LA-UR-05-5222

*Approved for public release;*
*distribution is unlimited.*

*Title:* MYLIBRARY AT LANL: PROXIMITY AND SEMI-METRIC NETWORKS FOR A COLLABORATIVE AND RECOMMENDER WEB SERVICE

*Author(s):*
Luis M. Rocha
Tiago Simas
Andreas Rechtsteiner
Mariella Di Giacomo
Rick E. Luce

*Submitted to:* ACM (Association for Computer Machinery)

Los Alamos
NATIONAL LABORATORY

Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36. By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

Form 836 (8/00)

# MyLibrary@LANL: Proximity and Semi-metric Networks for a Collaborative and Recommender Web Service

Luis M. Rocha[§], Tiago Simas[§], Andreas Rechtsteiner[†], Mariella Di Giacomo[‡], Rick Luce[‡]

[§]*School of Informatics and Cognitive Science Program Indiana University 1900 East Tenth Street Bloomington IN 47406, USA*

[†]*Modeling, Algorithms, and Informatics Group (CCS-3) Los Alamos National Laboratory MS B256, Los Alamos, New Mexico 87545, USA*

[‡]*Research Library Los Alamos National Laboratory Los Alamos, NM 87545, USA*

## Abstract

*We describe a network approach to building recommendation systems for a WWW service. We employ two different types of weighted graphs in our analysis and development: Proximity graphs, a type of Fuzzy Graphs based on a co-occurrence probability, and semi-metric distance graphs , which do not observe the triangle inequality of Euclidean distances. Both types of graphs are used to develop intelligent recommendation and collaboration systems for the MyLibrary@LANL web service, a user-centered front-end to the Los Alamos National Laboratory's (LANL) digital library collections and WWW resources.*

## 1. Introduction

The WWW is used today as a means to integrate and provide portals to many electronic information resources. In particular, it enables the creation of personalized and collaborative digital library services. Indeed, the WWW has changed the nature of scientific research by creating new expectations for libraries supporting research. Several digital library initiatives offer customized digital library environments, however, these services typically do not provide users with personalized and collaborative environments. *MyLibrary* at Los Alamos National Laboratory (LANL) provides scientists (as individuals or groups) with a personalized Web environment enhancing scientific collaboration independent of time and location. One of the unique characteristics of this capability is the ability to push recommendations to users and adapt the system further based on user interactions.

We have described some of the adaptive features of *MyLibrary@LANL* in other publications [1] [2]. In this paper we present a network analysis methodology to produce recommendation systems and enhanced collaboration in this particular web service. This methodology is clearly applicable to other types of web services beyond digital libraries, as also discussed in this paper.

## 2. MyLibrary@LANL

### 2.1 Description of the Service

MyLibrary@LANL is a user-centered front-end to the Los Alamos National Laboratory's (LANL) digital library collections and WWW resources. It supports a collection of personal links to a variety of information resources such as electronic journals, full-text content and bibliographic databases, reference materials, and generic web links. It can be customized to reflect specific disciplines and research needs. Users can select from a subset of over 5,800 electronic journals, 200+ electronic databases, and 400+ subject based web links from the Research Library's website. Users can also store and organize any general link to WWW content in this system. Users can access MyLibrary from any LANL computer[1].

This WWW service consists of three nested entities: (1) ***libraries*** (referred to in this article also as ***personalities***, as they are expected to be associated with a specific area of interest for each user), (2) ***folders***, used to specify sub-categories within a specific library, and (3) ***links*** which are organized within a folder. A library is associated with a given area of interest (e.g., physics, computer science, etc) and each library consists of one or more folders. A folder contains related types of links within a library and a link is simply a URL to a specific web page, document, file, or

[1] This service is not available outside LANL, though a functionally stripped down demo is available at http://mylibdemo.lanl.gov

resource. Libraries can also be shared amongst groups of users.

When a user creates a library, she chooses an interest topic from a finite list (e.g. Physics, Mathematics, Biology, Computer Science, General Web Links, etc.). We regard each library as a specific *personality* of a user. When a library is associated with a specific topic, it is automatically populated with a few links to relevant databases (in a databases folder) and e-journals (in a e-journals folder). Once a library has been established, new folders can be added to it, and links can be added to the folders. Link checking is run on a weekly basis and broken links are identified.

Users also have the option of adding their own preferred web links easily by clicking on an optional bookmarklet added to their browser's personal toolbar, giving users an easy way to add links to their libraries without any cutting and pasting.

### 2.2 Enabling Collaboration in MyLibrary

Digital libraries now offer opportunities for collaboration and communication that were not feasible in traditional libraries. The last generation of user interfaces to information systems (including databases, information retrieval applications and library catalogues) largely reflected single, generic user stereotypes. That is, the activities or behaviors of other users, and past behavior of users, have had almost no impact on the experience of any one user.

Today's web technology available to digital libraries allows us to consider new ways of working with library materials and well as new methods to analyze the use of digital resources. Rather than limiting the user to work in an isolated mode as a individual with generic capabilities, now we can enable users to work collaboratively and in a personalized manner when desired.

We have chosen to focus on collaboration among the users of our digital library via MyLibrary. Two types of collaboration are supported: direct and indirect. We define "direct collaboration" as instances where several users agree to work together as a defined group or team exploring and making use of digital library resources. Our direct collaboration features have been described elsewhere [3].

We define "indirect collaboration" as the anonymous utilization of the behavior of the user community for the potential benefit of any user. This indirect collaboration is instantiated by personalized recommendation systems [1] (or recommender systems e.g. [4], [5]) that we describe in the remainder of this article and which were developed by the *Active Recommendation Project*[2] (ARP) [6] for the LANL Digital Library. Here we describe ARP's work on the MyLybrary@LANL service.

[2] Http://arp.lanl.gov

## 3. Mathematical Background

### 3.1 Crisp and Fuzzy Relations

A relation $R$ between sets $X_1, X_2, ..., X_n$ is a subset of the Cartesian product of these sets: $R(X_1, X_2, ..., X_n) \subseteq X_1 \times X_2 \times ... \times X_n$. The elements of relations are $n$-tuples of the universal set $\mathbf{X} = X_1 \times X_2 \times ... \times X_n$: $\mathbf{x} = (x_1, x_2, ..., x_n)$. Relations can also be defined by a characteristic function which assigns a value of 1 to every element $\boldsymbol{x}$ of $\mathbf{X}$ belonging to the relation and a 0 to those elements not belonging to it:

$$R(x_1, x_2, ..., x_n) = \begin{cases} 1 & \text{iff}(x_1, x_2, ..., x_n) \in R \\ 0 & \text{otherwise} \end{cases}$$

The membership of $\mathbf{x}$ in $R$ signifies that the elements of $\mathbf{x}$ are related or associated to one another ([7] page 119). A relation between $n$ sets is generally referred to as a $n$-ary relation.

Naturally crisp relations can be generalized to fuzzy relations by allowing a degree of association between the elements of tuples $\mathbf{x}$. Thus, a fuzzy relation is a fuzzy set defined on $\mathbf{X}$.

$$R(\mathbf{x}) \in [0, 1], \quad \forall \mathbf{x} \in \mathbf{X}$$

The value of $R$ represents the degree of association or strength of the relation between the elements of $\mathbf{x}$.

### 3.2 Binary Relations and Graphs

Binary fuzzy relations, $R(X, Y)$, can be easily represented by matrices of dimension $n \times m$ where $n$ and $m$ are the number of elements of $X$ and $Y$ respectively.

Binary fuzzy relations can also be defined on a single set: $R(X, X)$. These relations define degrees of association between elements of the same set, and are referred to as *fuzzy directed graphs* (or weighted directed graphs) because they can be easily represented by diagrams where elements of $X$ are depicted by a single node, and the entries of $R$ denote weights of the edges amongst these nodes in a graph.

In this case, relations are represented by matrices of dimension $n.n$, that is, there are $n^2$ possible connections between the $n$ nodes of the directed graph.

A binary fuzzy relation $R(X, X)$ is *reflexive* iff $R(x, x) = 1$ for all $x \in X$. That is, if every element of $X$ is maximally associated with itself. A binary fuzzy relation $R(X, X)$ is *symmetric* iff $R(x, y) = R(y, x)$ for all $x, y \in X$. That is, if $x$ is associated with $y$ to a degree, then $y$ is associated with $x$ to the same degree, thus establishing *non-directed graphs* whose defining matrices are symmetrical relations requiring only $(n^2-n)/2$ elements to be defined.

A binary fuzzy relation $R(X, X)$ is (max-min) *transitive* iff:

$$R(x, z) \geq \max_{y \in X} \min[R(x, y), R(y, z)]$$

for all $x, z \in X$. This definition generalizes the crisp transitive property which requires that $(x, z)$ be related if $(x, y)$ and $(y, z)$ are related. It requires that $(x, z)$ be related at least to the maximum of the degrees of association obtained from performing the minimum of $(x, y)$ and $(y, z)$ for all other $y$ in $X$. That is, the weight of an indirect path between $x$ and $z$ through some $y$, is the smallest edge weight in the path ($x$ to $y$ and $y$ to $z$). Finally, the weight of the direct path between $x$ and $z$, must be larger or equal to largest of all indirect paths through each $y$.

### 3.3 Composition of Graphs and Transitive Closure

The composition of binary relations defined on the same set is performed by composition of the matrices that define them. The logical composition of matrices is performed in the same way as the numerical counterpart, except that multiplication and summation are substituted by Fuzzy Set Theory aggregation operations such as intersection and union – or more generally T-Norms and T-Conorms respectively [7]. The most commonly used aggregation operations for intersection and union are the minimum and maximum operations respectively. Thus, the standard composition of fuzzy graphs is referred to as the *max-min composition*:

$$R \circ R = \max_k \min(r_{ik}, r_{kj}) = r'_{ij}$$

where $r_{ij}$ denotes $R(x_i, x_j)$.

The *transitive closure* of a binary relation $R(X, X)$ is defined as the relation that is transitive, contains $R(X, X)$, and whose elements have the smallest possible membership weights that still allow the first two requirements to be met. Different types of transitive closures can be defined, based on the criteria for transitivity. Here we use max-min transitivity (as defined in 3.2). The algorithm to obtain the transitive closure $R_T$ of $R$ consists of the following three steps [7]:

1. $R' = R \cup (R \circ R)$.
2. If $R' \neq R$, make $R = R'$ and go back to step 1.
3. Stop: $R_T = R'$.

### 3.4 Similarity and Proximity Relations

A binary fuzzy relation $R(X, X)$ is a similarity relation if it is reflexive, symmetric, and transitive. These relations are also known as *equivalence relations*.

A binary fuzzy relation $R(X, X)$ is a proximity relation if it is reflexive and symmetric. These relations are also known as *compatibility relations*. The transitive closure of a proximity relation is a similarity relation.

## 4. Proximity Networks

### 4.1 Generic Proximity Probability Measure

In general, given a binary relation $R$ between sets $X$ (of $n$ elements $x$) and $Y$ (of $m$ elements $y$), we extract two complementary proximity relations: $XYP$ and $YXP$. $XYP(x_i, x_j)$ is the probability that both $x_i$ and $x_j$ are related in $R$ to the same element $y \in Y$. Conversely, $YXP(y_i, y_j)$ is the probability that both $y_i$ and $y_j$ are related in $R$ to the same element $x \in X$. The respective formulas are:

$$XYP(x_i, x_j) = \frac{\sum_{k=1}^{m} (r_{i,k} \wedge r_{j,k})}{\sum_{k=1}^{m} (r_{i,k} \vee r_{j,k})}; \quad YXP(y_i, y_j) = \frac{\sum_{k=1}^{n} (r_{k,i} \wedge r_{k,j})}{\sum_{k=1}^{n} (r_{k,i} \vee r_{k,j})} \quad \textbf{(1)}$$

### 4.2 Why Proximity Measures?

Other measures of probability can be used to discern closeness between elements of two sets in a binary relation. In information retrieval it is common to use conditional probabilities [9]. In this case, we do not obtain a proximity relation since conditional probabilities are not symmetric. For characterizing associations in relations extracted from databases, we prefer the proximity measure of association as defined in formula 1, over conditional probabilities, because the proximity measure is symmetric. Indeed, proximity intuitively captures the inverse of a distance, which is a symmetric concept. The idea of distance is important for the recommendation algorithms we develop below. Intuitively, we can clearly see that a large proximity value, requires a large value of both directions of conditional probability.

### 4.3 Capturing Knowledge in a Network

Proximity relations define graphs or networks of elements. We derive our proximity networks from the computation of the probability measures (1) on binary relations extracted from large collections of documents or records stored in databases. Such proximity graphs should be seen as associative knowledge networks that represent how often items co-occur in a large set of documents. As in any other co-occurrence method, the assumption is that items that frequently co-occur are associated with a common concept.

Notice that a graph of co-occurrence proximity allows us to capture network associations rather than just pair-wise co-occurrence. Therefore, we expect concepts or themes to be organized in more interconnected sub-graphs, or clusters of items in the proximity networks. In the next section we show how well proximity networks capture knowledge in different applications, before we describe their application to *MyLibrary*.

## 5. Evaluating Proximity Nets

### 5.1 Text-Mining with Proximity Networks

We used proximity networks in the Text Mining competition *BioCreAtIvE* (Critical Assessment of Information Extraction in Biology)[3]. We addressed Task 2, the problem of annotation of a protein with a node in the Gene Ontology (GO)[4] based on the text of a given document, and the selection of evidence text justifying the predicted annotation. We approached the task utilizing various combinations of two distinct methods [10]. The first method was an unsupervised algorithm for expanding words associated with GO nodes using the proximity networks, which we present here. For more details see [10].

The task was as follows: given a <protein, document, GO node identifier> triple, return the evidence text from the document supporting the annotation of the protein to that GO node. Each node in the Gene Ontology (GO) is associated with a textual label, intended to capture the meaning of the node. However, these labels tend to be relatively short and often the terms in a given label occur in many other labels throughout the GO. As such, the occurrence of an individual term that is part of a GO node label in a document may not be a sufficiently reliable indicator of the relevance of that GO node to the document. To address this, we utilized a method for expanding the set of terms associated with a given GO node.

After document pre-processing, we divided each document into paragraphs and calculated for each document a relation matrix of word occurrence in the paragraphs: $R$: $P \times W$, where $P$ is the set of all $m$ paragraphs in a document, and $W$ is the set of all $n$ words. This is a Boolean matrix ($r_{i,j} \in \{0, 1\}$) that specifies if a given word occurred at least once in a given paragraph.

From the $R$ matrices produced for each document, we calculated a word in paragraph proximity matrix, *WPP*, for each document, using the co-occurrence probability measure (1). *WPP* denotes the association strength between pairs of words ($w_i$, $w_j$) , based on how often they co-occur in the paragraphs of a given document. A value of $wpp$ ($w_i$, $w_j$) = 0.3, means that words $w_i$ and $w_j$ co-occur in the same paragraphs 30% of the time that either one of them occurs.

*WPP* is an associative network of words. Such a graph can also be understood as an associative knowledge structure that represents how words co-occur in a given document, and therefore as an associative model of the knowledge stored in each document in terms of its constituent words [2]. We expect concepts or themes to be organized in more interconnected sub-graphs, or clusters of words. Figure 1 depicts a sub-graph of the *WPP* for one of the BioCreAtIvE documents (JBC_1999/bc005868).

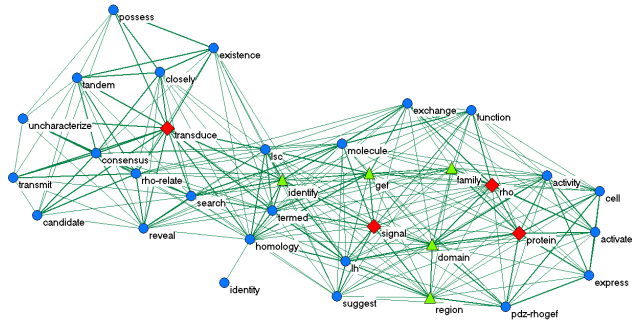


**Figure 1:** Subgraph of *WPP* network for example document, depicting all the word nodes highly associated with at least one of the words in the target GO annotation.

Next we set out to identify words associated with GO nodes. Using the GO nodes in the provided triples we retrieved the words from the GO node label. Let us refer to this set of words as $W_{GO}$ (the red diamond nodes in Figure 1, for GO node 0007266: 'transduce', 'signal', 'rho', 'protein'). For each document, we then retrieved a set of words highly associated with the words in $W_{GO}$ in the relevant *WPP* network. Specifically, we returned the top 5 to 10 additional words with largest average value of *WPP* to all the words in $W_{GO}$ (the green triangle nodes in Figure 1: 'identify', 'gel', 'family', 'domain', 'region'). The additional words thus discovered were used to expand $W_{GO}$. Let us refer to the expanded set of words as $W_{GOProx}$; the additional words are not found in the respective GO node label, but co-occur highly in a given document with the words in the GO node label.

We considered that the most appropriate evidence text (from a document) for a GO node, to the document's paragraph where we find most of the words in $W_{GOProx}$. Results were evaluated by professional annotators from the *European Bioinformatics Institute* (EBI). Our submission based on the word proximity network analysis, was the most successful of all other submissions from several research groups using many types of machine learning methods. Please refer to [10] for a full discussion of results.

### 5.2 Proximity-Based Retrieval

Here we present a web-search experiment which queries the *Altavista*™ search engine to produce proximity networks for two classes of items. We were given a set of known relations between two types of molecules that play an important role in the immune system: *Cytokines* and *Receptor Molecules*. Cytokines are basically signaling molecules: their levels affect immune response. But for this response to occur, Cytokines need to be recognized by Receptor Molecules.

[3] http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

[4] http://www.geneontology.org

Our task was to, using automatic search methods, discover as much as possible of this known relationship, without making use of any expert knowledge. We computed a proximity network of associations amongst the terms describing receptor molecules and cytokines in Table I using the *Altavista* search engine. To contrast with our approach, we also used [9] the *Pointwise Mutual Information - Information Retrieval* (PMI-IR) technique.

**Table I** : Known relationship between Cytokines and Receptor Molecules

| Cytokines | Receptor Molecules |
|---|---|
| IL-2 OR interleukin-2 | CD25 OR tac; CD122; CD132 OR "common gamma chain" |
| IL-3 OR interleukin-3 | CD123; beta |
| IL-4 OR interleukin-4 | CD132 OR "common gamma chain"; CD124 |
| IL-6 OR interleukin-6 | CD126; CD130 OR gp130 |
| IL-7 OR interleukin-7 | CD132 OR "common gamma chain"; CD127 |
| IL-8 OR interleukin-8 | CXCR1 OR CDw128a; CXCR2 OR CDw128b |
| IL-10 OR interleukin-10 | CDw210 |
| IL-12 OR interleukin-12 | CD212 |
| IL-13 OR interleukin-13 | CD132 OR "common gamma chain"; CD213a1; CD213a2 |
| IL-15 OR interleukin-15 | CD122; CD132 OR "common gamma chain" |
| GM-CSF | beta; CD116 |
| IFNgamma | CD119 |
| TNFalpha | CD120a; CD120b |
| MCP-1 | CCR2b |

For this problem, we used the proximity (1). Specifically for this problem:

$$prox(problem, choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(problem \text{ OR } choice_i)} \quad \textbf{(2)}$$

Where, in our example, a *problem term* is the union of all the terms associated with a Cytokine, and a *choice term* is the union of all the terms associated with a Receptor Molecule. This proximity measure between problem and choice terms is computed by counting the number of web pages indexed by *AltaVista* in which both terms co-occur within 10 words of one another (the numerator) over the number of web pages in which either one of the words occurs (the denominator). This was performed when *AltaVista* allowed the use of the NEAR operator, which allowed us to count the co-occurrence of terms in a web page within 10 words of each other.

The PMI-IR query we used is conditional probability as defined by formulae (2). In this case, given a problem term (the union of all terms associated with a Cytokine), we compute the score for all possible choice terms (the union of all terms associated with each Receptor Molecule) as follows:

$$score(choice_i) = \frac{hits(problem \text{ NEAR } choice_i)}{hits(choice_i)} \quad \textbf{(3)}$$

Turney defined more complex queries for PMI-IR involving the use of context words for reduction of the scope of web pages returned, as well as protection for antonyms. We have contrasted those with equivalent proximity formulations, but the precision and recall results were not much different from the proximity and score here discussed.

Using the known relation between Cytokines and Receptor Molecules defined in Table I, we computed the Precision and Recall measures for the proximity measure (2) and the PMI-IR score (3) obtained from querying AltaVista. In our case, Precision is the probability that an identified association is relevant, and Recall is the probability that an association has been identified given that it is relevant.

**Table II: Precision and Recall results**

| | Precision | Recall |
|---|---|---|
| Proximity (6) | 0.79 | 0.75 |
| PMI-IR Score (7) | 0.77 | 0.54 |

The precision and recall measures are depicted in Table II. As we can see, the proximity measure yielded the best combination of precision and recall values, while the PMI-IR score yields a n equivalent value of precision, but with considerably less recall. This exercise in retrieval using the proximity measure highlights the power of simple statistical measures of occurrence. Indeed, with a simple *AltaVista* query, we were able to discover close to 80% of the relevant associations between Cytokines and Receptor Molecules, without access to any expert information! Thus we can say that the proximity network obtained from querying AltaVista captures quite well expert knowledge associated with this immunology toy problem.

## 6. Mylibrary Network Extraction

### 6.1 Relating Links to Journals

The following data extraction procedures are performed weekly for the entire MyLibrary database, and for each link added to the service in real time. We extract all the non-default and non-system links from the MyLibrary database, where each link contains the following information:

user_id, library_id, library_name, link_id, link_url, link_name

for example:

'477','651','Physics','106195372','http://www.sciam.com/', 'Scientific American'

The ARP and MyLibrary teams keep compiling and updating a file associating specific scientific journals (identified by their International Standard Serial Number or ISSN) to a set of URLs where users can access journal articles. With this information we identify the subset of links in user libraries whose URL can be unequivocally related to the URL of a journal (a ISSN). From all the links in the MyLibrary database, about 61% are unequivocally associated with a scientific journal identified by a ISSN.

## 6.2 Relation Extraction

From the data extracted in 6.1, we produced the relation **PERSONALITY×ISSN** which is a binary relation $A$: $P \times I$ between the sets $P$ of $n_p$ personalities (or libraries) and $n_u$, and the set $I$ of $m$ ISSN. We treat $A$ as a crisp relation: $A(p_s, i_t) = a_{s,t}$ is 1 (True) if ISSN $i_t$ occurs in personality $p_s$, and 0 (false), otherwise.

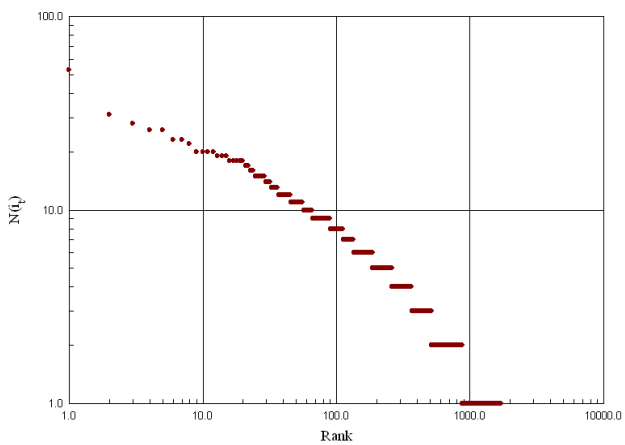


**Figure 2:** Rank distribution of ISSN frequency

We chose personalities over users as the unit of co-occurrence, because personalities tend to be thematically organized in the MyLibrary service, thus co-occurrence of ISSN in personalities is more of an indicator of a thematic association between ISSN than co-occurrence in users who may store very different topics in different libraries/personalities.

Given the relation $A$, we define $N(i_t)$ as the number of personalities in which ISSN $i_t$ occurs, and $m$ = 1702 as the total number of ISSN that occur in at least one personality. The ranked distribution of $N(i_t)$ is shown in figure 2. The top 12 most frequent ISSN are:

| | |
|---|---|
| 0031-9007– Physical review letters | (53) |
| 0556-2791– Physical review A General physics | (31) |
| 0556-2805– Physical Review B | (28) |
| 1095-3787; 1063-651X– Physical review E | (26) |
| 0021-9606– Journal of Chemical Physics | (26) |
| 0002-7863– J. American Chemical Society | (23) |
| 1089-5647– Journal of physical chemistry B | (23) |
| 0034-6861– Reviews of modern physics | (22) |
| 1089-5639– Journal of physical chemistry A | (20) |
| 0028-0836– Nature | (20) |
| 0036-8075– Science | (20) |
| 0027-8424 – PNAS | (20) |

We define $N(p_s)$ as the number of ISSN contained in personality $p_s$, and $n_p$ = 392 as the total number of personalities.

## 6.3 Proximity Network Extraction

To discern closeness between ISSN according to the personalities they occur in, using (1) we compute the *ISSN Personality Proximity* (*IPP*), obtained from $A$ by the following formula:

$$IPP(i_s, i_t) = \frac{\sum_{k=1}^{n_p} (a_{k,s} \wedge a_{k,t})}{\sum_{k=1}^{n_p} (a_{k,s} \vee a_{k,t})} = \frac{N_{\cap}(i_s, i_t)}{N(i_s) + N(i_p) - N_{\cap}(i_s, i_t)} \quad \textbf{(4)}$$

The proximity between two ISSN, $i_s$ and $i_t$, is the probability that both $i_s$ and $i_t$ co-occur in the same personality. $N_{\cap}(i_s, i_t)$ is the number of personalities both ISSN co-occur in. Thus, two ISSN are near if they tend to occur in many of the same personalities.

**Figure 3:** *IPP* network extracted from MyLibrary data, showing all edges with proximity larger than 0.3

The *IPP* defines an associative network of Journals(identified by ISSN). This network is a weighted, probabilistic graph, whose edges are the co-occurrence proximity values, as displayed in figure 3. For instance, the journals associated with the most frequent journal

(0031-9007--*Physical review letters*), with at least a proximity value of 0.3 are:

| | |
|---|---|
| 1095-3787 ; 1063-651X--Physical review E: | 0.4364 |
| 0556-2805--Physical Review B: | 0.3729 |
| 0034-6861--Reviews of modern physics: | 0.3636 |
| 0556-2791--Physical review A: | 0.3125 |

Figure 3 clearly shows two main clusters of nodes highly associated in the *IPP* network. The Principal Component Analysis (PCA) analysis of the *IPP* network revealed that the two first eigen-vectors (components) are very correlated with the two main clusters identified in figure 3. The first component refers to a set of journals related to Chemistry, Materials science and Physics. The second component refers to a set of journals related to Computer Science and Applied Mathematics. However, these groups are further separated and refined into more specific clusters as we consider more components.

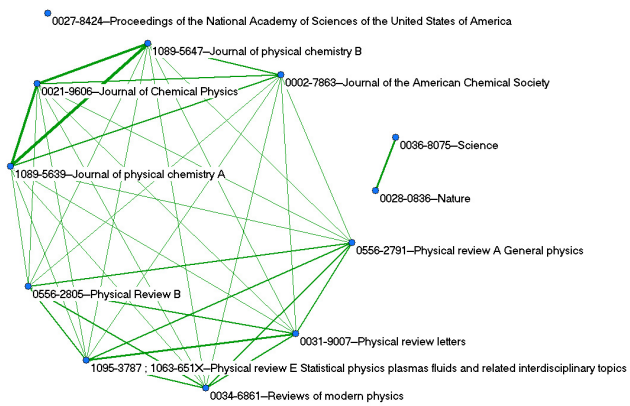


**Figure 4:** Sub-graph of *IPP* network with top 12 most frequent journals, showing edges larger than 0.2. Edge thickness denotes proximity strength.

The main clusters discovered in the *IPP* network highlight the reality of the research pursued at the Los Alamos National Laboratory. Indeed, being a nuclear weapons laboratory, much of its research is concerned with Materials Science and Physics on the one hand, and Simulation and Computer Science on the other. Thus, the *IPP* network captured the main communities of scientists (the users of MyLibrary) at Los Alamos.

Figure 4 depicts the sub-graph of the 12 most frequent journals and their associations. This sub-network shows that the top chemistry journals are strongly associated with one another, and so are the top physics journals. The two groups are then associated with one another with weaker edges. The more generalist and Biology-driven journals are separated from this main group.

Similarly, we obtain the complementary proximity measure from the same relation *A* between Personalities and ISSN: the *Personality ISSN Proximity* (PIP):

$$PIP(p_s, p_t) = \frac{\sum_{k=1}^{m} (a_{s,k} \wedge a_{t,k})}{\sum_{k=1}^{m} (a_{s,k} \vee a_{t,k})} = \frac{N_{\cap}(p_s, p_t)}{N(p_s) + N(p_t) - N_{\cap}(p_s, p_t)} \quad \textbf{(5)}$$

This proximity, defined on the set of Personalities *P*, captures the probability that two personalities $p_s$ and $p_t$ contain links to the same journal or ISSN. It depends on the sets of ISSN contained in each personality, and the intersection of these sets. $N_{\cap}(p_s, p_t)$ is the number of ISSN contained in both personalities. Two personalities are near if they tend to contain many of the same ISSN.

## 7. Proximity Networks for Recommendation

### 7.1 Recommending Journals in MyLibrary

With proximity networks, in addition to recommending items which are strongly associated with a single item, we can recommend items which are highly associated with a set of target items. For any given user personality (library) $p_u$, the set of unique ISSN contained in all its links, $I(p_u)$, is collected. For every ISSN $i_s \in I(p_u)$, we obtain all $i_t$ such that $IPP(i_s, i_t) \geq \alpha$, where $\alpha$ is a desired minimum value of proximity. This value specifies if the user gets journals more or less associated with the input ISSN. We use three different values of $\alpha$: 0.1 (Low), 0.2 (Medium), 0.3 (High) – giving users three different values of association.

This process yields all the ISSN $i_t$ which are associated to at least one of the $i_s \in I(p_u)$. But what we prefer to recommend are the ISSN $i_t$ which are associated with <u>all</u> $i_s \in I(p_u)$, or at least with as large a subset as possible. We can produce such a recommendation set of ISSN, $I_R$, in different ways. The most restrictive way is defined by formula (6):

$$I_R = \left\{ i_t : \underset{\forall i_s \in I(p_u)}{MIN} \left( IPP(i_s, i_t) \right) \geq \alpha \right\} \quad \textbf{(6)}$$

This formula requires that each recommended ISSN $i_t$ be associated with <u>every</u> input ISSN $i_s \in I(p_u)$ with a minimum value of proximity $\alpha$. A less restrictive procedure, which we currently use in MyLibrary, uses the average operator, rather than the minimum:

$$I_R = \left\{ i_t : \underset{\forall i_s \in I(p_u)}{AVG} \left( IPP(i_s, i_t) \right) \geq \alpha \right\} \quad \textbf{(7)}$$

As an example, we choose as input set $I(p_u)$ = {*Soft Computing*, *Fuzzy Sets and Systems, Networks*}. With $\alpha$ = 0.3, if we use the MIN operator we recommend the following journal:

1063-6706--IEEE transactions on fuzzy systems: 0.3333

If we use the AVG operator, we recommend the following journals:

1063-6706– IEEE transactions on fuzzy systems: 0.3778
1094-7167– IEEE intelligent systems: 0.3333
1089-778X– IEEE Trans. on evolutionary computation: 0.3333
0924-6495– Minds and machines: 0.3000

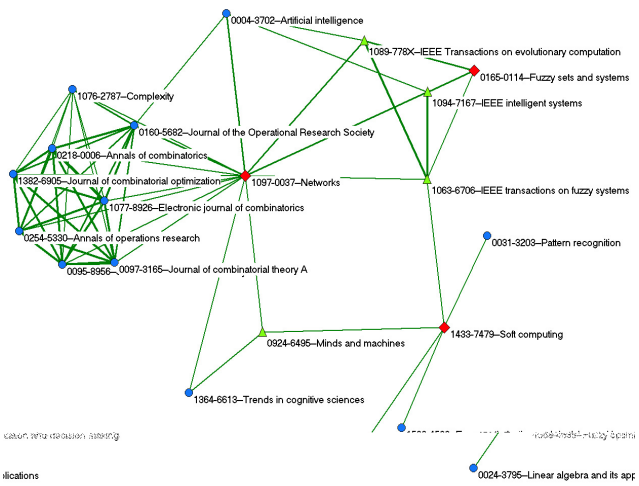


**Figure 5:** Subgraph of *IPP* network with all journals associated with at least one of the jornal target set (red diamonds) with proximity greater than 0.3. Green triangles denote the recommended journals.

Figure 5 depicts the subgraph of the *IPP* network with all nodes associated to example input set $I(p_u)$ with $\alpha \geq 0.3$. If we were to recommend all journals highly related to at least one of the journals in $I(p_u)$, several journals related only to one of the journals in the input set would be recommended. That would be the case of the entire cluster of combinatorics journals associated with the journal "Networks". Instead, we require that recommended journals be associated to the input set as a whole. If we use formula (6), only the journal "IEEE trans. On fuzzy systems" is recommended, since this is the only node in the network with edges to every node in the input set. If we use formula (7), we recommend 3 more journals whose nodes have edges to 2 out of 3 journals in the input set.

## 7.2 Recommending Other Users

Similarly to *IPP* ,the *PIP* proximity defines an associative network of user personalities. We can thus establish which personalities (libraries) are similar to those of a given user. Specifically, given personality $p_s$, we can retrieve the set, $P_R$, of personalities $p_t$ which are associated with $p_s$ with a value of proximity $PIP(p_s, p_t) \geq \alpha$.

This way, the *MyLibrary* service is also used recommend fellow users, thus establishing a collaboration tool. The ability to recommend fellow users is very useful at Los Alamos, especially for new hires and visitors who have difficulty in meeting fellow scientists with similar research interests.

# 8. Semi-metric Behavior

## 8.1 Identifying Transitive Associations

Proximity graphs as we have constructed them, capture associations amongst elements of a set, such as ISSN, which are directly measured. A high value of proximity means that two items tend to co-occur frequently in another set of objects (such as user libraries in *MyLibrary*). But what about items that do not co-occur frequently with one another, but do occur frequently with the same other elements? In other words, even if two items do not co-occur much, they may occur very frequently with a third (or more) item. Should we infer that the two items are associated via indirect associations, that is, from transitivity?

From the inverse of the generic proximity measures *XYP* and *YXP*, obtained from a relation *R* between sets *X* and *Y* using formulae (1), we compute a generic distance function among the elements of *X* or *Y*:

$$d_X(x_i, x_j) = \frac{1}{XYP(x_i, x_j)} - 1; \quad d_Y(y_i, y_j) = \frac{1}{YXP(y_i, y_j)} - 1 \quad \textbf{(8)}$$

$d_X$ and $d_Y$ are distance functions because they are nonnegative, symmetric, real-valued functions such that $d(x, x) = 0$ [12]. They define two weighted graphs $D_X$ and $D_Y$, which we refer to as *distance graphs*, whose vertices are all of the elements of *X* or *Y*, and the edges are the values of $d_X(x_i, x_j)$ and $d_Y(y_i, y_j)$, respectively. A small distance between elements implies a strong association between them.

The distance graphs *D* obtained from applying distance function *d* (formula 8) to proximity graphs, are not Euclidean because, for a pair of elements of *X* (or *Y*) $x_1$ and $x_2$, the triangle inequality may be violated: $d(x_1, x_2) \geq d(x_1, x_3) + d(x_3, x_2)$ for some element $x_3$. This means that the shortest distance between two elements in *D* may not be the direct edge but rather an indirect path. Such distance functions are referred to as semi-metrics [13].

Naturally, the distance graphs obtained from applying a distance function such as *d* can be made Euclidean. If we compute the transitive closure of the respective proximity, we would obtain a similarity relation on *X* or *Y*. The application of distance *d* (formula 8) to a similarity graph would result in a metric distance graph.

We have compiled evidence elsewhere [14] that those pairs of elements with larger semi-metric behavior (those which possess at least one indirect path between them whose distance is much shorter than the direct link) denote a *latent* association. That is, an association which is not grounded on direct evidence provided by the relation *R*, but rather implied by the overall network of associations in this relation. More

formally, when $d(x_i, x_j) << d(x_i, x_k) + \cdots + d(x_l, x_m) + \cdots + d(x_p, x_j)$, then the pair $(x_i, x_j)$ possesses a latent association in distance graph $D$.

What a latent association means depends on the semantics of the application. For instance, in graphs of keyword co-occurrence in documents, a latent association is associated with novelty and can be used to identify trends [14]. In the case of social networks such as terrorist networks [15], a latent association may identify pairs of people, groups, etc. for which we do not have direct evidence, in the available documents, that a real association exists, but who could easily be indirectly associated.

In the *MyLibrary* service, a latent association identifies journals that very few users have included in the same library, but which are nonetheless very strongly implied via indirect journals which people have included in the same libraries.

Clearly, semi-metric behavior (or latency) is a question of degree. For some pairs of vertices in a distance graph the indirect distance provides a much shorter short-cut, a larger reduction of distance, than for others. One way to capture this property of pairs of semi-metric vertices is to compute a *semi-metric ratio*:

$$s\left(x_i, x_j\right) = \frac{d_{direct}\left(x_i, x_j\right)}{d_{shortest}\left(x_i, x_j\right)} \qquad \textbf{(9)}$$

where $d_{shortest}$ is the shortest, direct or indirect, distance between the pair of elements. $s$ is positive and $\geq 1$ for semi-metric pairs. Often, the direct distance between two elements is $\infty$ because they do not relate to any common elements in the other set of relation $R$. For instance, a pair of ISSN that never co-occur in personalities. As a result, $s$ is also $\infty$ for these cases. Thus, $s$ is not capable of discerning semi-metric behavior for pairs that do not have an initial finite direct distance. To detect relevant instances of this infinite semi-metric reduction, we define the *below average ratio*:

$$b\left(x_i, x_j\right) = \frac{\overline{d_{x_i}}}{d_{shortest}\left(x_i, x_j\right)} \qquad \textbf{(10)}$$

where $\overline{d_{x_i}}$ represents the average direct distance from $x_i$ to all $x_k$ such that $d_{direct}(x_i, x_k) \geq 0$. $b$ is only applied to semi-metric pairs of elements $(x_i, x_j)$ where $d_{shortest}(x_i, x_j) < d_{direct}(x_i, x_j)$ and it measures how much the shortest indirect distance between $x_i$ and $x_j$ falls below the average distance of all elements $x_k$ directly associated with element $x_i$. Note that $b(x_i, x_j) \neq b(x_j, x_i)$. Of course, $b$ can also be applied to pairs with finite semi-metric reduction. $b > 1$ denotes a below average distance reduction.

## 8.2 Semi-metric Recommendation

From a recommendation standpoint, after calculating the semi-metric behavior of a distance graph, one is naturally interested in identifying the specific pairs of elements that are most semi-metric – those with higher latency. These are pairs of elements (e.g. journals) whose strong association is not picked by direct co-occurrence in a relation (e.g. in the journal/personality relation $A$), but is rather implied by the overall associative network obtained from the relation. We can say that a strong latent association between pairs of elements is a global property of the network. These are items which have not been directly associated in the data, but implied by the transitivity of the entire network of associations.

To compute these pairs, we compute the metric closure of the distance graph obtained from formula (8). By metric closure we mean that we calculate the shortest path between any pair of elements. To obtain the shortest distance between a pair of nodes of distance graphs $D$, one uses a (+, min) matrix composition of $D$ until closure is achieved[5]. We then compare the metric closure graph with the original distance graph using the semi-metric ratios (12) and (13), thus identifying the most semi-metric pairs in the distance graph. We notice that while recommendations issued based on proximity are grounded on directly observed co-occurrence, semimetric recommendations are not. Indeed, they are indirect, looser associations that we believe the users might be interested in. Therefore, they are recommended separately under a "you might also be interested in these items" heading.

The following journals are journals that did not co-occur at all with the journal "1097-6256--Nature neuroscience", but are indirectly related to it via a strong semi-metric path:

1065-9471--Human Brain Mapping
0278-0062--IEEE transactions on medical imaging
1053-8119--NeuroImage
0031-9155--Physics in medicine & biology
0096-3518--IEEE trans. acoustics speech and signal processing
0740-7467--IEEE acoustics speech and signal proc. magazine
1070-9908--IEEE signal processing letters

## 8.3 Evaluation of Semi-metric Recommendation

From *IPP*, we extracted the top 200 pairs of journal names with highest value of the semi-metric ratio $s$ and the top 200 pairs of journal names with highest parameter $b$. We then generated several random graphs with the same journal names as *IPP*. Specifically, we generated the following 10 random proximity graphs:

[5] Note that traditional algebraic matrix composition is (*, +).

- 4 graphs obtained from randomly shuffling the journal name labels.
- 3 random graphs with the same random distribution of weights a *IPP*. (Erlang-6 random distribution)
- 3 random graphs with uniform distribution of proximity weights.

From each of these 10 random graphs we extracted the top 20 pairs of journal names with highest semi-metric ration *s* and the top 20 pairs of journal names with highest *b*. A software application was developed to ask experts the relevance of pairs. The pairs were displayed to experts[6] by random sampling from the mixed set of 400 journal name pairs extracted from the semi-metric behavior of the real *IPP* and 400 from the semi-metric behavior of random graphs. Given a pair of journal names, experts were asked if the pair of journals was relevant to each other, not relevant, or they did not know the relevance.

Experts were asked about the relevance of 771 real pairs and 723 random ones. Of the real pairs, 512 (66.4%) were deemed relevant, only 71 (9.2%) irrelevant, and 188 (24.4%) were unknown. Of the random pairs, 161 (21.4%) were deemed relevant, 387 (51.4%) irrelevant, and for 205 (27.2%) the relevance was unknown. We notice that the amount of unknowns is very similar for both the real and the random set, which reflects the same amount of journals that our set of experts was unqualified to judge. But the number of positive responses for the real set is well above the number for the random set.

## 9. Future Directions

We have presented a novel recommendation methodology based on fuzzy graphs with probabilistic weights. We have shown that proximity networks can be useful to capture associative knowledge extracted from large collections of documents and web pages. Furthermore, when we convert them to distance graphs, we can uncover indirect or latent associations in the networks, useful for recommendation an data discovery purposes.

We plan to extend the MyLibrary WWW service to a tool that can be used in the WWW at large, and not just a specific digital library. We also plan to continue research in proximity networks and their semi-metric behavior. Specifically, by understanding the relationship between metric closure and transitive closure, and by studying in detail the network characteristics of the proximity networks we have extracted from real data.

[6] Experts consisted of 14 scientists at the Los Alamos National Laboratory.

## 10 References


[1] Rocha LM: **TalkMine: a Soft Computing Approach to Adaptive Knowledge Recommendation.** In *Soft Computing Agents: New Trends for Designing Autonomous Systems*. Edited by Vincenzo Loia, Salvatore Sessa. Physica-Verlag, Springer; 2001:89-116.

[2] Rocha LM: **Automatic Conversation Driven by Uncertainty Reduction and Combination of Evidence for Recommendation Agents.** In *Systematic Organization of Information in Fuzzy Systems*. Edited by Melo-Pinto P, Teodorescu HN, Fukuda T. IOS Press; 2003:249-265.

[3] Luce R, DiGiacomo M: **Personalized and collaborative digital library capabilities: responding to the changing nature of scientific research.** *Science and Technology Libraries* 2004.

[4] Balabanovi M, Shoham Y: **Content-based, collaborative recommendation.** *Communications of the ACM* 1997, **40:** 66-72.

[5] Herlocker JL, Konstan JA, Bouchers A, Riedl J: **An algorithmic framework for performing collaborative filtering.** Proc. of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, 230-237.

[6] Rocha LM: **TalkMine and the Adaptive Recommendation Project.** In: *Proceedings of the Association for Computing Machinery (ACM) - Digital Libraries 99, U.C. Berkeley, August 1999*; 242-243.

[7] Klir GJ, Yuan B: *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Upper Saddle River, NJ: Prentice Hall; 1995.

[8] SHENOI S, MELTON A: **Proximity relations in the fuzzy relational database model.** *Fuzzy Sets and Systems* 1989, **31:** 285-296.

[9] Turney PD: **Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL.** Proc. of the 12th European Conf. on Machine Learning (ECML-2001). 491-502, 2001.

[10] Verspoor K, Cohn J, Joslyn C, Mniszewski S, Rechtsteiner A, Rocha LM *et al*.: **Protein annotation as term categorization in the Gene Ontology using Word Proximity Networks.** *BMC Bioinformatics* 2004, In Press.

[11] Wall ME, Rechtsteiner A, Rocha LM: **Singular value decomposition and principal component analysis.** In *A Practical Approach to Microarray Data Analysis*. Edited by Berrar DP, Dubitzky W, Granzow M. Norwell,MA: Kluwer; 2003:91-109.

[12] Shore SD, Sawyer LJ: **Explicit Metrization.** *Annals of*

*the New York Academy of Sciences* 1993, **704:** 328-336.

[13] Galvin F, Shore SD: **Distance Functions and Topologies.** *American Mathematical Monthly* 1991, **98:** 620-623.

[14] Rocha LM: **Semi-metric Behavior in Document Networks and its Application to Recommendation Systems.** In *Soft Computing Agents: A New Perspective for Dynamic Information Systems*. Edited by V.Loia. IOS Press; 2002:137-163.

[15] Rocha LM. Proximity and Semi-Metric Analysis of Social Networks. LAUR 02-6557. 2002. Los Alamos National Laboratory Internal Report.