Automatic Conversation Driven by Uncertainty Reduction and Combination of Evidence for Recommendation Agents

Luis M. Rocha
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

Citation: Rocha, L.M. [2003]. "Automatic Conversation Driven by Uncertainty Reduction and Combination of Evidence for Recommendation Agents ". In: *Systematic Organization of Information in Fuzzy Systems*. NATO Science Series. P. Melo-Pinto, H.N. Teodorescu and T. Fukuda (Eds.) IOS Press, pp 249-265.

We present an adaptive recommendation system named *TalkMine*, which is based on the combination of Evidence from different sources. It establishes a mechanism for automated conversation between recommendation agents, in order to gather the interests of individual users of web sites and digital libraries. This conversation process is enabled by measuring the uncertainty content of knowledge structures used to store evidence from different sources. *TalkMine* also leads different databases or websites to learn new and adapt existing keywords to the categories recognized by its communities of users. *TalkMine* is currently being implemented for the research library of the Los Alamos National Laboratory under the *Active Recommendation Project* (http://arp.lanl.gov).

The process of identification of the interests of users relies on a process of combining several fuzzy sets into evidence sets, which models an ambiguous "and/or" linguistic expression. The interest of users is further fine-tuned by a human-machine conversation algorithm used for uncertainty reduction. Documents are retrieved according to the inferred user interests. Finally, the retrieval behavior of all users of the system is employed to adapt the knowledge bases of queried information resources. This adaptation allows information resources to respond well to the evolving expectations of users.

1 The Active Recommendation Project

The *Active Recommendation Project* (ARP), part of the Library Without Walls Project, at the Research Library of the Los Alamos National Laboratory 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. Typical databases are *SciSearch*® and *Biosis*®; the first contains articles from scientific journals from several fields collected by ISI (Institute for Scientific Indexing), while the second contains more biologically oriented publications. We do not manipulate directly the records stored in

these information resources, rather, we created a repository of XML (about 3 million) records which point us to documents stored in these databases [1].

1.1 Characterizing the Knowledge stored in an Information Resource

These matrices holding measures of closeness, formally, are proximity relations [2, 3] because they are reflexive and symmetric fuzzy relations. Their transitive closures are known as similarity relations (Ibid). The collection of this relational information, all the proximity relations as well as *A* and *C*, is an expression of the particular knowledge an information resource conveys to its community of users. Notice that distinct information resources typically share a very large set of keywords and records. 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 differently for distinct resources. Therefore, we refer to the relational information of each information resource as a *Knowledge Context* (More details in [4]).

In [1] we have discussed how these proximity relations are used in ARP. However, the ARP recommendation system described in this article (*TalkMine*) requires only the Keyword Semantic Proximity (*KSP*) matrix, obtained from *A* by the following formula:

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

The semantic proximity between two keywords, k_i and k_j, depends on the sets of records indexed by either keyword, and the intersection of these sets. $N(k_i)$ is the number of records keyword k_i indexes, and $N_{\cap}(k_i, k_j)$ the number of records both keywords index. This last quantity is the number of elements in the intersection of the sets of records that each keyword indexes. Thus, two keywords are near if they tend to index many of the same records. Table I presents the values of *KSP* for the 10 most common keywords in the ARP repository.

Table I: Keyword Semantic Proximity for 10 most frequent keywords

	cell	studi	system	express	protein	model	activ	human	rat	patient
cell	1.000	0.022	0.019	0.158	0.084	0.017	0.085	0.114	0.068	0.032
studi	0.022	1.000	0.029	0.013	0.017	0.028	0.020	0.020	0.020	0.037
system	0.019	0.029	1.000	0.020	0.017	0.046	0.022	0.014	0.021	0.014
express	0.158	0.013	0.020	1.000	0.126	0.011	0.071	0.103	0.078	0.020
protein	0.084	0.017	0.017	0.126	1.000	0.013	0.070	0.061	0.041	0.014
model	0.017	0.028	0.046	0.011	0.013	1.000	0.016	0.016	0.026	0.005
activ	0.085	0.020	0.022	0.071	0.070	0.016	1.000	0.058	0.053	0.021
human	0.114	0.020	0.014	0.103	0.061	0.016	0.058	1.000	0.029	0.021
rat	0.068	0.020	0.021	0.078	0.041	0.026	0.053	0.029	1.000	0.008
patient	0.032	0.037	0.014	0.020	0.014	0.005	0.021	0.021	0.008	1.000

From the inverse of *KSP* we obtain a distance function between keywords:

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

d is a distance function because it is a nonnegative, symmetric real-valued function such that $d(k, k) = 0$. It 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. Such measures of distance are referred to as semi-metrics [5].

1.2 Characterizing Users

Users interact with information resources by retrieving records. We use their retrieval behavior to adapt the respective knowledge contexts of these resources (stored in the proximity relations). But before discussing this interaction, we need to characterize and define the capabilities of users: our agents. The following capabilities are implemented in enhanced "browsers" distributed to users.

1. ***Present interests*** described by a set of keywords $\{k_1, \cdots, k_p\}$.
2. ***History of Information Retrieval (IR)***. This history is also organized as a knowledge context as described in 2.1, containing pointers to the records the user has previously accessed, the keywords associated with them, as well as the structure of this set of records. This way, we treat users themselves as information resources with their own specific knowledge contexts defined by their own proximity information.
3. ***Communication Protocol***. Users need a 2-way means to communicate with other information resources in order to retrieve relevant information, and to send signals leading to adaptation in all parties involved in the exchange.

Regarding point 2, the history of IR, 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, each enhanced browser allows users to define different "*personalities*", each one with its distinct history of IR defined by independent knowledge contexts with distinct proximity data (see Figure 1).

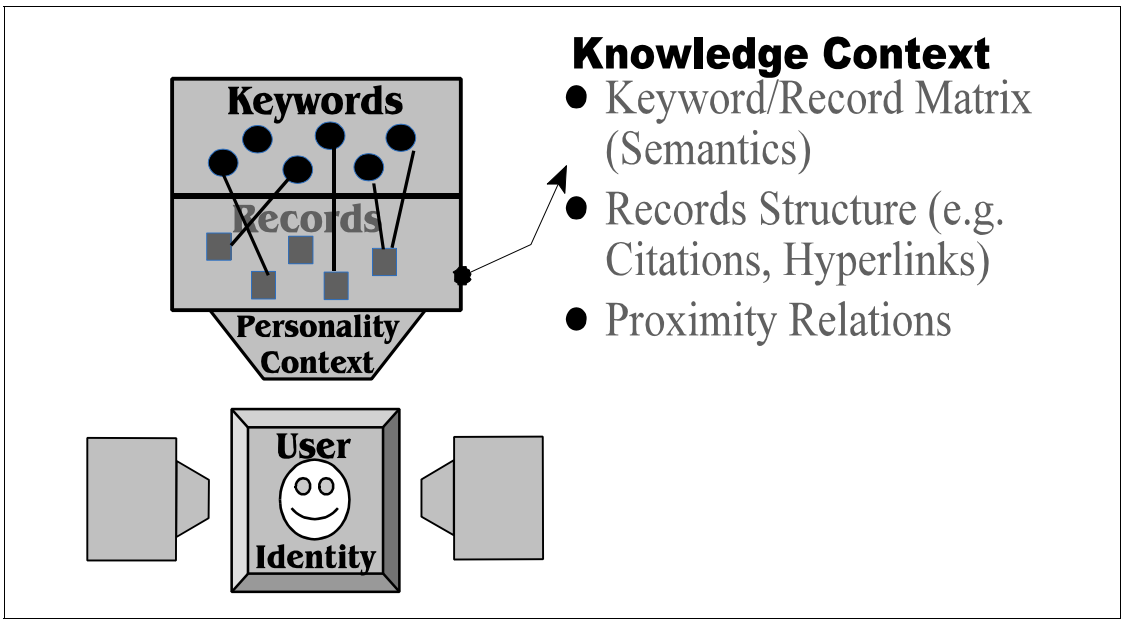

Figure 1: Each user can store different personalities in enhanced browsers. Each personality is stored as a knowledge context created from previous history of IR. The actual identity of the user can remain private.

Because the user history of IR is stored in personal browsers, information resources do not store user profiles. Furthermore, all the collective behavior algorithms used in ARP do not require the identity of users. When users communicate (3) with information resources, what needs to be exchanged is their present interests or query (1), and the relevant proximity data from their own knowledge context (2). In other words, users make a query, and then share the relevant knowledge they have accumulated about their query, their "world-view" or context, from a particular personality, without trading their identity. Next, the recommendation algorithms integrate the user's knowledge context with those of the queried information resources (possibly other users), resulting in appropriate recommendations. Indeed, the algorithms we use define a communication protocol between knowledge contexts, which can be very large databases, web sites, or other users. Thus, the overall architecture of the recommendation systems we use in ARP is highly distributed between information resources and all the users and their browsing personalities (see Figure 2).

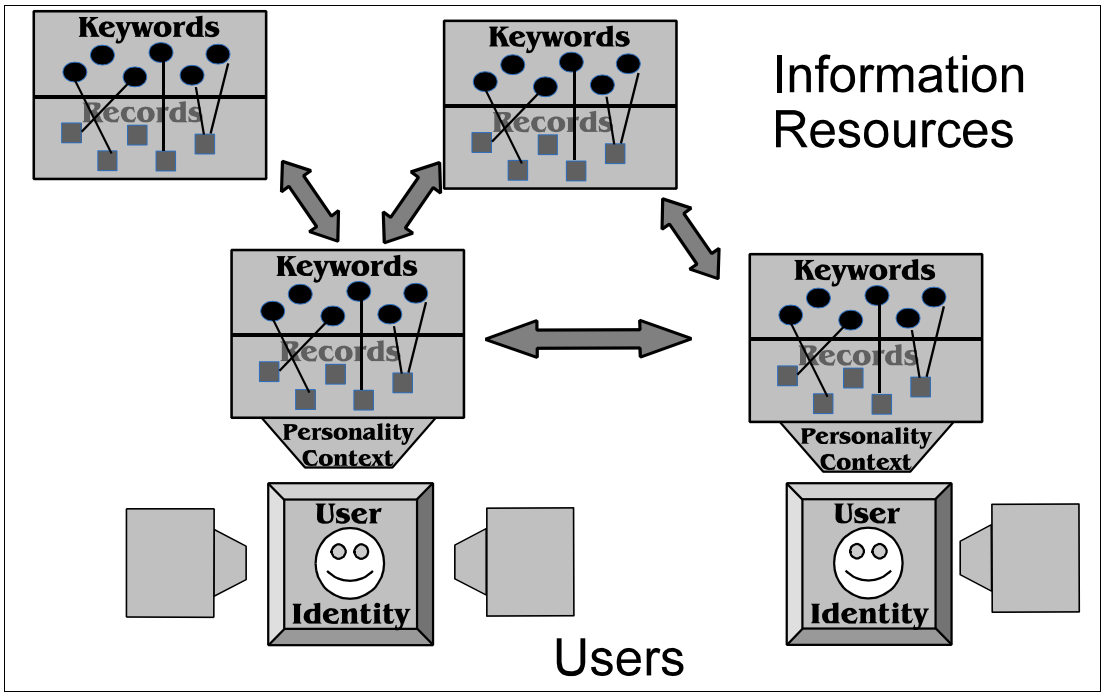

Figure 2: The algorithms we use in ARP define a distributed architecture based on communication between knowledge contexts from information resources and users alike.

The collective behavior of all users is also aggregated to adapt the knowledge contexts of all intervening information resources and users alike. This open-ended learning process [6] is enabled by the *TalkMine* recommendation system described below.

2 Categories and Distributed Memory

2.1 A Model of Categorization from Distributed Artificial Intelligence

TalkMine is a recommendation system based on a model of linguistic categories [7], which are created from conversation between users and information resources and used to re-combine knowledge as well as adapt it to users. The model of categorization used by *TalkMine* is described in detail in [6,7,8]. Basically, as also suggested by [9], categories are seen as representations of highly transient, context-dependent knowledge arrangements, and not as model of information storage in the brain. In this sense, in human cognition, categories are seen as linguistic constructs used to store temporary associations built up from the integration of knowledge from several neural sub-networks. The categorization process, driven by language and conversation, serves to bridge together several distributed neural networks, associating tokens of knowledge that would not otherwise be associated in the individual networks. Thus, categorization is the chief mechanism to achieve knowledge recombination in distributed networks leading to the production of new knowledge [6, 7].

TalkMine applies such a model of categorization of distributed neural networks driven by language and conversation to recommendation systems. Instead of neural networks, knowledge is stored in information resources, from which we construct the knowledge contexts with respective proximity relations described in section 1. *TalkMine* is used as a conversation

protocol to categorize the interests of users according to the knowledge stored in information resources, thus producing appropriate recommendations and adaptation signals.

2.2 Distributed Memory is Stored in Knowledge Contexts

A knowledge context of an information resource (section 1.1) is not a connectionist structure in a strong sense since keywords and records are not distributed as they can be identified in specific nodes of the network [10]. However, the same keyword indexes many records, the same record is indexed by many keywords, and the same record is typically engaged in a citation (or hyperlink) relation with many other records. Losing or adding a few records or keywords does not significantly change the derived semantic and structural proximity relations (section 1) of a large network. In this sense, the knowledge conveyed by such proximity relations is distributed over the entire network of records and keywords in a highly redundant manner, as required of sparse distributed memory models [11]. Furthermore, Clark [9] proposed that connectionist memory devices work by producing metrics that relate the knowledge they store. As discussed in section 1, the distance functions obtained from proximity relations are semi-metrics, which follow all of Clark's requirements [6]. Therefore, we can regard a knowledge context effectively as a distributed memory bank. Below we discuss how such distributed knowledge adapts to communities of users (the environment) with Hebbian type learning.

In the *TalkMine* system we use the KSP relation (formula (1)) from knowledge contexts. It conveys the knowledge stored in an information resource in terms of a measure of proximity among keywords. This proximity relation is unique to each information resource, reflecting the semantic relationships of the records stored in the latter, which in turn echo the knowledge of its community of users and authors. *TalkMine* is a content-based recommendation system because it uses a keywords proximity relation. Next we describe how it is also collaborative by integrating the behavior of users. A related structural algorithm, also being developed in ARP, is described in [1].

3 Evidence Sets: Capturing the Linguistic "And/Or" in Queries

3.1 Evidence Sets Model Categories

TalkMine uses a set structure named *evidence set* [7, 8, 12, 13], an extension of a fuzzy set [14] , to model of linguistic categories. The extension of fuzzy sets is based on the Dempster-Shafer Theory of Evidence (DST) [15], which is defined in terms of a set function m: $\mathcal{P}(X) \rightarrow [0,1]$, referred to as a *basic probability assignment*, such that $m(\varnothing) = 0$ and $\sum_{A \subseteq X} m(A) = 1$. $\mathcal{P}(X)$ denotes the power set of X, and A any subset of X. The value $m(A)$ denotes the proportion of all available evidence which supports the claim that $A \in \mathcal{P}(X)$ contains the actual value of a variable x. DST is based on a pair of nonadditive measures: *belief* (*Bel*) and *plausibility* (*Pl*) uniquely obtained from m. Given a basic probability assignment m, *Bel* and *Pl* are determined for all $A \in \mathcal{P}(X)$ by the equations:

$$Bel(A) = \sum_{B \subseteq A} m(B), \quad Pl(A) = \sum_{B \cap A \neq \varnothing} m(B)$$

the expressions above imply that belief and plausibility are dual measures related by: $Pl(A) = 1 - Bel(A^c)$, for all $A \in \mathcal{P}(X)$, where A^c represents the complement of A in X. It is also true that $Bel(A) \leq Pl(A)$ for all $A \in \mathcal{P}(X)$. Notice that "$m(A)$ measures the belief one commits exactly to A, not the total belief that one commits to A." [15, page 38] $Bel(A)$, the total belief committed to A, is instead given by the sum of all the values of m for all subsets of A.

Any set $A \in \mathcal{P}(X)$ with $m(A) > 0$ is called a *focal element*. A *body of evidence* is defined by the pair $(\mathcal{F}, m)$, where $\mathcal{F}$ represents the set of all focal elements in X, and m the associated basic probability assignment. The set of all bodies of evidence is denoted by $\mathcal{B}(X)$.

An *evidence set A* of X, is defined for all $x \in X$, by a membership function of the form:

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

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. Thus, evidence sets are set structures which provide interval degrees of membership, weighted by the probability constraint of DST. 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 subjective degree of belief on that membership (see Figure 3).

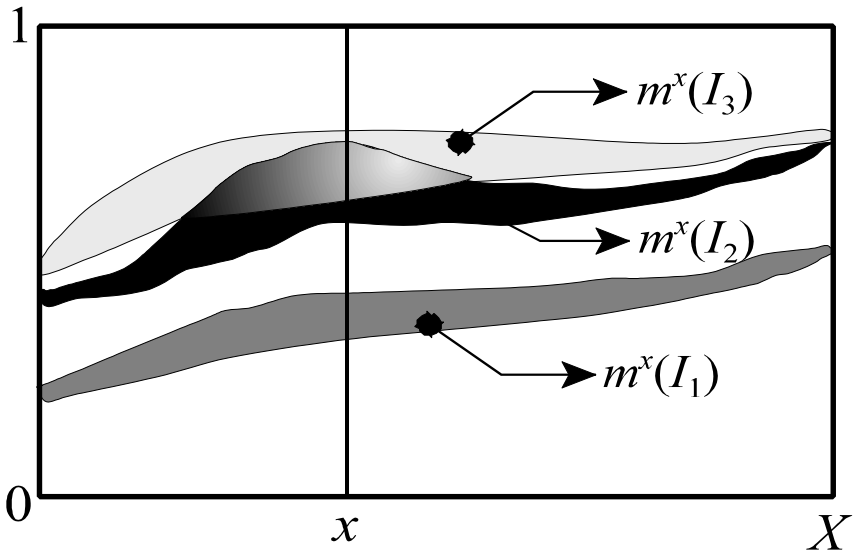

Figure 3: Evidence Set with 3 focal elements for each x.

Each interval of membership $\boldsymbol{I}_j^x$, with its correspondent evidential weight $m^x(\boldsymbol{I}_j^x)$, represents the degree of importance of a particular element x of X in category A *according* to a particular *perspective*. Thus, the membership of each element x of an evidence set A is defined by distinct intervals representing different, possibly conflicting, perspectives. This way, categories are modeled not only as sets of elements with a membership degree (or prototypicality [7]), but as sets of elements which may possess different interval membership degrees for different contexts or perspectives on the category.

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 [7, 8]. Intersection (Union) is based

on the minimum (maximum) operator for the limits of each of the intervals of membership of an evidence set. For the purposes of this article, the details of these operations are not required, please consult [7] for more details.

3.2 The Uncertainty Content of Evidence Sets

Evidence sets are set structures which provide interval degrees of membership, weighted by the probability constraint of DST. Interval Valued Fuzzy Sets (IVFS), fuzzy sets, and crisp sets are all special cases of evidence sets. The membership of an element x in a crisp set is perfectly certain: the element is either a member of the set or not. The membership of an element x in fuzzy set is defined as degree value in the unit interval; this means that the membership is *fuzzy* because the element is a member of the set with degree $A(x)$, and simultaneously, is also not a member with complementary degree $1\text{-}A(x)$. The membership of an element x in an IVFS is defined as an interval I contained in the unit interval; this means that the membership is both fuzzy and *nonspecific* [12, 16], because the element is a member of the set with a nonspecific degree that can vary in the interval I. Finally, membership of an element x in an evidence set is defined as a set of intervals constrained by a probability restriction; this means that the membership is fuzzy, nonspecific, and *conflicting*, since the element is a member of the set with several degrees that vary in each interval with some probability.

To capture the uncertainty content of evidence sets, the uncertainty measures of [17] were extended from finite to infinite domains [13]. The total uncertainty, $\boldsymbol{U}$, of an evidence set A was 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 [8, 13], where it was also proven that *IN* and *IS* possess good axiomatic properties wanted of information measures. *IF* is based on [18, 19] and [2] measure of fuzziness. *IN* is based on the Hartley measure [13], and *IS* on the Shannon entropy as extended by [17] into the DST framework. For the purposes of this article, all we need to know is that these measures vary in the unit interval, for full details see [13].

3.3 Obtaining an Evidence Set from Fuzzy Sets: The Linguistic "And/Or

Fundamental to the *TalkMine* algorithm is the integration of information from different sources into an evidence set, representing the category of topics (described by keywords) a user is interested at a particular time. In particular, as described below, these sources of information contribute information as fuzzy sets. This way, we need a procedure for integrating several fuzzy sets into an evidence set.

Turksen [20] proposed a means to integrate fuzzy sets into IVFS (or type-2 fuzzy sets). He later proposed that every time two fuzzy sets are combined, the uncertainty content of the resulting structure should be of a higher order, namely, the fuzziness of two fuzzy sets should be combined into the fuzziness and nonspecificity of an IVFS [16]. Turksen's Fuzzy Set combination is based on the separation of the disjunctive and conjunctive normal forms of logic compositions in fuzzy logic. A disjunctive normal form (DNF) is formed with the disjunction of some of the four primary conjunctions, and the conjunctive normal form (CNF) is formed with the conjunction of some of the four primary disjunctions, respectively: $A \cap B, A \cap \overline{B}, \overline{A} \cap B, \overline{A} \cap \overline{B}$ and $A \cup B, A \cup \overline{B}, \overline{A} \cup B, \overline{A} \cup \overline{B}$. In two-valued logic the CNF and DNF of a logic composition are equivalent: CNF = DNF. Turksen [20] observed that in fuzzy logic, for certain families of conjugate pairs of conjunctions and disjunctions, we have instead DNF $\subseteq$ CNF for some of the fuzzy logic connectives. He proposed that fuzzy logic compositions could be represented by IVFS's given by the interval [DNF, CNF] of the fuzzy set connective chosen [20].

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

$$IV^{\cup}(x) = \left[F_1(x) \underset{DNF}{\cup} F_2(x), F_1(x) \underset{CNF}{\cup} F_2(x) \right]$$
$$IV^{\cap}(x) = \left[F_1(x) \underset{DNF}{\cap} F_2(x), F_1(x) \underset{CNF}{\cap} F_2(x) \right] \quad \textbf{(3)}$$

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, with union and intersection operations chosen from the families of t-norms and t-conorms following the appropriate axiomatic requirements [2]. In *TalkMine* only the traditional maximum and minimum operators for union and intersection, respectively, are used. Clearly, all other t-norms and t-conorms would also work.

The intervals of membership obtained from the combination of two fuzzy sets can be interpreted as capturing the intrinsic nonspecificity of the combination of fuzzy sets with fuzzy set operators. Due to the introduction of fuzziness, the DNF and CNF do not always coincide. This lack of coincidence reflects precisely the nonspecificity inherent in fuzzy set theory: because we can arrive at different results depending on which normal form we choose, the combination of fuzzy sets is ambiguous. Turksen [16] suggested that this ambiguity should be

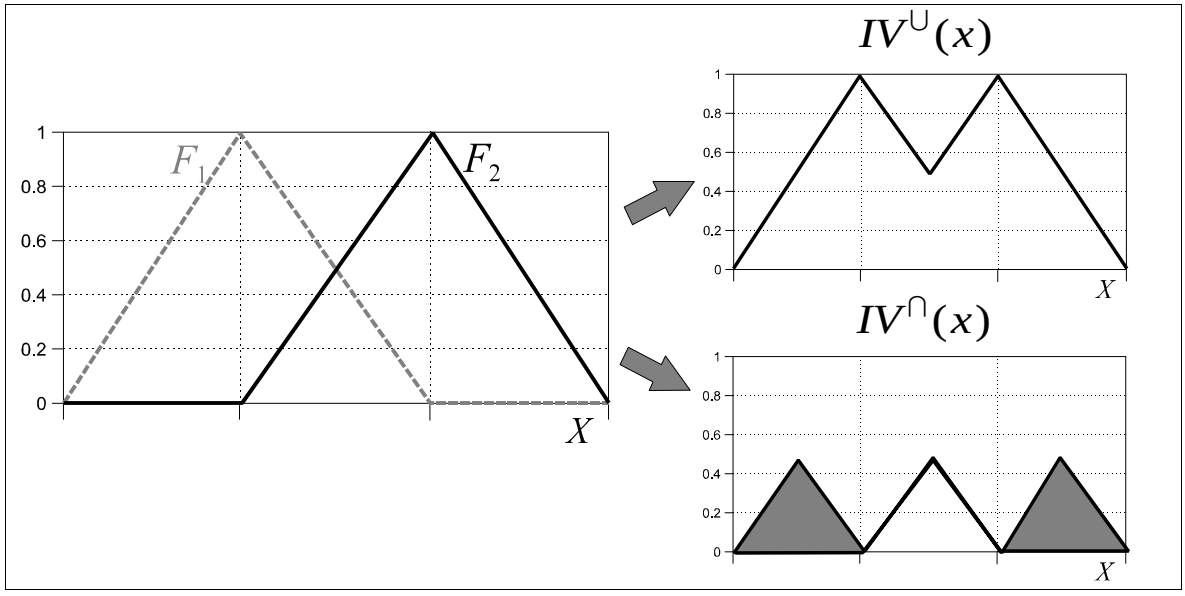

Figure 4: Combination of two fuzzy sets F_1 and F_2 into two IVFS according to formulae 3. The union IVFS, $IV^{\cup}$ is a fuzzy set since DNF and CNF coincide, which does not happen for $IV^{\cap}$.

treated as nonspecificity and captured by intervals of membership. In this sense, fuzziness "breeds" nonspecificity. Figure 4 depicts the construction of two IVFS from two fuzzy sets F_1 and F_2 according to the procedure described by formulae (3).

Formulae (3) constitute a procedure for calculating the union and intersection IVFS from two fuzzy sets, which in logic terms refer to the "Or" and "And" operators. Thus, $IV^{\cup}$ describes the linguistic expression "F_1 or F_2", while $IV^{\cap}$ describes "F_1 and F_2", – capturing both fuzziness and nonspecificity of the particular fuzzy logic operators employed. 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", can actually be seen 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 "and/or". 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:

$$ES(x) = \left\{ \left\langle IV^{\cup}(x), \min\left(m_1, m_2\right) \right\rangle, \left\langle IV^{\cap}(x), \max\left(m_1, m_2\right) \right\rangle \right\} \quad \textbf{(4)}$$

Because $IV^{\cup}$ is the less restrictive combination, obtained by applying the maximum operator, or suitable t-norm 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 (4) contains $IV^{\cup}$ with the lowest evidence, and $IV^{\cap}$ with the highest. Linguistically, it describes the ambiguity of the "and/or" 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". This introduces the third kind of uncertainty: conflict. Indeed, the ambiguity of "and/or" rests on the conflict between the interest in "and" and the interest in "or". This evidence set captures the three forms of uncertainty discussed in Section 2.3: fuzziness of the original fuzzy sets F_1 and F_2, nonspecificity of $IV^{\cup}$ and $IV^{\cap}$, and conflict between these two as they are included in the same evidence set with distinct evidential weights. Figure 5 depicts an example of the evidence set

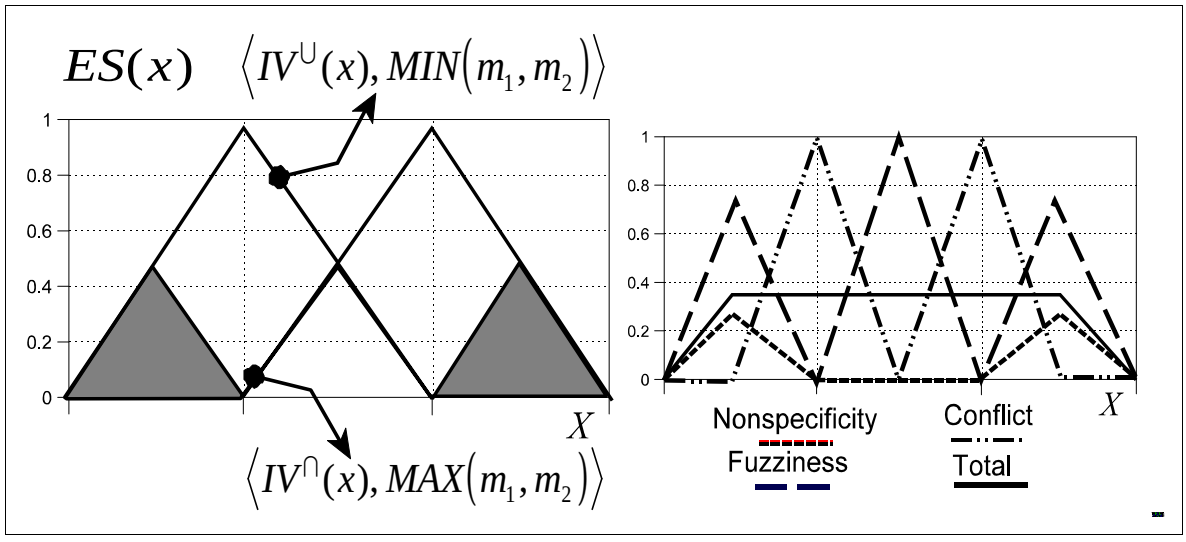

Figure 5: Evidence Set obtained from F_1 and F_2 and respective uncertainty content

obtained from F_1 and F_2, as well as its uncertainty content (fuzziness. Nonspecificity, and conflict).

Finally, formula (4) can be easily generalized for a combination of n fuzzy sets F_i with probability constrained weights m_i:

$$ES(x) = \left\{ \left\langle IV^{\cup}_{F_i/F_j}(x), \frac{\min\left(m_i, m_j\right)}{n-1} \right\rangle, \left\langle IV^{\cap}_{F_i/F_j}(x), \frac{\max\left(m_i, m_j\right)}{n-1} \right\rangle \right\} \quad \textbf{(5)}$$

This procedure can be used to combine evidence in the form of fuzzy sets from *n* weighted sources. It produces intervals obtained from the combination of each pair of fuzzy sets with a union and an intersection operator. Intersection is given the highest weight. The evidence set obtained is the ambiguous, common language, "and/or" for *n* items.

4 *TalkMine*: Integrating Several Sources of Knowledge via Conversation

4.1 Inferring User Interest

The act of recommending appropriate documents to a particular user needs to be based on the integration of information from the user (with her history of retrieval) and from the several information resources being queried. With *TalkMine* in particular, we want to retrieve relevant documents from several information resources with different keyword indexing. Thus, the keywords the user employs in her search, need to be "decoded" into appropriate keywords for each information resource. Indeed, the goal of *TalkMine* is to project the user interests into the distinct knowledge contexts of each information resource, creating a representation of these interests that can capture the perspective of each one of these contexts.

Evidence Sets were precisely defined to model categories (knowledge representations) which can capture different perspectives. As described in Section 1.2, the present interests of each user are described by a set of keywords $\{k_1, \cdots, k_p\}$. Using these keywords and the keyword distance function (2) of the several knowledge contexts involved (one from the user and one from each information resource being queried), the interests of the user, "seen" from the perspectives of the several information resources, can be inferred as an evidence category using (5).

Let us assume that *r* information resources R_t are involved in addition to the user herself. The set of keywords contained in all the participating information resources is denoted by $\mathcal{K}$. As described in Section 1, each information resource is characterized as a knowledge context containing a KSP relation (1) among keywords from which a distance function *d* is obtained (cfr. (2)). d_0 is the distance function of the knowledge context of the user, while $d_1 \ldots d_r$ are the distance functions from the knowledge contexts of each of the information resources.

4.1.1 Spreading Interest Fuzzy Sets

For each information resource R_t and each keyword k_u in the user's present interests $\{k_1, \cdots, k_p\}$, a *spreading interest fuzzy set* $F_{t,u}$ is calculated using d_t:

$$F_{t,u}(k) = \max\left[e^{\left(-a \cdot d_t(k,k_u)^2\right)}, e \right] \forall k \in R_t, t = 1 \ldots r, u = 1 \ldots p \qquad \textbf{(6)}$$

This fuzzy set contains the keywords of R_t which are closer than ϵ 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. $F_{t,u}$ represents the set of keywords of R_t which are near or very related to keyword k_u. Because the knowledge context of each R_t contains a different d_t, each $F_{t,u}$ will also be a different fuzzy set for the same k_u, possibly even containing keywords that do not exist in other information resources. There exist a total of $n = r.p$ spreading interest fuzzy sets $F_{t,u}$. Figure 6 depicts a generic $F_{t,u}$.

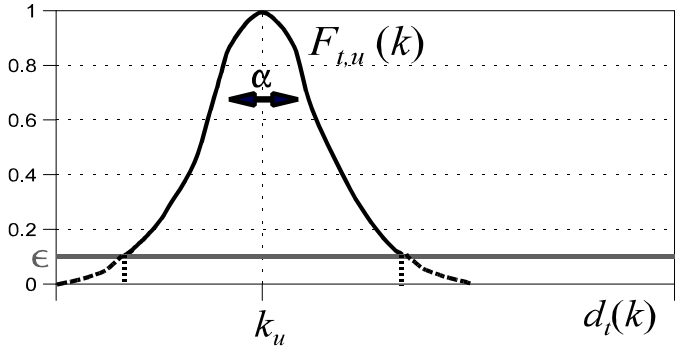

Figure 6: The exponential membership function of $F_{t,u}(k)$ spreads the interest of a user on keyword k_u to close keywords according to distance function $d_t(k)$ for each information resource R_t.

4.1.2 Combining the Perspectives of Different Knowledge Contexts on the User Interest

Assume now that the present interests of the user $\{k_1, \cdots, k_p\}$ are probabilistically constrained, that is, there is a probability weight associated with each keyword: $\mu_1,..., \mu_p$, such that $\mu_1 + ... + \mu_p = 1$. Assume further that the intervening r information resources R_t are also probabilistically constrained with weights: $\nu_1,..., \nu_r$, such that $\nu_1 + ... + \nu_r = 1$. Thus, the probabilistic weight of each spreading interest fuzzy set $F_i = F_{t,u}$, where $i = (t\text{-}1)p + u$, is $m_i = \nu_t.\mu_u$.

To combine the n fuzzy sets F_i and respective probabilistic weights m_i, formula (5) is employed. This results in an evidence set $ES(k)$ defined on $\mathcal{K}$, which represents the interests of the user inferred from spreading the initial interest set of keywords in the knowledge contexts of the intervening information resources. The inferring process combines each $F_{t,u}$ with the "and/or" linguistic expression entailed by formula (5). Each $F_{t,u}$ contains the keywords related to keyword k_u in the knowledge context of information resource R_t, that is, the perspective of R_t on k_u. Thus, $ES(k)$ contains the "and/or" combination of all the perspectives on each keyword $k_u \in \{k_1, \cdots, k_p\}$ from each knowledge context associated with all information resources 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 $m_1=\nu_1$ and $m_2=\nu_2$. $ES(k)$ is easily obtained straight from formula (4). This evidence set contains the keywords related to k_1 in R_1 "and/or" 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.

4.2 Reducing the Uncertainty of User Interests via Conversation

The evidence set obtained in Section 4.1 with formulas (5) and (6) is a first cut at detecting the interests of a user in a set of information resources. But we can compute a more accurate 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 [21] IR system, and extended to Evidence Sets by Rocha [6, 7].

In addition to the evidence set $ES(k)$ constructed in Section 4.1, a fuzzy set $F_0(k)$ is constructed to contain the keywords of the knowledge context R_0 of the user which are close to the initial interest set $\{k_1, \cdots, k_p\}$ according to distance function d_0. As discussed in Section 1, the user's history of IR is itself characterized as a knowledge context R_0 with its own KSP relation and derived distance function d_0. $F_0(k)$ is given by:

$$F_0(k) = \bigcup_{u=1}^{p} F_{0,u}(k) \quad \textbf{(7)}$$

where $F_{0,u}(k)$ is calculated using formula (6). $F_0(k)$ represents the perspective of the user, from her history of retrieval, on all keywords $\{k_1, \cdots, k_p\}$. Given $ES(k)$ and $F_0(k)$, for a default value of $\alpha=\alpha_0$, the algorithm for *TalkMine* is as follows:

1. Calculate the uncertainty of $ES(k)$ in its forms of fuzziness, nonspecificity, and conflict (see Section 3.2). If total uncertainty is below a pre-defined small value the process stops, otherwise continue to 2.
2. The most uncertain keyword $k_j \in ES(k)$ is selected.
3. If $k_j \in R_0$, then goto 4 (AUTOMATIC), else goto 6 (ASK).
4. If $F_0(k_j) > 0.5+\delta$, then goto 7 (YES).
5. If $F_0(k_j) \leq 0.5-\delta$, then goto 8 (NO), else goto 6 (ASK).
6. ASK user if she is interested in keyword k_j. If answer is yes goto 7 (YES), else goto 8 (NO).
7. An evidence set YES(k) is calculated using the procedure of section 4.1 for a single keyword k_j and all r information resources R_t. The spread of the exponential functions is controlled with parameter α so that answers to previous keywords k_j are preserved. $ES(k)$ is then recalculated as the evidence set union of YES(k) and $ES(k)$ itself.
8. An evidence set NO(k) is calculated as the complement of YES(k) used in 7. $ES(k)$ is then recalculated as the evidence set intersection of NO(k) and $ES(k)$ itself.
9. Goto 1.

The parameter δ controls how much participation is required from the user in this interactive process, and how much is automatically deduced from her own knowledge context used to produce $F_0(k)$. $\delta \in [0, 0.5]$; for $\delta = 0$, all interaction between user and information resources is mostly automatic, as answers are obtained from $F_0(k)$, except when $k_j \notin R_0$; for $\delta = 0.5$, all interaction between user and information resources requires explicit answers from the user. If the user chooses not to reply to a question, the answer is taken as NO. Thus, δ allows the user to choose how automatic the question-answering process of *TalkMine* is.

Regarding the change of spread employed in steps 7 and 8 for the construction of the YES(k) and NO(k) evidence sets. A list of the keywords the user (or $F_0(k)$ automatically) has responded YES or NO to is kept. The membership value of these keywords in the final $ES(k)$ produced must be 1 or 0, respectively. Thus, the union and intersections of $ES(k)$ with YES(k) and NO(k) in 7 and 8, must be defined in a such a way as to preserve these values. If the spread obtained with α_0 would alter the desired values, then a new α is employed in formula (6) so that the original values are preserved $\pm\epsilon$. Because of this change of spreading inference of the YES(k) and NO(k) evidence sets, the sequence of keywords selected by the question-answering process in step 2 affects the final $ES(k)$. That is, the selection of a different keyword may result in a different $ES(k)$.

The final $ES(k)$ obtained with this algorithm is a much less uncertain representation of user interests as projected on the knowledge contexts of the information resources queried, than the

initial evidence set obtained in Section 4.1. The conversation algorithm lets the user reduce the uncertainty from the all the perspectives initially available. The initial evidence set produced in Section 4.1 includes all associated keywords in several information resources. The conversation algorithm allows the user and her knowledge context to select only the relevant ones. Thus, the final $ES(k)$ can be seen as a low-uncertainty linguistic category containing those perspectives on the user's initial interest (obtained from the participating information resources) which are relevant to the user and her knowledge context [6, 7].

Notice that this category is not stored in any location in the intervening knowledge contexts. It is temporarily constructed by integration of knowledge from several information resources and the interests of the user expressed in the interactive conversational process. Such a category is therefore a temporary container of knowledge integrated from and relevant for the user and the collection of information resources. Thus, this algorithm implements many of the, temporary, "on the hoof" [9] category constructions as discussed in [6].

4.3 Recommending Documents

After construction of the final $ES(k)$, *TalkMine* must return to the user documents relevant to this category. Notice that every document n_i defines a crisp subset whose elements are all the keywords $k \in \mathsf{K}$ which index n_i in all the constituent information resources. The similarity between this crisp subset and $ES(k)$ is a measure of the relevance of the document to the interests of the user as described by $ES(k)$. This similarity is defined by different ways of calculating the subsethood [22] of one set in the other. Details of the actual operations used are presented in [7]. High values of these similarity measures will result on the system recommending only those documents highly related to the learned category $ES(k)$.

4.4 Adapting Knowledge Contexts

From the many $ES(k)$ obtained from the set of users of information resources, we collect information used to adapt the KSP and semantic distance of the respective knowledge contexts. The scheme used to implement this adaptation is very simple: the more certain keywords are associated with each other, by often being simultaneously included with a high degree of membership in the final $ES(k)$, the more the semantic distance between them is reduced. Conversely, if certain keywords are not frequently associated with one another, the distance between them is increased. An easy way to achieve this is to have the values of $N(k_i)$, $N(k_j)$ and $N_{\cap}(k_i, k_j)$ as defined in formula (1), adaptively altered for each of the constituent r information resources R_t. After $ES(k)$ is constructed and approximated by a fuzzy set $A(x)$, these values are changed according to:

$$N^t(k_i) = N^t(k_i) + w.A(k_i),\ t = 1\ldots r,\ k_i \in R_0 \cup R_1 \cup \ldots \cup R_r \qquad \textbf{(7)}$$

and

$$N^t_{\cap}\left(k_i, k_j\right) = N^t_{\cap}\left(k_i, k_j\right) + w.\min\left[A\left(k_i\right), A\left(k_j\right)\right],\ t = 1\ldots r,\ k_i, k_j \in R_0 \cup R_1 \cup \ldots \cup R_r \qquad \textbf{(8)}$$

where w is the weight ascribed to the individual contribution of each user. The adaptation entailed by (7) and (8) leads the semantic distance of the knowledge contexts involved, to increasingly match the expectations of the community of users with whom they interact. Furthermore, when keywords with high membership in $ES(k)$ are not present in one of the information resources queried, they are added to it with document counts given by formulas (7) and (8). If the simultaneous association of the same keywords keeps occurring, then an

information resource that did not previously contain a certain keyword, will have its presence progressively strengthened, even though such keyword does not index any documents stored in this information resource.

5 Collective Evolution of Knowledge with Soft Computing

TalkMine models the construction of linguistic categories. Such "on the hoof" construction of categories triggered by interaction with users, allows several unrelated information resources to be searched simultaneously, temporarily generating categories that are not really stored in any location. The short-term categories bridge together a number of possibly highly unrelated contexts, which in turn creates new associations in the individual information resources that would never occur within their own limited context.

Consider the following example. Two distinct information resources (databases) are searched using *TalkMine*. One database contains the documents (books, articles, etc) of an institution devoted to the study of computational complex adaptive systems (e.g. the library of the *Santa Fe Institute*), and the other the documents of a Philosophy of Biology department. I am interested in the keywords GENETICS and NATURAL SELECTION. If I were to conduct this search a number of times, due to my own interests, the learned category obtained would certainly contain other keywords such as ADAPTIVE COMPUTATION, GENETIC ALGORITHMS, etc. Let me assume that the keyword GENETIC ALGORITHMS does not initially exist in the Philosophy of Biology library. After I conduct this search a number of times, the keyword GENETIC ALGORITHMS is created in this library, even though it does not contain any documents about this topic. However, with my continuing to perform this search over and over again, the keyword GENETIC ALGORITHMS becomes highly associated with GENETICS and NATURAL SELECTION, introducing a new perspective of these keywords. From this point on, users of the Philosophy of Biology library, by entering the keyword GENETIC ALGORITHMS would have their own data retrieval system point them to other information resources such as the library of the *Santa Fe Institute* or/and output documents ranging from "The Origin of Species" to treatises on Neo-Darwinism – at which point they would probably bar me from using their networked database!

Given a large number of interacting knowledge contexts from information resources and users (see Figure 2), *TalkMine* is able to create new categories that are not stored in any one location, changing and adapting such knowledge contexts in an open-ended fashion. Open-endedness does not mean that *TalkMine* is able to discern all knowledge negotiated by its user environment, but that it is able to permutate all the semantic information (KSP and *d* described in Section 1) of the intervening knowledge contexts in an essentially open-ended manner. The categories constructed by *TalkMine* function as a system of collective linguistic recombination of distributed memory banks, capable of transferring knowledge across different contexts and thus creating new knowledge. In this way, *TalkMine* can adapt to an evolving environment and generate new knowledge given a sufficiently diverse set of information resources and users. Readers are encouraged to track the development of this system at http://arp.lanl.gov.

TalkMine is a collective recommendation algorithm because it uses the behavior of its users to adapt the knowledge stored in information resources. Each time a user queries several information resources, the category constructed by *TalkMine* is used to adapt those (cfr. Section 4). In this sense, the knowledge contexts (cfr. Section 1) of the intervening information resources becomes itself a representation of the knowledge of the user community. A discussion of this process is left for future work.

TalkMine is a soft computing approach to recommendation systems as it uses Fuzzy Set and Evidence Theories, as well as ideas from Distributed Artificial Intelligence to characterize

information resources and model linguistic categories. It establishes a different kind of human-machine interaction in IR, as the machine side rather than passively expecting the user to pull information, effectively pushes relevant knowledge. This pushing is done in the conversation algorithm of *TalkMine*, where the user, or her browser automatically, selects the most relevant subsets of this knowledge. Because the knowledge of communities is represented in adapting information resources, and the interests of individuals are integrated through conversation leading to the construction of linguistic categories and adaptation, *TalkMine* achieves a more natural, biological-like, knowledge management of distributed information systems, capable of coping with the evolving knowledge of user communities.

References

[1] Rocha, Luis M. and Johan Bollen (2001)."Biologically motivated distributed designs for adaptive knowledge management." In: *Design Principles for the Immune System and Other Distributed Autonomous Systems*. Cohen I. And L. Segel (Eds.). Santa Fe Institute Series in the Sciences of Complexity. Oxford University Press, pp. 305-334.

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

[3] Miyamoto, S. (1990). *Fuzzy Sets in Information Retrieval and Cluster Analysis.* Kluwer.

[4] Rocha, Luis M. (2002). “Combination of evidence in recommendation systems characterized by distance functions”. Proceedings of the FUZZ-IEEE 2002. In Press.

[5] Galvin, F. and S.D. Shore (1991)."Distance functions and topologies." *The American Mathematical Monthly.* Vol. 98, No. 7, pp. 620-623.

[6] Rocha, Luis M. (2001)."Adaptive Recommendation and Open-Ended Semiosis ." *Kybernetes.* Vol. 30, No. 5-6, pp. 821-851 2001.

[7] Rocha, Luis M. (1999)."Evidence sets: modeling subjective categories." *International Journal of General Systems.* Vol. 27, pp. 457-494.

[8] Rocha, Luis M. (1997). *Evidence Sets and Contextual Genetic Algorithms: Exploring Uncertainty, Context and Embodiment in Cognitive and biological Systems.* PhD. Dissertation. State University of New York at Binghamton. UMI Microform 9734528.

[9] Clark, Andy (1993). *Associative Engines: Connectionism, Concepts, and Representational Change.* MIT Press.

[10] van Gelder, Tim (1991)."What is the 'D' in 'PDP': a survey of the concept of distribution." In: *Philosophy and Connectionist Theory*. W. Ramsey et al. Lawrence Erlbaum.

[11] Kanerva, P. (1988). *Sparse Distributed Memory.* MIT Press.

[12] Rocha, Luis, M. (1994)."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, pp. 400-404.

[13] Rocha, Luis M. (1997b)."Relative uncertainty and evidence sets: a constructivist framework." *International Journal of General Systems.* Vol. 26, No. 1-2, pp. 35-61.

[14] Zadeh, Lofti A. (1965)."Fuzzy Sets." *Information and Control.* Vol. 8, pp. 338-353.

[15] Shafer, G. (1976). A Mathematical Theory of Evidence. Princeton University Press.

[16] Turksen, I.B. (1996)."Non-specificity and interval-valued fuzzy sets." *Fuzzy Sets and Systems.* Vol. 80, pp. 87-100.

[17] Klir, George, J. (1993)."Developments in uncertainty-based information." In: *Advances in Computers*. M. Yovits (Ed.). Vol. 36, pp. 255-332.

[18] Yager, R.R. (1979)."On the measure of fuzziness and negation. Part I: membership in the unit interval." *Int. J. of General Systems.* Vol. 5, pp. 221-229.

[19] Yager, R.R. (1980)."On the measure of fuzziness and negation: Part II: lattices." *Information and Control.* Vol. 44, pp. 236-260.

[20] Turksen, B. (1986)."Interval valued fuzzy sets based on normal forms." *Fuzzy Sets and Systems.* Vol. 20, pp. 191-210.

[21] Nakamura, K. and S. Iwai (1982)."A representation of analogical inference by fuzzy sets and its application to information retrieval systems." In: *Fuzzy Information and Decision Processes*. M.M. Gupta and E. Sanchez (Eds.). North-Holland, pp. 373-386.

[22] Kosko, B. (1992). *Neural Networks and Fuzzy Systems: A Dynamical Systems Approach to Machine Intelligence.* Prentice-Hall.