

Research Statement: Beyond Social Annotations

Introduction The links and relationships between online resources were once controlled by editors and authors alone. Today there are online tools, such as Delicious and Wikipedia, that enable users to markup any online resource or to contribute content to any topic. Collaborative tagging systems provide users with a mechanism to freely annotate pages, media files, or other objects. These annotations can be mined to discover relationships between or among users, resources, and tags [2]. My research focuses on assembling semantic similarity networks from user annotations. These similarity networks can in turn lead to improved Web search, recommendation, and navigation. I am also involved in the development of these types of applications leveraging socially induced semantic networks. Our group developed a playground to prototype these applications at GiveALink.org, of which I am the lead developer. Filippo Menczer and I wrote the NSF grant that funds this research.

Assembling Social-Semantic Networks: I have explored two user annotation representations for the purpose of assembling socially induced semantic networks. One is the organization of Web links in a personal hierarchy. Users employ a hierarchical representation when bookmarking favorite Web sites in their browsers. We introduced an entropy based similarity measure that exploits bookmark hierarchies for uncovering relationships between Web resources [11, 8]. An accompanying user study demonstrates that the resulting network captures meaningful relationships [7]. Another representation is the free-style tagging of resources with keywords. We studied several similarity measures based on this “folksonomy” representation to extract semantic relationships among tags and among resources. To address limitations in traditional evaluation techniques, we introduced a framework in which our similarity measures are grounded against reference datasets [4, 5]. This work has led to the design of a novel information-theoretic similarity measure outperforming all other measures in the literature for both hierarchical and flat tagging representations, and for both tags and resources [5]. I plan to continue this work by applying my similarity measures to larger systems, such as Delicious and Digg, improving upon their current search and browsing capabilities.

Scalability and Effectiveness: As online collaborative systems grow and evolve, it is important that semantic relationships be captured in an efficient way. We introduced several methods to aggregate social annotations into semantic similarity networks. We achieved scalability by an incremental aggregation algorithm that updates the network in real time. We explored the trade-off between efficiency and effectiveness, reaching an optimal balance by integrating a collaborative filtering component into our incremental approach [6, 4, 5]. The incremental approach has been deployed in our GiveALink platform allowing continuous real-world testing with an ever growing user base.

Web Interfaces: We explored the design of novel interfaces for improving Web navigation. My idea is to visualize socially induced semantic links to expand the user’s contextual view of a Web page. These connections are different from the embedded hyperlinks created by authors. We also explored the visualization of the relationships between the links within a page, for example the results from a Web search. Contemporary search engines return pages containing ranked lists. We found that users were able to find relevant information with fewer queries by extending the list view

with our network visualization [1]. I am currently involved in the integration of our interface with popular Web browsers to improve users' navigation and search experience.

Social Spam Detection: With the growing popularity of social tagging systems, “social spam” detection is becoming an important problem. My work has identified a number of features that are highly predictive of spam using a dataset from BibSonomy.org, a popular social bookmarking system. One feature examines the focus of the tags used to annotate a resource. Tags that are semantically dissimilar from one another are more likely to signal spam. Another feature is plagiarized content. Pages with content taken from another site, such as Wikipedia, are more inclined to be from spammers. Other suspicious features include ads, broken links, and automatically generated HTML. Combining these features has led to a spam detection accuracy above 98% [3].

Incentives for Annotation: As the democratization of the Web continues, I believe that increasing participation from as many diverse backgrounds as possible will only improve the usefulness of tagging systems [9, 10]. One way of creating an incentive is by entertainment through games. Similar to von Ahn's “Games with a Purpose,” an online game that introduces tags and annotations between Web resources will alleviate the sparsity problem common in collaborative systems. I am involved in the design of “tagging games” that induce users to find semantic connections between sites through tag sequences. Furthermore, taggers would benefit from more intelligent bookmark managers. Current tagging tools rely on lists and tag clouds for bookmark management, while browsers have until now imposed the hierarchical model discussed above. A more intelligent bookmark manager will draw upon the benefits of both modalities by providing a personal organization of a user's tags and resources. Beyond tag clouds, we envision an interface leveraging the socially induced relationships to make smart suggestions about relevant resources and annotations. I look forward to continuing this research to develop methods with the potential to increase online social participation.

References

- [1] J. Donaldson, M. Conover, B. Markines, H. Roinestad, and F. Menczer. Visualizing social links in exploratory search. In *Proc. Hypertext*, 2008.
- [2] B. Markines. Socially induced semantic networks and applications. Technical report, Indiana University Computer Science, 2009.
- [3] B. Markines, C. Cattuto, and F. Menczer. Social spam detection. In *Proc. AIRWeb*, 2009.
- [4] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proc. WWW*, 2009.
- [5] B. Markines and F. Menczer. A scalable, collaborative similarity measure for social annotation systems. Poster *Proc. Hypertext*, 2009.
- [6] B. Markines, H. Roinestad, and F. Menczer. Efficient assembly of social semantic networks. In *Proc. Hypertext*, 2008.

- [7] B. Markines, L. Stoilova, and F. Menczer. Bookmark hierarchies and collaborative recommendation. In *Proc. AAAI*, 2006.
- [8] B. Markines, L. Stoilova, and F. Menczer. Implicit tagging using donated bookmarks. In *Proc. WWW Workshop on Collaborative Web Tagging*, 2006.
- [9] H. Roinestad, J. Burgoon, B. Markines, and F. Menczer. Incentives for social annotation. Poster Proc. Hypertext, 2009.
- [10] H. Roinestad, J. Burgoon, B. Markines, and F. Menczer. Incentives for social annotation. Proc. of SIGIR 2009 Demonstrations, 2009.
- [11] L. Stoilova, T. Holloway, B. Markines, A. Maguitman, and F. Menczer. Givealink: Mining a semantic network of bookmarks for web search and recommendation. In *Proc. KDD Workshop on Link Discovery: Issues, Approaches and Apps.*, 2005.