FY02-CSSE-012
Proposal Title: **Identification of Interests, Trends and Dynamics in Document Networks**
Technical Category: Computer Science and Software Engineering (CSSE)
Principal Investigator Name: Luis Rocha
E-mail: rocha@lanl.gov

Luis Rocha

Modeling, Algorithms, and Informatics Group (CCS-3)

Identification of Interests, Trends and Dynamics in Document Networks

Abstract

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[1], preprints[2], internal reports[3], as well as databases of datasets used in scientific endeavors[4]. Each of these databases possesses several distinct relationships among documents and between documents and semantic tags or indices that 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 [Newman, 2000]. Furthermore, documents can be related to semantic tags such as keywords used to describe their content. Given these relations, we can compute distance functions amongst documents and/or semantic tags, thus creating associative networks between these items, which identify stronger or weaker co-associations.

This proposal aims to investigate the hypothesis that the metric behavior of the distance functions defining these associative networks, can be used as an indicator of the relevance of collections of documents, the interests of users who have selected certain sets of documents, the trends in communities associated with sets of documents, as well the dynamics of such networks in general. The hypothesis itself is based on empirical evidence gathered and discussed in the proposal. We are requesting funds to gather more empirical evidence, investigate adequate formalisms, validate the hypothesis, and build a recommendation system that makes use of results obtained.

The success of this research would have a strong impact on information retrieval and knowledge management. If the hypothesis is correct, we would be able to predict trends in a given community by the automatic analysis of the documents they produce . We would also gain another technique to identify the relevance of documents and the interests of users, which would need to be compared to existing methodology. This impact would be felt on LANL's current knowledge management initiatives, as well as externally, on advancing the study of DN in particular and social networks in general, as well as by producing better recommendation systems for the World Wide Web and digital libraries.

[1] Such as MEDLINE (http://www.nlm.nih.gov) and SciSearch @LANL (http://scisearch2.lanl.gov).

[2] Such as the e-Print Arxiv @ LANL (http://xxx.lanl.gov/).

[3] Such as LANL's Unclassified Publications (http://laup.lanl.gov:4003/htmls/repquery.html).

[4] Such as GenBank (http://www.ncbi.nlm.nih.gov/Genbank/) for Nucleic Acid Sequences and PROSITE (http://www.expasy.org/prosite/) for Protein Motif Libraries.

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 [Newman, 2000]. 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 [Rocha and Bollen, 2000].

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 records[5], 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 and among records: the *semantics* and the *structure* of the DN, respectively[6]. The structure of a DN is defined by the relations between documents in the document collection. In academic databases these relations refer to citations, while in the WWW to hyperlinks. In our case, we work with the citation structure of the 1996-1999 *SciSearch* records. The relation between records and keywords allow us to infer the semantic value of documents and the inter-associations between keywords. Naturally, semantics is ultimately only expressed in the brains of users who utilize the documents, but keywords are tokens of this ultimate expression, which we can infer from the relation between records and 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) record r_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. The 10 most common (stemmed) keywords in the ARP data set are listed in Table I. In subsequent sections we work only with the semantics of DN, though the structure or pragmatics could be studied in the same way.

Table I: 10 Most Common (stemmed) Keywords and their frequency

Frequency	Keyword
187705	Cell
150795	studi
149594	system
140738	express
127350	protein
124094	model
120215	activ
113740	human
112737	rat
112702	patient

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:

[5] Records contain bibliographical information about published documents. Records can be thought of as unique pointers to documents, thus, for the purposes of this proposal, the two terms are interchangeable.

[6] We can also extract a *pragmatics* of the DN from the relations among authors of documents, referred to as a collaboration network. For instance, Newman [2000] has studied such social networks from data extracted from another LANL database (the e-ArXiv pre-print database).

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

The semantic proximity[7] between two keywords, k_i and k_j, depends on the sets of documents indexed by either keyword, and the intersection of these sets. $N(k_i)$ is the number of documents 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 documents that each keyword indexes. Thus, two keywords are near if they tend to index many of the same documents. Table II presents the values of *KSP* for the 10 most common keywords in the ARP dataset.

Table II: 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 $d(k, k) = 0$ [Shore and Sawyer, 1993]. This distance function indicates how far, semantically, a keyword is from another in the set of keywords. This way, it defines a weighted graph *D* whose nodes are all of the keywords extracted from a given DN, and the edges are the values of *d*. Clearly, many other types of distance functions can be defined on the elements of a DN. All of the observations and hypothesis below would apply equally, but naturally, the conclusions drawn cannot be separated by how well, and how appropriately for a given application, a distance function is capable of discerning the elements of the set it is applied to. Thus, different distance functions applied to citation structures or collaboration networks, will require distinct semantic considerations than those used for keyword sets.

3. Semi-metric Behavior

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. Such measures of distance are referred to as semi-metrics [Galvin and Shore, 1991]. Indeed, given that most social and knowledge-derived networks possess Small-World behavior [Watts, 1999], we expect that nodes

[7] This measure of closeness, formally, is a proximity relation [Klir an Yuan, 1995; Miyamoto , 1990] because it is a reflexive and symmetric fuzzy relation. Its transitive closure is known as a similarity relation (Ibid).

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.

Most hypotheses are born out of anecdotal, often personal, evidence. The one put forward here is no exception. It arose from questioning what could one infer from the semi-metric behavior of the distance functions calculated from DN. Given a distance function, what can we say about a pair of highly semi-metric elements from a finite set? And what can we say about the set, from the pairs of highly semi-metric pairs it contains?

To construct an intuition to answer these questions, one needs to deal with very familiar examples. In this case, the author could think of no DN more familiar than the set of books cited by his own dissertation [Rocha, 1997]! A database similar (but much smaller) to the one used by ARP contains the relevant information. This database contains about 150 books, each indexed by the respective Library of Congress Keywords, for example:

Kearfott, R. Baker and Vladik Kreinovich (Editors). [1996]. *Applications of Interval Computations*. Kluwer.
Keywords: Optimization algorithms, Fuzzy logic, Uncertainty, Mathematics, Reliable Computation, Interval Computation.

Table III: Distance function for 5 keywords in the dissertation database

	Adaptive Systems	Evolution	Modeling systems	Complex Systems	Social Systems
Adaptive Systems	0.00	3.89	12.00	10.33	16.00
Evolution	3.89	0.00	21.50	4.22	35.00
Modeling Systems	12.00	21.50	0.00	5.75	10.00
Complex Systems	10.33	4.22	5.75	0.00	19.00
Social Systems	16.00	35.00	10.00	19.00	0.00

From this database, 86 keywords are extracted. A distance function d is calculated according to eq. 2. Table III shows the values of d for 5 of the keywords. One needs to note that this distance function is obtained from the relations extracted from a particular set of documents (in this case 150 books). Therefore, one should not expect these values to represent a universally accepted thesaurus or the associations one would anticipate from common sense knowledge. Indeed, this kind of distance is used to characterize particular information resources and users from the documents they contain or retrieve [Rocha, 2001]. In this case, the associative distances between keywords denote the way the dissertation set of books is related.

To discover the shortest distances between keywords using the distance metric, one uses a (+, min) matrix composition of D until closure is achieved[8]. In this case, the dimension of the graph is 8, that is, the longest path contains 8 nodes. Table IV shows the shortest distances for the same 5 keywords. We see for instance that the shortest indirect distance between MODELING SYSTEMS and EVOLUTION is 9.97, whereas the direct distance is 21.5. This means that the distance between the keyword pair MODELING SYSTEMS-EVOLUTION is semi-metric. This is not the case of the metric pair ADAPTIVE SYSTEMS-EVOLUTION, for which the shortest distance is the direct one.

Table IV: Shortest distance for 5 keywords in the dissertation database (semi-metric pairs shown in italics).

	Adaptive Systems	Evolution	Modeling systems	Complex Systems	Social Systems
Adaptive Systems	0.00	3.89	12.00	*8.11*	16.00
Evolution	3.89	0.00	*9.97*	4.22	*19.89*
Modeling Systems	12.00	*9.97*	0.00	5.75	10.00
Complex Systems	*8.11*	4.22	5.75	0.00	*15.75*
Social Systems	16.00	*19.89*	10.00	*15.75*	0.00

4. Characterizing Semi-metric Behavior

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*:

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

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

s is positive and ≥ 1 for semi-metric pairs. In our example, s(MODELING SYSTEMS,EVOLUTION) = 21.5/9.97 =2.157. This ratio is important to discover semi-metric behavior necessary for our hypothesis as discussed below, but 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:

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

rs compares the semi-metric distance reduction to the maximum possible distance reduction in graph D. $d_{\max}$ is the largest distance in the graph, and $d_{\min} = 0$ is the shortest distance.

Often, the direct distance between two keywords is ∞ because they do not index any documents in common. As a result, in closed graphs, 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*:

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

where $\overline{d_{k_i}}$ represents the average direct distance from k_i to all k_j such that $d_{direct}(k_{\setminus i}, k_j) \geq 0$. 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.

5. Analysis of a Collection of Documents: The Interests of the Collector

The semi-metric ratios were applied to graph D of the dissertation database, and the semi-metric pairs with higher ratios were identified. Table V lists the top 5 pairs for semi-metric ratio s. If we rank pairs for the relative semi-metric ratio rs, there is a slight ordering of the top as the pair EVOLUTION-DNA drops to rank 11 and the pair LIFE-COGNITION to 6th, while the pair EVOLUTION-CONTROL rises to rank 3 (from 6th) and the pair EVOLUTION-INFORMATION THEORY rises to 5th (from 20th).

Table V: Semi-metric pairs with highest s in dissertation database.

(k_i,k_j)	$s(k_i,k_j)$	$rs(k_i,k_j)$
ADAPTIVE SYSTEMS-COGNITION	6.39	0.84
EVOLUTION-CONSTRUCTIVISM	5.00	0.76
EVOLUTION-PSYCHOLOGY	5.00	0.73
EVOLUTION-DNA	4.69	0.64
LIFE-COGNITION	4.55	0.66

What is most interesting about these results is that these pairs denote the original contributions that were offered by the dissertation! Indeed, the dissertation was about using ideas and methodologies from Complex Adaptive Systems, Evolutionary Systems, and Artificial Life and apply them to Artificial Intelligence and Cognitive Science. In particular, the mathematical models (from Psychology) of cognitive categories were expanded using evolutionary ideas, by drawing an analogy with the symbolic characteristics of DNA. Furthermore, this framework was named Evolutionary Constructivism, a term that did not exist previously, but draws both from Evolutionary Theory and the Philosophy of Constructivism in Cognitive Science and Systems Theory.

To understand these results, we need to remember that the distance function d is derived from the finite set of books used in the dissertation. A high degree of semi-metricity for a keyword pair means that very few of the books

in the database are simultaneously indexed by these two keywords, but that there exists a strong indirect association between these keywords via some indirect path whose short distances require the existence of many related books for each keywords pair in the pathway. Thus, a keyword pair with high semi-metric behavior, implies an association that is a property of the specific collection of documents, but not one identifiable in many included documents, and rather constructed from an indirect series of strongly related documents. In other words, the highly semi-metric pairs represent associations that "were begging to be made, given this specific collection of documents. Indeed, the two pairs (EVOLUTION-CONTROL and EVOLUTION-INFORMATION THEORY) ranked in the top 5 for the relative semi-metric ratio, identify two associations that are certainly implied by the collection of books (given its large subsets of Cybernetics and Information Theory books), but which were not dealt with in this dissertation – offering some topics for other dissertations!

The below average ratio b, was also used to identify keyword pairs with infinite semi-metric reduction. Those are the pairs that do not index simultaneously a single book in the collection, but which are nonetheless indirectly strongly related. For this particular dataset, the pairs with highest values of b, did not seem to produce meaningful results. This could be because, being a small, tight collection of books, the relevant associations implied by the collections are already made at least by a small set of books producing a large, but finite, distance. Such situation is not expected to occur in larger, multi-authored collections unlike the dissertation one.

6. Analysis of Larger Datasets: Trends in Collections

The anecdotal analysis of the author's dissertation database served the purpose of creating an intuition of what semi-metric behavior may mean for DN, but to even build a hypothesis, other more "subject-independent" datasets need to be studied.

The same semi-metric behavior ratios were used to study the ARP dataset describe above. The distance function d (eq. 2) was calculated for the set of the 500 most common keywords, and the semi-metric ratios (formulas 3 to 5) were calculated for all keyword pairs. Table VI shows the top 5 keyword pairs ranked by highest values of s.

Table VI: Semi-metric pairs with highest s in ARP dataset.

(k_i,k_j)	$s(k_i,k_j)$	$rs(k_i,k_j)$
LEUKEMIA-MYOCARDI	272.20	0.4981
HORMON-THIN	214.08	0.9953
CARE-EXCIT	213.59	0.9953
GENE-EQUAT	205.76	0.9951
FILM-TRANSCRIPT	204.51	0.9951

To analyze these results, again, one must remember the original collection of documents, in this case, all the scientific articles published in journals indexed by ISI in SciSearch between the years of 1996 to 1999. A keyword pair with high semi-metricity, implies that while very few articles discuss the two topics together, a very large series of articles exists which creates an indirect pathway between these two keywords in D. To obtain a large semi-metric ratio, it is necessary that each link in the indirect pathway be defined by short distances, which in turn require the existence of many articles associated to both keywords in the link. Thus, a highly semi-metric keyword association implies that very few documents make that association, but that there are large sets of documents indirectly supporting it. In this sense, the existence of such support (particularly in scientific databases) may identify a trend that can be expected to be picked up.

While it is hard to understand all associations identified in such a dataset containing so many different topics, at least one association is observed in the data set which is meaningful to the author. The high semi-metricity of the GENE-EQUAT[9] pair may be a result of the trend observed in the late 1990's towards computational and mathematical biology as molecular biology started to move into a post-genome bioinformatics mode [Kanehisa, 2000]. Indeed, the analysis of the keyword pairs with infinite semi-metric reduction characterized by a high below average ratio b, seems to give further evidence for this claim, as the highest values of b are observed for the pairs EQUAT-MESSENGERRNA, EQUAT-TRANSCRIPT, and EQUAT-GENE-EXPRESS. These pairs associate the key word Equation with keywords that describe the chief technology that enabled the greatest advances in bioinformatics in the late 1990's and today: the Gene Expression Arrays that allow the rapid measurement in parallel of messenger RNA transcribed from DNA in the cell (the process of gene expression). As expected, ratio b is useful for larger datasets not collected by a single author.

[9] Notice that ARP keywords are stemmed to group different constructions of the same term: e.g. Equation and Equations.

In this case, it picked relevant associations that were not present in a single document but strongly implied by the overall collection.

7. Dynamics of DN and Other Proposed Developments

The hypothesis generated from the anecdotal evidence described above, is that high semi-metric keyword associations, discovered in the distance function of DN defined by eq. 2, can (1) capture the interests of a person associated with a given small collection of documents, and (2) be used to identify trends in large, multi-authored document collections.

Clearly, much more evidence is needed to support this hypothesis. We have already applied this study to other cases such as random distance graphs (very small relative semi-metric ratios), distance graphs built from word association norms used in psychological tests (very little semi-metric behavior), distance graphs built from web-site collective usage [Bollen et al, 1999] (similar results to section 5), etc. In addition to a more thorough analysis of the semi-metric behavior of different distance functions applied to different DN, we propose to study this hypothesis by (1) creating an experimental database interface to evaluate how well the interests of users are captured by semi-metricity, and (2) studying the trend dynamics in large collections of documents such as the ARP database.

7.1 Semi-metricity Recommender

LANL's research library is currently offering an ideal system to study how well semi-metric keyword associations capture the interests of a single person. The *MyLibrary* web portal (http://mylibrary.lanl.gov) allows LANL users to store links to publications in LANL's several digital libraries with scientific articles, as well as links elsewhere on the WWW. We obtained permission from the research library to develop a recommendation system [Rocha, 2001] to be added to *MyLibrary*.

We propose to issue recommendations to users of *MyLibrary* from pairs of highly semi-metric keywords. We will use the distance function defined by eq. 2 on the keywords extracted from the documents stored by each user of the *MyLibrary* system. Recommendations will be issued in the form of documents that users may be unaware of but which directly associate the semi-metric keyword pairs, and directly as keywords. We will collect data from user behavior to validate how useful the recommendations are, thus gathering evidence for the first aspect of the semi-metric hypothesis.

7.2 Dynamics of Trends

If high semi-metricity is indeed an indicator of trends, then, by analyzing how semi-metric behavior changes in time in a DN we should be able to distinguish real trends from indirect associations never picked up by authors of documents. As a trend develops in time, we can expect the semi-metric ratios to lower, as more and more documents are added which directly associate the previously highly semi-metric keyword pairs. For instance, if the high semi-metricity pairs of Bioinformatics keywords related to Gene Expression Arrays identified in 6 do indeed imply a trend phenomenon, we should expect to see lower values of semi-metricity for these concepts in those articles published in 2000 and beyond.

Thus, we propose to collect the SciSearch articles for years 2000 and beyond, and use them as a validation set for the trends picked in the ARP database. We will also conduct the semi-metric analysis year by year, rather than the whole interval from 1996 to 1999, as well as for much more than the top 500 most frequent keywords. A more detailed study such as this, will allow us to capture the change in semi-metric behavior of a large set of keywords through the years, thus gathering evidence for the second aspect of the semi-metric hypothesis to study the dynamics of trends in DN.

8. Work Plan and Expected Results

We intend to tackle this proposal with the following chronological milestones:

1. Familiarization with the mechanics of the *MyLibrary* system. Data gathering from this system.
2. Gathering of additional SciSearch documents for years after 1999. Addition of these documents to database. Extraction of document-keyword relation and computation of distance function.
3. Semi-metric analysis of other DN and other relations extracted from DN with different distance functions. Comparative study of distance functions and development of alternative semi-metric behavior measures.

4. Development and testing of prototype Semi-Metricity Recommender for *MyLibrary*.
5. Year-by-year analysis of semi-metric behavior in the ARP database.
6. Implementation of semi-metricity recommender. Gathering of user feedback and collective usage patterns.
7. Identification of trends and their dynamics in the ARP database.
8. Evaluation of hypothesis.

In the first year we expect to finalize 1 and 2, and start 3 which will be ongoing through the duration of the project. In the second year we will tackle 4, 5, and 6. In the third year we will continue 6, and complete 7 and 8.

9. References

Bollen, J., H. Vandesompel, L.M. Rocha [1999]."Mining associative relations from website logs and their application to contex-dependent retrieval using Spreading Activation." In: *Workshop on Organizing Web Space (WOWS), ACM Digital Libraries 99, August 1999, Berkeley, California.*

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

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

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

Newman, MJ [2000]."The structure of scientific collaboration networks." *Proc.Nat.Acad.Sci.* No. 98, pp. 404-409.

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. (http://www.c3.lanl.gov/~rocha/dissert.html)

Rocha, Luis M. [2001]."TalkMine: A Soft Computing Approach to Adaptive Knowledge Recommendation." In: *Soft Computing Agents: New Trends for Designing Autonomous Systems*. V. Loia and S. Sessa (Eds.). Springer-Verlag. In press. (http://www.c3.lanl.gov/~rocha/softrec.html)

Rocha, Luis M. and Johan Bollen [2000]."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. In Press. (http://www.c3.lanl.gov/~rocha/SFI99.html)

Shore, SD, Sawyer LJ [1993]."Explicit Metrization." *Annals of the New York Academy of Sciences*. v. 704 pp. 328-336 1993.

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