 1977, pages. 397-417.

[Baeze-Yates and Ribeiro-Neto, 1999] Baeze-Yates, R. and Ribeiro-Neto, B. (1999) *Modern Information Retrieval*, Imprint Addison-Wesley Longman, 1999

[Bannon and Kuuti, 1996] Bannon, L. and Kuuti, K. (1996) "Shifting Perspective on Organizational Memory From Storage to Active Remembering," in *Proceeding of the HICSS'96*, IEEE Computer Press, pages 156-167

[Bedard, 1991] Bedard, J. (1991) *Expertise and its Relation to Audit Decision Quality*, Contemporary Accounting Research, Fall, pages. 198-222,

[Bekkedahl, 1977] Bekkedahl. C.I. (1977) "Discipline and Profession of Naval Arms," *United States Naval Institute Proceedings*, Vol.13 No.891, May, 1977.

[Bennis and Biederman, 1997] Bennis, W. and Biederman, P. (1997) *Organizing genius: the secrets of creative collaboration*, Addison-Wesley, Reading, Mass.

[Bereiter and Scardamalia, 1993] Bereiter, C. and Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*, Chicago: Open Court.

[Berners-Lee, 1998] Berners-Lee, T. (1998) "Why RDF model is different from the XML model," available online: http://www.w3.org/DesignIssues/RDF-XML.html

[Bishop, 2000] Bishop, K., (2000) "Heads or Tales: Can Tacit Knowledge Really be Managed," in *Proceeding of ALIA 2000 Biennial Conference*, 23-26 October, 2000, Canberra, available online at http://www.alia.org.au/conferences/alia2000/proceedings/karen.bishop.html

[Bishr and Kuhn, 2000] Bishr, Y. and Kuhn, W. (2000) "Ontology-Based Modelling of Geospatial Information," in *the 3rd AGILE Conference on Geographic Information Science*, Finland, May 25- June 2, 2000, available from www.ida.liu.se/edu/fsgis/portal/downloads/Agile2003-modelAbst.pdf

[Brasethvik and Gulla, 2002] Brasethvik, T., and Gulla, J.A. (2002) "A Conceptual Modeling Approach to Semantic Document Retrieval," in *the 14th International Conference of Advanced Information System Engineering CAiSE'02*, 27-31, May, 2002, Toronto, Canada LNCS 2348, pages. 167-182

[Bray *et al.,* 2000] Bray, T., Paoli, J., Sperberg-McQueen, C.M. and Maler, E. (2000) *Extensible Markup Language (XML) 1.0* (Second Edition), W3C Recommendation, October 2000. http://www.w3.org/TR/2000/REC-xml-20001006.

[Brickley and Guha, 2000] Brickley, D. and Guha, R.V. (2000) "Resource Description Framework (RDF) Schema Specification 1.0", in *World Wide Web Consortium*, March 2000, http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

[Brickley, 2001] Brickley, D. (2001) "RDFWeb notebook: aggregation strategies," available online http://rdfweb.org/2001/01/design/smush.html

[Broadbent, 1997] Broadbent, M. (1997) "The Emerging Phenomenon of Knowledge Management," *Australian Library Journal*, Vol.46 No.1, 1997, pages.6-24.

[Broadbent, 1998] Broadbent, M. (1998) "The Phenomemnon of Knowledge Management: What Does it Mean to the Information Profession?" available online at http://www.sla.org/pubs/serial/io/1998/may98/broadben.html

[Bull *et al.* 2000] Bull, K. S., Montgomery, D., and Kimball, S. L. (2000) "Creating Community in the Classroom," in *Quality University Instruction Online: An Advanced Teaching Effectiveness Training Program--An Instructional Hypertext*, K. S. Bull, D. L. Montgomery, and S. L. Kimball, Eds, Stillwater, OK: Oklahoma State University

[Bullard *et al.*, 1995] Bullard, T., Capper, P., Hawes, K., Hill, R. and Tustin, C. (1995) "The Hunting of the Skills," in *ANZAM'95 Conference*, 3-6 December, 1995, Townsville, Australia.

[Busse *et al.,* 1999] Busse, S., Kutsche, R., Leser, U. and Weber, H. (1999) Federated Information Systems: Concepts, Terminology and Architectures Forschungsberichte des Fachbereichs Informatik 99-9, available online http://citeseer.nj.nec.com/busse99federated.html

[Bussell and Holter, 2002] Bussell, J. and Holter, J. (2002) "Do You Know Where Your Intellectual Capital Is?" *ITtoolbox Knowledge Management.* Available online at http://km.ittoolbox.com/documents/document.asp?i=1598

[Capper, 2000] Capper, P.(2000) "Understanding Competence in Complex Work Contexts," in *Competancy Based Education and Training: a world perspective*, Arguelles, A.and Gonczi, A.(eds.) Noriega Editores, Baldaras (Mexico), pages. 147-172

[Carey *et al.,* 1995] Carey, M.J., Haas, L.M., Schwarz, P.M., Arya, M., Cody, W.F., Fagin, R., Flickner, M., Luniewski, A.W., Niblack, W., Petkovic, D., Thomas, J., Williams, J.H., and Wimmers, E.L. (1995) "Towards Heterogeneous Multimedia Information Systems: The Garlic Approach," in *Research Issues in Data Engineering*, March 1995, Los Alamitos, Ca., USA, IEEE Computer Society Press, pages.124-131

[Carroll and Rosson, 1998] Carroll, J.M. and Rosson, M.B. (1998) "Network Communities, Community Networks," CHI 98 conference summary on Human factors in computing systems, pages.121-122, April 18-23, 1998, Los Angeles, California, United States.

[Choo, 2000] Choo, C.W. (2000) "Working With Knowledge: How Information Professionals Help Organizations Manage What They Know," *Library Management*, Vol.21 No.8, 2000

[Christophides, 2000] Christophides, V. (2000) "Community Webs (C-Webs): Technological Assessment and System Architecture" Research report, http://cweb.inria.fr/Resources/architecture3.pdf

[CIO, 2002] CIO (2002) "Knowledge Paradox - How to Manage Your Most Strategic Asset," CIO Information Network, available online at http://www.kmadvantage.com/ km_articles.htm

[Clancey, 1995] Clancey,W.J. (1995) "A boy scout, Toto, and a bird: How situated cognition is different from situated robotics," in *The "Artificial Life" Route to "Artificial Intelligence": Building Situated Embodied Agents*, L. Steels and R. Brooks, Eds, Hillsdale, NJ: Lawrence Erlbaum Associates.pages.227-236.

[Clark, 1999] Clark, J. (1999) "XSL Tranformations (XSLT)," W3C Recommendation November, 1999, http://www.w3.org/TR/xslt

[Clark *et al.,* 2000] Clark, P., Thompson, J., Holmback, H. and Duncan, L. (2000) "Exploiting a Thesaurus-Based Semantic Net for Knowledge –Based Search," in *Proceedings of 12th conference on Innovative Applications of AI (AAAI/IAAI'00)*, pages. 988-995

[Cohen and Levinthal, 1990] Cohen, W.M. and Levinthal, D.A., (1990) "Absorptive Capacity: A Perspective on Learning and Innovation," *Administrative Science Quarterly*, Vol.35 No.1, pages. 128-152

[Cohen *et al.,* 1998] Cohen, A.L., Maglio, P.P., and Barrett, R. (1998) "The Expertise Browser: How to Leverage Distributed Organizational Knowledge," presented at *The Workshop on*

Collaborative and Cooperative Information Seeking in Digital Information Environments at CSCW'98, 1998, Seattle, WA.

[Costello, 2000] Costello, D. (2000) "For Knowledge, Look Within - Businesses are discovering the value of internal infomediaries," *Knowledge Management Magazine*, September 2000, available on line at http://www.destinationkm.com/articles/default.asp?ArticleID=563

[Cross and Baird, 2000] Cross, R. and Baird, L., (2000) "Technology Is Not Enough: Improving Performance by Building Organizational Memory," *MIT Sloan Management Review,* Spring 2000, Vol.41 No.3, pages 69-78.

[Crossley, 1999] Crossley, M., Daview, N., McGrath, A., Rejman-Greene, M. (1999) "The Knowledge Garden," *BT Technology Journal*, January 1999, Vol.17 No.1, pages 76-84.

[Crowder *et al.,* 2002] Crowder, R., Hughes, G. and Hall, W. (2002) "An Agent Based Approach to Finding Expertise," in *Fourth International Conference on Practical Aspects of Knowledge Management*, 2-3 December, 2002, Vienna, Austria

[Cui *et al.,* 1999] Cui, Z., Tamma, V. and Bellifemine, F. (1999) "Ontology management in Enterprises," *BT Technology Journal*, October, 1999, Vol.17 No.4, pages 98-107

[Davenport and Prusak, 1997] Davenport, T. H. and Prusak, L. (1997) Working Knowledge: How Organizations Manage What They Know, Harvard Business School Press, Boston, MA

[Davenport and Prusak, 1998] Davenport, R. and Prusak, L. (1998) "Know What You Know," *CIO Magazine*, February 15, 1998, http://www.cio.com/archive/021598_excerpt.html

[Davenport and Probst, 2002] *Knowledge Management Case Book: Siemens best practises* (2nd edition), edited by Thomas H. Davenport and Gilbert J.B. Probst, 2002, Weinheim, Cambridge, Wiley-VCH.

[Davies *et al.,* 1998] Davies, N.J., Stewart, S. and Weeks, R. (1998) "Knowledge Sharing Agents over the World Wide Web," *British Telecom Technology Journal*, Vol.16 No.3, pages. 104-109.

[Davies *et al.,* 2002] Davies, J., Duke, A., and Stonkus, A. (2002) "OntoShare: Using Ontologies for Knowledge Sharing," in *Proceedings of WWW 2002 International Workshop on the Semantic Web*, Hawaii, May 7, 2002

[Dayal and Hwang, 1984] Dayal U, and Hwang, H.Y. (1984) "View definition and generalization for database integration in a multidatabase system," *IEEE Transactions on Software Engineering*, 1984; Vol.10 No.6, pages.628-645.

[Decker *et al.,* 1999] Decker, S., Erdmann, M., Fensel D., and Studer R. (1999) "Ontobroker: Ontology based Access to Distributed and Semi-Structured Information," in *Database Semantics - Semantic Issues in Multimedia Systems*, IFIP TC2/WG2.6 Eighth Working Conference on Database Semantics (DS-8), R. Meersman *et al.* (eds.), Kluwer, 1999, Pages 351--369.

[Deerwester *et al.,* 1990] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990) "Indexing by Latent Semantic Analysis," in *Journal of the American Society for Information Science*, Vol.41, No.6, pages.391-407, 1990

[Denning, 2000] Denning, S. (2000) "The Springboard: How Storytelling Ignites Action," in *Knowledge-Era Organizations*. October 2000 Butterworth-Heinemann, Boston, USA Edition Paperback, ISBN: 0750673559

[Dixon, 1999] Dixon, N.M. (1999) *The Organizational Learning Cycle: How Can We Learn Collectively*, Second edition, Gower, 1999

[D'Oosterlinck *et al.,* 2002] D'Oosterlinck, M., Freitag, H. and Graff, J. "SiemensIndustrialServices: Turning Know-how into results," in *Knowledge Management Case Book: Siemens best practises* (2nd edition) [Thomas H. Davenport and Gilbert J.B. Probst Weinheim (ed.)], Cambridge, Wiley-VCH, 2002.

[Drew *et al.,* 1996] Drew, R.; Dew, P.M.; Morris, D.T.; Leigh, C.M.; Curson, J.M. (1996) "The Virtual Science Park," in *British Telecommunications Engineering*, 14: 322-329.

[Dreyfus, 2001] Dreyfus, H. (2001) "A Phenomenology of Skill Acquisition as the basis for a Merleau-Pontian Non-representationalist Cognitive Science". Available from http://ist-socrates.berkeley.edu/~hdreyfus/pdf/MerleauPontySkillCogSci.pdf

[Duan 2002] Duan, X.Y. (2002) "Lexical Semantic Association Between Web Documents" MSc thesis, Leeds University, available from ftp://ftp.comp.leeds.ac.uk/scs/doc/theses/duan.pdf.gz

[Dyer, 2000] Dyer, G. (2000) "Collaboration is the path to raging knowledge" in *Computerworld*. Available from http://www.computerworld.com/computerworld/records/whitepapers/ragingdoc.pdf

[Efthimiadis, 1996] Efthimiadis, E. (1996) "Query Expansion," in *Annual Review of Information Science and Technology*, Vol.31, pages.121-187, 1996

[Egnor and Lord 2000] Egnor, D. and Lord, R. (2000) "Structured Information Retrieval using XML" *Working Notes of the ACM SIGIR Workshop on XML and Information Retrieval*, Athens, Greece. available from http://www.haifa.il.ibm.com/sigir00-xml/final-papers/Egnor/

[Ehrlich, 2003] Ehrlich, K. (2003) "Locating Expertise: Design Issues for an Expertise Locator System," in *Sharing expertise: beyond knowledge management*, edited by Mark S. Ackerman, Volkmar Pipek, and Volker Wulf. Cambridge, Mass.; London, MIT Press, 2003

[Eisenhart, 2002] Eisenhart, M., (2002) "The Human Side – A successful deployment of KM requires more than selecting tools," in *Knowledge Management Magazine*, March, 2002, Available on line at http://www.kmadvantage.com/docs/km_articles/The_Human_Side_-_A_Successful_Deployment_of_KM.pdf

[Engestrom and Engestrom, 1986] Engestrom, Y. and Engestrom, R. (1986) "Developmental Work Research: The Approach and Application in Cleaning Work," in *Nordisk Pedagogik*, Vol.6, No.1, pages 2-15

[Engestrom, 1992] Engestrom, Y. (1992) "Expertise as Mediated Collaborative Activity". In *Interactive Expertise: Studies in Distributed Working Intelligence*, University of Helsinki Department of Education Research Bulletin 83.

[Enkel *et al.,* 2002] Enkel, E. Heinold, P. Hofer-Alfeis, J and Wicki, Y. (2002) "The power of communities: How to build Knowledge Management on a corporate level using a bottom-up approach," in *Knowledge Management Case Book: Siemens best practises* (2nd edition), edited by Thomas H. Davenport and Gilbert J.B. Probst, 2002, Weinheim, Cambridge, Wiley-VCH.

[Fagrell and Ljungberg, 1999] Fagrell, H. and Ljungberg, F. (1999) "Exploring Support for Knowledge Management in Mobile Work," in *Proceedings 6th European Conference on Computer-Supported Cooperative Work*, Copenhagen, Denmark, pages 259-275

[Fallside, 2001] Fallside, D.C. (2001) XML Schema Part 0: Primer, W3C Recommendation, May 2001, http://www.w3.org/TR/2001/REC-xmlschema-0-20010502/

[Fensel *et al.,* 1999] Fensel, D., Angele, J., Decker, S., Erdmann, M., Schnurr, H.P., Staab, S., Studer, R., and Witt, A., (1999) "On2broker: Semantic-based access to information sources at the WWW," in *World Conference on the WWW and Internet (WebNet99).* 1999: Honolulu, Hawaii. Available from http://gunther.smeal.psu.edu/decker99onbroker.html

[Finley, 2001] Finley, M. (2001) FUTURE SHOES: "Expertise Management 101", February 16, 2001. Available at http://www.computeruser.com/articles/daily/7,5,1,0216,01.html

[Fitzpatrick, 2003] Fitzpatrick, G. "Emergent Expertise Sharing in a New Community", in *Sharing Expertise: Beyond Knowledge Management*, M. Ackerman and V. Wulf (eds), MIT pages 80-110

[Foner, 1997] Foner, L.N. (1997) "Yenta: A Multi-Agent Referral-Based Matchmaking System," in *Proceedings of the First International Conference on Autonomous Agents* (Agent'97), Marina del Rey, CA, February 1997, pages 301-307

[Frakes and Baeze-Yates, 1992] Frakes, W.B. and Baeze-Yates, R. (1992) *Information Retrieval Data Structures and Algorithms* Prentice-Hall, London, 1992

[Franz *et al.,* 2002] Franz, M., Freudenthaler, K. Kameny, M. and Schoen, S.(2002) "The development of the Siemens Knowledge Community Support" in *Knowledge Management Case Book: Siemens best practises* (2nd edition), edited by Thomas H. Davenport and Gilbert J.B. Probst, 2002, Weinheim, Cambridge, Wiley-VCH.

[French, 2001] French, J.C., Powell, A.L., Gey, F. and Perelman, N. (2001) "Exploiting A Controlled Vocabulary to Improve Collection Selection and Retrieval Effectiveness," in *Tenth International Conference on Information and Knowledge Management,* Atlanta, Georgia, November 5-10, 2001.

[Frensch and Sternberg, 1989] Frensch, P.A. and Sternberg, R. J. (1989) "Expertise and Intelligent Thinking: When is it Worse to Know Better," in Sternberg, R. (ed.). *Advances in the Psychology of Human Intelligence*, pages. 157-188

[Furnas *et al.,* 1983] Furnas, G.W., Landauer, T.K., Dumais, S.T. and Gomez, L.M.(1983) "Statistical Semantics: Analysis of the Potential Performance of Keyword Information Systems," in *Bell System Technical Journal*, Vol.62, No.6, 1983, pages.1753-1806.

[Gadamer, 1972] Gadamer, H.G. (1972). *Wahrheit und Methode*. Tübingen, Mohr.

[Gaines, 1995] Gaines, B. R. (1995) "The Collective Stance in Modeling Expertise in Individuals and Organizations," available online at http://ksi.cpsc.ucalgary.ca/articles/Collective/Collective2.html

[Garcia-Molina *et al.,* 1995] Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. and Widom, J. (1995) "Integrating and Accessing Heterogeneous Information Sources in TSIMMIS," in *Proceedings of the AAAI Symposium on Information Gathering*, pages. 61-64, Stanford, California, March 1995..

[Garvin, 1998] Garvin, D.A. (1998) "Building a Learning Organization". in *Harvard Business Review on Knowledge Management*. Harvard Business School Press, 1998. pages. 47-80.

[GMD-FIT, 1998] GMD-FIT.CSCW (1998) "The Social Web New Forms of Interaction in Virtual Environments: A Research Framework," available from http://orgwis.gmd.de/projects/SocialWeb

[Gibbert *et al.,* 2002] Gibbert, M., Jenzowsky, S. Jonczyk, C. Thiel, M. and Volpel, S. "ShareNet – the next generation knowledge management," in *Knowledge Management Case Book: Siemens best practises* (2nd edition), edited by Thomas H. Davenport and Gilbert J.B. Probst, 2002, Weinheim, Cambridge, Wiley-VCH.

[Gibson, 1996] Gibson, R. ed (1996) *Rethinking the Future.* Nicholas Brealey Publishing, London, 1996

[Goman, 2002] Goman, C. K. (2002) "Why People Don't Tell You What They Know: The Human Side of Knowledge Management," in *ASTD 2002 International conference and Exposition*, June 2-6, 2002, New Orleans, Louisiana, USA http://www.astd.org/astd2002/HandoutstoWeb/W312.pdf

[Gongla and Rizzuto, 2001] Gongla, P. and Rizzuto, C.R. (2001) "Evolving communities of practice: IBM Global Services experience," in *IBM Systems Journal* Vol. 40, Number 4, 2001, available from www.research.ibm.com/journal/sj/404/gongla.html

[Grant, 1996] Grant, R.M. (1996) "Toward a Knowledge-Based Theory of the Firm," in *Strategic Management Journal* 17, pages 109-122

[Gray, 2000] Gray, P. (2000) “Knowledge Management Overview”. Available from http://www.crito.uci.edu/itr/publications/pdf/km-overview-pgray.pdf

[Green and Gilhooly, 1992] Green, A. J. and Gilhooly, K. J. (1992). “Empirical advances in expertise research,” in, Keane, M T G & Gilhooly, K J (Eds), *Advances in the psychology of thinking*, Vol.1, pages 45-70. Hemel Hempstead: Harvester-Wheatsheaf.

[Guarino *et al.,* 1999] Guarino, N., Masolo, C., and Vetere, G. (1999) “Ontoseek: Content-based access to the web,” in *IEEE Intelligent Systems*, Vol.14, No.3, pages70-80

[Halpern, 1998] Halpern, J. Y. (1998) “A Computing Research Repository,” in *D-Lib Magazine* November 1998 ISSN 1082-9873 available at http://www.dlib.org/dlib/november98/11halpern.html

[Hammer *et al.,* 1995] Hammer J., Garcia-Molina H., Widom J., Labio W, and Zhuge Y.(1995) “The Stanford Data Warehousing Project,” in *IEEE Data Engineering Bulletin*, Vol.18, No.2, pages. 41--48, June 1995.

[Hansen, 1999] Hansen, M.T (1999). “The Search-Transfer Problem: The Role of Weak Ties in Sharing Knowledge across Organizational Subunits,” *Administrative Science Quarterly*, vol. 44, pages. 82-111.

[Hansen *et al.,* 1999] Hansen, M.T., Nohria, N. and Tierney, T. (1999) “What’s Your Strategy for Managing Knowledge?” in *Knowledge Management Yearbook 2000-2001* Cortada, J.W. and Woods, J.A., Eds, Oxford : Butterworth-Heinemann, 1999.pages 55-69

[Hasan, 2003] Hasan, M.M. (2003) “A Spreading Activation Framework for Ontology-enhanced Adaptive Information Access within Organisations,” in *Proceedings AAAI Spring Symposium on Agent Mediated Knowledge Management AMKM 2003*, Stanford University, California, USA.

[Heflin and Dale, 2002] Heflin, J. Volz, R. and Dale, J. (2002) “Requirements for a Web Ontology Language”. W3C Working Draft 07 March 2002 http://www.w3.org/TR/2002/WD-webont-req-20020307/

[Heimbigner and McLeod, 1985] Heimbigner, D., McLeod, D. (1985) “A Federated Architecture for Information Management,” in *ACM Transactions on Office Information Systems* Vol.3, No.3,: pages.253-278

[Hellström *et al.,* 2000] Hellström, T., Malmquist, U. and Mikaelsson, J. (2000) "Decentralizing Knowledge: Managing Knowledge Work in a Software Engineering Firm," in *Journal of High Technology Management Research,* Vol.2, No.3: Available on-line at http://www.viktoria.se/ results/result_files/131.pdf.

[Hersh *et al.,* 2000] Hersh, W.R., Price, S., and Donohoe, L. (2000) "Assessing thesaurus-based query expansion using the UMLS Metathesaurus," in *Proceedings of the 2000 Annual AMIA* Fall Symposium , pages.344-348.

[Hickins, 1999] Hickins, M. (1999). "Xerox Shares Its Knowledge," in *Management Review*, 1999, September, pages 40-46

[Hiemstra, 2000] Hiemstra, D. (2000) *Using Language Models for Information Retrieval*. PhD thesis, available online http://www.cs.utwente.nl/~hiemstra/papers/thesis.pdf

[Hill *et al.,* 1998] Hill, R., Bullard, T., Capper, P., Hawes, K. and Wilson, K. (1998) "Learning about learning organizations: Case studies of skill formation in five new Zealand organizations," in *The Learning Organization*. Vol.5, No.4, 1998, pages. 184-192

[Holloway, 2000] Holloway, P. (2000) "How to Capture and Deploy Tacit Knowledge in Your Organization," in *Braintrust* January 2000. Scottsdale. Available on-line at http://www.knowledgeharvesting.org/presentations.htm

[Horrocks, 2002] Horrocks, I. (2002) "DAML+OIL: a Reason-able Web Ontology Language," in *Proceedings of 8th International Conference on Extending Database Technology (EDBT 2002),* Prague, Czech Republic, March 25-27, 2002, pages. 2-13

[Horvath, 2000] Horvath, J. A. (2000) "Working with Tacit Knowledge" in: James W. Cortada and John A. Woods (Eds). *The Knowledge Management Yearbook 2000-2001*. Butterworth-Heinemann, 2000. pages. 34-51.

[Huang, 1998] Huang, A: (1998) "Intranets for Organizational Memory Building – An Exploratory Study," in: *Proceedings of Association for Information Systems*, Baltimore, USA, 14-16 August 1998, pages. 672-673

[Hunter and Lagoze, 2001] Hunter, J. and Lagoze, C. (2001) "Combining RDF and XML Schemas to Enhance Interoperability Between Metadata application profiles," in *10th international World Wide Web Conference*, HongKong, May 2001

[Huysman and de Wit, 2003] Huysman, M. and de Wit, D.(2003) “A critical evaluation of the practice of knowledge management,” in: Ackerman, S, V. Wulf, V. Pipek *Sharing expertise: beyond knowledge management* Cambridge MA, MIT Press pages 27-56

[Ide, 1971] Ide, E. (1971) “New Experiments in relevance feedback,” in G. Salton, editor, *The SMART Retrieval System*, pages 337-354. Prentice Hall, 1971

[Ishida, 1998] Ishida, T. (1998) *Community Computing: Collaboration over Global Information Networks* Chichester New York: Wiley, 1998

[Jim Eales, 2003] Jim Eales, R.T. (2003) “Supporting Informal Communities of Practice within Organizations” in *Sharing expertise: beyond knowledge management*, edited by Mark S. Ackerman, Volkmar Pipek, and Volker Wulf. Cambridge, Mass.; London, MIT Press, 2003

[Kanfer *et al.,* 1997] Kanfer, A., Sweet, J. and Schlosser, A. E. (1997) “Humanizing the net: Social navigation with a ‘know-who’ email agent,” in *Proceedings of the 3rd Conference on Human Factors and the Web*. Denver, Colorado, June 12, 1997

[Karvounarakis, *et al.,* 2000] Karvounarakis, G., Christophides, V. and Plexousakis, D.(2000) “Querying Semistructured (Meta)Data and Schemas on the Web: The case of RDF & RDFS,” Technical Report 269, ICS-FORTH, (2000). available at http://www.ics.forth.gr/proj/isst/RDF/rdfquerying.pdf

[Karypis and Han, 2000] Karypis, G. and Han, E. (2000) “Concept Indexing: A Fast Dimensionality Reduction Algorithm with Applications to Document Retrieval & Categorization,” CIKM 2000 Technical report available online http://www-users.cs.umn.edu/~karypis/publications/ir.html

[Kashyap *et al.,* 1995] Kashyap, V., Shah, K., Sheth, A. (1995) “Metadata for building the MultiMedia Patch Quilt,” in Jajodia, S., Subrahmanian, V.S. (eds) *Multimedia Database Systems: Issues and Research Directions.* Springer-Verlag: pages.297-319

[Kautz *et al.,* 1996] Kautz, H., Selman, B., and Milewski, A. (1996) “Agent Amplified Communication,” in *Proceedings of AAAI-96* (Portland, Oreg.). MIT Press, Cambridge, 1996, pages.3-9.

[Kautz *et al.,* 1997a] Kautz, H., Selman, B. and Shah, M. (1997) “Referral Web: Combining Social Networks and Collaborative Filtering,” in *Communications of the ACM*, Vol. 40, No. 3, pages.63-65.

[Kautz *et al.,* 1997b] Kautz, H., Selman, B. and Shah, M. (1997b) "The Hidden Web," in *AI Magazine*, Vol.18 No.2 1997 pages. 27-36

[Kautz and Selman, 1998] Kautz, H. and Selman, B. (1998) "Creating Models of Real-World Communities with ReferralWeb," in Working Notes of the *Workshop on Recommender Systems held in conjunction with AAAI-98*, Madison, WI, 1998, pages. 58 – 59.

[Khan 2000] Khan, L. (2000) *Ontology-based Information Selection* Ph.D. Dissertation, Department of Computer Science, University of Southern California, August 2000. available from http://www.utdallas.edu/research/esc/publications/lkhan_def.pdf

[Koskinen, 2001] Koskinen, K.U. (2001) "Tacit Knowledge as a Promoter of Success in Technology Firms," in *Proceedings of the 34th Hawaii International Conference on System Sciences,* January 3-6, 2002, Maui, Hawaii, Available on line at http://www.hicss.hawaii.edu/ HICSS_34/PDFs/DDOML10.pdf

[Krulwich and Burkey, 1996] Krulwich, B. and Burkey, C. (1996) "The ContactFinder: Answering bulletin board questions with referrals," in *Proceedings of the 1996 National Conference on Artificial Intelligence (AAAI-96)*, Portland, OR, vol.1, 1996, pages.10-15.

[Langford, 2002] Langford, C.H. (2002) "Measuring the Impact of University Research on Innovation," In Holbrook, J. Adam, and David A.Wolfe (eds). 2002 *Knowledge Clusters and Regional Innovation: Economic Development* in Canada. McGill-Queen's University Press. Montreal and Kingston

[Lau *et al.,* 1999] Lau, L.M.S., Curson, J., Drew, R., Dew, P.M. and Leigh, C., (1999) "Use of Virtual Science Park Resource Rooms to Support Group Work in a Learning Environment," in *Proceedings of GROUP'99 International ACM SIGGROUP Conference on Supporting Group Work*, November 14-17, 1999, Phoenix, Arizona, USA, pages 209-218, ACM press, 1999.

[Lau *et al.,* 2000] Lau, N.M.L., Lochovsky, F.H. and Karlapalem, K. (2000) "An Expertise Finding Agent. In International ICSC symposium on Multi-Agents and Mobile Agents," in *Virtual Organizations and E-Commerce, MAMA 2000*, December 11-13, Woologong, Australia, 2000, pages.11-15

[Lassila and Swick, 1999] Lassila, O. and Swick, R.R. (1999) "Resource Description Framework (RDF): Model and Syntax Specification". W3C Recommendation, February 1999, http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

[Leigh *et al.,* 1999] Leigh, C.M.; Curson, J.M.; Drew, R.S.; Dew, P.M. (1999) “Integrated Information Directory Services: Facilitating the transfer and exploitation of science and technology on the World Wide Web,” in *Geographical Information and Planning: European Perspectives* [Stillwell, JCH, Geertman, S and Openshaw, S (ed.)], pages 403-422, Springer, Berlin and London.

[Leont’ev, 1981] Leont’ev, A. N. (1981) *Activity, Consciousness and Personality*. Englewood Cliffs, Prentice Hall.

[Levy *et al.,* 1996] Levy, A., Rajaraman, A. and Ordille, J.(1996) “Querying Heterogeneous Information Sources Using Source Descriptions,” in *Proceedings of the twenty-second International Conference on Very Large Databases*, September 3-6, 1996, Mumbai, India, pages.251-262, 1996

[Li *et al.,* 2001] Li, B.C., Ghimire, B. and Batra, P. (2001) “Natural Language Processing” available from http://www.bridgeport.edu/sed/projects/449/Fall_2000/bingli/senior/research.doc

[Liao *et al.,* 1999] Liao, M., Hinkelmann, K., Abecker, A., and Sintek, M. (1999) “A Competence Knowledge Base System as Part of the Organizational Memory,” in *Proceedings of 5th Biannual German Conference on Knowledge-Based Systems – Survey and Future Directions*, March, 1999 in Würzburg, Germany, pages 125-137.

[Lindgren and Stenmark, 2002] Lindgren, R. and Stenmark, D. (2002). “Designing Competence Systems: Towards Interest-Activated Technology,”. in *Scandinavian Journal of Information Systems*. Volume 14, pages. 19-35.

[Litwin *et al.,* 1982] Litwin, W., Boudenant, J. Esculier, C., Ferrier, A., Glorieux, A. La Chimia, J., Kabbaj, K., Moulinoux, C., Rolin, P. and Stangret, C. (1982) “SIRIUS: Systems for Distributed Data Management,” In Schneider H-J(ed) *Distributed Data Bases*. North-Holland, Netherlands: 311-66

[Lyons, 2000] Lyons, K. L. (2000) “Using Patterns to Capture Tacit Knowledge and Enhance Knowledge Transfer in Virtual Teams,” in Malhotra, Y. (Ed.), *Knowledge Management and Virtual Organizations,* pages 124 – 143. Hershey, PA: Idea Publishing Group.

[Mahapatra and Chakrabarti, 2002] Mahapatra, P.K. and Chakrabarti, B. (2002) *Knowledge Management in Libraries*, New Delhi, Ess Ess 2002, viii, 288 p., ISBN 81-7000-331-8

[Marchant, 1989] Marchant, G. (1989) "Analogical Reasoning and Hypothesis Generation in Auditing," in *The Accounting Review* 64, July, pages.500-513

[Mattox *et al.,* 1999] Mattox, D., Maybury, M. and Morey, D. (1999) "Enterprise expert and knowledge discovery," in *Proceedings of the 8th International Conference on Human Computer Interaction (HCI International'99),* Munich, Germany, August. 23-27, 1999; pages.303-307

[Maurino *et al.,* 1995] Maurino, D.E., Reason, J, Johnston, N, and Lee, R.B. (1995). *Beyond Aviation Human Factors*. Aldershot, Avebury, 1995

[McDonald, 2000] McDonald, D. W. (2000) *Supporting Nuance in Groupware Design: Moving from Naturalistic Expertise Location to Expertise Recommendation.* University of California, Irvine. Ph.D. Thesis, 2000 available from http://www.ischool.washington.edu/mcdonald/papers/McDonald.Dissertation.final.pdf

[McDonald and Ackerman, 1998] McDonald, D. W., and Ackerman, M. S.(1998). "Just Talk to Me: A Field Study of Expertise Location," in *Proceedings of the ACM Conference on Computer-Supported Cooperative Work (CSCW '98),* pages 315-324

[McDonald and Ackerman, 2000] McDonald, D.W. and Ackerman, M.S. (2000) "Expertise Recommender: a Flexible Recommendation System and Architecture," in *Proceeding of the ACM 2000 Conference on Computer Supported Cooperative Work,* (CSCW'00), Philadelphia, PA, 2000, pages. 231-240.

[Mena *et al.,* 2000] Mena, E., Illarramendi, A., Kashyap, V. and Sheth, A. (2000) "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies," in *International journal on Distributed And Parallel Databases (DAPD)*, ISSN 0926-8782, Vol.8, No.2 pages.223-272, April 2000

[Michaelis, 1997] Michaelis, E.K. (1997) "The need for externally funded research" In *Planning for the Research Mission of Public Universities in the Twenty-first Century*. No 101, June 1997 available from http://merrill.ku.edu/publications/1997whitepaper/michaelis.html

[Middleton *et al.,* 2001] Middleton, S.E., De Roure, D.C., and Shadbolt, N.R (2001) "Capturing knowledge of user preferences: ontologies in recommender systems," in *Proceedings of First International Conference on Knowledge Capture*, pages 100-107, Victoria, British Columbia, Canada

[Mockus and Herbsleb, 2002] Mockus, A. and Herbsleb, J.D. (2002) "Expertise Browser: A Quantitative Approach to Identifying Expertise," in *International Conference on Software Engineering (ICSE'02),* May 19-25, Orlando, Florida, USA, pages. 503-512

[Monz and de Rijke, 2001] Monz, C. and de Rijke, M. (2001) "Introduction to Information Retrieval". Available online http://remote.science.uva.nl/~mdr/Teaching/IR/ESSLLI01/day-1.pdf

[Nakata *et al.,* 1998] Nakata, K., Voss, A., Juhnke, M., and Kreifelts, T., (1998) "Concept Index: Capturing Emergent Community Knowledge from Documents," in *the 7th Workshop on Designing Collective Memories*, Paris, 14. Sept. 1998

[Nickols, 2000a] Nickols, F. (2000) "The Knowledge in Knowledge Management," in *The Knowledge Management Yearbook, 2000-2001*, Editors, James W. Cortada and John A. Woods, Publisher, Butterworth-Heineman

[Nickols, 2000b] Nickols, F. (2000) "What Is in the World of Work and Working: Some Implications of the Shift to Knowledge Work," in *The Knowledge Management Yearbook, 2000-2001*, Editors, James W. Cortada and John A. Woods, Publisher, Butterworth-Heineman

[Nishida, 1995] Nishida, T. (1995) "The Knowledge Community: Towards Knowledge Level Communication" presented at *International Forum on the Frontier of Telecommunications Technology*, Ministry of Posts and Telecommunication, 1995

[Nonaka, 1994] Nonaka, I. (1994) "A dynamic theory of organizational knowledge creation," in *Organization Science*, Vol.5, No.1, page 14-37.

[Nonaka and Takeuchi, 1995] Nonaka, I. and Takeuchi, H. (1995) *The Knowledge-Creating Company: How Japanese Companies Create The Dynamics of Innovation,* Oxford University Press, New York

[Oakes and Rengarajan, 2002] Oakes, K. and Rengarajan, R. (2002) "The Hitachhiker's Guide to Knowledge Management," in *Training and Development*. Vol.56, No.6, pages 75-77, 2002

[O'Dell *et al.,* 1998] O.'Dell, C., Jackson, C., Grayson, J. (1998) *If Only We Knew What We Know: The Transfer of Internal Knowledge and Best Practice* , Free Press1998.

[O'Hara *et al.,* 2002] O'Hara, K., Alani, H. and Shadbolt, N. (2002) "Identifying Communities of Practice: Analysing Ontologies as Networks to Support Community Recognition," In *Proceedings of the 2002 IFIP World Computer Congress*, Montreal, Canada, August 2002.

[Olson and Shaffer, 2002] Olson, L. and Shaffer, R. (2002) "Expertise Management – and Beyond". White paper in RGS Associates. Available on-line at http://www.rgsinc.com/ publications/pdf/white_papers/Expertise_Management_and_Beyond.pdf

[O'Riordan and Sorensen, 1999] O'Riordan, C. and Sorensen, H. (1999) "Information Filtering and Retrieval: An Overview," Technical report available online http://citeseer.nj.nec.com/483228.html

[Papakonstantinou *et al.,* 1996] Papakonstantinou, Y., Abiteboul, S. and Garcia-Molina, H. (1996). "Object fusion in mediator systems," in *Proceedings of the Twenty-Second International Conference on Very Large Data Bases*, Bombay, India, 1996, pages 413—424.

[Pikrakis *et al.,* 1998] Pikrakis, A., Bitsikas, T., Sfakianakis, S., Hatzopoulos, M., DeRoure, D., Hall, W., Reich, S., Hill, G. and Stairmand, M. (1998) "MEMOIR - Software Agents for Finding Similar Users by Trails," in *Proceedings of the 3rd International Conference on the Practical Applications of Agents and Multi-Agent Systems (PAAM-98),* London, UK, 1998, pages 453-466

[Porter, 1980] Porter, M.F., (1980), "An algorithm for suffix stripping," in *Program*, Vol. 14, No.3, pages 130-137, Java version is available from http://www.tartarus.org/~martin/PorterStemmer/java.txt

[Rabarijaona *et al.,* 2000] Rabarijaona, A., Dieng, R., Corby, O., and Ouaddari, R. (2000) "Building and Searching XML-based Corporate Memory," in *IEEE Intelligent Systems and their Applications*, *Special Issue on Knowledge Management and the Internet,* May/June 2000, pages56-63

[Raeithel, 1993] Raeithel, A. (1993) "Activity Theory as a Foundation for Design," in R. Budde *et. al.*, (eds) *Software Development and Reality Construction*. Berlin, Springer, 1993

[Rahimi *et al.,* 1982] Rahimi, S.K., Spinrad, M.D., and Larson, J.A. (1982) "A Structural View of Honeywell's Distributed Database Testbed System: DDTS," in *Database Engineering Bulletin* Vol.5, No.4, pages.47-51, 1982

[Reed and DeFilippi, 1990] Reed, R. and DeFilippi, R.J., (1990) "Causal Ambiguity, Barriers to Imitation, and Sustainable Competitive Advantage," in *Academy of Management Review*, Vol. 15, No. 1.

[Reimer *et al.,* 2003] Reimer, U., Brockhausen, P., Lau, T. and Reich, J.R. (2003) "Ontology-based Knowledge Management at Work: The Swiss Life Case Studies," in *Towards The Semantic Web – Ontology-driven Knowledge Management*. Edited by John Davies, Dieter Fensel, and Frank van Harmelen, F. Publisher John Wiley and Sons Ltd, 2003

[Robertson, 1977] Robertson, S.E. (1977) "The Probabilistic Ranking Principle in IR," *Journal of documentation*, Vol.33, pages. 294-304, 1977

[Rocchio, 1971] Rocchio, J.J. (1971) "Relevance Feedback in Information Retrieval," In G. Salton, editor, *The SMART Retrieval System – experiments in Automatic Document Processing.* Prentice Haoo Inc. Englewood Cliffs, NJ, 1971

[Ruggles, 1998] Ruggles, R. (1998) "The state of the notion: knowledge management in practice," in *California Management Review*, Spring 1998, Vol. 40, (special) Iss.3 pages80-9

[Salas *et al.,* 1997] Salas, E., Cannon-Bowers, J. A. and Johnston, J. H. (1997). "How can you turn a team of experts into an expert team?: Emerging training strategies," in: Zsambok, C. E. and Klein, G. (eds.). *Naturalistic decision making*. Mahwah, NJ: Lawrence Erlbaum Associates.

[Salton *et al.,* 1975] Salton, B., Wong, A. and Yang, C.S. (1975) "A Vector Space Model for Information Retrieval," in *Communications of the ACM*, Vol.18, No.11, pages. 613-620, November 1975

[Salton and McGill, 1983] Salton, G. and McGill, M.J. (1983) *Introduction to Modern Information Retrieval*. McGraw-Hill Book Co., New York, 1983

[Salton and Buckley, 1988] Salton, G. and Buckley, C. (1988) "Term-weighting approaches in automatic text retrieval," in *Information Processing and Management.* Vol.25, No.5, pages.513-523, 1988

[Salton and Buckley, 1990] Salton, G. and Buckley, C. (1990) "Improving Retrieval Performance by Relevance Feedback," in *Journal of the American Society for Information Science*, Vol.41, No.4, pages 288-297.

[Scarbrough *et al.,* 1999] Scarbrough, H., Swan, J., and Preston, J. (1999) *Knowledge management: a literature review*; Imprint London: Institute of Personnel and Development, 1999

[Scarbrough, 1999] Scarbrough, H. (1999) "Network Nirvana: The management of knowledge in the postmodern organization," in *British Academy of Management Conference*, September 1999, Manchester

[Schön, 1983] Schön, D.A. (1983). *The Reflective Practitioner*. New York, Basic Books.

[Schwartz and Divitini, 2000] Schwartz, D. G. and Divitini, M. (2000) *Internet-based Organizational Memory and Knowledge Management* Hershey, Pa. Idea Group Publishing

[Seligman and Rosenthal, 2001] Seligman, L. and Rosenthal, A. (2001) *The Impact of XML on Databases and Data Sharing*, IEEE Computer

[Setzer, 2001] Setzer, V.W. (2001) "Data, Information, Knowledge and Competency," Available from http://www.ime.usp.br/~vwsetzer/data-info.html

[Shadbolt and O'Hara, 2003] Shadbolt, N. and O'Hara, K. (2003) "AKTuality: An Overview of the Aims, Ambitions and Assumptions of the Advanced Knowledge Technologies Interdisciplinary Research Collaboratio," AKT Advanced Knowledge Technologies selected papers, Avalable on line at http://www.aktors.org/publications/ selected-papers/01.pdf

[Shanteau, 1998] Shanteau, J. (1998) "Psychological Characteristics and Strategies of Expert Decision Makers," in *Acta Psychologica*, 68, pages. 203-215

[Sheth and Larson, 1990] Sheth, A.P. and Larson, J.A. (1990) "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases," in *ACM Computing Surveys*, Vol. 22/3.

[Sheth, 1998] Sheth, A. (1998) "Changing Focus on Interoperability in Information Systems: from System, Syntax, Structure to Semantics," in *Interoperating Geographic Information Systems*, M. F. Goodchild, M. J. Egenhofer, R. Fegeas, and C. A. Kottman (eds.), Kluwer, 1998

[Singhal, 2001] Singhal, A. (2001) "Modern Information Retrieval: A Brief Overview" available from http://singhal.info/ieee-2001.ps

[Skyrme and Farago, 1995] Skyrme, D. J. and Farago J. (1995). "The Learning Organisation," *Insight* No. 3: 1995 Available from. http://www.skyrme.com/insights/3lrnorg.htm

[Skyrme, 1998] Skyrme, D. J. (1998) "Developing A Knowledge Strategy" in the January 1998 edition of Strategy, the bi-monthly magazine of the Strategic Planning Society http://www.skyrme.com/pubs/knwstrat.htm

[Sparck-Jones, 1972] Sparck-Jones, K. (1972) "A Statistical interpretation of Term Specificity and its Application in Retrieval," in *Journal of Documentation*. Vol.28, pages.11-21, 1972

[Stenmark, 1999] Stenmark, D. (1999) "The Tacit Knowledge of Interests," in *Proceedings of the European Computer Supported Cooperative Work (ECSCW'99) Conference* Avalable on line at http://www.informatik.uni-bonn.de/~prosec/ECSCW-XMWS/FullPapers/Stenmark.rtf.

[Stenmark, 2001] Stenmark, D. (2001). "Leveraging Tacit Organisational Knowledge," in *Journal of Management Information Systems*, Special Winter Issue, Vol. 17, No. 3, pages. 9-24.

[Stenmark, 2002] Stenmark, D. (2002). "Information vs. Knowledge: The Role of intranets in Knowledge Management," in *Proceedings of HICSS-35,* Hawaii, January 7-10, 2002 Available: http://www.viktoria.se/results/result_files/183.pdf

[Stewart, 1997] Stewart, T. (1997) *Intellectual Capital: The New Wealth of Organizations* New York: Doubleday, 1997.

[Stuckenschmidt, 2000] Stuckenschmidt, H. (2000) "Using OIL for Intelligent Information Integration," in *Proceedings of the Workshop on Applications of Ontologies and Problem-*Solving Methods at ECAI, 2000

[Suseno, 2002] Suseno, Y. (2002) "Dispersed Knowledge: Management of International Knowledge-Intensive Organizations" in *The second European Academy of Management Conference*, May 9-11, 2002, Stockholm, Sweden

[Tan, 1997] Tan, S. (1997) "The Elements of Expertise," in *Journal of Physical Education, Recreation, and Dance*. Vol.68, No.2, pages 30-33

[Teece, 1987] Teece, D.J., (1987) "Profiting from technological innovation," In Teece, D.J. (Ed.): *The Competitive Challenge*. Ballinger. Cambridge.

[Temelkuran, 2003] Temelkuran, B. (2003) *Hap-Shu A Language for Locating Information in HTML Documents,* PhD thesis available at www.ai.mit.edu/people/jimmylin/papers/Temelkuran03.doc

[TKS, 2001] TKS (2001) "Power Your Portal with Real Brains – by tacit knowledge systems," a white paper published by Special Supplement to KM World July/August 2001 S18

[van Rijsbergen, 1990] C.J. van Rijsbergen. (1990) "The Science of Information Retrieval: Its Methodology and Logic," in *Proceedings of the European Summer School in Information Retrieval, 1990. Available at* http://dent.ii.fmph.uniba.sk/~kravcik/IR/Method.html,

[Vargas-Vera *et al.,* 2002] Vargas-Vera, M. Motta, E. Domingue, J. Lanzoni, M. Stutt, A. and Ciravegna, F. (2002) "MnM: Ontology Driven Semi-Automatic and Automatic Support for Semantic Markup," in *EKAW-2002, 13th International Conference on Knowledge Engineering and Management,* Spain 1-4 October 2002. Available at http://kmi.open.ac.uk/projects/akt/vargas-vera-etal.pdf

[Vdovjak and Houben, 2001] Vdovjak, R., and Houben, G. (2001) "RDF Based Architecture for Semantic Integration of Heterogeneous Information Sources," in *Workshop on Information Integration on the Web (2001)* pages.51-57

[Vivacqua and Lieberman, 2000] Vivacqua, A. and Lieberman, H. (2000) "Agents to Assist in Finding Help," in *ACM Conference on Computers and Human Interface (CHI-2000)*, Hague, Netherlands, April 2000.

[Vivacqua, 1999] Vivacqua, A. (1999) "Agents for Expertise Location," in *Proceedings of the AAAI Spring Symposium on Intelligent Agents in Cyberspace*, Stanford, CA, March 1999, pages. 9 – 13.

[Voss *et al.,* 1998] Voss, A., Guo, H., Hausen, H.L., and Kreifelts, T. (1998) "Agents for Collaborative Information Exploration," in *Proceeding CSCWIS 98, Third International Workshop on CSCW in Design* July 15-18, 1998, Tokyo

[Vygotsky, 1978] Vygotsky, L. S. (1978) *Mind in Society: The Development of Higher Educational Processes*. Cambridge, MA. Harvard University Press

[Wellins *et al.,* 1993] Wellins, R. S., Byham, W. C., and Wilson, J. M. (1993) *Empowered Teams: Creating Self-directed Work Groups that Improve Quality, Productivity, and Participation,* Jossey-Bass: San Francisco

[Wenger and Snyder 2000] Wenger, E.C. and Snyder, W.M. (2000) “Communities of practice: The organizational frontier,” in *Harvard Business Review*, Vol.78, Jan/Feb 2000, pages. 139-145.

[Widom, 1995] Widom, J. (1995) “Research Problems in Data Warehousing,” In *Proceedings of the 4th International Conference on Information and Knowledge Management*, pages.25-30, Nov. 1995

[Wiederhold, 1992] Wiederhold, G. (1992) “Mediators in the Architecture of Future Information Systems,” in *IEEE Computers*, Vol.25, No.3, pages.38-49, Mar. 1992

[Wilson and Fredericksen, 2000] Wilson, L.T. and Fredericksen, D. (2000) “Harvesting eProcess Know-how,” in *The Delphi eBusiness Summit 2000*, 9 May 2000, Coronado, California. Available from http://www.knowledgeharvesting.org/kho/presentations.htm

[Yimam-Seid, 1999] Yimam-seid, D. (1999) “Expert Finding Systems for Organizations: Domain Analysis and the DEMOIR approach,” in *Proceedings of the ECSCW’99 Workshop “Expertise Management”,* Kopenhagen, 1999

[Yimam-Seid, 2003] Yimam-Seid, D. and Kobsa, A. (2003) “Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach,” in *Journal of Organizational Computing and Electronic Commerce* Vol.13, No.1, 1-24.

[Yukawa and Kashara, 2001] Yukawa, T. and Kashara, K. (2001) “An Expert Recommendation System using Concept-based Relevance Discernment,” in *13th IEEE International Conference on Tools with Artificial Intelligence (ICTAI’01)* November 7-9, 2001, Dallas, Texas, pages. 257-264

[Zack, 1999] Zack, M.H. (1999) “Developing a Knowledge Strategy,” in *California Management Review*, Vol. 41, No. 3, Spring, 1999, pages. 125-145

[Zack, 2002] Zack, M.H. (2002) “A Strategic Pretext for Knowledge Management,” in *Proceedings of the third European Conference on Organizational Knowledge, Learning, and Capabilities (OKLC 2002),* 5-6 April, 2002, Athens, Greece

[Zuboff, 1988] Zuboff, S. (1988). *In the Age of the Smart Machine: The Future of Work and Power*. New York Basic Books

Appendix A Survey of Experts Finding Systems Within Academia

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					
University of Oxford - Templeton College	Directory of Experts (http://www.templeton.ox.ac.uk/directoryofexperts/)		Research areas; Experts name		Contact details; Publications; Biography; Special interests	Keywords	Order by subject and surname		
University of Wisconsin-Madison	Experts Database (http://experts.news.wisc.edu/)	1997		Keyword; Expert name	Contact details, Area of expertise	Keywords	Order by surname	Experts fill in the registration form	Every year staff are reminded to update their listings
Massachusetts Institute of Technology - MIT Sloan School of Management	Expertise Guide (http://mitsloan.mit.edu/news/expertise_guide.php)		Geography, Research areas, Industry application		Contact details; Biography; Expertise; Publications	Keywords	Order by subject and surname		
University of South California	USC Experts Directory (http://uscnews3.usc.edu/experts/)		Subjects; Language	Keyword, Expert name	Contact details; Biography; Expertise; url of homepage	Keywords	Order by subject and surname		
University of California	Science Experts (http://ucsdnews.ucsd.edu/Science_Experts/)			Keyword	Contact details; Research interests; Professional society; Expertises; Publications; url of homepage	Subjects; Short description	Order by surname		
University of Aberdeen	Directory of Experts (http://www.abdn.ac.uk/experts/)		Subjects; Research areas	Keyword	Contact details; Areas of expertise; Biography;	Keywords; Subjects	Order by subject and surname		

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					
University of California	Agriculture, Natural Resources and Food Safety Topic Experts (http://danr.ucop.edu//news/askuc/AskAg.htm)		Research areas		Contact details; Expertise description	Short description	Order by surname		
Expertise Ireland	(http://www.expertiseireland.com/wcHome.aspx?WCI=htmMain)	June 2003	Classification; Institution	Keyword in particular field (e.g. publication)	Contact details; Research interests; Consultancy history; Patents	Keywords	Order by institution, department, surname	Experts fill in the registration form	Depend on individuals
University of Hull	Resource & Expertise Database (http://www.red.hull.ac.uk/)	2001	Subjects	Keyword;	Contact details; Expertise	Keywords; Short description	Order by surname; last updated	Manually collecting information by asking academics what skills they have	Currently 2 years out of date, it is the research and enterprise officer's duty to remind each expert to update the information
University College London	Experts on line (http://www.ucl.ac.uk/media/ucl-experts/)			Keyword; surname (restricted in department)	Contact details; Expertise (a few keywords)	Keywords	Order by surname		
University of Bristol	Directory of Experts (http://www.bris.ac.uk/media/experts/)		Expert name; Controlled keyword	Keyword	Contact details; Membership; Keyword; Area of expertise; experience	Keywords; Short description	Order by surname		

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					
University of Westminster	Experts Database (http://www.wmin.ac.uk/experts/)	May, 2002		Expert name; Keyword; Publication	Contact details; Educational qualification; Biography; Research area; Publications; Grants; PhD students currently supervised	Short descriptions	Order by surname		
University of Manchester	Nanotechnology expertise database (http://www.business.man.ac.uk/nano/database/themeselect.php)	March, 2003	Subjects		Contact details; Research interests; url of homepage	Keywords	Order by surname	Support team gather information from internet	By experts themselves and also checked by the maintenance member
University of London – Institute of Education	(http://ioewebserver.ioe.ac.uk/ioe/cms/get.asp?cid=1977)	August 2001		Expert name; Keyword	Contact details; responsibilities; Interests; Publications; Key phrases	Keywords; Short description	Order by surname	Experts fill in the registration form	Database administrator send requests periodically
Coventry University	Experts' Directory (http://www.coventry.ac.uk/cms/jsp/polopoly.jsp?d=304)	1999	Subjects	Keyword	Areas of Expertise; Contact details; Professional bodies; Major grants, Contracts and consultancies; publications; Keywords	Keywords	Order by subjects and surname	Experts fill in the registration form	Maintenance member send email to remind experts to update their information every 3 months

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					
University of Pretoria	Guide to Expertise (http://www.up.ac.za/expertise/expsearc.phtml)	1998		Keyword; Subject; Expert name	Contact details; Field of expertises	Keywords	Order by surname	Experts fill in the registration form	Depend on individuals
University of Bath	Directory of Expertise (http://www.bath.ac.uk/expertise/)	May 2003	Subjects	Name; Keyword	Contact details; Qualification; Expertise; Major grants; Publications	Keywords; Short description	Order by surname	Input by the research staff themselves, or by departmental coordinators.	Depend on individuals
University of Leicester	LEXICAN (http://www.le.ac.uk/press/experts/database/index.html)	2002	Subjects	Keyword; Expert name;	Contact details; Research specification; Subject area; url of homepage	Keywords and Subjects	Order by surname	Experts fill in the registration form	Annual review and remind experts to update their information
University of Cambridge	Expertise Directory (http://www.clo.cam.ac.uk/expertise/index.htm)	January, 2003		Keyword; Expert name; (Can be restricted in department)	Contact details; Research overview; Projects; Publications; Patents			Experts fill in the registration form	Depend on individuals
City of London Cass Business School	Cass Experts Online (http://bunhill.city.ac.uk/research/faculty.nsf/httpFacultyRecordsByName?openview&collapseview)	1998	Experts name; Subjects; Language	Keyword	Contact details; Qualification background; Research area; Publications; Projects; PhD students currently supervised		Relevance; last/first modified; return of maxim results	Collected centrally from academic Curriculum Vitaes (CVs)	Depend on individuals

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					
University of Dundee	Directory of Experts (http://www.dundee.ac.uk/pressoffice/mediaguide02/intro.htm)		Subjects; Expert name		Contact details; Expertise	Keywords	Order by subject and surname		
Edge Hill University	Directory of Expertise (http://info.edgehill.ac.uk/DirofExp/)		Research areas	Expert name; Keyword	Contact details; Expertise	Keywords	Order by subject and surname		
Cardiff University	Directory of Expertise (http://www.cf.ac.uk/news/experts/)	November 2001		Keyword; Expert name	Contact details; Area of expertise; Language	Keywords	Order by surname	Experts fill in the registration form	Depend on individuals
Massey University	EXPERTISE Directory (http://www-db.massey.ac.nz/mp_prod/muppit.cgi?use_config=wwp_expertise.config&query_id=20256)		Classification	Keywords; Experts name	Contact details; Expertise; Consultancy experience; Professional interests	Keywords	Order by surname		
Pennsylvania State University – College of Medicine	Faculty Research Expertise Database (http://www.hmc.psu.edu/fred/)	April 2003 (still in building)	Classification	Keyword	Educational background; Expertise; Research interests; Contact details	Classification term; Short description	Order by surname	Support team gather information from national library of medicine, USA	Every two weeks by one support team member performs publications downloading

Academia	System name (URL)	Creation time	Facility to find experts		Expert profile	Expertise representation	Order of listed experts	Data collection	Update mechanism
			Browse	Search					

University of Edinburgh	Directory of Experts (http://129.215.67.86/Media_Directory/)			Keyword;
				Expert name;
				Subject;
					Contact details;
					Expertise;
					Keyword
						Keywords	Order by surname		

University of Leeds	The University of Leeds Publications Database (http://ulpd.leeds.ac.uk))	1999	Classification	Keyword;
				Expert name
					Contact details;
					Area of expertise;
					Publications;
					Projects
						Classification terms	Order by surname	Collected by department administrator	Depend on individuals

The survey covers 27 universities on their expertise matching systems. Most of them are UK universities, some of them are US universities. From the survey it can be seen that expertise matching systems are not mature enough due to the following limitations.

- **Data collection** depends heavily on experts or administrative officers. The common way for collecting data of each expert is providing a registration form to experts who will fill in the forms themselves. Some systems rely on the support team members to collect data from different departments and manually input this data into a database.

- **Store and retrieval** the experts' information is stored in relational database, or LDAP directory. One system stores the experts' information into Excel spreadsheet. Experts can be retrieved through browsing the simple subject tree or through keywords searching. The fields for searching experts are normally "surname", "expertise description", "department". Very few systems can provide searching publication facility.

- **Difficulties in maintaining the up-to-date information** Again, this is the experts' duty to make sure that the new information is added to the expertise matching systems. This update process cannot be guaranteed although in a few systems the support team member(s) remind the experts to do so periodically (every 3 months, every year, etc).

- **Output of the retrieval** Nearly all the systems can only list experts according to alphabetical order of their surname, except in Cass Experts Online, the experts are ranked according to their relevance. The quality of output presentation varies in these systems. More than 1/3 systems provide only contact information and a few keywords as expertise description. Some systems provide very detailed information of each expert (such as Cass Experts Online), however, since this information has to be manually input by experts, not all the experts fill in every section in the registration form.

Appendix B Difference between Data Retrieval and Information Retrieval

	Data Retrieval (DR)	*Information Retrieval* (IR)
Matching	Exact match	Partial match, best match
Inference	Deduction	Induction
Model	Deterministic	Probabilistic
Classification	Monothetic	Polythetic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Items wanted	Matching	Relevant
Error response	Sensitive	Insensitive

Match: checking whether an item is or is not present in the file. In information retrieval this may sometimes be of interest but more generally to find those items which partially match the request and then select from those a few of the best matching ones.

Inference: The inference used in data retrieval is of the simple deductive kind, that is, aRb and bRc then aRc. In information retrieval it is far more common to use inductive inference; relations are only specified with a degree of certainty or uncertainty and hence the confidence in the inference is variable. This distinction leads one to describe data retrieval as deterministic but information retrieval as probabilistic.

Classification: In DR we are most likely to be interested in a monothetic classification, that is, one with classes defined by objects possessing attributes both necessary and sufficient to belong to a class. In IR such a classification is one the whole not very useful, in fact more often a polythetic classification is what is wanted. In such a classification each individual in a class will possess only a proportion of all the attributes possessed by all the members of that class. Hence no attribute is necessary nor sufficient for membership to a class.

Query language: The query language for DR will generally be of the artificial kind, one with restricted syntax and vocabulary, in IR natural language is preferred although there are some notable exceptions.

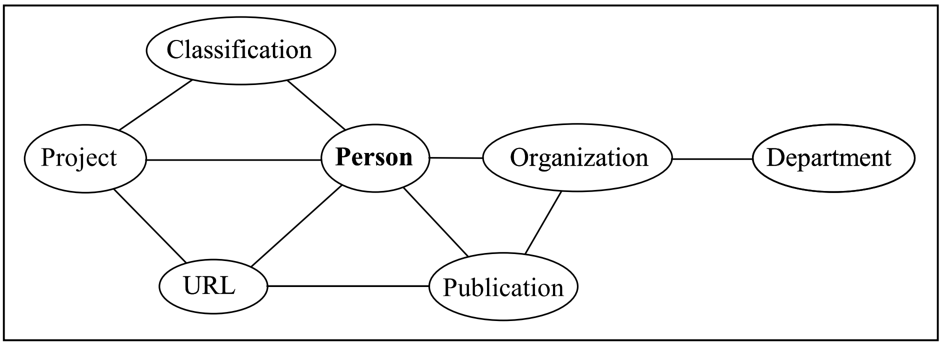
Query specification: In DR the query is generally a complete specification of what is wanted, in IR it is invariably incomplete.

Items wanted: This last difference arises partly from the fact that in IR we are searching for relevant documents as opposed to exactly matching items. The extent of the match in IR is assumed to indicate the likelihood of the relevance of that item.

Error Response: One simple consequence of this difference is that DR is more sensitive to error in the sense that, an error in matching will not retrieve the wanted item which implies a total failure of the system. In IR small errors in matching generally do not affect performance of the system significantly.

Source: http://www.dcs.gla.ac.uk/~iain/keith/data/pages/2.htm

Appendix C The ULPD Data Model



The ULPD data model is a person-centric data model to ensure that there is a relationship from each relevant entity to the person entity. Thus, each time users search for information they will get the related person as well. In this model *Person* entity is the centre of the model; the other entities are all connected to the *Person* entity directly or indirectly. *Organization* describes the organization that the person belongs to; *Organization* is in turn connected with the *Department*. *Publication* refers to the person's publication; *URL* refers to the person's homepage address; *Project* refers to all the projects the person is working on or has completed before; *Projects* and *Publications* are linked to the particular field of research terms in the *Classification*.

Appendix D Testing Results of Finding Similar Experts Using Vector Space Model

Experts are randomly selected from 10 research groups in the School of Computing, the similarity between each expert and others are calculated. It is found that in most cases, the experts with similar research interest (in italic) can be identified from a thousands people.

Group 1: Vision group

1.000 Dr R.D. Boyle --- Computing
0.530 Prof D.C. Hogg --- Computing
0.499 Dr A.J. Bulpitt --- Computing
0.461 Dr K.C. Ng --- Music
0.393 Prof A.J. Daly --- Institute for Transport Studies
0.359 Mr E.S. Atwell --- Computing
0.352 Prof A.G. Wilson --- Geography
0.352 Dr D.P. Watling --- Institute for Transport Studies
0.339 Miss S.A. Smith --- Development Nursing Policy and Practice
0.334 Prof M.J. Kirkby --- Geography

Group 2: Multimedia imaging

1.000 Dr K.C. Ng --- Music
0.511 Dr D.G. Cooper --- Music
0.461 Dr R.D. Boyle --- Computing
0.275 Dr V.A.F. Gammon --- Education
0.274 Prof D.C. Hogg --- Computing
0.210 Dr G.R. Rastall --- Music
0.203 Mr D. Lindley --- English
0.203 Prof J.G. Rushton --- Music
0.186 Prof P.M. Dew --- Computing
0.172 Dr A.J. Bulpitt --- Computing

Medical Imaging

1.000 Dr A.J. Bulpitt --- Computing
0.582 Prof D.C. Hogg --- Computing
0.499 Dr R.D. Boyle --- Computing
0.382 Dr N.D. Efford --- Computing
0.351 Miss S.A. Smith --- Development Nursing Policy and Practice
0.339 Prof M.J. Kirkby --- Geography
0.338 Prof A.J. Daly --- Institute for Transport Studies
0.308 Prof A.G. Wilson --- Geography
0.307 Prof C.M. Snowden --- Electronic and Electrical Engineering
0.293 Dr J.E.J. Staggs --- Fuel and Energy

Group 3: Natural language processing

1.000 Mr E.S. Atwell --- Computing
0.359 Dr R.D. Boyle --- Computing
0.335 Dr D.C. Souter --- Computing
0.326 Prof D.C. Hogg --- Computing
0.283 Dr M.D. Brown --- Mechanical Engineering
0.273 Prof A.J. Daly --- Institute for Transport Studies
0.271 Dr L.J. Cameron --- Education

0.258 Dr D.P. Watling --- Institute for Transport Studies
0.257 Dr M. Bygate --- Education
0.256 Prof M.J. Kirkby --- Geography

Group 4: Qualitative Spatial Reasoning

1.000 Prof A.G. Cohn --- Computing
0.868 Dr B. Bennett --- Computing
0.300 Prof M.C. Clarke --- Geography
0.246 Dr J. Mason --- Sociology and Social Policy
0.239 Prof D.C. Hogg --- Computing
0.239 Dr I.J. Turton --- Geography
0.239 Mr P.L. Mott --- Computing
0.230 Dr P.M. Hill --- Computing
0.220 Dr G.P. Clarke --- Geography
0.194 Mr S.A. Roberts --- Computing

Logic Programming
1.000 Dr P.M. Hill --- Computing
0.311 Mr P.L. Mott --- Computing
0.245 Dr B.M. Smith --- Computing
0.232 Dr L.G. Proll --- Computing
0.230 Prof A.G. Cohn --- Computing
0.164 Dr P. Brna --- Computer Based Learning Unit
0.157 Dr B. Bennett --- Computing
0.133 Prof M.E. Dyer --- Computing
0.132 Dr I.J. Turton --- Geography
0.121 Prof P.M. Dew --- Computing

Group 5: Database integration

1.000 Mr S.A. Roberts --- Computing
0.507 Dr J.E. McCormack --- Computing
0.416 Dr J. Hogg --- Geography
0.384 Dr N.D. Efford --- Computing
0.366 Prof M.C. Clarke --- Geography
0.352 Dr M.H. Birkin --- Geography
0.344 Mr P.L. Mott --- Computing
0.333 Dr G.P. Clarke --- Geography
0.328 Dr S.J. Carver --- Geography
0.325 Prof A.G. Wilson --- Geography

1.000 Mr P.L. Mott --- Computing
0.344 Mr S.A. Roberts --- Computing
0.311 Dr P.M. Hill --- Computing
0.239 Prof A.G. Cohn --- Computing
0.201 Dr B. Bennett --- Computing
0.165 Dr M.J. Carter --- Leeds University Business School
0.143 Prof G. Birtwistle --- Computing
0.119 Dr A.J. Maule --- Leeds University Business School
0.119 Mr E.S. Atwell --- Computing
0.118 Prof A.J.E. Anning --- Education

Group 6: Scientific Computing

Unstructured Adaptive Mesh Algorithms
1.000 Prof M. Berzins --- Computing
0.650 Dr P.K. Jimack --- Computing
0.368 Prof P.M. Dew --- Computing
0.282 Mr R. Fairlie --- Computing
0.282 Dr A.S. Tomlin --- Fuel and Energy
0.245 Mr J.R. Davy --- Computing
0.242 Prof M.E. Dyer --- Computing
0.230 Prof D.C. Hogg --- Computing
0.220 Prof P.H. Gaskell --- Mechanical Engineering
0.205 Prof S.N. Lane --- Geography

Parallel computing
1.000 Dr P.K. Jimack --- Computing
0.650 Prof M. Berzins --- Computing
0.277 Prof P.M. Dew --- Computing
0.232 Mr J.R. Davy --- Computing
0.207 Dr P.C. Brooks --- Mechanical Engineering
0.180 Prof M.E. Dyer --- Computing
0.168 Dr K.W. Dalgarno --- Mechanical Engineering
0.168 Dr I.J. Turton --- Geography
0.163 Dr G.D. Halikias --- Electronic and Electrical Engineering
0.148 Dr R. Hardy --- Geography

Group 7: Visualization

1.000 Dr K.W. Brodlie --- Computer Science
0.677 Dr J. Wood --- Computing
0.255 Dr P. Brna --- Computer Based Learning Unit
0.251 Prof P.M. Dew --- Computing
0.198 Prof J.A. Self --- Education
0.162 Dr J.B.C. Whitaker --- Chemistry
0.159 Mr R. Fairlie --- Computing
0.159 Prof D.C. Hogg --- Computing
0.152 Dr S.J. Carver --- Geography
0.143 Dr R.M. Pilkington --- Computer Based Learning Unit
0.137 Dr I.J. Turton --- Geography

Group 8: Theoretical Computer Science

Algorithms and Complexity
1.000 Prof M.E. Dyer --- Computing
0.383 Prof P.M. Dew --- Computing
0.380 Mr J.R. Davy --- Computing
0.340 Dr L.G. Proll --- Computing
0.246 Dr B.M. Smith --- Computing
0.245 Dr G.D. Halikias --- Electronic and Electrical Engineering
0.242 Prof M. Berzins --- Computing
0.218 Dr M. Kara --- Computing
0.198 Prof M.C. Clarke --- Geography
0.193 Dr I.J. Turton --- Geography

Formal methods
1.000 Prof G. Birtwistle --- Computing
0.188 Mr E.S. Atwell --- Computing
0.148 Dr M. Bygate --- Education
0.143 Mr P.L. Mott --- Computing
0.130 Dr A.S. Fowkes --- Institute for Transport Studies
0.118 Dr A.K.H. Holzenburg --- Biochemistry and Molecular Biology
0.114 Prof P.M. Dew --- Computing
0.113 Dr R.M. Drummond-Brydson --- Materials
0.113 Prof A.G. Cohn --- Computing
0.109 Dr T.F. Burgess --- Leeds University Business School

Informatics
Virtual Environment
1.000 Prof P.M. Dew --- Computing
0.751 Mr J.R. Davy --- Computing
0.435 Prof D.C. Hogg --- Computing
0.431 Dr K. Djemame --- Computing
0.431 Dr I.J. Turton --- Geography
0.394 Dr P. Brna --- Computer Based Learning Unit
0.383 Prof M.E. Dyer --- Computing
0.379 Prof J.A. Self --- Education
0.373 Prof A.J. Daly --- Institute for Transport Studies
0.372 Dr J.M. Curson --- Geography

Group 9: Transport Scheduling

1.000 Prof A. Wren --- Computing
0.925 Dr R.S. Kwan --- Computing
0.805 Dr S. Fores --- Computing
0.734 Mrs M.E. Parker --- Computing
0.710 Dr A.S.K. Kwan --- Computing
0.428 Dr L.G. Proll --- Computing
0.302 Dr B.M. Smith --- Computing
0.298 Prof P.W. Bonsall --- Institute for Transport Studies
0.265 Dr N.J. Ward --- Psychology
0.264 Mr J.D. Shires --- Institute for Transport Studies

Group 10: Computer Based Learning

1.000 Dr P. Brna --- Computer Based Learning Unit
0.658 Prof J.A. Self --- Education
0.527 Dr R.M. Pilkington --- Computer Based Learning Unit
0.394 Prof P.M. Dew --- Computing
0.336 Dr D. Goodley --- Sociology and Social Policy
0.332 Prof R.K.S. Taylor --- Continuing Education
0.332 Dr E.J. Foster --- Education
0.332 Prof D.C. Hogg --- Computing
0.329 Dr K.P. Forrester --- Continuing Education
0.306 Dr J.M. Curson --- Geography

Appendix E Background Knowledge of RDF/S

RDF

RDF stands for Resource Description Framework. It is a foundation for processing metadata that provides interoperability between applications that exchange machine-understandable information on the Web, it defines a mechanism for describing resources that makes no assumptions about a particular application domain, nor defines (a priori) the semantics of any application domain. The definition of the mechanism should be domain neutral, yet the mechanism should be suitable for describing information about any domain.

Basic RDF Model

The RDF data model is a model for representing named properties and property values. It is a syntax-neutral way of representing RDF expressions. I.e. two RDF expressions are equivalent if and only if their data model representations are the same.

The basic data model consists of three object types – resources, properties and statements

Resources

All things being described by RDF expressions are called resources, for example

- an entire Web page; such as the HTML document "http://www.w3.org/Overview.html"
- a part of a Web page; e.g. a specific HTML or XML element within the document source.
- a whole collection of pages; e.g. an entire Web site.
- an object that is not directly accessible via the Web; e.g. a printed book.

Resources are identified by a resource identifier. A resource identifier is a URI plus an optional anchor id. (see [URI]). Anything can have a URI; the extensibility of URIs allows the introduction of identifiers for any entity imaginable.

Properties

A property is a specific aspect, characteristic, attribute, or relation used to describe a resource. Each property has

- a specific meaning, defines its permitted values,
- the types of resources it can describe, and
- its relationship with other properties (see RDF Schema specification).

RDF properties also represent relationships between resources and an RDF model can therefore resemble an entity-relationship diagram. In object-oriented design terminology, resources correspond to objects and properties correspond to instance variables.

Statements

An RDF statement is a specific resource together with a named property plus the value of that property for that resource. These three individual parts of a statement are called, respectively,

- the subject,
- the predicate, and
- the object (i.e., the property value) can be another resource or it can be a literal; i.e., a resource (specified by a URI) or a simple string or other primitive datatype defined by XML.

Consider as a simple example the sentence:

"Ora Lassila is the creator of the resource http://www.w3.org/Home/Lassila."

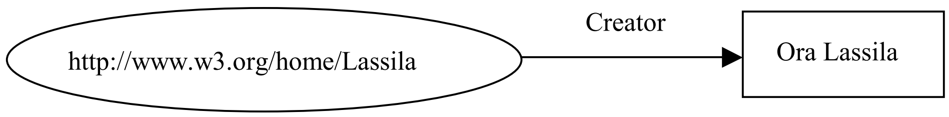
This sentence has the following parts:

Subject (Resource)	http://www.w3.org/Home/Lassila
Predicate (Property)	Creator
Object (literal)	"Ora Lassila"

Using directed labeled graphs (also called "nodes and arcs diagrams") in which

- the nodes (drawn as ovals) represent resources,
- arcs represent named properties, and
- nodes that represent string literals will be drawn as rectangles.

The sentence above would thus be diagrammed as:



In the example above

Ora Lassila is the creator of the resource http://www.w3.org/Home/Lassila.

The RDF/XML representation becomes:

```
<rdf:RDF>
  <rdf:Description about="http://www.w3.org/Home/Lassila">
    <s:Creator>Ora Lassila</s:Creator>
  </rdf:Description>
</rdf:RDF>
```

Note the use of XML namespaces in the RDF/XML representation. RDF uses the XML namespace facility to avoid confusion between independent -- and possibly conflicting -- definitions of the same term. Namespaces are simply a way to tie a specific use of a word in context to the dictionary (schema) where the intended definition is to be found.

RDF Schema

Schemas and Namespaces

It is crucial that both the writer and the reader of an RDF statement understand the same meaning for the terms used, such as Creator, approvedBy, Copyright, etc. or confusion will result.

Meaning in RDF is expressed through reference to a schema. A schema defines the terms that will be used in RDF statements and gives specific meanings to them. A schema is the place where definitions and restrictions of usage for properties are documented.

In RDF, each predicate used in a statement must be identified with exactly one namespace, or schema. However, a Description element may contain statements with predicates from many schemas.

Core Classes in RDFS

- *rdfs:Resource:* All things described by RDF are called resources, and are members of the class rdfs:Resource.
- *rdfs:Class:* This corresponds to the generic concept of a type or category of resource. RDF class membership is used to represent types or categories of resource. Two classes may happen to have the same members, while remaining distinct resources.
- *rdf:Property:* rdf:Property represents those resources that are RDF properties.
- *rdf:Statement:* The class of RDF statements.

Core Properties in RDFS

- *rdfs:subClassOf:* The rdfs:subClassOf property represents a specialization relationship between classes of resource. The rdfs:subClassOf property is transitive.
- *rdf:type:* The rdf:type property indicates that a resource is a member of a class. When a resource has an rdf:type property whose value is some specific class, we say that the resource is an instance of the specified class. The value of an rdf:type property will always be a resource that is an instance of rdfs:Class. The resource known as rdfs:Class is itself a resource of rdf:type rdfs:Class.
- *rdfs:range:* An instance of rdf:Property that is used to indicate the class(es) that the values of a property will be members of. The value of an rdfs:range property is always a Class. The rdfs:range property can itself be used to express this: the rdfs:range of rdfs:range is the class rdfs:Class. This indicates that any resource that is the value of a range property will be a class. The rdfs:range property is only applied to properties. This can also be represented in RDF using the rdfs:domain property. The rdfs:domain of rdfs:range is the class rdf:Property. This indicates that the range property applies to resources that are themselves properties.

- *rdfs:domain:* An instance of rdf:Property that is used to indicate the class(es) that will have as members any resource that has the indicated property. The rdfs:domain of rdfs:domain is the class rdf:Property. This indicates that the domain property is used on resources that are properties. The rdfs:range of rdfs:domain is the class rdfs:Class. This indicates that any resource that is the value of a domain property will be a class.

Other Important Classes and Properties

- *rdfs:subPropertyOf:* The property rdfs:subPropertyOf is an instance of rdf:Property that is used to specify that one property is a specialization of another. Sub-property hierarchies can be used to express hierarchies of range and domain constraints.
- *rdfs:label:* The rdfs:label property is used to provide a human-readable version of a resource's name.
- *rdfs:comment:* The rdfs:comment property is used to provide a human-readable description of a resource. A textual comment helps clarify the meaning of RDF classes and properties. Such inline documentation complements the use of both formal techniques (Ontology and rule languages) and informal (prose documentation, examples, test cases). A variety of documentation forms can be combined to indicate the intended meaning of the classes and properties described in an RDF Schema.

Multilingual documentation of schemas is supported at the syntactic level through use of the xml:lang language tagging facility. Since RDF schemas are expressed as RDF graphs, vocabularies defined in other namespaces may be used to provide richer documentation.

Source: http://bioserv.cis.nctu.edu.tw/bio/book/RDF.htm

Appendix F Expertise Model for Expertise Matching in Academia (Represented in RDFS)

```
<?xml version= “1.0”?>

<rdf : RDF xml:lang= “en”
        xmlns:rdf= “http://www.w3.org/1999/02/22-rdf-syntax-ns#”
        xmlns:rdfs= “http://www.w3.org/2000/01/rdf-schema#”
        xmlns= “ ”>

<rdfs:Class rdf:ID = “Organization”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Department”>
        <rdfs:subClassOf  rdf:resource = “#Organization”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Institute”>
        <rdfs:subClassOf  rdf:resource = “#Organization”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Research_Group”>
        <rdfs:subClassOf  rdf:resource = “#Organization”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Person”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Employee”>
        <rdfs:subClassOf  rdf:resource = “#Person”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Student”>
        <rdfs:subClassOf  rdf:resource = “#Person”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Researcher”>
        <rdfs:subClassOf  rdf:resource = “#Employee”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Lecturer”>
        <rdfs:subClassOf  rdf:resource = “#Employee”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Senior_Lecturer”>
        <rdfs:subClassOf  rdf:resource = “#Lecturer”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Assistant”>
        <rdfs:subClassOf  rdf:resource = “#Employee”/>
```

```
</rdfs:Class>

<rdfs:Class rdf:ID = “Teaching_Assistant”>
        <rdfs:subClassOf  rdf:resource = “#Assistant”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Research_Assistant”>
        <rdfs:subClassOf  rdf:resource = “#Assistant”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Research Fellow”>
        <rdfs:subClassOf  rdf:resource = “#Employee”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Professor”>
        <rdfs:subClassOf  rdf:resource = “#Employee”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “PhDStudent”>
        <rdfs:subClassOf  rdf:resource = “#Student”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Publication”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Book”>
        <rdfs:subClassOf  rdf:resource = “#Publication”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Journal”>
        <rdfs:subClassOf  rdf:resource = “#Publication”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “OnlinePublication”>
        <rdfs:subClassOf  rdf:resource = “#Publication”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Article”>
        <rdfs:subClassOf  rdf:resource = “#Publication”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “TechnicalReport”>
        <rdfs:subClassOf  rdf:resource = “#Article”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “JournalArticle”>
        <rdfs:subClassOf  rdf:resource = “#Article”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “ArticleInBook”>
        <rdfs:subClassOf  rdf:resource = “#Article”/>
</rdfs:Class>
```

```
<rdfs:Class rdf:ID = “ConferencePaper”>
        <rdfs:subClassOf  rdf:resource = “#Article”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “WorkshopPaper”>
        <rdfs:subClassOf  rdf:resource = “#Article”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Project”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Research_Topic”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Classification”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdfs:Class rdf:ID = “Expertise”>
        <rdfs:subClassOf  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Resource”/>
</rdfs:Class>

<rdf:Property rdf:ID = “first_name”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “last_name”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “homepage”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “email”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “Pub_title”>
        <rdfs:domain  rdf:resource = “#Publication”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “Proj_title”>
        <rdfs:domain  rdf:resource = “#Project”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>
```

```
<rdf:Property rdf:ID = “Pub_abstract”>
        <rdfs:domain  rdf:resource = “#Publication”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “Proj_abstract”>
        <rdfs:domain  rdf:resource = “#Project”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “YearOfPub”>
        <rdfs:domain  rdf:resource = “#Publication”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “Start_date”>
        <rdfs:domain  rdf:resource = “#Project”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “End_date”>
        <rdfs:domain  rdf:resource = “#Project”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “author_of”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “#Publication”/>
</rdf:Property>

<rdf:Property rdf:ID = “works_on”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “#Project”/>
</rdf:Property>

<rdf:Property rdf:ID = “supervises”>
        <rdfs:domain  rdf:resource = “#Employee”/>
        <rdfs:range  rdf:resource = “#PhDStudent”/>
</rdf:Property>

<rdf:Property rdf:ID = “memberOf”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “#Organization”/>
</rdf:Property>

<rdf:Property rdf:ID = “Org_name”>
        <rdfs:domain  rdf:resource = “#Orgnization”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “description”>
        <rdfs:domain  rdf:resource = “#Expertise”/>
        <rdfs:domain  rdf:resource = “#Research_Topic”/>
        <rdfs:range  rdf:resource = “http://www.w3.org/2000/01/rdf-schema#Literal”/>
</rdf:Property>

<rdf:Property rdf:ID = “has_expertise”>
        <rdfs:domain  rdf:resource = “#Person”/>
        <rdfs:range  rdf:resource = “#Expertise”/>
</rdf:Property>

<rdf:Property rdf:ID = “Relate_to”>
        <rdfs:domain  rdf:resource = “#Expertise”/>
        <rdfs:domain  rdf:resource = “#Research_Topic”/>
        <rdfs:range  rdf:resource = “#Classification”/>
</rdf:Property>

<rdf:Property rdf:ID = “researchInterest”>
        <rdfs:domain  rdf:resource = “#Student”/>
        <rdfs:range  rdf:resource = “#Research_Topic”/>
</rdf:Property>
```

Appendix G Questionnaire for the First Experiment – Compare Extended Expertise Matcher with Current One

Pre-Experiment Questionnaire

Please tell us about your background by answering these questions.

1. What is your name and email address?

………………………………………………………………………………………

2. Which year are you in your PhD study?

………………………………………………………………………………………

3. What is your age? (please tick relevant box)

10-20 □

20-30 □

30-40 □

above 40 □

4 How often do you use search tools? (please tick relevant box)

Daily □

A few times a week □

A few times a month □

Rare □

Never □

5 Do you know how to use operators (AND, OR, NOT) when you search something?

………………………………………………………………………………………

6 How much do you know about searching a database? (please tick relevant box)

Nothing □

a little □

quite a lot □

Post-Experiment Questionnaire

1 What keyword(s) did you search on (Please write down any operators you have used in Search 1 e.g. virtual or working)?

…………………………………………………………………………………………

2 Did any of the 3 searches bring up your supervisor?

- which search was it? (please tick relevant box)

  Search 1 □ Search 2 □ Search 3 □

- How far down the list was your supervisor placed in each search?

  Search 1……………

  Search 2…………….

  Search 3…………….

- What do you think of the ranking of the results on Searches 2 and 3, are they helpful in finding your supervisor or potential supervisor(s)? (Please give a brief explanation of the reason for your response)

  …………………………………………………………………………………………

  …………………………………………………………………………………………

  …………………………………………………………………………………………

- Look at the details page of each expert retrieved by each search, how many people could be your potential supervisor for your research?

  Search 1…………………………………………………………………………….

  Search 2…………………………………………………………………………….

  Search 3…………………………………………………………………………….

3 Which form of query do you think is more convenient (with or without operators e.g. AND, OR, NOT)? (please tick relevant box)

With □ Without □

4 Look at the order of the results given by Search 2 and Search 3. Are they different? How different (please tick relevant box) ?

identical □ similar □ half □ few □ no □

5 If there are some differences in Search 2 and Search 3. Find the people in the top 10 results who are ranked differently in Search 2 and Search 3, by looking at the details of each person which result would you say has the more appropriate rank? (Please give a brief explanation of the reason for your response)

Search 2 □ Search 3 □

……………………………………………………………………………………………

………..…………………………………………………………………………………

6 What keyword(s) did you add in (Please write down any operators you have used in Search 1 e.g. virtual or working or environment)?

……………………………………………………………………………………………

……………………………………………………………………………………………

7 Did any of the 3 searches bring up your supervisor?

- Which Search was it? (please tick relevant box)

  Search 1 □ Search 2 □ Search 3 □

- How far down the list was your supervisor placed in each search?

  Search 1……………

  Search 2…………….

  Search 3…………….

- What do you think of the ranking of the results on Searches 2 and 3, are they helpful in finding your supervisor or potential supervisor(s)? (Please give a brief explanation of the reason for your response)

  ……………………………………………………………………………………………

  …………..………………………………………………………………………………

  ……………………………………………………………………………………………

- Look at the details page for top ten people in each search, how many people could be your potential supervisor for your research?

  Search 1…………………………………………………………………………….

  Search 2…………………………………………………………………………….

  Search 3…………………………………………………………………………….

8 Which form of query do you think is more convenient (with or without operators e.g. AND, OR, NOT)? (please tick relevant box)

With □ Without □

9 Look at the order of the results given by Search 2 and Search 3. Are they different? How different (please tick relevant box) ?

identical □ similar □ half □ few □ no □

10 If there are some differences in Search 2 and Search 3. Find the people in the top 10 results who are ranked differently in Search 2 and Search 3, by looking at the details of each person which result would you say has the more appropriate rank? (Please give a brief explanation of the reason for your response)

Search 2 □ Search 3 □

…………………………………………………………………………………………

…………..……………………………………………………………………………

…..………………..……………………………………………………………………

11 Which search do you think is the easiest one to help you find the most suitable supervisor? (Please give a brief explanation of the reason for your response)

Search 1 □ Search 2 □ Search 3 □

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

…………………………………………………………………………………………

Appendix H Questionnaire for Evaluation the Expertise Locator against the Extended Expertise Matcher

Pre-Experiment Questionnaire

1. Name: ________________________ email: _________________________________

2. Did you want to know who had the expertise in your preferred research area when you applied to be a PhD student at the School of Computing, University of Leeds?

Yes □ No □

If no, why?

____

____

3. Did you choose your supervisor(s) yourself when you applied to be a PhD student?

Yes □ No □

If no, why?

You were not asked to choose yourself □

You did not mind who would be your supervisor. □

Other reasons:

____

4. Where did you seek the information used to locate your potential supervisor(s)?

a. The homepage of each member of staff □

b. The website of each research group □

c. The technical reports from "Research Report Series" on the School website □

Please indicate below if you also searched other information resources

____

____

____

5. Did you find sufficient information you required from a single data source (e.g., from a personal homepage)?

Yes □ No □

If yes, which data source did you look at?

____

If no, why?

____

____

6. How long did it take you to locate the supervisor(s) in your preferred research area?

Less than 30 minutes □

Less than 1 hour □

Less than 2 hours □

Less than 4 hours □

Longer than this □

7. How easy was it to find the people who have expertise in your preferred research area?

very easy □ easy □ O.K. □ difficult □ very difficult □

Please give a brief explanation of the reason for your response:

____

____

____

8. How long have you been a PhD student? Have you done the literature review in your specific area?

____

____

9. Indicate below how useful the following types of information were to you when choosing your supervisor(s)?

	very useful	useful	not useful
Research interests	□	□	□
Research group	□	□	□
Position	□	□	□
Publications	□	□	□
Projects	□	□	□
PhD students	□	□	□
Teaching activities	□	□	□
Affiliations	□	□	□
Biography	□	□	□

Post-Experiment Questionnaire

1. For the first search, how many potential supervisors were displayed on the left-hand side of the page?

____

2. Look at the detail publication and project information page of each potential supervisor, do you think it provides enough information you need? (please give a brief explanation for your response)

Yes □ No □

____

____

3. How many potential supervisors did you finally accept? How far down the list were the ones which you accepted (state position)?

____

4. Was the name of your real supervisor(s) in the final list of accepted potential supervisor? How far down the list was your supervisor placed in the left list?

____

5. Did you agree with the ranking of the results when you viewed the detail pages of the potential supervisors?

Agree □ Partially agree □ Disagree □

please give a brief explanation for your response

____

____

6. For the second search, how many research areas you have accepted? (Please tick relevant box)

1 □ 2 □ 3 □ 4 □

7. For each research area you have accepted how many potential supervisors were listed?

No.1_______ No.2_______ No.3_______ No.4_______

8. Look at the details page for your potential supervisor, are you satisfied with the content in it?

(1- very satisfied; 5 – not satisfied at all)

1 □ 2 □ 3 □ 4 □ 5 □

In more detail, which of following information is useful?

Personal (contact) information □

Homepage □

Research Interest □

Research Group □

Publication □

Project □

What else information do you think should be included?

____

____

Do you agree with the ranking of the results after you viewed the detail pages of the potential supervisors? (please give a brief explanation for your response)

Agree □ Partially agree □ Disagree □

Comparing these pages to those personal detail pages in Search 1, which one do you prefer?

Search 1 □ Search 2 □

9. How many potential supervisors were accepted? What was the position of the accepted potential supervisor(s)?

research area 1: 1 □ 2 □ 3 □ 4 □

research area 2: 1 □ 2 □ 3 □ 4 □

research area 3: 1 □ 2 □ 3 □ 4 □

10. Was the name of your real supervisor(s) in the final list of accepted potential supervisor(s)?

Yes □ No □

11. Looking at the two sets of results obtained from Search 1 and Search 2, how different were they (please tick relevant box)?

identical □ similar □ half □ quite different □ totally different □

If the two sets of results were not identical, which one was more appropriate?

Search 1 □ Search 2 □

12. Comparing the two searches, which Search do you find more useful? (Please give a brief explanation of the reason for your response)

Search 1 □ Search 2 □ they are the same □

13. Do you think you have the ability to assess the potential supervisors' expertise and why?

14. If you could change something about the system, what would you change?

15. Overall what do you think of the two searches so far?

16. If these two searches are available when you apply for PhD study in the University of Leeds, will you use one of them to search for your potential supervisor?

17. Do you have any other comments?

Appendix I: Computing Techniques Used in Geoinformatics

(1) DATA EXTRACTION AND STATISTICS

   a. Information retrieval
   b. Data mining
   c. Spatial data analysis
   d. Knowledge discovery tools
   e. Geostatistics
   f. Data quality

(2) MODELLING, MAPPING AND PATTERN RECOGNITION

   a. Simulation
   b. Modelling of complex systems
   c. Process-based modelling
   d. Statistical modelling (predictive and descriptive)
   e. Diagnostics and pattern recognition
   f. Cellular automata
   g. Artificial intelligence
      i. Intelligent agents
      ii. Expert systems
      iii. Neural networks
      iv. Fuzzy computing
      v. Advanced numerical algorithms
      vi. Smart spatial analysis

(3) COMPUTING ENVIRONMENTS

   a. Grid-based processing
   b. Computer architecture and design
   c. Distributed computing environments
      i. Distributed GIS environments
      ii. Collaborative spatial decision making
   d. Problem solving environments

(4) VISUALISATION

a. Interactive visualisation

b. Virtual reality

c. Virtual environments

d. Multimedia

e. GIS

(5) KNOWLEDGE MANAGEMENT

a. Knowledge discovery

b. Spatial decision support systems

Appendix J: Keywords Associated with the Six Items in the Geography Classification

1. *Population & Migration* - residential developments, census, ethnic minority, populations migration dynamics, transnationalism, diaspora, family migration, estimation and projection demography, population policy, mortality

2.1 *Urban or regional geography* – urban, rural, land use, landscape, location planning, local, national, growth predict, communities, houses consumer, society, urban consumption, retailing, countryside, policies, global, cities, transportation planning, community planning, economy, space

2.3 *Economic geography* - retailing store, network, expansion, market saturation, competition, firms grocery, spatial, monopolies, duopolies, floorspace, financial transactions, deals, M&A, service, partner, organizations, health care, international business, globalisation, economic development, regional analysis

3.1 *Water policy and development* – flood, water resources, water deficits, environment, water policy, resource management

3.2 *Sustainable development and resource geography* – sustainability, sustainable, water management, agricultural systems, conservation, policy, pollution control, wilderness, conservation, environmental development, sustainability strategies and indicators, energy analysis, renewable energy

3.3 *Global environmental history and change* – deforestation, desertification, wilderness, climate change, climate modelling, atmosphere, cloud, physics, marine, ice, soil, hydrochemistry of upland ecosystems, pollution ecology, ozone depletion

Appendix K: Important Areas of Current Research in Geoinformatics

- Acquisition of digital geodata in the field and in the laboratory
- Global positioning systems and navigation systems
- Analysis and evaluation of remotely sensed data
- Databases, metadata databases, methods databases and models databases
- Geographical information systems, environmental information systems
- Development of open, interoperable systems
- Improvement of the usability of geosoftware
- Multimedia applications in the geosciences
- Digital cartography systems
- 3D-visualization, VR (virtual reality) - developments
- Decision-support systems
- Numerical simulation models and prognosis models for spatial data
- Data processing which supports local, regional and national planning
- Data processing which supports landscape planning and studies of climate suitability
- Artificial neural networks and fuzzy set theory for natural resource studies
- GIS and public health
- Simulation population

Source: **http://castafiore.uni-muenster.de/vorlesungen/Geoinformatics/frames/fsteuer.htm**