LA-UR- 02-154 c.1

Approved for public release;
distribution is unlimited.

Title: COMBINATION OF EVIDENCE IN RECOMMENDATION SYSTEMS CHARACTERIZED BY DISTANCE

Author(s): Luis M. Rocha, CCS-3, LANL

Submitted to. 2002 World Congress on Computational Intelligence, 2002 IEEE International Conference on Fuzzy Systems

Combination of Evidence in Recommendation Systems Characterized by Distance Functions

Luis M. Rocha
Complex Systems Modeling
Los Alamos National Laboratory, **MS B256**
Los Alamos, **NM 87545**, U.S.A.
E-Mail: rocha@lanl.gov
WWW: http://www.c3.lanl.gov/~rocha

Abstract - Recommendation systems for different Document Networks (DN) such as the World Wide Web (WWW), Digital Libarries, or Scientific Databases, often make use of distance functions extracted from relationships among documents and between documents and semantic tags. For instance, documents in the WWW are related via a hyperlink network, while documents in bibliographic databases are related by citation and collaboration networks. Furthermore, documents can be related to semantic tags such as keywords used to describe their content, The distance functions computed from these relations establish associative networks among items of the DN, and allow recommendation systems to identify relevant associations for individual users. The process of recommendation can be improved by integrating associative data from different sources. Thus we are presented with a problem of combining evidence (about associations between items) from different sources characterized by distance functions. In this paper we summarize our work on (1) inferring associations from semi-metric distance functions and (2) combining evidence from different (distance) associative DN.

1. Recommendation in Document Networks

The prime example of a Document Network (DN) is the World Wide Web (WWW). But many other types of such networks exist: bibliographic databases containing scientific publications, preprints, internal reports, as well as databases of datasets used in scientific endeavors. Each of these databases possesses several distinct relationships among documents and between documents and semantic tags or indices that classify documents appropriately.

DN typically function as information resources for communities of users who query them to obtain relevant information for their activities. Resources such as the Internet, Digital Libraries, and the like have become ubiquitous in the past decade, demanding the development of new techniques to cater to the information needs of communities of users. These techniques come from the field of Information Retrieval, and are typically known as Recommender Systems e.g. [6] [5] [3] [16].

The algorithms we have developed in this area integrate evidence about the association amongst elements of DN, amongst users, and about the interests of individual users and their communities. In particular, a soft computing algorithm (*TalkMine*) has been created to integrate such evidence and also adapt DN to the expectations of their users [15]. The process of integration of knowledge in *TalkMine* requires the construction of distance functions on DN that characterize the associations amongst their components. Below we discuss how such distance functions are used to characterize DN and for recommendation.

2. Distance Functions in Document networks

2.1 Harvesting Relations from Document Networks

For each DN we can identify several distinct relations among documents and between documents and semantic tags used to classify documents appropriately. For instance, documents in the WWW are related via a hyperlink network, while documents in bibliographic databases are related by citation and collaboration networks [113. Furthermore, documents can be related to semantic tags such as keywords used to describe their content. Although all the technology and the hypothesis here discussed would apply equally to any of these relations extracted from DN, let us exemplify the problem with the datasets we have created for the ***Active Recommendation Project*** (ARP) (http://arp.lanl.gov), part of the Library Without Walls Project, at the Research Library of the Los Alamos National Laboratory [1;8].

ARP is engaged in research and development of recommendation systems for digital libraries, The information resources available to ARP are large databases with academic articles. These databases contain bibliographic, citation, and sometimes abstract information about academic articles. One of the databases we work with is ***SciSearch***, containing articles from scientific journals from several fields collected by ISI ***(Institute for Scientific Indexing).*** We collected all ***SciSearch*** data from the years of **1996** to **1999.** There are **2,915,258** documents, **from** which we extracted **839,297** keywords **(semantic** tags) that occurred at least in **two** distinct documents. We have compiled relational information between records and keywords. This relation allows us to infer the semantic value of documents and the inter-associations between keywords. Such semantic relation is stored as a very sparse Keyword-Record Matrix A. Each entry $a_{i,j}$ in the matrix is boolean and indicates whether keyword k_i indexes **(1)** document'd_j or not (0). The sources of keywords are the terms authors and/or editors chose to categorize (index) documents, as well as title words.

2.2 Computing Associative Distance Functions

To discern closeness between keywords according to the documents they classify, we compute the Keyword ***Semantic Proximity*** (*KSP*), obtained from A by the following formula:

$$KSP\left(k_i,k_j\right)=\frac{\sum_{k=1}^{m}\left(a_{i,k}\wedge a_{j,k}\right)}{\sum_{k=1}^{m}\left(a_{i,k}\vee a_{j,k}\right)}=\frac{N_{\cap}\left(k_i,k_j\right)}{N_{\cup}\left(k_i,k_j\right)}= \quad (1)$$

The semantic proximity[1] between two keywords, k_i and k_j, depends on $N_{\cap}(k_i, k_j)$, the number of documents both keywords index, and $N_{\cup}(k_i, k_j)$, the number of documents either keyword indexes. Two keywords are near if they tend to index many of the same documents. Table I lists the values of *KSP* for the 10 most common keywords in the ARP dataset. From the inverse of ***KSP*** we obtain a distance function between keywords:

$$d\left(k_i,k_j\right)=\frac{1}{KSP\left(k_i,k_j\right)}-1 \quad (2)$$

d is a distance function because it is a nonnegative, symmetric real-valued function such that $d(k, k) = 0$ [20]. It defines a weighted, non-directed ***distance graph D*** whose nodes are all of the keywords extracted from a given DN, and the edges are the values of ***d***.

Table I: **KSP** for **10** most frequent keywords

cell	studi	system	express	protein	model	activ	human	rat	patient
1.00	0.02	0.02	0.16	0.08	0.02	0.09	0.11	0.07	0.03
0.02	1.00	0.03	0.01	0.02	0.03	0.02	0.02	0.02	0.04
0.02	0.03	1.00	0.02	0.02	0.05	0.02	0.01	0.02	0.01
0.16	0.01	0.02	1.00	0.13	0.01	0.07	0.10	0.08	0.02
0.08	0.02	0.02	0.13	1.00	0.01	0.07	0.06	0.04	0.01
0.02	0.03	0.05	0.01	0.01	1.00	0.02	0.02	0.03	0.01
0.09	0.02	0.02	0.07	0.07	0.02	1.00	0.06	0.05	0.02
0.11	0.02	0.01	0.10	0.06	0.02	0.06	1.00	0.03	0.02
0.07	0.02	0.02	0.08	0.04	0.03	0.05	0.03	1.00	0.01
0.03	0.04	0.01	0.02	0.01	0.01	0.02	0.02	0.01	1.00

3. METRIC BEHAVIOR

The distance function d (eq. 2) is not an Euclidean metric because it may violate the triangle inequality: $d(k_1, k_2) \leq d(k_1, k_3) + d(k_3, k_2)$ for some keyword k_3. This means that the shortest distance between two keywords may not be the direct link but rather an indirect pathway in *D*. Such measures of distance are referred to as semi-metrics [2]. Indeed, given that most social and knowledge-derived networks possess Small-World behavior [22], we expect that nodes which tend to be clustered in a local neighborhood of related nodes, have large distances to nodes in other clusters. But because of the existence of "gateway" nodes relating nodes in different clusters (the small-world phenomenon), smaller indirect distances between nodes in distinct clusters, through these "gateway" nodes, are to be expected.

Clearly, semi-metric behavior is a question of degree. For some pairs of keywords, 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 keywords is to compute a ***semi-metric ratio:***

$$s\left(k_i,k_j\right)=\frac{d_{direct}\left(k_i,k_j\right)}{d_{indirect}\left(k_i,k_j\right)} \quad (3)$$

s is positive and ≥ 1 for semi-metric pairs. Given that larger graphs tend to show a much larger spread of distance, *s* tends to increase with the number of keywords. Therefore, to be able to compare semi-metric behavior between different DN and their respective different sets of keywords, a ***relative semi-metric ratio*** is also used:

$$rs\left(k_i,k_j\right)=\frac{d_{direct}\left(k_i,k_j\right)-d_{indirect}\left(k_i,k_j\right)}{d_{\max}} \quad (4)$$

rs compares the semi-metric distance reduction to the maximum possible distance reduction, $d_{\max}$, in graph *D*.

Often, the direct distance between two keywords is ∞ because they do not index any documents in common. As a result, *s* and ***rs*** are also ∞ for these cases. Thus, ***s*** and ***rs*** are not capable of discerning the degree of semi-metric behavior for pairs that do not have a finite direct distance. To detect relevant instances of this infinite semi-metric reduction, we define the below ***average*** ratio:

$$b\left(k_i,k_j\right)=\frac{\overline{d_{k_i}}}{d_{indirect}\left(k_i,k_j\right)} \quad (5)$$

where $\overline{d_{k_i}}$ represents the average direct distance from k_i to all k_j such that $d_{direct}(k_i, k_j) \geq 0$. *b* measures how much an indirect distance falls below the average distance of all keywords directly associated with a keyword. Of course, b can also be applied to pairs with finite semi-metric reduction.

We have used these three measures of semi-metric behavior to analyze several types of DN [17]. We have shown that $s(k_i, k_j)$ and $rs(k_i, k_j)$ are useful to infer the interests of a user associated with a collection of documents. Specifically, given a collection of documents a user has retrieved, this measure identifies pairs of keyterms highly correlated with the interests of the user as implied by the collection, but which tend not to be simultaneously present in many documents. In other words, it identifies pairs of keyterms which represent well the entire collection (by being highly indirectly associated in the collection

[1] This measure of closeness, **formally,** is **a proximity** relation [4;9] because it is **a** reflexive and symmetric **fuzzy** relation. Its transitive closure is known **as** a **similarity** relation **(Ibid).**

of documents), but not many individual documents in the collection. Such pairs are properties of the network, but not of individual documents. This is clearly an important piece of knowledge to allow us to recommend to users documents which are similar to their interests implied by the entire collection of documents they have retrieved, but which may not be similar to individual documents in the collection.

We have also shown that $s(k_i, k_j)$, $rs(k_i, k_j)$, and $b(k_i, k_j)$ are useful to identify trends in large collections of documents associated with many authors and/or users. When we deal with large DN such as the ARP database discussed above, the derived distance function, reflects myriad associations amongst keywords from a very heterogeneous collection of documents. Instead of a smaller collection associated with a particular user, we deal with a collection of documents from multiple authors and/or users. In this case, the semi-metric behavior measures pick up pairs of keyterms which tend not to co-occur in the same documents, but are nonetheless highly indirectly associated in the distance graph D. We have shown that often, high semi-metric behavior can be used to predict where a given community is moving thematically. Specifically, high semi-metric pairs of keyterms are good predictors that in subsequent years, individual documents will appear which use those pairs directly.

Finally, we have shown that the behavior of $s(k_i, k_j)$, $rs(k_i, k_j)$, and $b(k_i, k_j)$ allows us to characterize the type of DN. By analyzing the semi-metric behavior of a DN, we can infer if it is a collection of documents with many authors/users or if it is more thematically coherent and thus associated with a single user or very coherent community.
For details on these results please refer to [17], here we discuss how to integrate distance information from different sources to improve recommendation.

4. Information Resources and Users

4.1 Knowledge Context

Clearly, many other types of distance functions can be defined on the elements of a DN. Distance functions applied to citation structures or collaboration networks, will require distinct semantic considerations than those used for keyword sets. In any case, we characterize an information resource with sets of these distance functions. Indeed, the collection of all relevant associative distance functions from a DN, is an expression of the particular knowledge it conveys to its community of users as an information resource.

Notice that distinct information resources typically share a very large set of keywords and documents. However, these are organized differently in each resource, leading to different collections of relational information. Indeed, each resource is tailored to a particular community of users, with a distinct history of utilization and deployment of information by its authors and users. For instance, the same keywords will be related to different sets of documents in distinct resources. Therefore, we refer to the relational information of each information resource as a ***Knowledge Context*** [15]. More specifically, we characterize an information resource R with a structure named Knowledge Context:

$$KN_R = \{\mathcal{X}, \mathcal{R}, \mathcal{D}\} \qquad (7)$$

Where $\mathcal{X}$ is a set of available sets of elements X_i, e.g. $\mathcal{X} = \{K, D, U\}$, where K is a set of keyterms, D a set of documents, and U a set of users. $\mathcal{R}$ is a set of available relations amongst the sets in $\mathcal{X}$, e.g. $\mathcal{R} = \{C(D, D), A(K, D)\}$, where C denotes a citation relation between the elements of the set of documents, and A a semantic relation between documents and keyterms, such as the keyterm-record matrix defined in section 2.1. Finally, $\mathcal{D}$ is a set of distance functions built from some subset of relations in $\mathcal{R}$, e.g. $\mathcal{D} = \{d_k\}$, where d_k is the distance between keyterms such as the one defined by formula (2).

4.2 Agent Recommendation Architecture

In our architecture of recommendation [16], users are also characterized as information resources, where $\mathcal{X}$ may contain, among other application-specific elements, the sets of documents previously retrieved by the user and their associated keyterms. Notice that the same user may query information resources with very distinct sets of interests. For example, one day a user may search databases as a biologist looking for scientific articles, and the next as a sports fan looking for game scores. Therefore, the ARP architecture allows users to define different ***"personalities"***, each one with its distinct history of information retrieval defined by independent knowledge contexts.

The analysis of distance functions as mentioned in section 3, provides a baseline recommendation feature [17]. Indeed, given each knowledge context of a user or a larger information resource, we can infer what are the important associated topics and trends . But these knowledge contexts and respective distance functions can additionally be used in integrative algorithms useful for fine tuning the present interests of users, as well as adapt all the knowledge contexts accessed according to user behavior, Such recommendation algorithms, instantiate an automated conversation fabric amongst a population of users and a set of information resources [15]. Each user accesses the set of information resources via a browser that functions as an agent for the user as it engages in automated conversations with the agents of other users and the information resources [18] [16]. This process relies on the integration of evidence about the interests of users implied by distinct distance graphs as discussed below.

5. Evidence From Different Knowledge Contexts

5.1 Describing User Interest with Evidence Sets

Humans use language to communicate categories of objects in the world. But such linguistic categories are notoriously context-dependent[7] [14], which makes it harder for computer programs to grasp the real interests of users. In information retrieval we tend to use keyterms to describe the content of documents, and sets of keyterms to describe the present interests of a given user at a particular time (e.g. a web search).

One of the advantages of using the knowledge contexts in our recommendation architecture is that the same keyterms can be differently associated in different information resources. Indeed, the distance functions of knowledge contexts allow us to regard these as connectionist memory systems [15] [16]. This way, the same set of keyterms describing the present interests (or search) of a user, is associated with different sets of other keyterms in distinct knowledge contexts. Thus, the interests of the user are also context-dependent when several information resources are at stake.

In this setting, the objective of a recommendation system that takes as input the present interest of a user, is to select and integrate the appropriate contexts, or perspectives, from the several ways the user interests are constructed in each information resource. We have developed an algorithm named *TalkMine* which implements the selective communication fabric necessary for this integration [14] [15] [16].

TalkMine uses a set structure named ***evidence set*** [12] [14], an extension of a fuzzy set [25], to model the interests of users defined as categories, or weighted sets of keyterms. Evidence sets are set structures which provide interval degrees of membership, weighted by the probability constraint of the Dempster-Shafer Theory of Evidence (DST) [19]. They are defined by two complementary dimensions: membership and belief. The first represents an interval (type-2) fuzzy degree of membership, and the second a degree of belief on that membership. Specifically, an ***evidence set*** A of X, is defined for all $x \in X$, by a membership function of the form:

$$A(x) \rightarrow (\mathcal{F}^x, m^x) \text{ E } \mathcal{B}[0, 1]$$

where $\mathcal{B}[0, 1]$ is the set of all possible bodies of evidence $(\mathcal{F}^x, m^x)$ on $\mathcal{I}$, the set of all subintervals of [0,1]. Such bodies of evidence are defined by a basic probability assignment m^x on $\mathcal{I}$, for every x in X.

Each interval of membership I_j^x represents the degree of importance of a particular element x of X (e.g. a keyterm) in category A (e.g. the interests of a user) ***according*** to a particular ***perspective*** (e.g. a particular database), defined by evidential weight $m^x(I_j^x)$. Thus, the membership of each element x of an evidence set A is defined by distinct intervals representing different perspectives.

The basic set operations of complementation, intersection, and union have been defined and establish a belief-constrained approximate reasoning theory of which fuzzy approximate reasoning and traditional set operations are special cases [13] [14]. Measures of uncertainty have also been defined for evidence sets. The total uncertainty of an evidence set A is defined by: $U(A) = (IF(A), IN(A), IS(A))$. The three indices of uncertainty, which vary between 1 and 0, IF (*fuzziness*), *IN* (***nonspecificity***), and ***IS*** (***conflict***) were introduced in [13]. IF is based on [23] [24] and Klir and Yuan [4] measure of fuzziness. IN is based on the Hartley measure, and ***IS*** on the Shannon entropy as extended by Klir (1993) into the DST framework.

5.2 Inferring User Interest in Different Knowledge Contexts

Fundamental to the *TalkMine* algorithm is the integration of information from different knowledge contexts into an evidence set, representing the category of topics (described by keywords) a user is interested at a particular time. Thus, the keywords the user employs to describe her interests or in a search, need to be "decoded" into appropriate keywords for each information resource: the perspective of each knowledge context.

The present interests of each user can be described by a set of keywords $P^u = \{k_1, \cdots, k_p\}$. Using these keywords and the keyword distance function (2) of the several knowledge contexts involved, we want to infer the interests of the user as "seen" from the several knowledge contexts involved.

Let us assume that r knowledge contexts R_t are involved in addition to one from the user herself. The set of keywords contained in all the participating knowledge contexts is denoted by $\mathcal{K}$. d_0 is the distance function of the knowledge context of the user, while $d_1 .. d_r$ are the distance functions **from** each of the other knowledge contexts. For each knowledge context R_t and each keyword k_u in the user's $P^u = \{k_1, \cdots, k_p\}$, a ***spreading interest fuzzy set*** $F_{t,u}$ is calculated using d_t:

$$F_{t,u}(k) = \max\left[e^{\left(-\alpha . d_t (k, k_u)^2\right)}, \varepsilon \right] \forall k \in R_t, t = 1 \ldots r, u = 1 \ldots p \quad \textbf{(8)}$$

This fuzzy set contains the keywords of R_t which are closer than E to k_u, according to an exponential function of d_t. $F_{t,u}$ spreads the interest of the user in k_u to keywords of R_t that are near according to d_t. The parameter α controls the spread of the exponential function. Because the knowledge context R_t contains a different d_t, each $F_{t,u}$ is also a different fuzzy set for the same k_u, possibly even containing keywords that do not exist in other knowledge contexts. There exist a total of $n = r.p$ spreading interest fuzzy sets $F_{t,u}$ given r knowledge context and p keyterms in the user's present interests.

5.3 The Linguistic "And/OR" Combination

Since each knowledge context produces a distinct fuzzy set, we need a procedure for integrating several of these fuzzy sets into an evidence set to obtain the integrated representation of user interests we desire. We have proposed such a procedure [16] based on Turksen's [21] combination of Fuzzy Sets into Interval Valued Fuzzy Sets (IVFS). Turksen proposed that fuzzy logic compositions could be represented by IVFS's given by the interval obtained from a composition's Disjunctive Normal Form (DNF) and Conjucntive Normal Form (CNF): [DNF, CNF]. We note that in fuzzy logic, for certain families of conjugate pairs of conjunctions and disjunctions, DNF $\subseteq$ CNF.

Using Turksen's approach, the union and intersection of two fuzzy sets F_1 and F_2 result in the two following IVFS, respectively:

$$IV^{\cup}(x) = \left[F_1(x) \underset{DNF}{\cup} F_2(x), F_1(x) \underset{CNF}{\cup} F_2(x)\right]$$
$$IV^{\cap}(x) = \left[F_1(x) \underset{DNF}{\cap} F_2(x), F_1(x) \underset{CNF}{\cap} F_2(x)\right] \quad \textbf{(9)}$$

where, $A \underset{CNF}{\cup} B = A \cup B$, $A \underset{DNF}{\cup} B = (A \cap B) \cup (A \cap \overline{B}) \cup (\overline{A} \cap B)$, $A \underset{CNF}{\cap} B = (A \cup B) \cap (A \cup \overline{B}) \cap (\overline{A} \cup B)$,and $A \underset{DNF}{\cap} B = A \cap B$, **for any two** fuzzy sets A and B.

Formulae(9) constitute a procedure for calculating the union and intersection IVFS from two fuzzy sets. $IV^{\cup}$ describes the linguistic expression "F_1 or F_2", while IF"describes "F_1 and F_2", – capturing both fuzziness and nonspecificity of the particular fuzzy logic operators employed, as Turksen suggested[16]. However, in common language, often "and" is used as an unspecified "and/or". In other words, what we mean by the statement "I am interested in x and y", is more correctly understood as an unspecified combination of "x and y" with "x or y". This is particularly relevant for recommendation systems where it is precisely this kind of statement from users that we wish to respond to.

One use of evidence sets is as representations of the integration of both $IV^{\cup}$ and $IV^{\cap}$ into a linguistic category that expresses this ambiguous "andor". To make this combination more general, assume that we possess an evidential weight m_1 and m_2 associated with each F_1 and F_2 respectively. These are probabilistic weights ($m_1 + m_2 = 1$) which represent the strength we associate with each fuzzy set being combined. The linguistic expression at stake now becomes "I am interested in x and y, but I value x more/less than y". To combine all this information into an evidence set we use the following procedure:

$$ES(x) = \left\{\left\langle IV^{\cup}(x), \min(m_1, m_2)\right\rangle, \left\langle IV^{\cap}(x), \max(m_1, m_2)\right\rangle\right\} \quad \textbf{(10)}$$

Because $IV^{\cup}$ is the less restrictive combination, obtained by applying the maximum operator to the original fuzzy sets F_1 and F_2, its evidential weight is acquired via the minimum operator of the evidential weights associated with F_1 and F_2. The reverse is true for $IV^{\cap}$. Thus, the evidence set obtained from (10) contains $IV^{\cup}$ with the lowest evidence, and $IV^{\cap}$ with the highest. Linguistically, it describes the ambiguity of the "andor" by giving the strongest belief weight to "and" and the weakest to **"or"**. It expresses: "I am interested in x and y to a higher degree, but I am also interested in x or y to a lower degree".

Finally, formula (10) can be easily generalized for a combination of n fuzzy sets F_i with probability constrained weights m_i:

$$ES(x) = \left\{\left\langle IV^{\cup}_{F_i/F_j}(x), \frac{\min(m_i, m_j)}{n-1}\right\rangle, \left\langle IV^{\cap}_{F_i/F_j}(x), \frac{\max(m_i, m_j)}{n-1}\right\rangle\right\} \textbf{(11)}$$

In ***TalkMine***, this formula is used to combine the n spreading interest Fuzzy Sets obtained from r knowledge context and p keyterms in P^u as described in section 5.2. The resulting evidence set $ES(k)$ defined on $\mathcal{K}$, represents the interests of the user inferred from spreading the initial interest set of keywords in the intervening knowledge contexts using their respective distance functions. The inferring process combines each $F_{t,u}$ with the "and/or" linguistic expression entailed by formula (11). Each $F_{t,u}$ contains the keywords related to keyword k_u in the knowledge context R_t, that is, the perspective of R_t on k_u. Thus, $ES(k)$ contains the "andor" combination of all the perspectives on each keyword k_u E $\{k_1, \cdots, k_p\}$ from each knowledge context R_t.

As **an** example, without loss of generality, consider that the initial interests of an user contain one single keyword k_1, and that the user is querying two distinct information resources R_1 and R_2. Two spreading interest fuzzy sets, F_1 and F_2, are generated using d_1 and d_2 respectively, with probabilistic weights m_1=v$_1$ and m_2=v$_2$., say, with $m_1 > m_2$ to indicate that the user **trusts** R_1 more than R_2. $ES(k)$ is easily obtained straight from formula **(10).** This evidence set contains the keywords related **to** k_1 in R_1 "andor" the keywords related to k_1 in R_2, taking into account the probabilistic weights attributed to R_1 and R_2. F_1 is the perspective of R_1 on k_1 and F_2 the perspective of R_2 on k_1.

6 DISTANCE FUNCTIONS IN RECOMMENDATION SYSTEMS

The evidence set obtained in Section **5.3** with formulas (10) and (11) is a first cut at detecting the interests of a user in a set of information resources, We can compute **a** more tuned interest set of keywords using an interactive conversation process between the user and the information resources being queried. Such conversation is an uncertainty reducing process based on Nakamura and Iwai's [10] IR system, which we extended to Evidence Sets [14] [16] with *TalkMine*.

TalkMine is then an algorithm for obtaining a representation of user interests in several information resources (including other users). It works by combining the perspectives of each information resources on the user interests into an evidence set, which is fine-tuned by an automated conversation process with the user's agent/browser [16] , The combination of perspectives is based on evidence sets, and uses the semi-metric distance functions described in this article. The importance of such semi-metric distance functions is thus described in this article, as they allow us to both analyze Document Networks for interests and trends (sec. 3), as well as offer an avenue to combine user interests in distinct information resources (sec. 5).

References

[1] Bollen, J. and Rocha, L. M., "An adaptive systems approach to the implementation and evaluation of digital library recommendation systems," *Research and Advanced Technology for Digital Libraries: 4th European Conference, ECDL 2000 Lectures Notes in Computer Science.* Springer-Verlag, 2000,356-359.

[2] Galvin ,F. and Shore ,S. D., "Distance Functions and Topologies," *American Mathematical Monthly,* vol. 98, no. 7, pp. 620-623, 1991.

[3] Herlocker, J. L., Konstan, J. A., Bouchers, A., and Riedl, J., "An algorithmic framework for performing collaborative filtering," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* **1999** ACM Press, 1999,230-237.

[4] Klir, G. J. and Yuan, B., *Fuzzy Sets and Fuzzy Logic: Theory and Applications* Upper Saddle River, NJ: Prentice Hall, 1995.

[5] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J., "GroupLens - Applying Collaborative Filtering to Usenet News," *Communications of the ACM,* vol. 40, no. 3, pp. 77-87, 1997.

[6] Krulwich, B. and Burkey, C., "Learning user information interests through extraction of semantically significant phrases," *Proceedings of the AAAI Spring Symposium on Machine Learning in Information Access* 1996.

[7] Lakoff, **G.,** *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind* University of Chicago Press, 1987,

[8] Luce, R., "Evolution and scientific literature: towards a decentralized adaptive web," *Nature: Web Debates,* no. May 10,2001,2001.

[9] Miyamoto, **S.,** *Fuzzy Sets in Information Retrieval and Cluster Analysis* Kluwer Academic Publishers, 1990.

[10] Nakamura, Kiyohiko and Iwai, Sosuke, "Representation of Analogical Inference by Fuzzy **Sets** and its Application to Information Retrieval System," *Fuzzy **Inf** and Decis Processes* 1982, 373-386.

[11] Newman, M. E., "The structure of scientific collaboration networks," *Proc. Natl. Acad. Sci. U.S.A,* vol. 98, no. 2, pp. 404-409, Jan. 2001.

[12] Rocha, L. M., "Cognitive categorization revisited: extending interval valued **fuzzy** sets as simulation tools concept combination," *Proc. of the **1994** Int. Conference of NAFIPS/IFIS/NASA* **IEEE** Press, 1994,400-404.

[13] Rocha, L. M., "Relative uncertainty and evidence sets: A constructivist framework," *International Journal of General Systems,* vol. 26, no. 1-2, pp. 35-61, 1997.

[14] Rocha, L. M., "Evidence sets: Modeling subjective categories," *International Journal of General Systems,* vol. 27, no. 6, pp. 457-494, 1999.

[15] Rocha, L. M., "Adaptive recommendation and open-ended semiosis," *Kybernetes,* vol. 30, no. 5-6, pp. 821-851,2001.

[16] Rocha, L. M., "TalkMine: a Soft Computing Approach to Adaptive Knowledge Recommendation," in Vincenzo Loia and Salvatore Sessa (eds.) ***Soft** Computing Agents: New Trends for Designing Autonomous Systems* Physica-Verlag, Springer, 2001, pp. 89-116.

[17] Rocha, L. M., "Semi-Metric Behavior in Document Networks and Adaptive Recommendation Systems," *Journal of Soft Computing,* 2002. In Press.

[18] Rocha, L. M. and Bollen, J., "Biologically motivated distributed designs for adaptive knowledge management," in Segel, L. A. and Cohen, I. (eds.) *Design Principles for the Immune System and other Distributed Autonomous Systems* Oxford University Press, 2001, pp. 305-334.

[19] Shafer, **G.,** *A Mathematical Theory of Evidence* Princeton University Press, 1976.

[20] Shore, S. D. and Sawyer, L. J., "Explicit Metrization," *Annals of the New York Academy of Sciences,* vol. **704** pp. 328-336, 1993.

[21] Turksen, I. B., "Non-specificity and interval-valued fuzzy sets," *Fuzzy Sets and Systems,* vol. 80, no. 1, pp. 87-100, May 1996.

[22] Watts, **D.,** *Small Worlds: The Dynamics of Networks between Order and Randomness* Princeton University Press, 1999.

[23] Yager ,R. R., "Measure of Fuzziness and Negation .1. Membership in the Unit Interval," *International Journal of General Systems,* vol. **5**, no. 4, pp. 221-229, 1979.

[24] Yager ,R. R., "On the Measure of Fuzziness and Negation .2. Lattices," *Information and Control,* vol. 44, no. 3, pp. 236-260, 1980.

[25] Zadeh, L. A., "Fuzzy Sets," *Information and Control,* vol. 8 pp. 338-353, 1965.