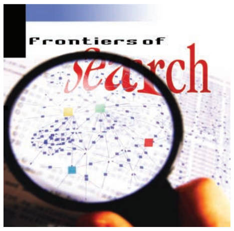
Searching Association Networks for Nurturers



Studying the evolution of association networks offers insights that researchers can use to develop new forms of Web information retrieval and improve searches. In addition to finding nurturers, this work can be applied to targeted recommendations, human resource management, and social network analysis.

Bharath Kumar Mohan
Indian Institute of Science

Searching the Web involves more than sifting through one huge graph of pages and hyperlinks. Specific association networks have emerged that serve domain-specific queries better by exploiting the principles and patterns that apply there.

A technique that searches these association networks and finds nurturers—early adopters—in them can be particularly effective. New nodes not only emerge around these nurturers, but also become important in the network. Finding nurturers can improve Web search, especially when answering *sticky* queries—persistent searches for which a user expects frequent fresh updates.

SEARCH IN SPECIAL CONTEXTS

One mantra summarizes progress in search engines: Keep queries simple and yet give fast and precise answers. Increasingly, search engines suggest query improvements to users (Google Suggest: www.google.com/webhp?complete=1&hl+en; search.yahoo.com); attempt to boost certain results based on user and community feedback (eurekster.com); or request the user to impose contexts on queries such as the Web, images, groups, news, directories, location, literature, or even people. Search contexts allow using customized rules and heuristics to improve results. For example, a search that seeks to

- "find Microsoft employees" can best be answered through a *people search vertical* such as ZoomInfo (zoominfo.com);
- "buy digital cameras" can best be answered by a shopping site that not only extracts summaries and photographs but also uses parameters like price, brand, and reviews to rank the results; and
- "rank research papers on Web search" can be done best by CiteSeer (citeseer.ist.psu.edu) or Google Scholar (scholar.google.com), based on the citations papers receive.

In each of these *search verticals*, exploiting the knowledge and insights specific to the domain improves results. ZoomInfo extracts affiliation information for people by mining their mentions on the Web and presents them all in one view. Comparison shoppers use product catalogs, sales figures, and reviews to rank products and opinions. CiteSeer uses citations to rank publications.

In this evolving path, researchers understand the user's query and its context, and use principles underpinning that domain to provide the best results. From a search engine's viewpoint, the Web provides an overlay of many association networks, each of which can answer different queries well.

Figure 1 shows some special association networks that can be extracted from the Web. Corporations, which hold and manage a few of these, still use them to enhance results from their search sites. A comparison shopper might seem to have nothing in common with a news aggregation service or scientific literature, but all three show a graph with entities and relationships between them. A comparison shopper draws upon a network of related products and the people who review them. A news aggrega-

Published by the IEEE Computer Society

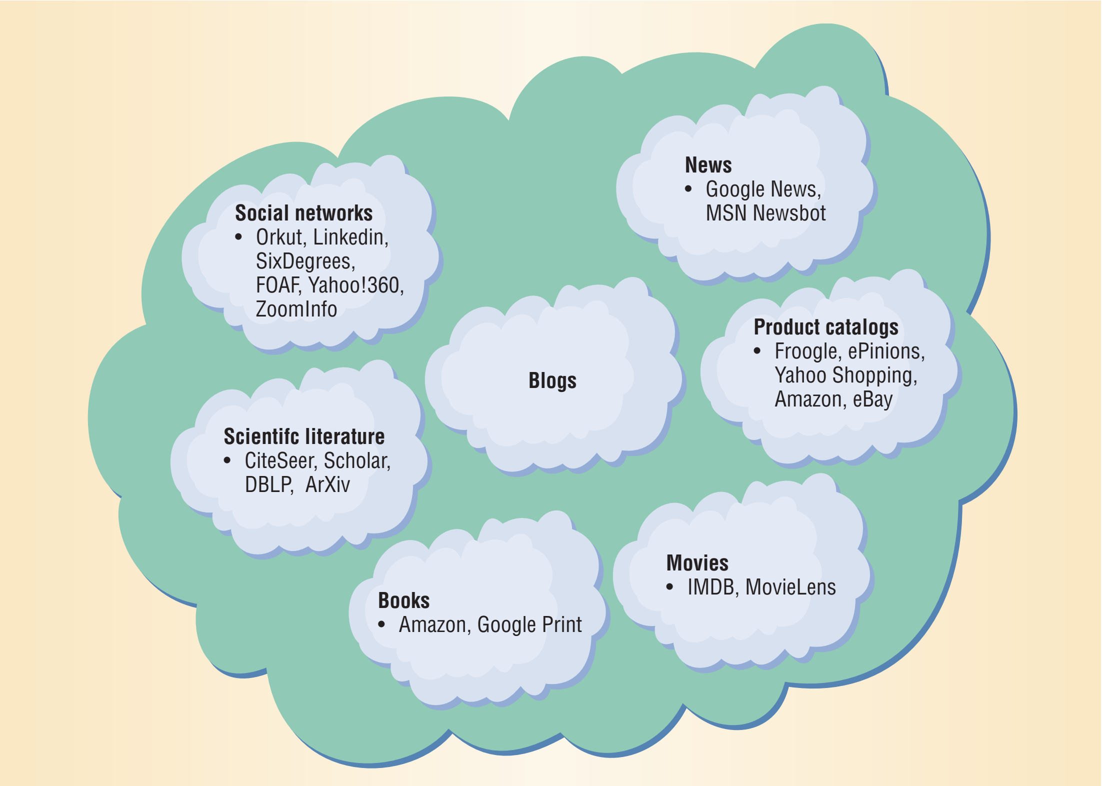

Social networks
• Orkut, Linkedin, SixDegrees, FOAF, Yahoo!360, ZoomInfo
News
• Google News, MSN Newsbot
Blogs
Product catalogs
• Froogle, ePinions, Yahoo Shopping, Amazon, eBay
Scientifc literature
• CiteSeer, Scholar, DBLP, ArXiv
Movies
• IMDB, MovieLens
Books
• Amazon, Google Print

Figure 1. Web association networks. Laid atop the Web, these networks provide access to aggregated information about the products and services available online.

tor finds similarities between news articles from different sources and organizes them into stories and topics hierarchically. Graphs of scientific literature portray coauthorship among researchers and citations between publications. To handle queries better, the basic page-hyperlink-page association network on the Web is specialized to more specific entities and relationships.

While these networks have widened the scope of Web search and improved its accuracy, they still fall short when handling sticky searches. These include the following:

- blogs on best practices in programming and design,
- promising and upcoming technologies and products,
- news scoops,
- recent graduates in the Web search area, and
- new movies.

Link-analysis-based ranking techniques[1,2] can't be used for such queries because fresh artifacts in association networks seldom have many in-links at first. The past record of their early adopters in adopting good content can assist in ranking fresh content. Consider the following "successful" early adopters:

- *bloggers* that trigger widespread posts and reactions to best practices,
- *hub pages* that link to new technologies and products that become famous later,
- *news sources* that trigger major exposés,
- *professors* who have nurtured the best students in Web search, and
- *good reviewers* whose positive opinions have preceded hit movies.

When new blogs, products, news scoops, graduates, or movies appear, they can be ranked based on the quality of their early nurturers. Blogs written by people who have a history of starting extensive discussions can be ranked higher than others. Movies that have gotten good reviews from critics whose previous reviews have heralded other hits can be thrust into the limelight.

The artifacts that good nurturers adopt rank higher than others. Nurturers are not ranked based on the number of artifacts that link to them but by the performance of their adoptees. Insights into the nurturers in an association network lie in its evolution, not in a static snapshot.

NURTURERS IN ASSOCIATION NETWORKS

Nurturing, a pervasive mammalian trait, naturally extends to most association networks that involve humans. Until now, nurturing has been only a soft phenomenon studied by sociologists. The increased availability of digital and online data about associations lets researchers experiment with algorithms to gain insight into such phenomena. In any given association network in which *association events* happen at discrete instances, many provide contexts for nurturing. Consider the following:

- *Slashdot endorsement*. Slashdot was not the first site to link to Firefox, but the publicity Firefox received from the association surely helped it become popular quickly.
- *A VC seed-funding a new startup*. This event has a high nurturing value if the startup's valuation increases rapidly after the funding.
- *A blogger writing about a topic*. Kim Cameron has nurtured the "Laws of Identity" topic if it later becomes the buzz in blog circles.

A nurturer need not always be the innovator or originator. The evangelist who adopts a prodigal idea and launches it on its way to success also can be a nurturer, as the "Nurturers in the Web" sidebar shows.

The effect and value of such nurturing is perceived only in the *postassociative period*—after the association event. An entity can be nurtured by many over time, and the earliest nurturers do not necessarily provide the greatest benefit.

While it's possible to go wrong in predicting that any single event will have a nurturing effect, an entity can be deemed a good nurturer if it can repeat its impact. A one-off Slashdot-Firefox event might not prove that Slashdot is a nurturer, but if Slashdot provides the same level of recognition to many other sites, the case for its nurturing ability grows proportionately stronger. Devising a good measure of success and a way to credit those who nurture will reveal the best nurturers in each association network.

Nurturers in the Web

Finding what makes a Web site like Slashdot (slashdot.org) different from CNN (cnn.com) can be informative. Although the two sites might have comparable Web rankings, a news item initially posted on Slashdot is not as well known and neither is the person who posts it. The community rates the news as it moves up in popularity and becomes the buzz. Slashdot's success has not been in reporting breaking news on current affairs or highly credible information, but in posting almost exclusively fresh news that no other dissemination channel has access to. The key to this lies in the common people who provide Slashdot's information.

On a Web site like CNN, news often goes through many rounds of editing, must be approved by authorized people, and can deal only with acceptable topics. When linking to a page, an article on CNN essentially focuses on the already famous.

In contrast, early adopter Slashdot provides a site at which people often post or learn about something new and interesting. Slashdot even has a phenomenon named after it: the Slashdot effect (http://en.wikipedia.org/wiki/Slashdot_effect). This term refers to the sudden burst of Internet traffic that a site reported on Slashdot often receives. This traffic can, ironically, flood the site and make it temporarily inaccessible.

Slashdot does not spotlight the same Web page all the time. Given Slashdot's nature, the focus will change to another Web page sooner or later. Occasionally, a Web page will sustain this initial focus because other sites find it interesting and link to it, thereby increasing its rank. This linking surge makes the Web page appear higher in search engine results, which in turn increases its popularity. A site like Slashdot can thus serve as a launch pad that nurtures Web pages and makes them popular.

While Google (google.com) generates numerous external referrals to most Web pages on the Internet, the probability of a given page being accessed is directly proportional to its popularity, given that more popular results appear higher in the search results. Consequently, Google can make famous sites more famous, while a site like Slashdot nurtures little-known sites through word-of-mouth. Slashdot is indeed a nurturer in the Web.

But it is not alone. Freshmeat (freshmeat.org), ZDNet (zdnet.com), and many other sites also nurture less-well-known sites. In the age of blogging, even the blogs that some celebrities maintain have a Slashdot effect: Their recommendation of a Web site often leads to increased linking to that site in other blogs, Web-based bookmarks, and personal homepages.

NURTURER DISCOVERY FRAMEWORK

Researchers can mine every association to extract indicators that hint at the extent of nurturing activity. They can use rules specific to the association network to identify nurturers, nurturees, and the significance of the event to the parties involved. All association events have their time stamps as well.

Association significance

The set of nurturers and nurturees could be bipartite (bloggers and blog topics, reviewers and movies) or unipartite (coauthorship networks in which one author nurtures another). A single association event can have many pairs of nurturers and nurturees, and these will differ according to their context.

On the Web, when a page P puts up a link to another page Q, a crawler normally detects it quite soon, say at time t. (The crawl time might not be the most accurate measure of the event's time, but it's the best we can get. It helps to use an intelligent crawler.) The search engine's ranking algorithm then computes a certain weighted vote x that P gives to Q. P is a nurturer, Q the nurturee, and the event has a significance of x. The weighted vote might be as simple as the inverse outdegree of P, or an algorithm like PageRank could compute it.[1]

In the blog context, when blogger B writes a blog pertaining to a topic T, possibly inferred using techniques such as named-entity extraction and inverse-document frequency,[3] with an estimated importance x given to T in the blog (using a Term-Frequency-Inverse-Document-Frequency-like measure), B becomes the nurturer, T the nurturee, and x the event's significance.

In scientific coauthorship, every publication has a coauthor list of, say, size n. Any author can be a potential nurturer for the other. The significance of the association between every pair can be taken as $s/{}^{n}C_2$, since there are ${}^{n}C_2$ possible associations, with *s* denoting the significance of the publication, the number of citations it received, or the impact factor of the conference or journal in which it appeared.

Nurturing influence

An entity can be nurtured by several nurturers, with the significance of nurturing higher early in the entity's life cycle. For example, a venture company's seed funds are more useful while bootstrapping a company than when investing the same amount in the post-IPO phase.

In an association event, the *nurturing influence* a nurturer has on a nurturee depends on the significance of the association event, its earliness in the nurturee's life span, and the nuturer's nurturing history up to that point.

If, in an association event *c*, *p* is the nurturer and *q* the nurturee, we could use models like

$$ {}^{p}ni_{c}^{q} = \frac{{}^{p}sig_{c}^{q}}{sig^{q}} $$

where ${}^{p}sig_{c}^{q}$ is the significance of the event *c* to *q* involving *p*, or

$$ {}^{p}ni_{c}^{q} = \frac{{}^{p}sig_{c}^{q}}{(sig^{q})^{0.5}} $$

In general,

$$ {}^{p}ni_{c}^{q} = \frac{{}^{p}sig_{c}^{q}}{f(sig^{q})} $$

where *f* decides the bias toward earliness that the nurturing influence gets.

Consider a publication with three coauthors, {*a*, *b*, *c*}. In practice, *a* could be a student, *b* another student, and *c* a professor. From the preceding measure, both *b* and *c* will have the same nurturing influence on *a*, which does not reflect reality. We can expect that *c*'s nurturing influence on *a* is much greater than *b*'s. To introduce a bias, the individual nurturing influence is weighted by the nurturer's estimated nurtureship—a measure of that person's nurturing history up to this point. Thus,

$$ {}^{b}ni_{c}^{a} = N_{b} * \frac{{}^{p}sig_{c}^{q}}{f(sig^{q})} $$

and

$$ {}^{c}ni_{c}^{a} = N_{c} * \frac{{}^{p}sig_{c}^{q}}{f(sig^{q})} $$

For now, we assume that N_b and N_c, the nurtureships of *b* and *c*, can be computed. In general,

$$ {}^{b}ni_{c}^{a} = g(N_{b}) * \frac{{}^{p}sig_{c}^{q}}{f(sig^{q})} $$

In the Firefox example, suppose Firefox received the following links at different times:

- Mozilla links to Firefox at t_1, with a vote of 10. Mozilla has a nurtureship of 30.
- A blogger links to Firefox at t_2, with a vote of 1. The blogger has a nurtureship of 1.
- Slashdot reports Firefox at t_3, with a vote of 5. Slashdot has a nurtureship of 150.

At times t_1, t_2, and t_3, Firefox has a cumulative significance of 10, 11, and 16, respectively. Assuming $f(sig^q) = sig^q$ and $g(N_b) = N_b$, the nurturing influence of Mozilla on Firefox is 30 × 10/10 = 30, the blogger's influence is 1 × 1/11 = 0.09, and Slashdot has an infuence of 150 × 5/16 = 46.875. We expect Slashdot to have a greater nurturing influence on Firefox given its history.

Tributes

In Indian tradition, upon completion of his studies, a disciple customarily offers a tribute to his teacher, or *guru*, thanking that person for the nurturing provided. This is termed *Gurudakshina*. Here, we extend this principle to credit nurturers for their influence on their nurturees. A nurturee, on being successful, pays tribute to all contributing nurturers in proportion to their nurturing influences.

Likewise, every time Firefox receives new votes from other sites, it acknowledges its nurturers by paying tribute to them. Suppose that Firefox received a new mention from Google, with a vote of 20. This success is credited to all the past nurturers in proportion to their nurturing influences. Thus, from this event, Mozilla receives 20 × 30/ (30 + 0.09 + 46.875) = 7.8 credits, and the blogger and Slashdot receive 0.02 and 12.18 credits, approximately. These tributes all derive from the Google linking event alone. Mozilla would also receive tributes from the blogger and Slashdot linking events.

If *q* gets some new success *s* in the association network, the tribute given to a past nurturer *p* ($p \neq q$) is

$$ {}^{q}t_{s}^{p} = s * \frac{{}^{p}ni^{q}}{\sum_{r} {}^{r}ni^{q}} $$

A nurturee pays tribute to all contributing nurturers in proportion to their nurturing influences.

where $^{p}ni^{q}$ is the cumulative nurturing influence p has had on q over all past association events. Note that q does not give tribute to itself. The nurturee can choose to give away the entire success s or a function $h(s)$ as suits the domain.

Nurtureship

Every nurturer collects the tributes from all its nurturees, accumulating them to make its nurtureship, N_p. The nurturees pay tributes iteratively as events unfold, with the updated values used in computing nurturing influences thereafter. The computational cost of paying tribute is linear in the number of nurturers for the nurturee.

Mining among associations in bibliographic databases remains the best context in which to look for nurturers in research.

BEST NURTURERS IN RESEARCH

Collaboration in research involves people with different traits. A few in the group might be young and energetic, willing to put in the extra hours to make an idea or a system work. On the other hand, experienced people can give direction on the most important aspects of the innovation and provide appropriate feedback on its capabilities and limitations.

The right mix of enthusiasm and wisdom culminates in a publication—a reviewed and accepted statement of the research effort. Thus, in the research community, experienced researchers seek enthusiastic, hard-working, intelligent students, while students in turn seek good nurturers through whom they can learn and also benefit by being bootstrapped into a good research network. This results in a social selection of sorts that tests both the experienced researcher's ability to identify talent and the young researcher's ability to identify the best nurturers.

Selection process

Although the experienced rely upon selection processes that test different aspects of students' mental prowess, the students rely on university rankings and professors' publication records and funding health. Word-of-mouth gains importance, with the professors valuing the recommendations made by other trusted sources and the students relying on past students' assessments and anecdotes.

The only way to make a professor's task easier is to evolve a global system of trust and evaluation—which may be impractical. However, students can benefit through statistical approximations that characterize their experiences during and after their association with professors.

Given the availability of associations (publications with coauthor information) and citations (a measure of checking how popular and influential a publication has been) in bibliographic databases that also document the time of publication, researchers can use statistical techniques to mine for such social traits.

Although nurturing could happen within the confines of a classroom or even through well-written books, mining among associations in bibliographic databases remains the best context in which to look for nurturers in research because

- publishing is the de facto standard for evaluating good research;
- scientific reporting is best taught hands on—senior collaborators typically give direction on the innovation's most important aspects, provide appropriate feedback on its capabilities and limitations, and contrast it with other progress in the area;
- contributors toward a research project often become coauthors of the subsequent publication; and
- bibliographic databases are well documented and already used for extensive analysis of the research's impact—in addition to its use by the student community, the coauthorship graph formed in scientific literature provides an excellent testbed for studying the feasibility of finding nurturers.

The Digital Bibliography and Library Project (www.informatik.uni-trier.de) provides digital information on major computer science journals and publications, indexing more than 520,000 articles. DBLP citations can also be found for a subset of the articles indexed—mainly from the SIGMOD anthology. The DBL browser offers an interface to access the compressed database that contains the article information.

We used the nurturer heuristic to perform a DBLP analysis in two experiments. One experiment measured the success of researchers by the normalized sum of their publications, with every published author receiving a credit of $1/n$ if there were n coauthors. The other experiment measured the success of researchers by the number of citations their publications received.

Customizations to the nurturer heuristic

When customizing the heuristic, my colleagues and I debated on the measures of success. In the first experiment, a publication with n coauthors will give each coauthor a success of 1/*n*. We considered having a bias for first author, second author, and so on, but realized the difficulty of achieving a reasonable breakdown. Moreover, computer science does not universally follow the convention of ordering author names based on each one's contribution.

In the second experiment, we first computed the citations an article received, then attributed the success to each author as citations/*n*. The citation measure shows a clear bias toward older publications because new papers would not have reached their complete "citation potential" yet.

Next, we focused on nurturing influence. Most publication counts grow linearly over time. However, their nurtureship can grow faster. A good professor often continues to have good students, while his past students keep publishing too—all of which lets that professor's nurtureship grow quadratically.

Using

$$ {}^{b}ni_c^a = g(N_b) * \frac{{}^{p}sig_c^q}{f(sig^q)} $$

to compute nurturing influences, $g(N_b) = N_b$ and $f(sig^q) = sig^q$ can cause the bias toward big nurturers to outweigh earliness. To correct this, we use $g(N_b) = N_b^{0.5}$ and $f(sig^q) = sig^q$.

We consider researchers who authored many publications on their own early in their careers. These self-made researchers probably received little or no nurturing from others. They thus pay less tribute to their associates, as follows: $h(s) = s \times ed_p$, where $0 \le ed_p \le 1$, and ed_p is indicative of *p*'s early dependence. Thus,

$$ ed_p = 1 - \frac{\sum_c \frac{1}{n_c} * \frac{1}{sig_p^c}}{\sum_c \frac{1}{sig_p^c}} $$

where *c* is a publication, n_c the number of coauthors in *c*, and sig_p^c the cumulative success *p* has achieved after *c*. For example, consider the first three publications of Sergey Brin:

1. Sergey Brin, "Near Neighbor Search in Large Metric Spaces," VLDB, 1995.
2. Sergey Brin, James Davis, and Hector Garcia-Molina, "Copy Detection Mechanisms for Digital Documents," SIGMOD, 1995.
3. Sergey Brin, Rajeev Motwani, and Craig Silverstein, "Beyond Market Baskets: Generalizing Association Rules to Correlations," SIGMOD, 1997.

After the first publication appears, Brin's success rises to 1. For the second publication, Garcia-Molina is a bigger nurturer than Davis given his past nurtureship. Garcia-Molina's nurturing influence on Brin is

$$ (\text{association} = \frac{1}{3}) \times $$
$$ (\text{Brin's success' inverse after this publication} = \frac{1}{4/3}) \times $$
$$ (\text{Garcia-Molina's nurtureship} = 25^{0.5}) $$

Davis's nurturing influence on Brin is

$$ \frac{1}{3} \times \frac{1}{4/3} \times (\text{Davis's nurtureship} = 1^{0.5}) $$

Brin also has a nurturing influence on Davis and Garcia-Molina, albeit a small one. Brin's success for this second paper, provides no credit to anyone else because the first paper had only one author, so Brin has no past associates.

For the third paper, Brin's success, 1/3, will be given to Davis and Garcia-Molina. His early dependence is

$$ 1 - \left[\frac{1 + (1/3 \times 4/3) + (1/3 \times 5/3)}{(1 + 4/3 + 5/3)} \right] = \frac{1}{2} $$

Hence, Davis and Garcia-Molina receive only 1/2 × 1/3 as tribute, proportionate to their nurturing influence on Brin.

Next, we calculate Motwani's nurturing influence on Brin as

$$ (\text{association} = \frac{1}{3}) \times $$
$$ (\text{Brin's inverse success} = \frac{1}{5/3}) \times $$
$$ (\text{Motwani's nurtureship} = 16^{0.5}) $$

So too with Silverstein. At this point, Garcia-Molina and Motwani are Brin's most significant nurturers.

If Brin becomes successful, Garcia-Molina and Motwani receive the most credit, while Davis and Silverstein receive proportionately less.

Nurturer ranking

We have applied this nurturer heuristic to the DBLP to generate author/nurturer rankings based on both publication count and citations, as well as drill-down lists of each author/nurturer showing a ranked list of their respective nurturees or collaborators. Another publication[4] covers in greater detail some of the subtleties involved in finding nurturers.

The Theoretical Computer Science Genealogy Project (sigact.acm.org/genealogy) offers a list of professors and their students in the field of theoretical computer science. This list can be used as a starting point for validating these results. The list shows a large set of viable nurturer and nurturee pairs. However, the data set does not allow detection of some other nurturing types—for example, many professors nurture students into successful careers in industry.

This list's main use is for finding researchers who have nurtured other careers in research. Refining the list to mention nurturers in specific areas of computer science also can be done easily by using a smaller set of conferences and journals relevant to the area.

Studying the evolution of association networks offers insights that can be used for developing new forms of information retrieval and for improving searches. Finding nurturers in research is just one application. Our DBLP case study revealed something new about the way the computer science community operates. The design principles in the algorithm helped differentiate between junior and senior researchers, identify the more influential nurturer between two collaborators, and accumulate credit iteratively. We expect new insights when applying these techniques to other data sets and expect this work to yield useful applications in information retrieval, Web search, targeted recommendations, human resource management, and social network analysis. n

Acknowledgments

This research was supported by the Infosys Fellowship program. We thank the DBLP for its service to the computer science community and also for making its data sets available for analyses.

References

1. S. Brin et al., *The Page Rank Citation Ranking: Bringing Order to the Web*, tech. report 1999-66, Stanford Digital Libraries, 1999; http://dbpubs.stanford.edu:8090/pub/1999-66.
2. J. Kleinberg, "Authoritative Sources in a Hyperlinked Environment," *Proc. 9th ACM-SIAM Symp. Discrete Algorithms*, ACM Press, 1998, pp. 668-677.
3. D. Gruhl et al., "Information Diffusion through Blogspace," *Proc. 13th Conf. World Wide Web*, ACM Press, 2004, pp. 491-501.
4. M.B. Kumar and Y.N. Srikant, "The Best Nurturers in Computer Science Research," CSA, IISc tech. report, 2004; http://archive.csa.iisc.ernet.in/TR/2004/10/.

***Bharath Kumar Mohan** is a doctoral candidate at the Indian Institute of Science, Bangalore. His research interests include information retrieval, social network analysis, and social software. He received an MS in computer science from the Indian Institute of Science. Contact him at mbk@csa.iisc.ernet.in.*