

# Collaborative Expertise Retrieval: A Referral Approach to Finding Distributed Experts

Weimao Ke

Laboratory of Applied Informatics Research  
University of North Carolina at Chapel Hill  
wke@unc.edu

Javed Mostafa

Laboratory of Applied Informatics Research  
University of North Carolina at Chapel Hill  
jm@unc.edu

## ABSTRACT

We live in a networked environment, where expertise and computing powers are highly distributed. A distributed approach to the retrieval of distributed expertise appears to be reasonable. We propose an agent simulation framework where distributed agents, representatives of information consumers, providers (experts), and referrers, learn to collaborate with each other for finding the experts. Two fundamental information organization operations, namely, clustering and classification, will be used to organize information items and to label information needs within each agent. The organized/indexed information is then mapped to the agent's perception of the society (neighbors) reinforced through machine learning. We reason why this approach is desirable and propose the investigation of: 1) whether information organization at individual levels can help expertise retrieval at the collective level; and 2) to what extent learning can facilitate the adaptive building of an efficient agent network for the finding of expertise. The proposed approach is presented as a conceptual framework. However, potentially, the implementation of the approach will provide guidance on new information and expertise retrieval models that utilize the huge distributed informational and computational resources on the Web and beyond the Web.

## Categories and Subject Descriptors

H.3.3 [Information storage and retrieval]: Information Search and Retrieval—*Search process*

## General Terms

Algorithms, Measurement, Performance, Experimentation

## Keywords

Information retrieval, expertise retrieval, referral system, information filtering, information organization, agent, P2P

## 1. INTRODUCTION

We live in a distributed networked environment. In reality, we have different expertise, share information with each other, and ask trusted peers for information/opinions on various questions. The World Wide Web is a good example of information distribution, where Web sites serve narrow

information topics and tend to form communities through hyperlink connections. Using distributed nodes to share the computational burden and to collaborate in retrieval operations appears to be reasonable.

Research on network sciences has discovered that the planet we live on is a small world with six degrees of separation [10]. That is, there are only about 5.5 social connections between any two persons in the huge population of billions. This small world phenomenon also appears in various types of networks such as the World Wide Web [1]. One implication of this is that information or expertise might be only a few degrees (connections) away from the one who needs it. This provides the potential for distributed algorithms to traverse such a network and to find what is desired efficiently. However, the question is how we, or automatic software agents on behalf of us, can learn to find shortcuts to peers that have desired expertise.

In this work, we will propose a novel model for collaborative expertise retrieval through referrals. The task is to route (refer) information needs to experts (information providers) that have relevant information resources or expertise to satisfy the needs. We will propose the use of multi-agent simulations for the study of this distributed model and investigate: 1) whether and how information organization at individual levels can help expertise retrieval at the collective level; and 2) to what extent learning can facilitate the adaptive building of an efficient agent network for the finding of expertise. Potentially, the findings of this work will provide guidance on new information and expertise retrieval models that utilize the huge distributed computational and informational resources on the Web and beyond the Web [5].

## 2. RELATED WORK

Expertise Retrieval (ER) is an emerging area related to IR which recognizes the fact that individuals have distributed collections of information and expertise. In other words, it is unrealistic to assume a global collection of information at one place. Prior to the retrieval of information is the need for finding the expert(s) who potentially has the relevant information [5].

Our primary focus will be on the automatic referral for finding experts in a distributed networked environment. The paper will discuss research on Information Organization (IO) for the mapping of information needs and expertise at the individual levels. Additionally, it will utilize Machine Learning (ML) algorithms to adaptively reinforce local connections and to collectively optimize the global referral network for the efficient retrieval of expertise. Finally, it will suggest

the use of multi-agent technologies to simulate a distributed network for conducting experimental studies to examine the questions discussed earlier.

## 2.1 Distributed Info and Expertise Retrieval

Distributed IR has become a fast-growing research topic in the last few years. Recent distributed IR research has been focused on intra-system retrieval fusion, cross-system communication, decentralized P2P network, and distributed information storage and retrieval algorithms [2]. Research also concentrated on decentralized genetic algorithms for feature selection and distributed solutions for intelligent information collection and filtering.

Referral systems for expertise retrieval have attracted increasing research attention. Kautz et al. (1997) observed that much valuable information was not kept on-line for issues such as privacy. Nonetheless, this hidden information is potentially accessible through personal referrals in a social network [5]. The fact that people shared pointers to experts through word-of-mouth motivated researchers to study automated expertise retrieval systems based on referral chains [3]. The REFERRALWEB was one of the early expertise retrieval systems that demonstrated promising results on referral accuracy and responsiveness [5]. In related works, software agents were used to traverse social connections for the finding of experts in an autonomously distributed manner [3, 15].

In recent years, research examined application of distributed methods to finding expertise for information filtering [12]. It was shown that, by using acquaintance lists and other collaboration strategies, learning helped distributed agents seek expertise more effectively without consuming too much communication resources. Different learning algorithms were proposed to enable effective and efficient collaboration of experts in distributed environments [6]. These studies examined various parameters for collectively finding experts on processing information. Research showed promising results on small networked communities and called for closer scrutiny on scalability.

McDonald and Ackerman (2000) acknowledged the important roles played by information mediators in organizational settings. They recognized the potential for locating expertise by making referrals and proposed an expertise retrieval system based on a range of collaborative recommendation models and behaviors [9]. The EXPERTISE RECOMMENDER demonstrated the flexibility and usefulness of such a system for automatically locating experts in a work setting. Zhang and Ackerman (2005) studied various strategies for social network based search and found that social characteristics, in addition to graph characteristics, had important impacts on the searching process [16]. It also demonstrated the usefulness of simulation in distributed expertise retrieval research.

Research on distributed methods has drawn controversies over issues such as privacy and security. Some researchers reasoned that because of no centralized database, a distributed approach is more fault tolerant and less vulnerable to attacks and privacy leak [12, 3]. Others tended to disagree and argued that a distributed architecture may deploy personal information and opinions to each user, "risking exposure of information to every peer" [13, p. 317]. These questions, together with many other challenges in distributed retrieval and filtering, require continued examination.

## 2.2 Information Organization

Information organization is an important step toward the effective and efficient retrieval/filtering of information items. Automatic methods for information organization has been useful in centralized information retrieval operations. Although rarely discussed in distributed IR literature, it is potentially useful by building distributed indexes of expertise. We argue that, without information organization at the individual levels, it will be very difficult, if not impossible, to build orderly referral chains to expertise at the collective level.

Humans understand the world through the process of organizing concepts into an ordered group of categories. Clustering and classification, as information organization mechanisms, involve the aggregation of like-entities [8]. While clustering organizes information by grouping similar or related entities together and derives patterns (concepts) from data, text categorization, or classification, is to label texts with concepts from an existing set [14].

Text clustering and classification are fundamental functions of Information Retrieval (IR) and can be applied to various information management processes such as indexing and filtering [14]. Automatic clustering and classification, as applied in automatic information extraction and knowledge discovery, have been important research topics in Machine Learning (ML) and IR [7, 14].

The usefulness of automatic information organization for information retrieval and filtering has been extensively studied [11, 14]. Research examined the impact of information organization on automatic filtering of information, in which document classification served as an intermediate stage [11]. The proper use of classification reduced the memory size for information representation but maintained a level of granularity that could be accurately mapped to information needs.

Likewise, by making individual sets of expertise in order through information organization, peers will facilitate the distributed construction of referral chains to desired expertise. We reason that it is the ability of conceptual abstraction supported by information organization that will enable agents to understand each other's expertise. This makes possible an efficient referral network that has been demonstrated in human societies [5].

## 3. STUDY PROPOSAL

A multi-agent framework is useful for studying complex social and information systems. By definition, an agent is a computer program capable of autonomous action to meet its designed objectives in certain environment [4]. In multi-agent systems, agents are treated as distributed peers that have scattered intelligence and can collaborate with each other to do certain tasks. Research on information retrieval has relied upon multi-agent technologies for better understanding of collective retrieval operations in distributed environments [12, 6]. This framework also responds to the increasing computational demands for retrieval and offers a great potential for scalability.

We propose the use of multi-agent simulations for the study of expertise retrieval in a distributed networked environment. It involves referrals of information needs to experts that have matched information resources or expertise to satisfy the needs. We present the conceptual model below and elaborate on the major components of the model.

Assume that agents, representatives of information consumers, providers (experts), and referrers, live in an  $n$  dimensional space. An agent’s location in the space is determined by the expertise it has. Therefore, finding experts for an information need is to route the need/query to agents in the *relevant* expertise space. To simplify the discussion, assume all experts can be characterized using two dimensions (features). Figure 1 visualizes a 2D representation of the conceptual model.

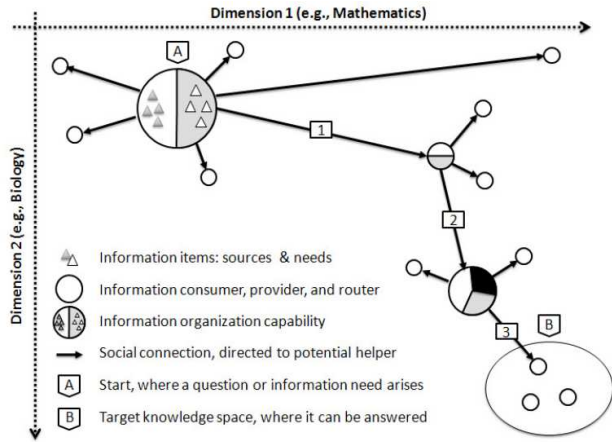


Figure 1: Agent collaboration network

As shown in Figure 1, the expertise space has two dimensions, e.g., Mathematics and Biology. Suppose Agent A has a need for expertise which is highly related to “mathematics” and “biology.” That means it can be answered by agents in the space B. The problem becomes how agents in the connected society can collectively point to the right direction and find out a shortcut to the space B. We decide that, in order for us to scrutinize the dynamics of referral traffics, only one copy of each query will traverse in the network. In other words, each agent will *only* forward a query to *one* chosen neighbor. They forward it from one to another until it reaches the destination.

### 3.1 Organization: Index Expertise and Needs

In the expertise space, direction matters. Pointing to the right direction means the agents have to have some ability to differentiate items on certain dimensions. For instance, one should be able to tell if a query is related to mathematics or not in order to route the query properly on that dimension. This is essentially an information organization problem, which involves clustering and classification.

Firstly, an agent needs to derive patterns or concepts from information it already has through clustering. This will provide the basis of each agent’s knowledge and enable the labeling of information needs. Now, when a query is posed to it, the agent will be able to tell what the query is about and assign a label to it. The label associated with the query serves as a clue for the potential referral direction.

To be realistic, each agent only has a tiny fraction of the global expertise. Hence, its information organization capability, i.e., clustering and classification, is constrained. Individual agents do not know all the dimensions of the space. In other words, each can only label the query in terms of a few dimensions—the extreme case is that one can only do *binary*

classification on *one dimension*. Nonetheless, the diversity of the agent community will help if they work collaboratively. Potentially, they will refine the referral direction collectively until the query reaches the targeted expertise space.

Given that limited information within each agent, many widely appreciated classification methods, such as k Nearest Neighbor (kNN) and Support Vector Machine (SVM), require a fair amount of training data and are therefore not applicable [14]. For this study, we will use a simple centroid-based approach that has produced competitive results on a benchmark collection in a similar context [6].

### 3.2 Mapping Indexed Needs to Neighbors

Pointing to the right direction also requires that each agent knows which neighbor(s) should be contacted given the direction it has labeled. Therefore, there needs to be a mechanism of mapping a labeled query to a potential *good* neighbor. By *good neighbor*, we mean agents on a short path to the targeted expertise space. Sometimes, a neighbor might have the expertise to answer the query directly; sometimes, that means the neighbor can forward the query to another agent potentially closer to the desired expertise space.

Initially, of course, an agent knows nothing about its neighborhood and has to explore by trying randomly. Overtime, the agent will learn from interactions with its neighbors and get a better sense of who has (or has connections to) what expertise. In other words, a ranking function for each label can be learned from the history of interactions, which is used to predict good neighbors in the future.

In previous research [6], we explored the use of a reinforcement learning algorithm called Pursuit Learning for the automated referral to expertise on information retrieval tasks. By rewarding *successful* collaborations and penalizing *failures*, the algorithm enabled distributed agents to find experts effectively and efficiently.

Keep in mind that each agent only knows about a limited number of neighbors for a couple of reasons: first, it is rarely possible for the agents to remember every other in a huge network given their memory constraints; second, if an agent does remember all the others, it will take forever for the learning to progress—there are simply too many neighbors to explore.

### 3.3 Network Topology and Distributions

In the networked environment of agents, it matters how expertise is distributed and how agents connect to each other. If expertise is uniquely distributed among the agents—i.e., each agent has a unique set of information items—then the retrieval of expertise from a huge network is like finding a needle in the haystack. However, in reality, we have overlapped expertise among each other. It becomes easier to find one or some of the experts given an information need.

Another important variable in this study is the network topology, e.g., the size of the network, the in-/out-degree distributions, average path length, etc. This study assumes that every agent is connected to the network and, furthermore, the network is a small world. This is to make sure that, theoretically, all agents are only a few degrees from each other and any query can be potentially answered after a short traverse in the network.

Now the research question becomes: Given that there is indeed a shortcut (or shortcuts) to a desired expertise space,

how can agents find the shortcut(s) collectively? Additionally, as the small world phenomenon exists in a variety of networks, we do not need to arbitrarily construct such a network topology; it is something already there in reality. Seen in this light, this is a variable that we have already known and controlled.

Experimental simulations of the proposed model can be run on a scholarly communication dataset. For example, given a collection of scholarly publications, we can create a community of agents to represent scholars who authored the publications, which in turn serve as their individual expertise. A co-authorship network, presumably a small world, can be derived from the data to initialize the referral neighborhood. The simulative task for the agents could be: When assigned a paper to “read,” to find the experts (ideally the authors) to help interpret the work.

To find expertise, each agent will only forward an information need (query) to one neighbor and so forth. If an agent has been contacted for a query, it will not be involved in this query again. In addition, a constraint on the maximum involvement, i.e., the maximum number of agents to be involved in each query, will be applied to all the tasks. This ensures that a query will not trouble the network for ever. We studied the maximum involvement as a variable within a small number of agents and plan to examine its impact on a large-scale agent community [6].

#### 4. THOUGHTS ON EVALUATION

The previous sections discuss major components of the conceptual model, which involve potential independent and control variables. The dependent variables of this study are effectiveness and efficiency of expertise retrieval. We need to evaluate how accurate the found experts are or how relevant the retrieved expertise is. In addition, efficiency is also important as the model aims not to overload the network. It turns out that the efficiency evaluation will involve a couple of levels.

At the individual agent (computing node) levels, efficiency is about how fast agents can perform information organization and machine learning. It is true that these functions will consume a certain amount of computational resources. Particularly, many clustering algorithms are algorithmically complex and computationally intensive. However, this is not a primary concern in the evaluation of efficiency for the following reasons. First, each agent only runs clustering once in order to initialize its indexing space. In real situations, this can be done when a computer idles. Although classification and machine learning are needed for each query, they require less computing power than clustering does.

More importantly, the objective of the study is to take advantage of individual computing powers for indexing and learning in order to minimize network traffics for distributed expertise retrieval. Presumably, these computing resources, distributed in the network, have not been sufficiently utilized.

#### Summary

In this paper, we argued that a distributed architecture is desirable for the retrieval of distributed expertise in a networked environment. We proposed an automatic referral system for finding experts and elaborated on a conceptual model that can be studied using multi-agent simula-

tions. In the model, information organization (IO) operations, namely, clustering and classification, will be used to organize information items (expertise) and to label information needs within each agent. The indexed information will then be mapped to the agent’s perception of the society (neighbors) reinforced through machine learning (ML). We walked through the rationale of this model and presented initial thoughts on how to evaluate the system. Potentially, the findings of this work will provide guidance on new information and expertise retrieval models that utilize the huge distributed informational and computational resources on the Web and beyond the Web.

#### 5. REFERENCES

- [1] R. Albert and A.-L. Barabasi. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.
- [2] J. Callan, F. Crestani, and M. Sanderson. SIGIR 2003 workshop on distributed information retrieval. *SIGIR Forum*, 37(2):33–37, 2003.
- [3] L. N. Foner. Yenta: a multi-agent, referral-based matchmaking system. In *AGENTS ’97*, pages 301–307, New York, NY, USA, 1997. ACM.
- [4] N. R. Jennings and M. J. Wooldridge, editors. *Agent technology: foundations, applications, and markets*. Springer-Verlag, Secaucus, NJ, USA, 1998.
- [5] H. A. Kautz, B. Selman, and M. A. Shah. The hidden web. *AI Magazine*, 18(2):27–36, 1997.
- [6] W. Ke, J. Mostafa, and Y. Fu. Collaborative classifier agents: studying the impact of learning in distributed document classification. In *JCDL ’07*, pages 428–437, 2007.
- [7] K. Knight. Mining online text. *Commun. ACM*, 42(11):58–61, 1999.
- [8] D. D. Lewis. *Text representation for intelligent text retrieval: a classification-oriented view*, pages 179–197. Hillsdale, NJ: Lawrence Erlbaum, 1992.
- [9] D. W. McDonald and M. S. Ackerman. Expertise recommender: a flexible recommendation system and architecture. In *CSCW ’00*, pages 231–240, New York, NY, USA, 2000.
- [10] S. Milgram. Small-world problem. *Psychology Today*, 1(1):61–67, 1967.
- [11] J. Mostafa, S. Mukhopadhyay, W. Lam, and M. Palakal. A multilevel approach to intelligent information filtering: Model, system, and evaluation. *ACM Trans. Inf. Syst.*, pages 368–399, 1997.
- [12] S. Mukhopadhyay, S. Peng, R. Raje, J. Mostafa, and M. Palakal. Distributed multi-agent information filtering - a comparative study. *JASIST*, 56(8):834–842, 2005.
- [13] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen. *Collaborative filtering recommender systems*, pages 291–324. Springer, Heidelberg, 2007.
- [14] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, 2002.
- [15] B. Yu and M. P. Singh. Searching social networks. In *AAMAS ’03*, pages 65–72, 2003.
- [16] J. Zhang and M. S. Ackerman. Searching for expertise in social networks: a simulation of potential strategies. In *GROUP ’05*, pages 71–80, NY, USA, 2005. ACM.