In: Proc. 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI'05),
IEEE Press. IEEE Press, pp. 565-571.

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

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

[†]School of Informatics and Cognitive Science Program, Indiana University
1900 East Tenth Street, Bloomington, IN 47401, USA, rocha@indiana.edu
[‡]Center for Genomics and Bioinformatics, Indiana University, Bloomington, IN 47401, USA
[§]Research Library, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

## Abstract

*We describe a network approach to building recommendation systems for a Web 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 digital library collections and Web resources.*

## 1. Introduction

The Web is used today as a means to integrate many electronic information resources. In particular, it enables the creation of personalized and collaborative digital library services. Indeed, the Web 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 the Los Alamos National Laboratory (LANL) provides scientists 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 [10]. In this paper we present a network analysis methodology to produce recommendation systems and enhanced collaboration in this particular web service. This methodology is 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 LANL's digital library collections and Web resources. It supports a collection of personal links to a variety of information resources such as electronic journals, full-text content and bibliographic databases. 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, 400+ subject based web links from the Research Library's website, as well as any generic link to Web content. Users can access MyLibrary from any LANL computer[1].

This Web service consists of three nested entities: (1) **libraries**, (2) **folders**, and (3) **links**. A library (also referred to in this article as a **personality**) is associated with a given area of interest for each user, e.g., physics, computer science, etc. Each library consists of one or more (subcategory) folders that contain related types of links (URLs to specific web resources). 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 two folders with links to relevant databases

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

and e-journals, respectively. Once a library has been created, new folders and links can be added to it. Link checking is run on a weekly basis and broken links are identified.

### 2.2. Enabling Collaboration in MyLibrary

Digital libraries (DL) now offer opportunities for collaboration and communication that were not feasible in traditional libraries. The last generation of DL interfaces largely reflected single, generic user stereotypes. That is, the activities or behaviors of users have had almost no impact on the experience of any one user. In contrast, today's web technology allows us to consider new ways of working with DL. Rather than limiting the user to work in an isolated mode as a individual with generic capabilities, we can now 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 the MyLibrary service. Two types of collaboration are supported: direct and indirect. *Direct collaboration* refers to situations where several users agree to work together as a defined group exploring and making use of digital library resources. Our direct collaboration features have been described elsewhere [6].

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 [10] (or recommender systems e.g. [3]) that we describe in the remainder of this article and which were developed by the *Active Recommendation Project* (ARP) [7] for LANL's Digital Library. Here we describe ARP's work on the MyLybrary@LANL service.

## 3. Proximity Networks

### 3.1. Background on Fuzzy Graphs

A $n$-ary relation, $R$, between $n$ sets $X_1, X_2, \cdots, X_n$, assigns a value, $r$, to elements, $\mathbf{x} = (x_1, x_2, \cdots, x_n)$, of the Cartesian product of these sets: $X_1 \times X_2 \times \cdots \times X_n$. The value $r$ signifies how strongly the elements $x_i$ of the $n$-tuple $\mathbf{x}$ are related or associated to one another ([5] page 119). When $r \in [0, 1]$, $R$ is known as a *fuzzy relation* [5], and when $n = 2$ as a *binary relation*. 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. Here, we work with binary relations which relate pairs of items such as users and journals, keyterms and documents, etc.

Binary fuzzy relations defined on single set, $R(X, X)$, are also known as *fuzzy graphs* (a kind of weighted graphs). The composition of fuzzy graphs is performed by the (fuzzy) logic composition of their matrices in much the same way as the algebraic composition of matrices, except that multiplication and summation are substituted by fuzzy set aggregation operations such as intersection and union or more generally T-Norms and T-Conorms respectively [5]. 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*, where $r_{ij}$ denotes $R(x_i, x_j)$, the weight of the edge between vertices $x_i$ and $x_j$:

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

The *transitive closure* of a fuzzy graph $R(X, X)$ is defined as the graph that is transitive, contains $R(X, X)$, and whose edges have the smallest possible 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. A fuzzy graph $R(X, X)$ is (max-min) *transitive* iff:

$$r_{ik} \geq \max_{\forall x_j \in X} \min[r_{ij}, r_{jk}], \forall_{x_i, x_k \in X}$$

This definition generalizes the crisp transitive property which requires that $(x_i, x_k)$ be related if $(x_i, x_j)$ and $(x_j, x_k)$ are related. It requires that the weight of an indirect path between $x_i$ and $x_k$ through some $x_j$, is the smallest edge in the path ($x_i$ to $x_j$ or $x_j$ to $x_k$). Finally, the weight of the edge between $x_i$ and $x_k$, must be larger or equal to the largest of all indirect paths through each $x_j$. The algorithm to obtain the transitive closure $R^T$ of $R$ is [5]:

1. $R' = R \cup (R \circ R)$
2. If $R' \neq R$, make $R = R'$ and go back to step 1.
3. Stop: $R^T = R'$

$R(X, X)$ is a *similarity* (or equivalence) graph/relation if it is reflexive ($R(x, x) = 1$), symmetric ($R(x, y) = R(y, x)$), and transitive. $R(X, X)$ is a *proximity* (Compatibility) graph/relation if it is reflexive and symmetric. The transitive closure of a proximity graph is a similarity graph.

### 3.2. Generic Proximity Measure

Our approach is based on a probabilistic proximity measure computed from binary relations between any two sets of items (e.g. keywords and documents). Given a generic binary relation $R$ between sets $X$ (of $n$ elements $x$) and $Y$ (of $m$ elements $y$), we extract two complementary proximity graphs: $XYP$ 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:

$$xyp(x_i, x_j) = \frac{\sum_{k=1}^{m}(r_{ik} \wedge r_{jk})}{\sum_{k=1}^{m}(r_{ik} \vee r_{jk})}$$

$$yxp(y_i, y_j) = \frac{\sum_{k=1}^{n}(r_{ki} \wedge r_{kj})}{\sum_{k=1}^{n}(r_{ki} \vee r_{kj})}$$

Other measures of probability can be used to capture a degree of association or closeness between elements of two sets in a binary relation. In information retrieval it is common to use conditional probabilities [11]. In that case, we do not obtain a proximity relation since conditional probabilities are not symmetric. For characterizing closeness in relations, we prefer our proximity measure because it is symmetric. Indeed, proximity intuitively captures the inverse of a distance, which requires symmetry. As we discuss below, the idea of distance is important for our recommendation algorithms. We note that a large value of proximity requires a large value of both directions of conditional probability.

### 3.3. Capturing Knowledge in a Network

Proximity relations are fuzzy graphs which we can think of as networks of elements. We derive our proximity networks from the computation of the probability measures of section 3.2 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 [9, 10]. As in any other co-occurrence method, the assumption is that items that frequently co-occur are associated with a common concept understood by the community of users and writers of the documents.

Notice that a graph of co-occurrence proximity allows us to capture network associations rather than just pairwise co-occurrence. In other words, we expect concepts or themes to be organized in more interconnected sub-graphs, or clusters of items in the proximity networks. Indeed, we have successfully used proximity networks in several knowledge extraction and literature mining applications, such as the Text Mining competition *BioCreAtIvE* (Critical Assessment of Information Extraction in Biology) [4]. Our submission based on the word proximity network analysis, was one of the most successful submissions from several research groups using many types of machine learning methods. Please refer to [12] for a full discussion of results. Here, we restrict our discussion to the application of proximity networks to *MyLibrary*.

## 4. MyLibrary Network Extraction

### 4.1. Relation Extraction

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. Using a continuously updated file associating scientific journals (identified by their *International Standard Serial Number* or ISSN) to a set of URLs where users can access journal articles, we identify the subset of links in user libraries/personalities 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.

| ISSN | Journal Name | $N(i_t)$ |
|---|---|---|
| 0031-9007 | Physical review letters | 53 |
| 0556-2791 | Physical review A | 31 |
| 0556-2805 | Physical Review B | 28 |
| 1095-3787 | Physical review E | 26 |
| 0021-9606 | J. of Chem. Physics | 26 |
| 0002-7863 | J. American Chemical Soc. | 23 |
| 1089-5647 | Journal of physical chem. B | 23 |
| 0034-6861 | Reviews of modern physics | 22 |
| 1089-5639 | Journal of physical chem. A | 20 |
| 0028-0836 | Nature | 20 |
| 0036-8075 | Science | 20 |
| 0027-8424 | PNAS | 20 |

**Table 1. top 12 most frequent ISSN**

From this data we produce the relation PERSONALITY $\times$ ISSN which is a binary relation $A : P \times I$ between the sets $P$ of $n_p$ personalities (or libraries), and the set $I$ of $m$ ISSN which occur in at least one personality. In the particular dataset used below, $n_p = 392$ and $m = 1702$. We treat $A$ as a crisp relation: $a(p_s, i_t) = a_{s,t} = 1$ (True) if ISSN $i_t$ occurs in personality $p_s$, and 0 (false), otherwise. We also define $N(i_t)$ as the number of personalities in which ISSN $i_t$ occurs and $N(p_s)$ as the number of ISSN contained in personality $p_s$. The top 12 most frequent ISSN are listed in table 1. 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.

### 4.2. Proximity Network Extraction

To discern closeness amongst ISSN according to the personalities they occur in, we compute the *ISSN Personality Proximity* ($IPP$), from relation $A$ using the proximity formula of section 3.2. The proximity between two ISSN is the probability that both co-occur in the same personality. Thus, two ISSN are near if they tend to occur in many of the same personalities.

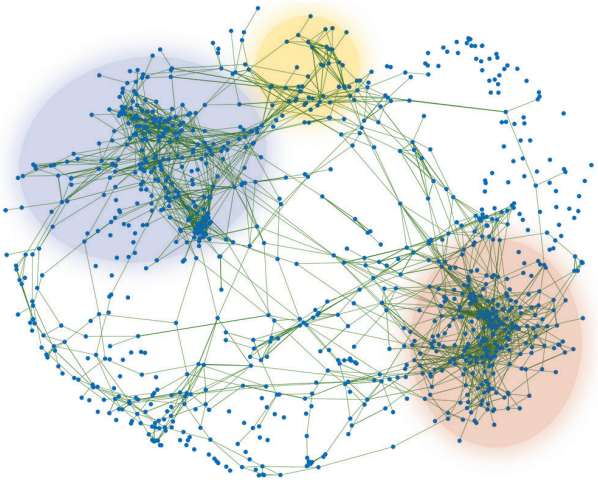

**Figure 1.** $IPP$ **network showing edges with** $ipp \geq 3$**.**

$IPP$ is an associative network of Journals (identified by ISSN). This network, displayed in figure 1, is a weighted, probabilistic graph, whose edges are the co-occurrence proximity values[2]. For instance, the journals associated with the most frequent journal ("*Physical Review letters*"), with a proximity value of $ipp \geq 0.3$ are listed in table 2.

Figure 1 clearly shows two main clusters of nodes highly associated in the $IPP$ network. The Principal Component Analysis (PCA) analysis of this network revealed that the two first eigen-vectors (components) are very correlated with the two main clusters identified. The first component refers to a set of journals related to "Chemistry, Materials science and Physics" (left). The second component refers to a set of journals related to "Computer Science and Applied Mathematics" (right). However, these groups are further separated and refined into more specific clusters as we consider more components. A smaller third cluster refers to "Bioinformatics and Computational Biology" (top).The main clusters discovered in the $IPP$ network highlight the reality of the research pursued at LANL. 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.

2 All network figures depicted were drawn using the Fruchterman-Reingold algorithm in Pajek [1]

| ISSN | Journal Name | $ipp$ |
|---|---|---|
| 1095-3787 | 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 |

**Table 2. Journals most associated with PRL**

Figure 2 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; generalist journals are separated from this main group.

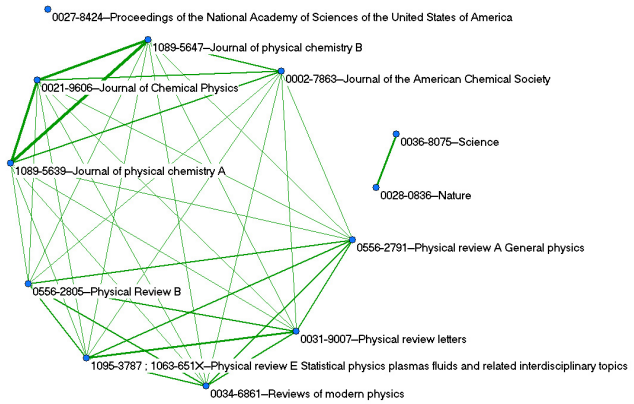


**Figure 2. Sub-graph of** $IPP$ **network with top 12 most frequent journals, showing edges with** $ipp \geq 0.2$**. Edge thickness denotes proximity strength.**

We also compute the complementary proximity network from the same relation $A$ between Personalities and ISSN: the *Personality ISSN Proximity* ($PIP$). This network, defined on the set of Personalities $P$, captures the probability that two personalities contain links to the same journal or ISSN. Two personalities are near if they tend to contain many of the same ISSN.

## 5. Network Recommendation

### 5.1. Recommending Journals

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. In MyLibrary, for any given user personality $p_u$, the set of unique ISSN contained in all its links, $I(p_u)$, is collected. Then, 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 with 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 all or most $i_s \in I(p_u)$. We can produce such a recommendation set of ISSN, $I_R$, in different ways. The most restrictive way is computed using the minimum operator:

$$I_R^{MIN} = \left\{ i_t : \min_{i_s \in I(p_u)} (ipp(i_s, i_t)) \geq \alpha \right\}$$

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

$$I_R^{mean} = \left\{ i_t : \frac{\sum_{i_s \in I(p_u)} (ipp(i_s, i_t))}{|I(p_u)|} \geq \alpha \right\}$$

As an example, consider a user personality $p_u$ with three journals: $I(p_u)$ = *Soft Computing*, *Fuzzy Sets and Systems*, and *Networks*. With $\alpha = 0.3$, if we compute $I_R^{MIN}$ on $IPP$ we would recommend a single journal for this personality $p_u$: *IEEE transactions on fuzzy systems* with minimum proximity = 0.3333. If we compute $I_R^{mean}$ we would recommend four journals (mean proximity shown): *IEEE transactions on fuzzy systems* (0.3778), *IEEE intelligent systems* (0.3333), *IEEE Trans. on evolutionary computation* (0.3333) and *Minds and machines* (0.3).

Figure 3 depicts the subgraph of the $IPP$ network with all journals associated to at least a journal in our example's 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 $I_R^{MIN}$, 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 $I_R^{mean}$, we recommend three more journals with edges (in the sub-graph) to 2 out of 3 journals in the input set.

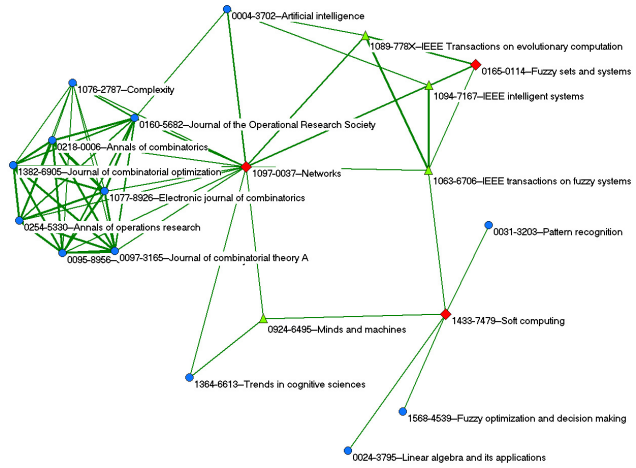


**Figure 3. Sub-graph of $IPP$ network with all journals associated with at least one of the journals in the target set (diamonds) with $ipp \geq 0.3$. Triangles denote the recommended journals with $I_R^{mean}$.**

### 5.2. Recommending other Users

Similarly to $IPP$, the $PIP$ proximity defines an associative network of user personalities. We can thus establish which personalities 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 to 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.

## 6. Semi-metric Behavior

### 6.1. 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 the formulae of section 3.2, we compute generic distance functions among the elements of $X$ and $Y$:

$$d_X(x_i, x_j) = \frac{1}{xyp(x_i, x_j)} - 1$$

$$d_Y(y_i, y_j) = \frac{1}{yxp(y_i, y_j)} - 1$$

$d_X$ and $d_Y$ are distance functions because they are non-negative, symmetric, real-valued functions such that $d(x, x) = 0$ [2]. They define weighted graphs $D_X$ and $D_Y$, which we refer to as *distance graphs*, whose vertices $x_i$ or $y_i$ are the elements of $X$ or $Y$, and the edges are the values $d_X(x_i, x_j)$ and $d_Y(y_i, y_j)$, respectively. A small distance between elements implies a strong association between them. These distance graphs are not in general 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 $D_X$ or $D_Y$ may not be the direct edge but rather an indirect path. Distance functions that violate the triangle inequality are referred to as *semi-metrics* [2].

We have compiled evidence elsewhere [9] 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 $d(x_i, x_j) \ll d(x_i, x_k) + \cdots + d(x_l, x_m) + \cdots + d(x_p, x_j)$, then the edge $(x_i, x_j)$ possesses a latent association in distance graph $D$. We have shown elsewhere that in graphs of keyword co-occurrence in documents, a latent association is associated with novelty and can be used to identify trends [9]. In the case of social networks [8], a latent association identifies 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 in the $D_I$ distance graph obtained from $IPP$ identifies journals that very few users have included in the same library/personality, but which are nonetheless very strongly implied via indirect journals which people have included in the same personalities.

Clearly, semi-metric behavior (or latency) is a question of degree. For some pairs of vertices in a distance graph an indirect path may provide a much shorter indirect short-cut, a shorter distance, than for others. To measure a degree of semi-metric behavior we have introduced the *semi-metric* and *below average ratios* [9]:

$$s(x_i, x_j) = \frac{d_X(x_i, x_j)}{\underline{d}(x_i, x_j)}$$

$$b(x_i, x_j) = \frac{\overline{d_{x_i}}}{\underline{d}(x_i, x_j)}$$

where $\underline{d}(x_i, x_j)$ is the shortest, direct or indirect, distance between $x_i$ and $x_j$ in distance graph $D_X$, and $\overline{d_{x_i}}$ is the mean direct distance from $x_i$ to all other $x_k \in X$ such that $d_X(x_i, x_k) \geq 0$. $s$ is positive and $> 1$ for semi-metric pairs. $b$ is only applied to semi-metric pairs of elements $(x_i, x_j)$ where $0 < \underline{d}(x_i, x_j) < d_X(x_i, x_j)$ and it measures how much the shortest indirect distance between $x_i$ and $x_j$ falls below the average distance of $x_i$ to all its directly associated elements $x_k$. The below average ratio is designed to capture semi-metric behavior of pairs $(x_i, x_j)$ which do not have a finite direct distance $d_X(x_i, x_j)$. Note that $b(x_i, x_j) \neq b(x_j, x_i)$. $b > 1$ denotes a below average distance reduction (see [9] for more details).

## 6.2. Semi-metric Recommendation

From a recommendation standpoint, one is naturally interested in identifying the specific pairs of elements that are most semi-metric. In MyLibrary, these are pairs of journals or users whose association is not picked by direct co-occurrence in the journal/personality relation $A$ (section 4.1, but is rather implied (as a global property) by the proximity networks obtained from the relation. 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 relevant distance graph. By *metric closure* we mean that we calculate the shortest distance between any pair of elements in a distance graph $D$. To do this we use a $(+, min)$ matrix composition of $D$ until closure is achieved, producing a metric closure distance graph $D^{mc}$. Using $D$ and $D^{mc}$, we identify the most semi-metric pairs in the $D$ using the semi-metric ratios of section 6.1.

We notice that while recommendations issued based on proximity are grounded on directly observed co-occurrence, semi-metric 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 "*Nature neuroscience*", but are indirectly related to it via a strong semi-metric path in the distance graph $D_I$ obtained from the $IPP$ proximity network of section 4.2 (ranked by most semi-metric first): "*Human Brain Mapping*", "*IEEE transactions on medical imaging*", "*NeuroImage*", "*Physics in medicine & biology*", "*IEEE trans. acoustics speech and signal processing*", "*IEEE acoustics speech and signal proc. magazine*", and "*IEEE signal processing letters*".

### 6.3. Evaluation of Semi-metric Recommendation

From $D_I$, 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 10 random graphs with the same journals as $D_I$. Specifically, we generated 4 graphs obtained by randomly shuffling the labels of the vertices of $D_I$, 3 random graphs with the same weight distribution as $D_I$ (Erlang-6 random distribution), and 3 random graphs with uniform distribution of distance weights.

From each of these 10 random graphs we extracted the top 20 pairs of journal names with highest semi-metric ratio $s$ and the top 20 pairs of journal names with highest $b$. A software application was developed to ask experts (14 scientists at LANL) the relevance of pairs. The pairs were displayed by random sampling from the mixed set of 400 journal name pairs extracted from the semi-metric behavior of the real $D_I$ and 400 from the semi-metric behavior of random graphs. Given a pair of journal names, experts were asked if a person who would be interested in one of them, would also be interested in the other. In other words, if the the pair of journals were very related to each other, not related, or they did not know about the subject matter of the journals.

Experts were asked about the relatedness of 771 real pairs and 723 random ones. Of the real pairs, 512 (66.4%) were deemed related, only 71 (9.2%) unrelated, and 188 (24.4%) were unknown. Of the random pairs, 161 (21.4%) were deemed related, 387 (51.4%) unrelated, 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.

## 7. 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 Web service into a tool that can be used in the WWW at large, and not just a specific digital library. We will continue work on the validation of our recommendation methodology against other methods. 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.

## References


[1] V. Batagelj and A. Mrvar. Pajek - program for large network analysis. *Connections*, 21(2):47–57, 1998.

[2] F. Galvin and S. Shore. Distance functions and topologies. *American Mathematical Monthly*, 98:620–623, 1991.

[3] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, 2004.

[4] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6:S1, 2005.

[5] G. J. Klir and B. Yuan. *Fuzzy Sets and Fuzzy Logic: Theory and Applications*. Prentice Hall, 1995.

[6] R. Luce and M. DiGiacomo. Personalized and collaborative digital library capabilities: responding to the changing nature of scientific research. *Science and Technology Libraries*, 24(1-2):135–152, 2004.

[7] L. M. Rocha. Talkmine and the adaptive recommendation project. In *DL99: Proc. 4th ACM conf. on Digital libraries*, pages 242–243. ACM Press, 1999.

[8] L. M. Rocha. Proximity and semi-metric analysis of social networks. Technical report, Los Alamos National Laboratory: LAUR 02-6557, 2002.

[9] L. M. Rocha. Semi-metric behavior in document networks and its application to recommendation systems. In V. Loia, editor, *Soft Computing Agents: A New Perspective for Dynamic Information Systems*, pages 137–163. IOS Press, 2002.

[10] L. M. Rocha. Automatic conversation driven by uncertainty reduction and combination of evidence for recommendation agents. In H. F. T. MeloPinto, P; Teodorescu, editor, *Systematic Organisation of Information in Fuzzy Systems*, volume 184, pages 249 – 265. I O S PRESS, 2003.

[11] P. D. Turney. Mining the Web for synonyms: PMI–IR versus LSA on TOEFL. *LNCS*, 2167:491–502, 2001.

[12] K. Verspoor, J. Cohn, C. Joslyn, S. Mniszewski, A. Rechtsteiner, L. M. Rocha, and T. Simas. Protein annotation as term categorization in the gene ontology using word proximity networks. *BMC Bioinformatics*, 6 Suppl 1:S20, 2005.