Creating Models of Real-World Communities with ReferralWeb

Henry Kautz
AT&T Laboratories
kautz@research.att.com

Bart Selman
Cornell University
selman@cs.cornell.edu

Most recommender systems generate recommendations based on anonymous opinions. This may be fine if the recommendation is about sometime minor, such as which movie to see or which CD to buy. However, when a person is making critical decisions about his life or his business, he does not want anonymous opinions. Instead, he wants the advice of a trusted expert. If he does not personally know such an expert, then he wants a personal reference to such an expert via a series of trusted friends and colleagues.

When such a "referral chain" can be found, then the person with the question has a good reason to believe the expert's advice, and the expert has reason to respond to the questioner in a trustworthy manner.

Finding experts by referral chaining can be time-consuming process, particularly if a person is not well-connected himself. We have created the ReferralWeb system (Kautz, Selman, and Shah 1997) as a tool to aid in finding experts on arbitrary topics. The set of all possible referral chains in a community is called a "social network" (Milgram 1967, Wasserman and Galaskiewicz 1994). ReferralWeb lets a user visualize and search such a social network.

How can one generate a social network based on a community of interest? Systems such as Firefly (http://www.firefly.net) and 6DOS (http://www.q-and-a.com/) require everyone to explicitly register with the system and to provide information about their interests and colleagues. Again, this may be fine if you really want to be restricted to finding experts among the set of people who are highly motivated to join that particular online community. In general, however, you want to find experts who are members of existing "real world" communities. Furthermore, the best, busiest experts are probably the least likely ones to bother to register with any kind of "expert locator" service.

Therefore it is necessary to generate social networks without the explicit cooperation of any of the people who appear in the networks. ReferralWeb does this by mining information from publicly available documents.

We are experimenting with different ways to determine the "links" between people using public documents, using different kind of (off line) spiders to populate social networks.

- In scientific communities, the "co-author" relationship is a reliable indicator of connectedness. One of our spiders downloads and parses bibliographic databases. Our current demonstration system contains a network of about 10,000 computer science researchers in the areas of AI, natural language processing, and theory.
- In corporate settings, relationships can be determined from databases of the organizational structure, past project teams, and internal publications.
- For the "rest of the world", we have built spiders that determine relatedness based on frequency of co-occurrence of names in the entire WWW. The spiders perform selective web-crawling to generate lists of potential pairs, and then query full-text Web indexes such as Altavista to gather precise statistics.
- Email logs provide a very good indicator of relatedness (Schwartz and Wood 1993). However, because of privacy concerns we do not use such records. In fact, our experiments with an early version of our system that did use email information (Kautz, Selman, and Milewski 1993) ended when the test subjects refused to allow any program to analyze their email, despite the elaborate security measures we had built into the system.

In addition to revealing "who knows who", a useful model of a community should include "who knows what". Manually-entered profiles of expertise are inevitably incomplete and out-of-date. Therefore, ReferralWeb also automates the process of gathering information on each individual's areas of expertise. We have implemented two different strategies for finding experts. The first is to generate a database of expertise at the same time the social network model is created. For example, in the case of the network of computer scientists mentioned above, the expertise database associates individuals with the titles of all papers they have written. There are currently about 32,000 papers listed in the system. The SMART information retrieval engine (Buckley 1985) can be used to answer such queries as, "Find an expert on error correcting codes who is within 3 links of myself".

The second strategy is to actively search the webs for experts at the moment the user queries the system. This strategy is slower, but more general and always up to date. ReferralWeb converts a query like the one just mentioned into a set of queries to Altavista, each of which is the conjunction of the name of person and the query topic. Altavista can determine how frequently the name and the topic are associated on the web, and these statistics are normalized and then used as an indicator of expertise.

A demonstration version of ReferralWeb can be accessed by following links from

http://www.research.att.com/~kautz.

The demo includes the computer science researcher network described above. One use is of this version of the system is as a "reviewer finder" for people organizing workshops and editing journals. We are also developing versions for use within AT&T that will be targeted at improving intra-corporate communication.

References

Buckley, C. (1985). Implementation of the SMART information retrieval system. Department of Computer Science, Cornell University, TR85-686.

Kautz, H., Selman, B., and Milewski, A. (1993). Agent Amplified Communication.

Kautz, H., Selman, B. and Shah, M. (1997). The Hidden Web. *AI Magazine*, 18(2) 1997, 27 — 36.

Schwartz, M. F. and Wood, D. C. M. (1993). Discovering shared interests using graph analysis. *Comm. ACM*, 36(8) 1993, 78–89. *Proceedings of AAAI-96.* MIT Press, Cambridge, MA, 1996, 3–9.

Wasserman, S. and Galaskiewicz, J., Eds. (1994). *Advances in Social Network Analysis.* Sage Publications, Thousand Oaks, CA, 1994.