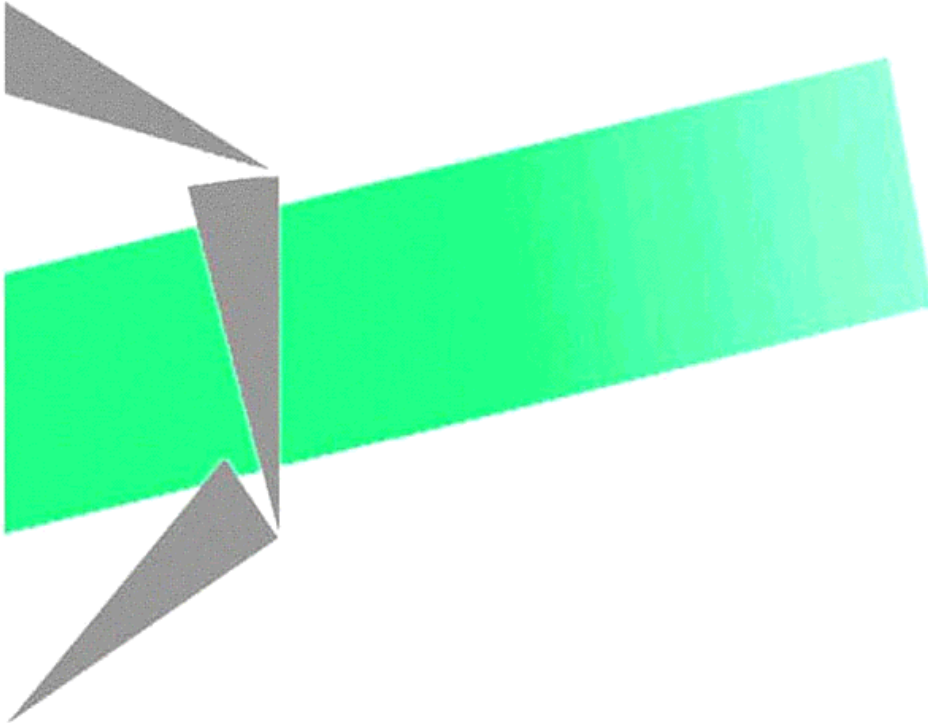


# Les cahiers du laboratoire Leibniz



## **Building Virtual Communities for Information Retrieval**

Daniel MEMMI, Olivier NEROT

Laboratoire Leibniz-IMAG, 46 av. Félix Viallet, 38000 GRENOBLE, France -

ISSN : 1298-020X

n° 64

Nov. 2002

Site internet : <http://www-leibniz.imag.fr/LesCahiers/>



# Building Virtual Communities for Information Retrieval

Daniel Memmi <sup>1</sup> & Olivier Nérot <sup>2</sup>

<sup>1</sup> LEIBNIZ-IMAG  
46 avenue Felix Viallet  
38000 Grenoble (France)  
memmi@imag.fr

<sup>2</sup> AMOWEBA SAS  
41 rue de Cronstadt  
75015 Paris (France)  
olivier@amoweba.com

## Abstract

The search for relevant information is often hindered by the initial difficulty in formulating precise requests, and because much knowledge is actually tacit and thus not easily accessible. Asking for human assistance is the usual response to these problems, but one can develop computer systems to help locate the right persons in the search for information. We will describe the structure and functioning of a collaborative, distributed search system designed to emulate the information-gathering functions of social communities. Such systems can be used to create virtual communities as well as to improve information retrieval. Potential benefits and likely problems of this approach will also be discussed.

## Keywords

information retrieval, peer-to-peer, collaborative search, distributed computing, social networks, knowledge management.

## 1. INTRODUCTION

In our knowledge-based economy, information retrieval methods have become essential to manage an ever-growing information store, notably on the Web. Information Retrieval is indeed a mature domain (Salton & McGill 1983)(Manning & Schütze 1999), offering well-known techniques and applications, such as search engines. It is also a domain where progress now consists mostly in small incremental steps, improving classical performance indicators of recall and precision by a few percentage points only.

One of the main reasons for such limited search performance is poor request formulation. When starting a search process, users often submit vague, incomplete or inaccurate requests. Because of their initial lack of knowledge about the subject matter, users do not come up easily with established terms and expressions or their common variants, and search results therefore prove disappointing.

Of course, methods have been devised to tackle the common problem of poor initial requests (Salton & McGill 1983). Using a thesaurus to expand or rephrase a request with related terms is one possible answer, but the search might still remain too general. Relevance feedback is more focused, but usually requires one or two interactive passes. The user must indicate whether the results of an initial request are relevant or not, so that the request may be modified accordingly by the system.

Another fundamental reason for the limited usefulness of search techniques is that much information is in fact tacit (Polanyi 1966). Indexing and retrieval techniques can only deal with explicitly formulated knowledge, to be found in databases or textual documents in electronic form. But know-how, professional skills and expertise are tied up with individuals and their interactions, and are not publicly accessible. Tacit knowledge is unfortunately both very common and socially important.

So how do human beings solve their information needs? More often than not, the initial source of information is not explicit documents, but other human beings. When most of us set about some inquiry, we usually start by asking more knowledgeable people for advice about how and where to find relevant information. We ask for help in formulating the right requests with the correct vocabulary, and in locating the most likely sources of information, whether human or textual.

In short, human beings use interest communities and social networks to find the information they want. Modeling or building such communities and networks should therefore help in the search for relevant knowledge, and several research projects have been developed in this direction.

We will then show in this paper how to design a collaborative search system modeling and emulating the informative function of social networks. By profiling computer users and clustering them according to their interest centers, the system will create a virtual community and use it for information retrieval purposes. Requests will be expanded, relations established and documents shared so as to promote a better exchange of knowledge between participants.

We will actually describe a particular system, called Human Links, but we will try to discuss it in a generic fashion. As the approach shows potential for variations and improvements, we will emphasize the overall method and the associated issues more than specific technical details.

## 2. SYSTEM OVERVIEW

The Human Links software, a collaborative and distributed search engine, has been developed by a recently created French start-up: Amoweba. For more details about the company and its products, please see the company's Web site:

<http://www.amoweba.com>

### 2.1 System Outline

The basic concept of Human Links is to establish a *peer-to-peer* network of registered users to help them in their search for relevant information. The peer-to-peer (P2P) approach consists in creating direct links between computer users without necessarily going through a central server. Unlike a classical server/client architecture, operations and data are as distributed as possible among a network of users, and a particular piece of data could be anywhere in the network.

A peer-to-peer architecture offers potentially unlimited computing power and data storage, contributed by the many computers linked together in a common network. The approach also provides an answer to the updating problem of centralized indexing servers, which cannot keep track of the constant additions, deletions or modifications to Web pages and documents. Of course, adequate software is then necessary to co-ordinate operations throughout the network. This is the general model behind famous file-sharing software such as Napster, Gnutella or KaZaA (see references), which have been very successful in spite of legal difficulties about data copyright.

The goal of Human Links is not simply to share files, however, but to help find all kinds of information, through human contact or by answering formal requests. There are two main modes of operation for this system: it can be employed to locate knowledgeable users on a given subject, or to answer imprecise requests by augmenting them with expert knowledge found in the user network.

The two modes have in common the profiling of the system's users so as to create virtual communities sharing similar interests. These communities are essential to the functioning of the system, for the fundamental idea underlying Human Links is that social links are a good way to locate relevant information. In the case of tacit knowledge, this might indeed offer the only possible access to information. In other words, information retrieval should start first with a search for the right persons, before trying to launch a more classical search for documents.

From a technical point of view, this results in a distributed and collaborative system, typical of the peer-to-peer approach. The advantage is a potentially more powerful and robust functioning, as data and operations could be distributed throughout the network of registered computers. On the other hand, this approach might cause serious privacy and security problems, which will have to be addressed.

## **2.2 Similar Work**

Human Links may be compared to related projects. The main idea of linking people with similar interests is to be found in several systems, which differ in their primary motivation and their definition of similarity.

Using bookmarks to regroup people is now quite common, probably because it is both reasonably informative and fairly easy to do. Systems such as Grassroots (Kamiya et al. 1997), SiteSeer (Rucker & Polanco 1997) or kMedia (Takeda et al. 2000) consider not only a user's set of bookmarks, but also their folder structure to form interest communities. Other systems use different sources of information to compute a similarity between users. For example Referral Web (Kautz et al. 1997) looks at bibliography databases, but one could also use e-mail messages or browsing patterns to establish individual profiles for comparison.

Community organization or browsing recommendations, and not information retrieval per se, however, seem to be the primary motivation of those systems, whereas Human Links was designed to be a practical search engine. In this respect, the closest system is probably Opencola (see reference), which also performs a distributed search through a network of various sources. The problem of tacit information is also a well-known issue in knowledge management, but is usually not discussed sufficiently in computer science.

## **3. SYSTEM DESCRIPTION**

Users of Human Links register at first by downloading the system software on their individual computer, which must of course be connected to the Internet (or to an intranet). Human Links will then work as an addition (*plug-in*) to common browsers (e.g. Explorer or Netscape). The software will operate both locally and by co-ordinating distributed operations and data transfers over the network of connected users.

### **3.1 User Profiling**

The first stage is to establish locally a characteristic user profile by examining a user's files. This could be done by inspecting various types of documents: e-mail messages, attached documents, Web pages, text files, etc., but we have chosen user bookmarks (i.e. favorite Web pages) as a simple solution. One may suppose that a user's bookmarks are characteristic enough of his interests, at least as a first approximation.

#### *3.1.1 Vectorization*

The Web pages corresponding to the bookmark pointers are gathered as an unstructured set (the bookmark folder structure is not used in the present version). Each Web page is

represented by a numerical vector according to the vector-space model of document processing (Salton & McGill 1983). Common function words (articles, particles, auxiliaries...) with little semantic content are first discarded. Occurrences of more significant words (mostly nouns and verbs) are counted, after a simple stemming process has reduce related words to a common form (term). The frequency of terms in a page is weighted to give more importance to the more discriminating terms in the corpus of pages (classical TFIDF measure) and this value for each term produces a vector for the page.

This vectorization stage is standard in document indexing for information retrieval, and there exist many variants. Term frequency is usually the basic measure for vector components, but more sophisticated properties are also available. Information-theoretic measures may be used, but they are more expensive computationally. For common profiling purposes, simple indexing schemes such as TFIDF seem good enough. This approach is also language-independent and does not requires an initial dictionary.

### *3.1.2 Clustering*

These vectors are now clustered with a variant of the well-known k-means algorithm (Anderberg 1973)(Lebart et al. 2000), so as to reveal groups of pages corresponding to the user's interest centers. The clusters obtained may be compared to the bookmark folders (if any) to ascertain their relevance: clusters should correspond to user-defined folders. This unsupervised classification is necessary because a user may have totally different interests (let's say computing, politics and sports) that should be kept separate for further processing. The centroid of each cluster will then be used as a prototype vector for an interest center, and a user profile will consist of several such prototypes.

The classification of bookmarks is a first benefit for the user. It can be used to organize a set of bookmarks, improve or modify a prior folder structure, or simply to visualize this organization on an interactive map (see below). The clustering algorithm may also be applied to other document types, and the resulting map can be very useful.

## **3.2 Collaborative Search**

Initial profiling can be done locally on the individual user's computer, but the next stages must involve interactions between different computers on the network. This is typical of a peer-to-peer approach and offers potentially much more computing power than on a central server.

### *3.2.1 Virtual communities*

Individual user profiles may be compared in order to regroup users with similar interest centers so as to form communities, but in fact communities remain virtual and are implicitly revealed by the proximity of requests to particular users. Note that users in a community are seen only through their relevant interest center, and users may well belong

to several communities. For the sake of clarity, we call "contact" a user identified by one of its interest centers, and communities consist of contacts rather than users.

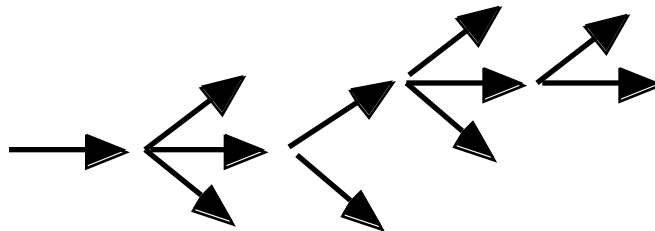
Communities may also be represented by vectors: the centroid of a cluster of users is the prototype for the community. As a matter of fact, all items in this system (documents, interest centers, contacts, communities, requests...) are represented by lexical vectors, which will make comparisons possible between various items.

Virtual communities can now be used to foster social interactions. Members of a given community may communicate with one another by e-mail, instant messaging ("chat" such as ICQ) or any other means. A request for information can then be discussed with other knowledgeable human beings. Of course the privacy of individuals must be protected: their consent is required before disclosing personal information about them (such as their name and address).

### 3.2.2 Search process

These communities can also support an automatic collaborative and distributed search process. A novice about some subject domain will find it hard to formulate pertinent, focused requests, but given an imprecise request, the system will be able to find expert users with similar interests. Their bookmarks or documents might be relevant for a novice and expert interest prototypes help to formulate more focused requests.

An initial request may consist of keywords as well as particular interest centres. The request is sent to other users with a profile similar to the request. If the similarity is good enough, documents will be sent back, and the new profile may be used to focus the request, and so on from user to user. Each user may lead to several others, so that the search graph looks like a tree (Fig. 1). For a more detailed description of this process, see (Spalanzani 2002).



**Figure 1. Search graph.**

This process propagates forward and backward through the network to gather relevant documents or items. Requests jump forward from one user to another with similar interests, but care must be taken to limit the depth and breadth of the search. To ensure this, a propagation parameter is set before launching a search; this may be seen as a kind of diffusion energy associated with a request, which is progressively used up during the forward pass. The backward pass starts when this energy has been exhausted, and search depth has been limited to 7 jumps anyway.

Requests also keep a trace of the path followed, to avoid potential loops and to allow a distributed backward pass without central supervision. Answers to a request are sent back directly to the caller, so that it does not matter if an individual computer is switched off between the forward and the backward pass (or otherwise becomes inaccessible). Duplicate documents are then discarded and answers sorted by order of relevance.

### **3.3 Interactive Map**

The graphical interface is essential to the system's usefulness. Classical information retrieval systems (using requests and lists of documents) are not intuitive enough for inexperienced users, and we have devoted great care to the design of the interface.

#### *3.3.1 Map components*

The main component of the interface is a map (Figs. 2 and 3) on which are displayed interest centers, users, virtual communities, documents and requests. Those different items can be shown on the same map because they are all represented by vectors. This common representation format makes it possible to compare very different objects. Note, however, that the two-dimensional map is a projection from a multi-dimensional space, with some inevitable loss of information. One must also suppose that all items belong to the same vector-space, i.e. the space of possible lexical terms.

Such a map is much more intuitive, flexible and easier to use than a set of folders or a category tree. Documents are more or less distant from other items on the map, and do not have to be classified into exclusive binary categories. Beside the map, the interface also shows a list of interest centers and a list of documents returned.

If the vector-space is defined by reduced dimensions of the kind computed by factor analysis from original lexical features, terms can also be placed within the same space and on the same map. Principal Component Analysis (Jolliffe 1986)(Lebart et al. 2000) is probably the best known dimensionality reduction technique, and one version of the system employs it to represent all items. We used a neural network technique to compute the new dimensions, an approach of interest in itself (Delichère & Memmi 2002).



### *3.3.2 User interaction*

This is an interactive map: the user can move items on the map (with the mouse) and this action will change the underlying representation. For example the user may want to move a document closer to an interest center, and the document vector will be modified accordingly. Interest centers may also be placed within a cloud of documents, superseding the system's classification. Requests may also be defined by their placement on the map.

One can define an influence zone around items (roughly by drawing on the map a circle around an item). When an item is displaced and modified, all items within its zone of influence are also visibly displaced on the map and modified accordingly. This influence may be designed as decreasing with the distance to the central item; gravity would be a good metaphor for this diminishing effect.

Such interactive modifications pose serious updating and coherence problems, made even worse by the two-dimensional projection, and this capability should be designed with care and used with caution. There is a trade-off between modification power and coherence, we have experimented with several variants, and we think more work should be done in this area. Nonetheless, giving the user some control over the representation of objects in the system seems important, and poses interesting theoretical questions.

Lastly, for this interactive map to be acceptable in practice, the user should not have to wait for results and modifications to be displayed. This is a very important practical constraint on the profiling and clustering algorithms, which must work online. Incremental versions of the main algorithms might have to be developed to speed up computation.

## **3.4 First Results**

All the features of Human Links presented here have been implemented. The system works as described, and all modules (profiling, clustering, visualization, search and retrieval) have been tested separately and together. Tests have been run with thousands of participants (at most 8500 so far), but we do not have yet the experience of real-life, everyday use with thousands of participants. We are confident about the operation of the system, but we are not quite sure about how participants will choose to use it. For example will they emphasize human communication or formal requests? It is too early for us to tell, but this is an intriguing question.

Security concerns are also worth mentioning. Systems managers are often uneasy with peer-to-peer, notably because of the danger of security breaches (by viruses or hackers) in a distributed computation network. Company firewalls try to limit access from outside, and we had some difficulties at first to make Human Links work in such environments.

One conclusion might be that such a system is easier to use (or to recommend) over an intranet within a huge organization than at large via the Internet. We are indeed heading in this direction for future versions of the software.

## **4. DISCUSSION**

This collaborative approach to information retrieval raises several questions. The underlying philosophy is that the best way to go about a request for information is to do a computer-aided search for the right people. This is closer to the way human beings operate, but human factors must then be considered and evaluated.

### **4.1 Privacy Problems**

Privacy is an important issue. To ensure the ethical and social acceptability of such a system, the right of participants to remain anonymous must be guaranteed. The name and address of each participant should be hidden from others (by using a coded identifier for example). Before disclosing somebody's address, he must have granted his permission. Experts in particular are busy people who do not want to be flooded with personal requests from novices.

Protection against unwanted information (spamming) is a related issue. Belonging to a community can be beneficial, but may also bring unwanted attention. Profiling decreases to some extent the risk of receiving irrelevant information, but care must be taken to protect members of the network from obtrusive messages. Information push should be limited and very specific. Using mostly expert information could help; experts can be detected automatically by their richer connection pattern, in the same way that Google (see reference) rates Web pages.

### **4.2 Actual Benefits**

One should be clear about the real benefits and limitations of such a system. This approach attempts to assist in the search for explicit information by exploiting tacit knowledge and human expertise. But whatever the particular source chosen (Web bookmarks in this case), the system can only capture expertise having left computer traces. This is obviously a very limited view of human knowledge.

Although people deal more and more with computers and computer-controlled devices (e.g. cash machines and credit-card readers) in their professional and daily life, a major part of social interactions does not take place by computer. Face-to-face meetings and phone calls still represent the bulk of human contacts, and voice is more important than texts. The kind of system we have been describing can only make use of computerized interactions and is clearly geared toward textual documents.

Extensions would be possible, however. Because more and more social activity is mediated or recorded by computers, more tacit information could be captured, which does not have to be in written form. In fact, many firms have started exploiting purchase patterns, bank withdrawals, transportation flows... Such data-mining is constantly increasing, both within firms and among the general public. This raises serious ethical issues, but the fact is that more and more tacit information becomes available, which could be gathered and exploited.

## **5. CONCLUSION**

We have described a collaborative and distributed search system developed to help find knowledge which happens to be shared among various human beings. It is thus possible to locate information which might not be found otherwise in electronic documents or databases. The Human Links system will help a novice user to formulate more focused requests by exploiting the profiles of more expert users in a network of registered participants, and will facilitate direct contact between participants with similar interests. The system shows how to make valuable tacit knowledge more accessible.

This approach can be seen as an attempt to fulfill several goals at the same time. Its primary objective was to locate relevant documents, even with poor initial requests, a classical goal of information retrieval. But this is achieved by creating virtual communities, which could then give rise to real human relationships among the participants. In this way is also modeled to some extent the information-gathering function of human social networks, an interesting goal in itself. The computer system may thus fulfill some important social needs as well as more technical goals.

More generally, this system should be placed within the framework of recent trends in knowledge management. Collaborative work and knowledge-sharing over electronic networks, the development of virtual communities, the increasing use of outsourcing and distributed organizations facilitated by the Internet, clearly require appropriate software tools. Promoting electronic contacts and information exchange in a natural way, Human Links is a promising contribution to this growing domain.

### **Acknowledgments**

We would like to thank the whole development team of the Human Links project. The system described here would not have become a reality without their various individual contributions.

## REFERENCES

- Anderberg M.R. (1973) *Cluster Analysis for Applications*, Academic Press.
- Delichère M. & Memmi D. (2002) Neural dimensionality reduction for document processing, *ESANN'02*, Bruges.
- Delichère M. & Memmi D. (2002) Analyse factorielle neuronale pour documents textuels, *TALN'02*, Nancy.
- Gnutella, <http://www.gnutella.com>
- Google, <http://www.google.com>
- Jolliffe I.T. (1986) *Principal Component Analysis*, Springer Verlag.
- Kamiya K., Roscheisen M. & Winograd T. (1997) Grassroots: a system providing a uniform framework for communicating, structuring and organizing people, *Proceedings of WWW-6*.
- Kautz H., Selman B. & Shah M. (1997) Referral Web: combining social networks and collaborative filtering, *Communications of the ACM*, 40 (3).
- KaZaA, <http://www.kazaa.com>
- Lebart L., Morineau A. & Piron M. (2000) *Statistique Exploratoire Multidimensionnelle*, Dunod.
- Manning C.D. & Schütze H. (1999) *Foundations of Statistical Natural Language Processing*, MIT Press.
- Napster, <http://www.napster.com>
- Opencola, <http://www.opencola.com>
- Polanyi M. (1966) *The Tacit Dimension*, Routledge & Kegan Paul.
- Rucker J. & Polanco M.J. (1997) SiteSeer: personalized navigation for the Web, *Communications of the ACM*, 40 (3).
- Salton G. & McGill M. (1983) *Introduction to Modern Information Retrieval*, McGraw-Hill.
- Spalanzani A. (2002) Partage des connaissances et système peer-to-peer : le système Human Links, *AIM'02*, Hammamet.
- Takeda H., Matsuzuka T. & Taniguchi Y. (2000) Discovery of shared topics networks among people, *PRICAI'00*.

**Le laboratoire Leibniz est fortement pluridisciplinaire. Son activité scientifique couvre un large domaine qui comprend aussi bien des thèmes fondamentaux que des thèmes très liés aux applications, aussi bien en mathématiques qu'en informatique.**

**Les recherches sur les Environnements Informatiques d'Apprentissage Humain et la didactique des mathématiques ouvrent cette pluridisciplinarité sur les sciences humaines, elles jouent un rôle particulier en favorisant les coopérations entre différentes composantes du laboratoire.**

- \* mathématiques discrètes et recherche opérationnelle
- \* logique et mathématique pour l'informatique
- \* informatique de la connaissance
- \* EIAH et didactique des mathématiques

*Les cahiers du laboratoire Leibniz* ont pour vocation la diffusion des rapports de recherche, des séminaires ou des projets de publication réalisés par des membres du laboratoire. Au-delà, Les cahiers peuvent accueillir des textes de chercheurs qui ne sont pas membres du laboratoire Leibniz mais qui travaillent sur des thèmes proches et ne disposent pas de tels supports de publication. Dans ce dernier cas, les textes proposés sont l'objet d'une évaluation par deux membres du Comité de Rédaction.

### **Comité de rédaction**

- \* mathématiques discrètes et recherche opérationnelle  
Gerd Finke, Andrés Sebõ
- \* logique et mathématique pour l'informatique  
Ricardo Caferra, Rachid Echahed
- \* informatique de la connaissance  
Yves Demazeau, Daniel Memmi,
- \* EIAH et didactique des mathématiques  
Nicolas Balacheff, Jean-Luc Dorier, Denise Grenier

Contact Gestion & Réalisation : Jacky Coutin  
Directeur de la publication : Nicolas Balacheff  
ISSN : 1298-020X - © laboratoire Leibniz