

ExpertiseNet: Relational and Evolutionary Expert Modeling

Xiaodan Song¹, Belle L. Tseng², Ching-Yung Lin¹, and Ming-Ting Sun¹

¹ Department of Electrical Engineering,
University of Washington, Box 352500, Seattle, WA 98195, USA
{song, cylin, sun}@ee.washington.edu

² NEC Labs America, 10080 N. Wolfe Road, SW3-350, Cupertino, CA 95014, USA
belle@sv.nec-labs.com

Abstract. We develop a novel user-centric modeling technology, which can dynamically describe and update a person's expertise profile. In an enterprise environment, the technology can enhance employees' collaboration and productivity by assisting in finding experts, training employees, etc. Instead of using the traditional search methods, such as the keyword match, we propose to use relational and evolutionary graph models, which we call ExpertiseNet, to describe and find experts. These ExpertiseNets are used for mining, retrieval, and visualization. We conduct experiments by building ExpertiseNets for researchers from a research paper collection. The experiments demonstrate that expertise mining and matching are more efficiently achieved based on the proposed relational and evolutionary graph models.

1 Introduction

Finding experts in an enterprise is a challenging problem. It is important for enterprises to understand and leverage their most valuable assets – employees' minds and knowledge bases. For instance, when a service-personnel faces problems at a customer's site, it would be desirable to have some technology to help identify the best experts for solving the problem within the company, and thus the solution may only be a phone call away. In another scenario, if employees' skills can be explicitly described, a manager can easily assemble a team for a new project, or an HR person can suggest what kind of skills an employee should learn.

Currently in an enterprise, the expertise information of employees is usually established by manual updates. However, this results in serious drawbacks. For instance, employees may not invest the necessary efforts in creating rich profiles, or they may not keep the information up-to-date as their interests, responsibilities, and expertise change. Many employees just provide a few keywords, without any timing information and the relationships between different expertises. Also, it is often difficult to differentiate who are the more suitable experts from people with similar expertise skills.

A number of automatic expert finding prototypes have been reported in literature. Liu *et al.* [6] and Mockus *et al.* [9] proposed to solve the expert-mining problem by using traditional information retrieval techniques. In their approaches, a person's expertises are described in terms of a vector without relational or timing information.

HelpNet uses the user-filled information to provide the expertise profile [7]. *Expert-Locator* uses a representative collection of users' technical documents to build expertise indices [17]. Recently, NASA's *Expert Finder* [1][16] uses the name-entity extraction to process employees' published resumes and documents as well as corporate newsletters. It identifies keywords to create expertise profiles. *Expert Finder* presents a ranked list of employees that best matches the query. *I-Help* [2], an agent-based system, models a user's characteristics so that it can assist him/her in identifying a peer who can help. To select the most appropriate peer for a particular request, it uses a matchmaking agent that communicates with the personal agents of other users, by accessing various kinds of vector-based user information.

To capture the relationships between entities, link analysis has been studied for some time [18]. For instance, the Google search engine [10] and Kleinberg's HITS algorithm [5] both use link analysis on the web to find "hubs" and "authorities". The success of these approaches, and the discovery of widespread network topologies with nontrivial properties, had led to a flurry of research on introducing link analysis into the traditional information retrieval area [4]. In the user modeling area, a related technology – collaborative filtering – is used to help people make choices based on other people's opinions [11].

The goal of our paper is to develop a user-centric modeling technology, which can dynamically describe and update a person's expertise profile. The most important feature of our approach is to use expertise graphs for user modeling. A graph model is built to describe the relationship as well as the temporal evolution of the person's expertises. We refer to this expertise profile represented as graphs as an ExpertiseNet. Figure 1 shows an illustration of the proposed scheme. Both text content and citation linkages of papers are analyzed to extract the relation and the evolution information of the person's expertises. Exponential random graph models (ERGM) [15] and stochastic actor-oriented models [14], which deal with the relations between the entire network and individual entities, are incorporated to describe the structural properties and dynamics. Also, in this paper, we propose to add the citation linkages to boost the expertise classification accuracy. After we get the ExpertiseNets, expert mining and matching can be performed accurately and easily. To the best of our knowledge, our work is the first to explicitly use relational and evolutionary graph models for user profiling.

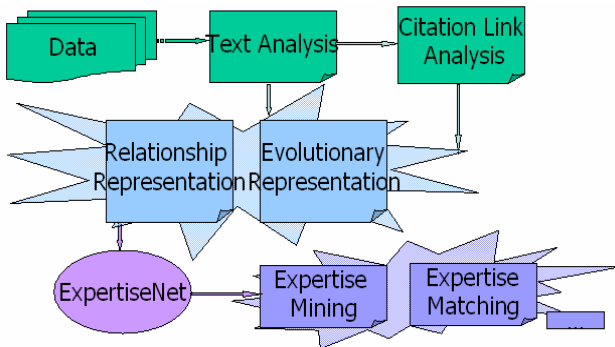


Fig. 1. Illustration of the ExpertiseNet building and expertise mining process

The main contributions of this paper are summarized as follows:

- ✧ We propose a scheme to represent user profiles by graphs.
- ✧ We propose an ExpertiseNet that provides relational and evolutionary information for user modeling.
- ✧ We incorporate the exponential random graph models and evolutionary social network models into the user model analysis.
- ✧ We propose to boost the text classification accuracy by including the citation linkages in the classification algorithm.

The rest of the paper is organized as follows. In Section 2, we present the processes to build relational and evolutionary ExpertiseNets. We discuss the expertise mining and matching in Section 3. In Section 4, we show the experimental results of mining and matching. Finally, conclusions and future work are addressed in Section 5.

2 ExpertiseNet

In this paper, we use a dataset, Cora Research Publication Corpus [8], to simulate the information we may get in real applications. This set includes 24,778 authors with more than 50,000 research papers and about 715,000 citation links. The dataset also provides bibliographic information for each paper, which includes the authors' names, affiliations, email addresses, paper title, abstract, and references. These papers were pre-classified into a hierarchical ontology consisting of 69 expertise categories [8].

2.1 Classifying a Person's Expertises

We use a text classification algorithm to detect and extract expertise information of each person. Among the state-of-the-art classification algorithms, we choose the boosting algorithm [13], which is evaluated as one of the most accurate text classification algorithms [12]. This algorithm finds a "strong hypothesis" by combining many "weak" hypotheses for fusing features in different forms.

Previous techniques extract expertise using the words from the title and the abstract of one's publications. After a pre-processing stage which excludes "stop words" (common words that do not really provide any meaning to the document, such as "the", "it", "or"), and conducts "stemming" which collapses multiple forms of a term into the same term, we can get the terms as the features. In our work, the citation linkages are also used to form a feature vector (citation ratio) for classification as defined below.

$$P(X = x) = c_x / R \quad (1)$$

where X is a variable which indicates the citation information for each publication, x represents one of the 69 categories, c_x is the number of the citations belonging to category x , and R is the number of the references in the publication. The intuition of incorporating this feature is that the paper from one category tends to cite the papers in the same category. This relational structure is useful to describe the strength of a person's expertise on specific categories by his/her publications. In Section 4, we show that this feature can significantly increase the classification accuracy of a person's expertise. After document classification, we have the information about the number of published papers in each category for one person, which represents the expertise skills of the person, and is used for building ExpertiseNet for that person.

2.2 Relational ExpertiseNet

The relational ExpertiseNet of a specific person is formulated as a directional graph $G(V, E)$, where V represents the node set, and E represents the edge set. Two nodes v_i and v_j are adjacent, if edge $e_{ij} = (v_i, v_j)$ or $e_{ji} = (v_j, v_i)$ is in the set of edges E . In the relational ExpertiseNet, each node represents an expertise. The strength of each expertise is defined as:

$$s_i = p_i \quad (2)$$

where s_i represents the strength of the expertise i for the person, and p_i is the number of publications in category i . Edges represent the relationship between expertise nodes.

The correlations of a person's published papers and all papers in the research society decide the weights of the edges in the graph model. Citation linkages provide solid evidence of correlations among different expertise. When the citation linkages are not available, the correlation can be explored by the text similarity analysis. The process is illustrated in Figure 2, and explained as follows.

For each paper, the dataset contains citation linkages that include both the information of how this paper cites other papers (out-direction), and how other papers cite this paper (in-direction). From the publications of a person P , it is reasonable to infer that his/her expertise X is influenced by Y if papers in category X cite papers in category Y . For example, in Figure 2, paper #1 is one of the papers in Natural Language Processing (NLP). It cites paper #5, #6, #7 (out-direction citation linkages) in NLP, Machine Learning (ML), and Information Retrieval (IR) respectively. Papers #2, #3, #4 in NLP, NLP, and IR, cite paper #1 (in-direction citation linkages). We can infer that for this person, his/her NLP expertise is influenced by NLP, ML, and IR, and at the same time, affects IR. With this consideration, the strengths of the edges of the relational ExpertiseNet are determined by:

$$e_{A \rightarrow B} = \left(\sum_{i=1}^K n_{iAB} / N_i \right) / K \quad (3)$$

where $e_{A \rightarrow B}$ represents the strength of the edge from expertise A to expertise B , K is the total number of the publications for the person, n_{iAB} represents, the number of papers in category A cited by the papers in category B , and N_i represents the number of citations in paper i .

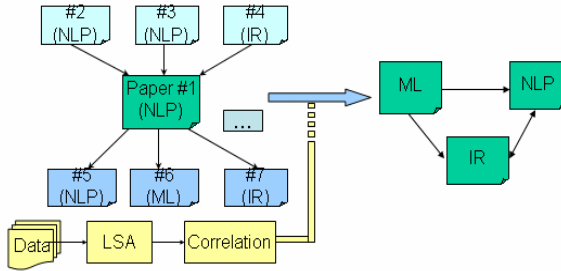


Fig. 2. Building the relational ExpertiseNet from text and citation linkages

For those people without papers, ExpertiseNet will be built upon the documents they generated. Latent Semantic Analysis (LSA) [3] is applied on the occurrence matrix which includes the frequency of occurrence of the words in the document. LSA is a method for extracting and representing the contextual meaning of words. It has been used as a technique to measure the coherence of texts. By comparing the vectors formed by the keywords of two documents in a high-dimensional semantic space, this method provides a characterization of the degree of semantic relatedness between documents. LSA starts with a sparse matrix in which the rows represent the documents and the columns represent the keywords. The values in the matrix are proportional to the number of times the keywords appear in the document, where rare terms are upweighted to reflect their relative importance. The correlation matrix is calculated to determine the edges of the relational ExpertiseNet.

An observation of the ExpertiseNet profile is obtained from the above analysis. Then, the exponential random graph model (or p^* model) [15] is used to estimate an underlying distribution to describe the relational ExpertiseNet. The strong point of this statistical model is that they can represent structural tendencies (such as transitivity which represents the number of transitive patterns) that define complicated dependence patterns not easily modeled by deterministic models. Given a set of n nodes, let Y denote a random graph on those nodes, and y denotes a particular graph on those nodes. Then

$$P_{\theta}(Y = y) = \frac{\exp(\theta^T s(y))}{c(\theta)} \quad (4)$$

where θ is an unknown vector of parameters, $s(y)$ is a known vector of graph statistics on y (Density (defined by the out-degrees), reciprocity (defined by the number of reciprocated relations), transitive triads (defined by the number of a set of edges $\{(i \rightarrow j), (j \rightarrow k), (i \rightarrow k)\}$), and the strengths of the nodes as well as the edges as considered in this paper), $c(\theta)$ is a normalization term. This probabilistic expression has advantages on describing the insights of the network, thus can help to describe the evolution of Expertise Nets.

2.3 Evolutionary ExpertiseNet

In the evolutionary ExpertiseNet, the dynamics and the evolution of expertises are explored. Two tasks are performed: evolution segmentation (detect changes between expertise cohesive sections) and expertise tracking (keep track of the expertise similar to a set of previous expertises). The strength of the nodes as well as the structure of the network is considered in the evolution segmentation. Temporal sliding windows are applied. An exponential random graph model is estimated from the data in each window, *i.e.*, time period. We can obtain a series of parameters which indicate the network configurations. Then the change points of an evolutionary ExpertiseNet are determined by:

$$\sum_{k=1}^M |\theta_{t,k} - \theta_{t-1,k}| > th \quad (5)$$

where $\theta_{t,k}$ indicates the parameters of the exponential random graph model at time t , M represents the number of parameters, th is a threshold. The goal is to find all t satisfy (5).

The evolution segments are obtained from these change points. Based on the correlation between the citations and the development of related research areas, the tracking edges are determined by

$$e_{A_{t-1} \rightarrow B_t} = \left(\sum_{i=1}^K n_{iA_{t-1}B_t} / N_{i,t} \right) / K_t \quad (6)$$

where all variables have the same meaning as in (3), except only those papers at time t are considered.

We borrow the ideas from dynamic actor-oriented social networks [14] to describe the dynamics of the evolutionary ExpertiseNet. The change in the network is modeled as the stochastic result of network effects (density, reciprocity, etc.). The network evolution is modeled by continuous time Markov chain models, whose parameters are estimated from the observed data by the Markov chain Monte Carlo procedures [14]. The transition probability is formulized as:

$$p_{ij}(\beta, y) = \frac{\exp\left(\sum_{k=1}^L \beta_k s_{ik}(y)\right)}{c(\beta, y)} \quad (7)$$

where β_k represents statistical parameters indicating the strength of the corresponding effect $s_k(y)$ (In our model, we consider density, reciprocity, and transitivity), $c(\beta, y)$ is a normalization term, L represents the number of parameters.

3 Expertise Matching and Mining Based on ExpertiseNets

After obtaining the ExpertiseNets (profiles) of individuals, we can perform expertise mining and matching. We define two kinds of mining. The first one is to get a list of people who have the expertise of interest and if certain relational patterns between their expertises exist. The second one is the evolutionary expertise mining, for finding persons who have certain evolutionary expertise patterns. The searching results are ranked by the strength of the linkage in relational or evolutionary expertise pattern.

Instead of using traditional vector-based matching algorithms, ExpertiseNet provides a novel way to search persons with similar evolutionary and relational information of the expertises. This distance function to compare different relational ExpertiseNets in order to differentiate persons is defined as follows:

$$\text{dist}(G_1, G_2) = \sum_{k=1}^M |\theta_{1,k} - \theta_{2,k}| \quad (8)$$

where G indicates graphs, θ indicates the parameters of the exponential random graph models, and M represents the number of parameters to describe one graph. For evolutionary ExpertiseNets, the distance function is formulated as:

$$\text{dist}(G_1, G_2) = \sum_{k=1}^L |\beta_{1,k} - \beta_{2,k}| \quad (9)$$

where β are statistical parameters in the actor-oriented model.

4 Experimental Results

In this section, we discuss the implementations and show experimental performance of the relational and evolutionary ExpertiseNets.

4.1 Classification Results

We use text classification to automatically extract the strength of each person’s expertise in different topic categories. Figure 3 shows the classification accuracy vs. the use of other labeled data over the whole dataset. The highest curve illustrates the performance of our method. We can achieve about 5% improvement, which is considered significant in the text classification area.

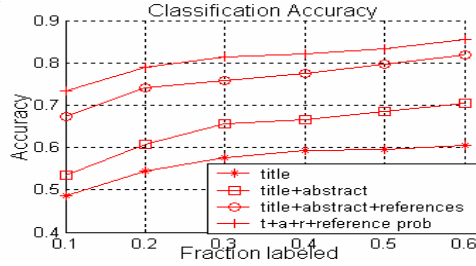


Fig. 3. Incorporation of citation ratio as a feature boosts classification accuracy

4.2 ExpertiseNet Profiles and Expert Searching

ExpertiseNet provides a dynamic model of semantics evolution in which expertise as well as inter-expertise relationships exhibit. In this experiment, we show two mining processes which are based on the relational and evolutionary ExpertiseNets. In the first experiment, by placing a query term “machine leaning – planning”, we get a list of persons with expertise of Machine Learning and Planning (Figure 4 (a)). Figure 4(b) shows another example, where “machine leaning $\rightarrow$ planning” is set as the input. A list of people with machine learning expertise in an earlier stage and planning in a later stage shows up.

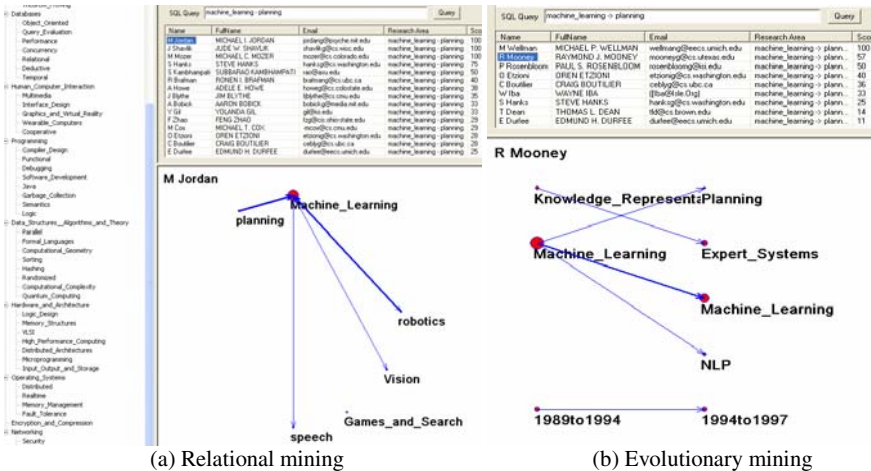


Fig. 4. Our expertise mining system allows relational and evolutionary mining that results in a ranked people list, along with visualizations of ExpertiseNets and the 69 categories from Cora

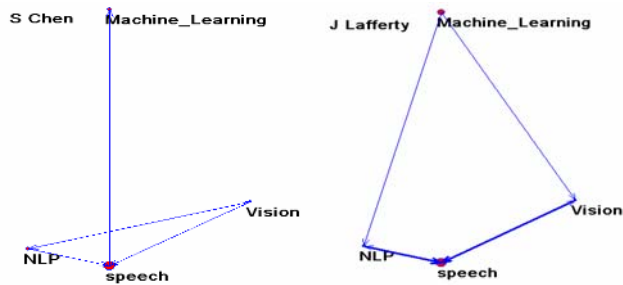


Fig. 5. Two persons with the same expertise vector can be differentiated by ExpertiseNet

In order to validate our model, we searched each person’s website to get more information from one’s resume, editorships, publications, etc. We randomly chose 50 persons and did this validation manually. It shows that the ExpertiseNet models can represent accurate information about the persons. Two examples are illustrated in Figure 5 and Figure 6. In Figure 5, there are two persons with the same expertise, “Prof. Chen”(<http://www.cise.ufl.edu/~suchen/DrChen.htm>) and “Prof. Lafferty” (<http://www-2.cs.cmu.edu/~lafferty>). By vector-based representation, it is difficult to differentiate them. However, it is easy to illustrate the differences by relational ExpertiseNets. From their websites, we know that Prof. Chen has background in speech/signal processing, while Prof. Lafferty is mainly interested in machine learning and statistical methods. This information validates our models: Speech contributes more to other expertises for Prof. Chen, and machine learning contributed more to other expertises for Prof. Lafferty.

Figure 6 illustrates the necessity of using evolutionary ExpertiseNets. In this example, we try to distinguish two professors. We can see that Prof. Langley mainly contributed in machine learning, and later, he applied this expertise to other areas, while Prof. Durfee’s expertise on planning, agents, robotics seems to be developed at the same time, and interacts to each other a lot. This conclusion is also validated by the information from their websites.

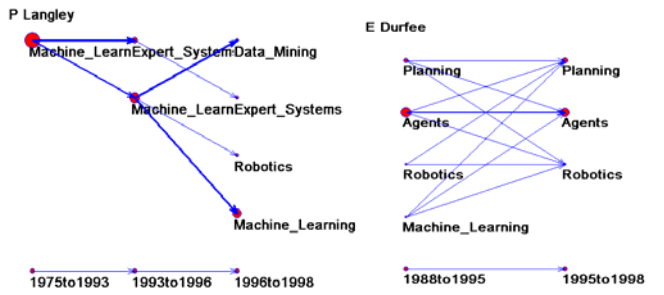


Fig. 6. Two persons’ expertise evolutions are accurately captured by the ExpertiseNets

4.3 Expertise Matching

Figure 7 shows the expertise matching result for finding persons with similar expertise and relationship among expertise. Here we use Professor “Michael Jordan”

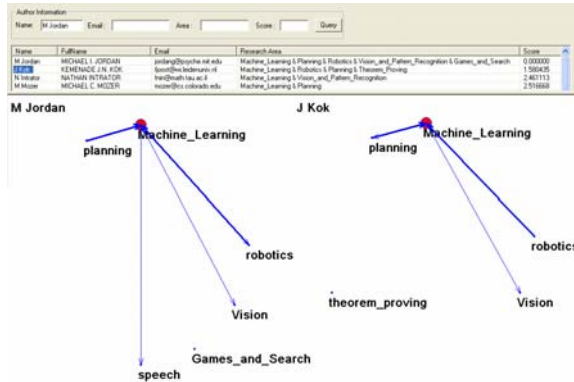


Fig. 7. A system for searching similar persons based on the ExpertiseNets

as the query input (<http://www.cs.berkeley.edu/~jordan/>). We can get a list of ranked experts according to the similarity to “Michael Jordan” in terms of ExpertiseNet.

To complement quantitative evaluation, we have set up an online evaluation system as well as a demo system which allow interactive exploration of expert searching and matching (http://students.washington.edu/~xdsong/index_files/Demos/expertisenet.htm). In the online evaluation system, experts in the database are allowed to evaluate their own ExpertiseNet as well as the expert searching and matching results.

5 Conclusions and Future Work

In this paper, we propose a relational and evolutionary graph model called ExpertiseNet for expert modeling, which discovers both the relational and evolutionary information. Citation linkages and textual data are used to build the ExpertiseNets. The temporal evolution and relational information of expertises are explicitly described in the ExpertiseNet graph profiles that are used for mining, retrieval, and visualization. The parameters from ERGM and dynamic actor-oriented graph models are used to describe the structural properties and dynamics of the ExpertiseNet. We also introduce a new feature that boosts the expertise text classification accuracy by approximately 5%. Manual validations demonstrate the effectiveness of the proposed relational and evolutionary ExpertiseNets on applications of expert mining and matching. So far, the studies are complemented by qualitative evaluation. We have set up an online evaluation system for further quantitative evaluation.

Both the relational and the evolutionary ExpertiseNets are scalable since when new data arrive, only parts of the models need to be updated with the new info. For relational ExpertiseNet, only the related nodes and edges need to be updated without changing of other nodes and edges in the graph. For evolutionary ExpertiseNet, new information from the recent time period only affects the last level, or generates another level in the graph.

To the best of our knowledge, we are the first to explicitly use relational and evolutionary graph models for user profiling. Our ongoing work includes extending this graph-based user modeling frame work for user behavior modeling, and finding and predicting people’s information dissemination behaviors and roles in various events.

Acknowledgments

The authors would like to thank: NEC Laboratories America for their support and Yi Wu for her valuable discussions.

References

1. Becerra-Fernandez, I.. Searching for Experts with Expertise-Locator Knowledge Management Systems. ACL 01 Workshop Human Language Technology and Knowledge Management: New York, (2001), 9-16
2. Bull, S., Greer, J., McCalla, G., Kettel, L., Bowes, J.. User Modelling in I-Help: What, Why, When and How. In Proc. of the 8th International Conference on User Modeling, Sonthofen, Germany, July, (2001), 117-126
3. Deerwester, S., Dumais, S. T., Furna, G. W., Landauer, T. K., Harshman, R.. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), (1990), 391-407
4. Henzinger, M. Hyperlink Analysis on the Web. *IEEE Internet Computing*, 5(1): (2001), 45-50
5. Kleinberg, J.. Authoritative sources in a hyperlinked environment, In Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, (1998), 604-632
6. Liu, P., Curson, J., Dew, P. M.. Exploring RDF for Expertise Matching within an Organizational Memory. In Proc. of the 14th International Conference on Advanced Information Systems Engineering: (2002), 100-116
7. Maron, M. E., Curry, S., Thompson, P.. An Inductive Search System: Theory, Design and Implementation. *IEEE Transaction on Systems, Man and Cybernetics*, 16(1) (1986), 21-28
8. McCallum, A. K., Nigam, K., Rennie, J., Seymore, K.. Automating the construction of internet portals with machine learning. *Information Retrieval*, 3(2): (2000), 127-163
9. Mockus, A., Herbsleb, J. D.. Expertise Browser: a quantitative approach to identifying expertise," In Proc. of the 24th International Conference on Software Engineering: (2002), 503-512
10. Page, L., Brin, S., Motwani, R., Winograd, T.. The PageRank Citation Ranking: Bringing Order to the Web. Stanford Digital Libraries Working Paper, (1998)
11. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., Riedl, J.. Grouplens: An open architecture for collaborative filtering of netnews. In Proc. of the ACM Conference on Computer Supported Cooperative Work: (1994), 175-186
12. Sebastiani, F.. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1), (2002), 1-47
13. Schapire, R. E., Singer, Y.. BoosTexter: A boosting-based system for text categorization, *Machine Learning*, 39(2/3): (2000), 135-168
14. Snijders, T. A. B.. Models for Longitudinal Network Data. To appear as Chapter 11 in P. Carrington, J. Scott, & S. Wasserman (Eds.), *Models and methods in social network analysis*. New York: Cambridge University Press (in press), (2004)
15. Snijders, T. A. B.. Markov Chain Monte Carlo Estimation of Exponential Random Graph Models. *Journal of Social Structure*, 3(2), (2002)
16. Staab, S.. Human language technologies for knowledge management. *Intelligent Systems*, 16(6): (2001), 84-94
17. Steeter, L. A., Lochbaum, K. E.. An Expert/Expert Locating System based on Automatic Representation of Semantic Structure. In Proc. of the Fourth IEEE Conference on Artificial Intelligence Applications: San Diego, CA, (1988), 345-349
18. Wasserman, S., Faust, K., Iacobucci, D.. *Social Network Analysis: Theory and Methods*. Cambridge University Press: Cambridge, (1995)