ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

An Evaluation of Web Mining Application in Social Networks

Sreedhar Appalabatla, Dr. Naveen Kumar, Dr. Mungamuru Nirmala

Research Scholar, Dept. of Computer Science and Engineering, University of Allahabad, India,

Professor, Dept. of Computer Science and Engineering, University of Allahabad, India.

Assistant Professor, Dept of Computing, Adama Science and Technology University, Adama , Ethiopia

ABSTRACT: Social Networks have experienced a meteoric rise recently. They provide a number of functionalities such as network of friends or business contacts listings, content-sharing, profile surfing, discussion and messaging tools. Interoperability among Social Networks being a key challenge, the Google-powered Open-Social alliance has partly solved it and unveiled a new breed of strategies to gather data from Social Network users. In this paper, we discuss about on the Open-Social functionality and combine it with filtering and ranking algorithm to enhance email management. We analyze the traffic-weighted Web host graph obtained from a large sample of real Web users. A number of interesting structural properties are revealed by this complex dynamic network, some in line with the well-studied Boolean link host graph and others pointing to important differences.

KEY WORDS: Open Social Network Dataset (OSND), Latent Semantic Analysis (LSA), Web traffic, Web host graph.

I. INTRODUCTION

Social Networks provide the means to explicitly create and manage connections based on information gathered and stored in user profiles. Social Networks and Semantic Social Networks [7] have emerged as a second generation of the mailing lists, Usenet, bulletin boards online communities, providing a number of services such as network of friends or business contacts listings, content-sharing, profile surfing, discussion and messaging tools. They are also part of the recently created new breed of user generated content aware technologies which have been encompassed by the “Web 2.0” buzzword umbrella and have turned up to provide a huge amount of metadata and information about the user as a particular entity.

However, these applications are not addressing fundamental problems of information overload, such as email hoarding or lack of management, but contributing to increase the burden. On the other hand, efforts such as [5] and [6] are under way to examine email filtering and ranking based on social networks. In addition, semantic technologies are evolving to a more mature state in which ontology [1], its backbone technology; provide a formal representation of a domain. The shift enabled by the use of machine understandable ontology can outperform the current endeavors that require finding data spread out across the Web or dynamically drawing inferences which are continually hampered by their reliance on adhoc data frameworks. A Google-powered Open Social based strategy in the context of Social Network user information is presented.

Open Social is an application programming interface to build social applications across the Web, in other words, a common set of APIs for social applications across multiple websites. Open Social is currently being developed by Google in conjunction with members of the web community. The ultimate goal is for any social website to be able to implement the APIs and host 3rd party social applications. There are many websites implementing Open Social, including Engage.com, Friendster, hi5, Hyves, imeem, LinkedIn, MySpace, Bebo, Ning, Oracle, orkut, Plaxo, Salesforce.com, Six Apart, Tianji, Viadeo, and XING [3]. Open Social is not a social network itself; rather it is a set of three common APIs that allow developers to access the following core functions and information on social networks:

- People and Friends data API
- Activities data API
- Persistence data API

The Open Social Network Dataset (OSND) is a lightweight ontology used for collaborative data filtering and rating in which we follow an integrated approach of combining three types of techniques for improving its construction from the tag sets gathered from the afore mentioned Web 2.0 social networks such as Engage.com, Friendster, hi5 etc., The three techniques we are applying are as follows:

- Applying the Vector Space Model:
- Using Latent Semantic Analysis (LSA)
- Validating the set of terms pertaining to the OSND with online lexical resources, such as Wordnet1.

II. RELATED WORK

Many studies have used Web crawlers to reveal important insights on the large-scale *structure* of the Web graph, such as the "bow-tie" model, the presence of self-similar structures and scale-free distributions, and its small-world topology [2, 4, 1, 6, 5]. While these insights have informed the design of a variety of applications such as crawlers and caching proxy servers, structural analysis has seen its greatest application in ranking pages returned by search engines. In particular, the well-known PageRank and HITS algorithms are able to use the pattern of links connecting pages to rank them without needing to process their contents; these algorithms have inspired a vast amount of research into ranking algorithms based on link structure [8, 9].

The structural properties of the link graph extend to the host graph, which considers the connectivity of entire Web servers rather than individual pages [10]. The earliest efforts have used browser logs to characterize user navigation patterns, time spent on pages, bookmark usage, page revisit frequencies, and overlap among user paths [11, 12]. The most direct source of behavioral data comes from the logs of Web servers, which have been used for applications such as personalization and improving caching behavior [13]. Because search engines serve a central role in users' navigation, their log data is particularly useful in improving results based on user behavior [14, 15].

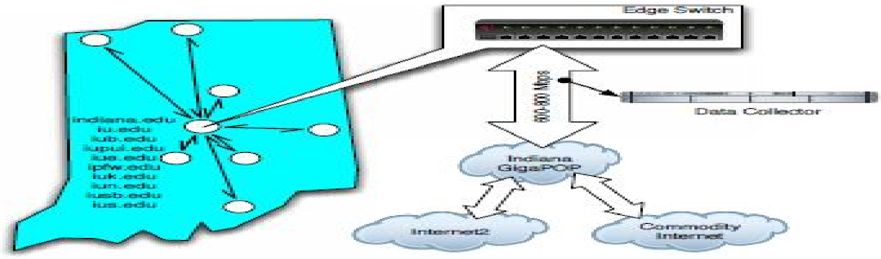


Figure 1 Sketch of University's Internet connectivity.

Ranking Web pages and sites is one of the most critical tasks of any search engine. The last decade brought terrific advances in Web search technology, owing in large part to the development of sophisticated ranking techniques. While modern search engines have likely refined and improved on Page Rank, in addition to combining it with many other criteria, it remains a reference tool for the study of the Web as a complex dynamic network, as well as for the engineering of improved ranking functions.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 6, June 2014

III. EVALUATION

In principle it is possible to capture the entire URLs of the referring and requested pages with our experimental setup, and to build a weighted link graph with pages as nodes. This is indeed our goal. In this Paper, however, we report on an initial stage in which we focus on the host graph. One reason is that this is more feasible with our current storage and computing resources, and indeed necessary to tune our collection and analysis algorithms; another is that the host graph already reveals several interesting insights about Web traffic. The web host graphs are stored as sparse connectivity matrices for analysis in Matlab. Node size is proportional to the log of the traffic to each site, and edge thickness is proportional to the log of the number of clicks on links between two sites.

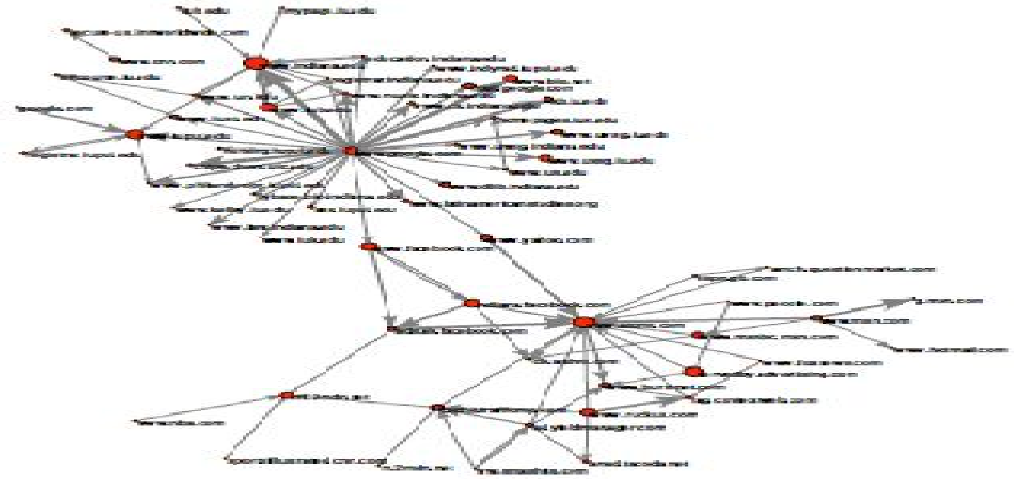


Figure 2.Visualization of the most requested hosts and clicked links.

A. *Structural Properties*

The click data was collected over a period. Table 1 offers a view of a small portion of the resulting weighted host graph, consisting of the most popular destination sites and the most clicked links between them.

		FULL		HUMAN	
		Number	Percent	Number	Percent
Requests	With empty referrer to unknown destination total	2,632,399,381 232,147,862 12,884,043,440	20.4% 1.8%	490,290,850 2,078,725 907,196,059	54.0% 0.2%
Hosts	Referring destination total	5,151,634 7,026,699 7,595,907	67.8% 92.5%	2,199,307 3,743,074 4,031,842	54.5% 92.8%
Edges		37,537,685		10,790,759	

Table 1: Summary Statistics of the FULL and HUMAN host graphs

We first report on general properties of this data and on the structure of the weighted host graph. Each human page click involves an average of 14.2 HTTP requests for embedded media files, style sheets, script files, and so on. One notable observation is that a majority of human-generated clicks do not have a referrer page, meaning that users type the URL directly, click on a bookmark, or click on a link in an email.

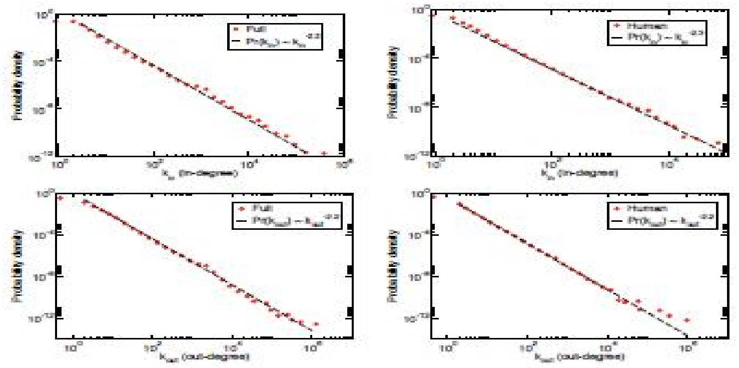


Figure 3: Distributions of in-degree (top) and out-degree (bottom) for the FULL (left) and HUMAN (right) host graphs.

The first question about the host graph reconstructed from our sample of traffic is whether it recovers the well-known topological features of the link graphs built from large-scale crawls [2, 4, 6]. The most stable signature of the Web graph is its scale-free in-degree distribution, which many studies consistently report as being well fitted by a power law $Pr(K_{in})$ $K_{in}$with exponent. As shown in Figure 3 the behavior is recovered from the FULL host graph (= 2.2 ± 0.1); although Web traffic may not follow on every link, it produces a picture of the Web that is topologically consistent with those obtained from large-scale crawls. The power-law in-degree distribution in the HUMAN host graph has a slightly larger exponent = 2.3±0.1. This hints at an important caveat. While the structure of the traffic-induced and crawler-induced networks may be similar, they are based on very different sampling procedures, each with its own biases.

One cannot compare the two networks directly on a node-by-node basis. To illustrate this point, nodes are sampled from the HUMAN graph and compared their in-degree with that given by a search engine (via the Yahoo API). As evident from the scatter plot in Figure 3, the correlation is weak (Pearson's R = 0.26 on the log-values), and we cannot assume proportionality. If one conjectures a power-law scaling $K_{in} \sim \widehat{K}_{in}^{\eta}$where $\widehat{K}_{in}$ is the in-degree obtained from crawl data, we see that a sub linear bias < 1 fits the data better than proportionality 1. While we cannot say that such a power-law scaling is the most appropriate model of the relationship, this does highlight a sample bias whereby the in-degree of popular nodes is underestimated by a greater amount than that of low-degree nodes. The lack of proportionality explains the higher exponent in the power-law distribution of in-degree. Assuming again that kin and ˆkin are deterministically related by the power formula conjectured above, it follows immediately that $Pr(K_{in})\mathrm{d}K_{in} = Pr(\widehat{K}_{in})\mathrm{d}\widehat{K}_{in}$. Therefore

$$\begin{aligned} Pr(K_{in})dK_{in} &\sim K_{in}^{-\gamma}dK_{in} \sim \widehat{K}_{in}^{-\eta\gamma}\, d\left(\widehat{K}_{in}^{\eta}\right) \\ &\sim K_{in}^{-\eta\gamma+\eta-1}d\widehat{K}_{in} \sim \widehat{K}_{in}^{-\gamma}\, d\widehat{K}_{in} \end{aligned}$$

and thus the K_{in} exponent changes to

$$\gamma = (\hat{\gamma}-1)/\eta + 1 > \hat{\gamma} \text{ if } \eta<1$$

The difference between our network representation of the Web host graph and that obtained from crawls, of course, is that we have weighted edges telling how many times links between hosts are clicked. For weighted networks, the notion of degree is generalized to that of strength, defined as the sum of the weights over incoming or outgoing links:

$$S_{in}(j) = \sum_i W_{ij} \; S_{out}(i) = \sum_j W_{ij}$$

WhereW_{ij} is the weight of edge (i, j), i.e. the number of clicks on the link from host i to host j. Note that because $S_{in}(j)$ represents the total number of times that site j is visited, this is what we refer to by the less formal term traffic.

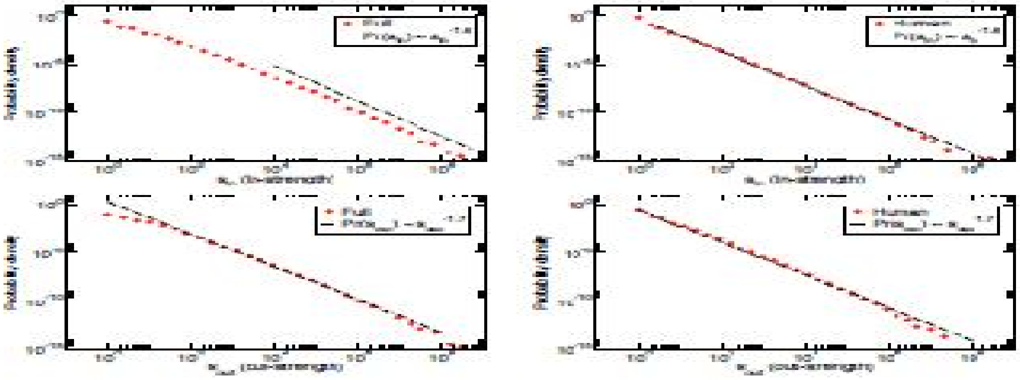
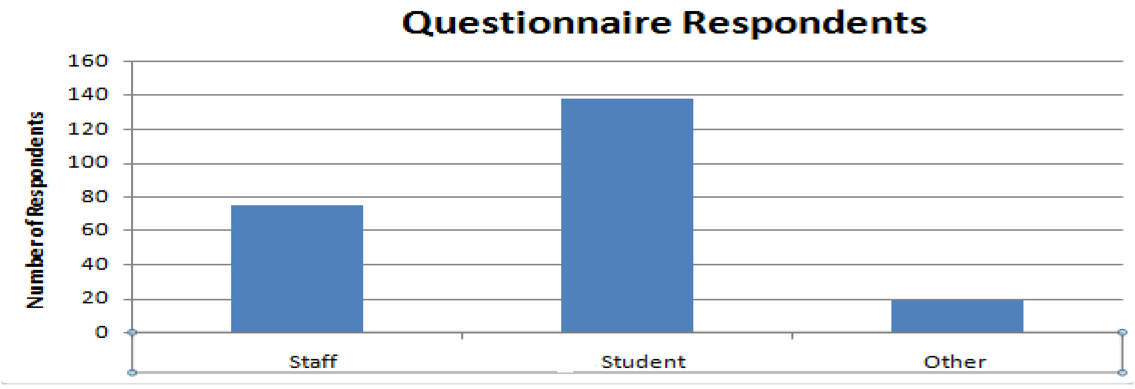


Figure 4: Distributions of in-strength (top) and out-strength (bottom) for the FULL (left) and HUMAN (right) host graphs.

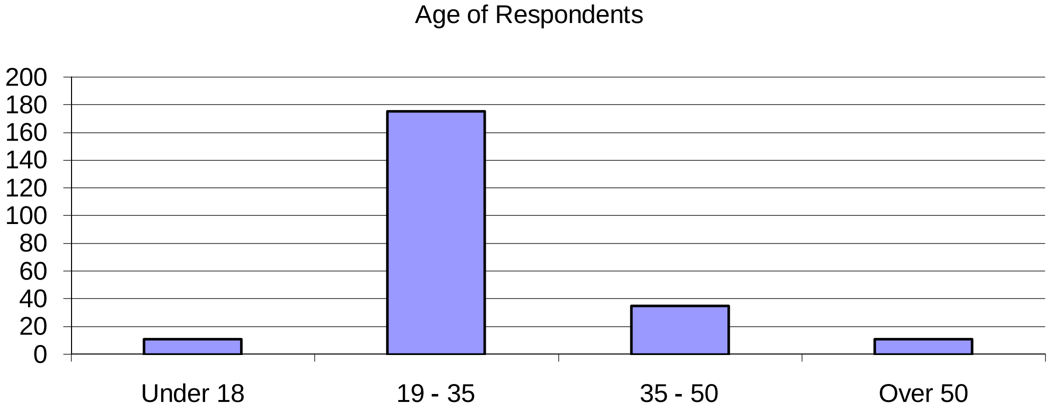
IV. RESULTS AND DISCUSSIONS

All comment sections have been left how they were entered on the web, most spelling errors. This will allow you to see the responses that are received; only a few irrelevant results have been removed. In total 232 people responded to the questionnaire, the majority being students making up 60% of the total respondents. 32% of respondents were staff and the remainder (8%) in the other category. The other category contained the following responses. Govt Agent, PhD students Research student, university alumni, site visitor, university Graduate applicant, Graduate, Parent of student, Prospective pg student etc

A. *Respondent Category*

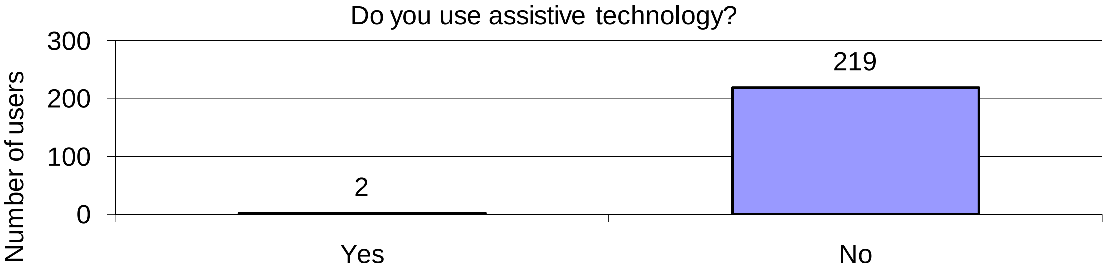


B. *Respondent Age*



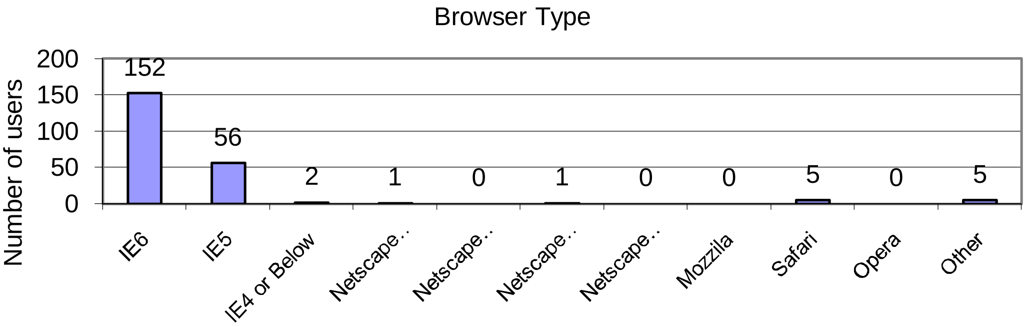
The majority of respondents (75%) were aged between 19 and 35, the second largest groups were aged between 35 and 50.

C. Assistive Technology:



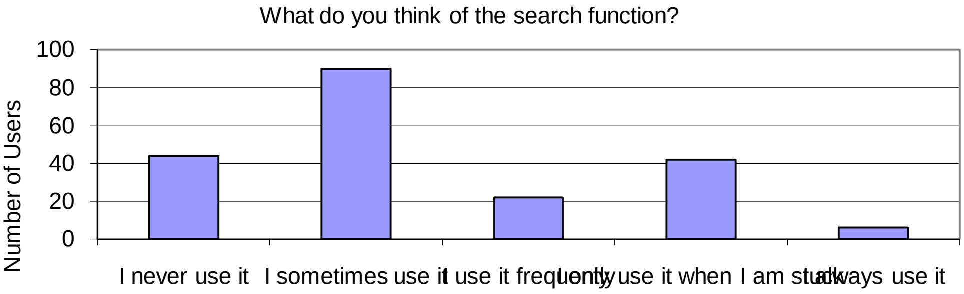
99% of respondents do not use assistive technology to access the Loughborough site, those who said they did offered poor explanations such as mouse.

D. Browser Software



Internet Explorer is the most popular browser software used to navigate our site.

E. Search Facility



The search function is not used at all by 22% of respondents, 43% use it sometimes but only 14% use it frequently or always use it.

V. CONCLUSIONS AND FUTURE WORK

An effort has been made to analyze Open social networks. The initial of my research is focused and presented in this paper in brief. Researchers have been quick to recognize that structural analysis of the Web can become far more useful when combined with *behavioral* data. Some paths through the Web are used far more heavily than others, and a variety of behavioral data sources exist that can allow researchers to identify these paths and improve Web models accordingly. The earliest efforts have used browser logs to characterize user navigation patterns, time spent on pages, bookmark usage, page revisit frequencies, and overlap among user paths Because search engines serve a central role in users' navigation, their log data is particularly useful in improving results based on user behavior. .However, these applications are not addressing fundamental problems of information overload, such as email hoarding or lack of management, but contributing to increase the burden.

REFERENCES

[1]. Fensel, D, Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce, Springer-Verlag. 2002.

[2]. University of California Police Department, You've Got Spam: How to Avoid Unwanted Email.

[3]. OpenSocial – Google Code official web site. http://code.google.com/apis/opensocial/

[4]. Joseph S. Kong, Behnam A. Rezaei, Nima Sarshar, and Vwani P. Roy chowdhury, Let Your Cyber Alter Ego Share Information and Manage Spam, 2005.

[5]. Golbeck, J. and Hendler, J. Reputation Network Analysis for Email Filtering, Proceedings of Conference on Email and Anti-Spam. Mountain View, California, USA, 2004.

[6]. Ankolekar A, Krötzsch M, and Vrandecic, D, 2007, The two cultures: Mashing up web 2.0 and the semantic web, Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07, ACM Press, New York, NY, 825-834.

[7]. Gomez, J.M. Colomo, R. Ruiz, B Garcia : A Semantics-based Social Network for Software Project, International Journal of Information Technology and Management, Special issue: Work Change in the Era of ICTs. 2007.

[8]. Deerwester, S. Dumais, Furnas, G. W. Landauer, T. K. Harshman, R, Indexing by Latent Semantic Analysis, Journal of the Society for Information Science 41, Issue 6. Pp 391-407. 1990.

[9]. Marsh, S. (1994), Formalizing Trust as a Computational Concept, PhD thesis, Department of Mathematics and Computer Science, University of Sterling.

[10]. Heyman, P. Garcia-Molina, H, Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems, Technical Report Stanford University, 2006.

[11]. Gomez, J.M. Colomo, R. Alor-Hernandez, G. Posada- Gomez, R. Garcia, A Search in the Eye of the Beholder: Using the Personal Social Dataset and Ontology-guided Input to Improve Web Search Efficiency, Proceedings of the 5th IEEE Latin-American Web Conference (LA-WEB07), Santiagode Chile, Chile. October, 31- November, 2nd 2007.

[12]. Christian Bird, Alex Gourley, Perm Devanbu, Michael Gertz, Anand Swaminathan. Mining Email Social Networks. MSR'06, May 22–23, 2006, Shanghai, China.

[13]. Ziegler, Cai-Nicolas, Georg Lausen (2004), Spreading Activation Models for Trust Propagation, Proceedings of the IEEE International Conference on E-Technology.

[14]. www.microsoft.com/mscorp/safety/technologies/senderid/default.mspx

[15]. Richardson, Matthew, Rakesh Agrawal, Pedro Domingos. (2003) “Trust Management for the Semantic Web,” Proceedings of the Second International Semantic Web Conference. Sanibel Island, Florida.