On Webfeed Aggregators and Social Navigation

Brian M. Dennis
Computer Science Department, McCormick School of Engineering
New Media Program, Medill School of Journalism
Northwestern University
bmd@cs.northwestern.edu

September 22, 2004

On the Web, more and more frequenlty updating information sources publish changes in a small set of ad hoc, de facto, XML based formats which I'll collectively term *webfeeds*. These webfeeds enable applications, *webfeed aggregators*, to automatically monitor, retrieve, and present the dynamic information. This makes it easy for users to track a large number of web information sources.

Webfeed aggregators, increasing in popularity, are simplistic in helping end users make sense of a pool of feeds, being at the sophistication of e-mail or USENET news readers. The goal of my NusEye project is to experiment with the application of social network analysis techniques to support social navigation [1] in helping users makes sense of a torrent of information

Paralleling the growth of weblogs as a self publishing platform, has been the rising popularity of attendant content syndication and aggregation tools. At the producer end, most weblog tools support de facto syndication standards (RSS [2], Atom [3]) out of the box. For content consumers, aggregation applications and services ease the task of staying on top of a large number of weblogs. The syndication mechanisms are not strictly weblog based either. Traditional news producers such as The New York Times, Yahoo/AP News, and ABC also participate in this ecology. As well, different forms of dynamic information, such as the results of standing search queries or wiki page changes, are being published as webfeeds. Current trends seem indicate these formats becoming dominant for distributing change information on the Web, and thus aggregators becoming central in monitoring and managing dynamic information on the Web.

Webfeed aggregation will become an increasingly interesting venue for CSCW researchers as more users and more content sources join the webfeed ecology. Social network analysis and social navigation techniques will be useful in helping users deal with information overload in this environment.

Applying Social Network Analysis Techniques to Public Blogrolls

Recently, Dave Winer, arguably the father of the current syndication infrastructure, implemented a Web service that collects feed subscription information from self identifying users. Called "Share Your OPML" [4] (SYOPML), after the file format for describing subscriptions, the service has induced hundreds of users to upload lists of their subscribed feeds. A metaindex of the lists that are publicly available is published by the SYOPML service. I have written a simple robot that downloads the SYOPML index, checks for new and updated subscriptions, and maintains a database of these public subscriptions. The results of these crawls will be made available to other researchers.

A key observation is that a large number of syndicated content readers are openly providing subscription information. SYOPML not only has a number of

weblog publishers but a significant number of "just plain folks". This is in contrast to other popular socially based applications such as e-mail, instant messaging, or USENET, where it is extremely difficult to get social relation information without corporate assistance. The effect is not limited to SYOPML. The popular Web hosted aggregator Bloglines allows users to make their blogrolls publicly available and many users have done so. Similarly, the `del.icio.us` social bookmarking service [5] allows users to subscribe to other users and *tags*, ad hoc labelings of urls. `del.icio.us` subscriptions are publicly available and easily discoverable.

Subjecting such public data to social network analysis techniques can lead to a better understanding of how subscribers, feeds, and content interrelate. Such analysis can also support providing social clues within webfeed aggregators. In contrast to easily calculated statistics (e.g. node degree distributions) that lead to simplistic results (e.g. the graph exhibits a power law of coefficient gamma), network analysis can provide subtler insights. Using the collected subscription data, I have generated corresponding social network data and examined it through a combination of standard network analysis tools (Pajek [6]) and hand written code.

The SYOPML data can be used to construct what is known as an affiliation network. In an affiliation network, nodes are separated into two classes, actors and events. For my purposes, subscribers are the actors and news feeds the events. Under this definition, the network graph is bipartite. Sociology researchers such as Wellman [7] , Borgatti and Everett [8], Faust [9] and others have developed a wealth of techniques for interpreting and analyzing such affiliation networks.

Faust describes a number of centrality measures appropriate for analyzing affiliation networks, including degree, eigenvector, closeness, betweenness, and flow betweenness. Besides centrality measures, the network can be examined for particular roles and positions. Roles are patterns of network structure that actors are embedded in. Positions are partitions of the graph nodes based upon similiarity of nodes' local network structure.

A challenge for this work is that the network analysis software known to me for discovering roles and positions (UCINET, BLOCKS, Pajek) is unusable for graphs on the scale of our network: over 900 subscribers, and over 29000 webfeeds as of this writing. Note that such numbers are relatively small for a reasonably popular Web based service.

To surmount these obstacles, I have taken two approaches to make the analysis feasible. First, I have employed the CLUTO [10] clustering toolkit to find interesting partitions of the network nodes, looking for significant network positions. Each node is assigned a number of network based attributes such. Then this data is processed using the CLUTO toolkit. CLUTO uses optimization based clustering in partitional, agglomerative, and graph based schemes. Since CLUTO is designed for high dimensional datasets and large numbers of instances, the toolkit generates partitions on a reasonable time scale.

Second, examining the 1-mode versions of the actors and events network indicates the domination of a small number of feeds becomes apparent. The top twenty feeds, ranked by number of subscribers, reach over ninety percent of the community. Removing these twenty feeds allows tools like Pajek to start revealing hidden clusters within the networks.

Our initial efforts indicate that fine grained social clues can be teased out of the network data. How might these clues be employed for improved aggregation services? This information can be used to augment content based analysis of feed. I have prototyped some visualizations of content clustering applied to a collection of news feeds. This is done using CLUTO and pycluster [11], a toolkit which implements a technique that lends itself to visual display, Self-Organizing Maps [12] . The clustering results were annotated with information from our network analysis of the SYOPML affiliation network.

Despite the fact that members of this community rarely interact with each other directly, there is a social structure of syndication feeds that can be discerned. This structure can provide social navigation clues to assist users in dealing with information overload. Such clues could be incorporated into feed ag-

gregators and increasingly important "meta-weblog services", which appear to use rather unsophisticated network analysis techniques.

NusEye

In a recent paper [13], authored with my graduate student Azzari Caillier Jarrett, the application of social network analysis, along with graph visualization and interaction, for navigating syndicated web content, a.k.a webfeeds, was presented. A key approach was to apply network analysis to content item and content source relationsips in addition to analysis of traditional human networks. Three social networks were used to to generate interactive graphs in particularly useful visual styles. The results of a small initial user study were presented indicating initial promise for our approach.

One serious criticism of our approach is that network visualization may not be the most effective means by which to present social navigation clues. First, as Herman [14] et. al. point out, network visualization is appropriate for certain cognitive tasks, which may not be applicable in this domain. Network visualizations can also take up significant screen real estate, which is an issue in that we intended these displays to be secondary information displays. Further, for complex networks the visualizations are difficult to construct such that they are comprehensible.

We are committed to further experiments to determine the feasibility of using such network displays because, despite our minimal efforts, users were favorably attracted to such visualizations. Thus, if the above criticisms are invalid or can be overcome, we may see a high rate of adoption. In short, we would not have an initial selling hurdle to engage users.

In parallel with these efforts, we are building hosted analysis services that actually deliver their results as webfeeds. The proposition is that users provide us with a blogroll, our service analyzes the blogroll individually, and within the social context of other blogrolls we have access too. We then provide a custom webfeed for the user to monitor. This approach provides a number of benefits:

- We know that every webfeed aggregator understands webfeeds. Thus our services will automatically support any aggregator, regardless of platform.
- Services deploy incrementally, upgrade in place and instantaneously propagate to all users.
- We can do content analysis of the subscribed webfeeds alongside the social network analysis.
- We can provide web based collaboration tools, ala `del.icio.us` centered around webfeed subscriptions.
- Due to centralization, we can easily track the popularity and usage of various features that are provided to end users.

In short, we are building infrastructure that allows us to run interesting user experiments with social navigation mechanisms in the webfeed ecology. We anticipate network analysis techniques to be a core part of the backend of our services.

The only major limitation of this approach is the very limited "user interface" that webfeed formats provide. While many desktop aggregators embed a Web browser control, across a broad range of platforms, we can probably only rely on controlled rendering of text, images and links. In fact, links may be our only means of interacting with the user, although there is some hope that usage of JavaScript may be viable. Still, Web applications such as GMail, Flickr (acknowledging the heavy use of Flash in Flickr), and `del.icio.us` indicate that popular, interesting, and social applications can be constructed out of those limited elements.

At worst, we will discover that desktop level interaction is needed to make the use of these social mechanisms viable. Our suspicion though is that a small amount of social navigation information, straightforwardly displayed can have significant utility and impact.

Conclusion

Webfeed aggregation will become an increasingly interesting venue for CSCW researchers as more

users and more content sources join the webfeed ecology. Social network analysis and social navigation techniques will be useful in helping users deal with information overload in this environment. My NusEye project has made an initial foray into adding social navigation to webfeed aggregation, and we will be continuing to pursue this work. I would be glad to join with other CSCW researchers to share ideas, techniques, and approaches.

References

[1] K. Höök, D. Benyon, and A. Munro, eds., *Designing Information Spaces: The Social Navigation Approach*. Spinger-Verlag, 2003.

[2] D. Winer, "RSS 2.0 specification." `http://blogs.law.harvard.edu/tech/rss`, Dec. 2003.

[3] M. Nottingham, "The atom syndication forma 0.3 (pre-draft)." `http://www.mnot.net/drafts/draft-nottingham-atom-format-02.html`, Dec. 2003.

[4] D. Winer, ""Share Your OPML!" a commons for sharing outlines, feeds, taxonomy." `http://feeds.scripting.com/`, Jan. 2004.

[5] J. Schachter, "`del.icio.us`, a social bookmarking service." `http://del.icio.us/about`, 2003.

[6] V. Batagelj and A. Mrvar, "Pajek - program for large network analysis," *Connections*, vol. 21, pp. 47 – 57, Jan. 1999.

[7] B. Wellman and S. Berkowitz, eds., *Social Structures: A Network Approach*, ch. Thinking Structurally. Cambridge University Press, 1988.

[8] S. Borgatti and M. Everett, "Network analysis of 2-mode data," *Social Networks*, vol. 19, pp. 243 – 269, Aug. 1997.

[9] K. Faust, "Centrality in affiliation networks," *Social Networks*, vol. 19, pp. 157 – 191, Apr. 1997.

[10] G. Karypis, "Cluto: Software package for clustering high dimensional data." http://www-users.cs.umn.edu/ karypis/cluto/, Nov. 2003.

[11] M. de Hoon, S. Imoto, and S. Miyano, "The C clustering library." http://bonsai.ims.u-tokyo.ac.jp/ mdehoon/software/cluster/software.htm, 2004.

[12] T. Kohonen, *Self Organizing Maps*. Spring-Verlag, 1997.

[13] B. M. Dennis and A. C. Jarrett, "Nuseye: Visualizing network structure to support navigation of aggregated content," in *Proceedings of HICSS-38*, 2005.

[14] I. Herman, G. Melancon, and M. S. Marshall, "Graph visualization and navigation in information visualization: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, pp. 24 – 43, jan 2000.