Visualizing and Verifying Directed Social Queries

Angus Graeme Forbes, Saiph Savage, and Tobias Höllerer

Fig. 1. Detail of the interactive verification visualization. Here we see the friends, keywords, and topics correlated to the selected *like*, "house music."

Abstract—We present a novel visualization system that automatically classifies social network data in order to support a user's directed social queries and, furthermore, that allows the user to quickly verify the accuracy of the classifications. We model a user's friends' interests in particular topics through the creation of a crowd-sourced knowledge base comprised of terms related to user-specified semantic categories. Modeling friends in terms of these topics enables precise and efficient social querying to effectively fulfill a user's information needs. That is, our system makes it possible to quickly identify friends who have a high probability of being able to answer particular questions or of having a shared interested in particular topics. Our initial investigations indicate that our model is effective at correlating friends to these topics even without sentiment or other lexical analyses. However, even the most robust system may produce results that have false positives or false negatives due to inaccurate classifications stemming from incorrect, polysemous, or sparse data. To mitigate these errors, and to allow for more fine-grained control over the selection of friends for directed social queries, an interactive visualization exposes the results of our model and enables a human-in-the-loop approach for result analysis and verification. A qualitative analysis of our verification system indicates that the transparent representation of our shared-interest modeling algorithm leads to an increased effectiveness of the model.

Index Terms—Interactive verification, user classification, shared-interest modeling, topic modeling, social network visualization.

1 Introduction

An increasing number of people use social networks as a general-purpose social tool to fulfill their information needs. For instance, as discussed recently in [10], individuals often use their status message as a platform from which to pose questions to and to solicit information from their community of friends. These questions are often directed to a subset of the user's friends who could potentially help in answering a particular question or who share a particular topic interest. Facebook, Twitter, Google+, and other social networking sites all let users define collections of people in various ways. With Facebook and Google+, for instance, a user can use these friend collections to specify different social interactions [1]. For example, a user might create groups of friends related to school, to social clubs, and to work, and then specify different online social interactions with each of these groupings. The ability to predefine collections of people is a useful tool that enables a user to direct their social interactions appropriately. However, manually classifying friends into collections is time consuming, and even if users were to finish this exhaustive categorization task, the predefined lists might not cover all of the user's intended social interactions, especially as social groupings dynamically shift. For example, a user might be hosting an event related to a particular theme and seeking to invite only the friends that are interested in that theme. In this case, the user's predefined lists might be too coarse or otherwise inappropriate for this particular task.

Many social network sites make it possible to conduct ad hoc searches to obtain a list of friends that are in someway related to particular keywords or phrases. Although these lists may be used to target relevant friends, they do not necessarily allow the user to get an overall understanding of which friends could be good candidates for an intended social interaction. A method for more intelligently determining appropriate friend lists has been presented in [12], utilizing data mining techniques to find other Twitter users with similar interests. A recent project by [3] introduces a system to recommend friends with shared interests for sharing web content, and [2] also presents a system that helps the user to create custom, on-demand groups in online social networks. Though these system, and others like it, are useful for recommending users relevant to a particular information need, they in large function as a "black box," and do not expose how each recommendation was determined to be relevant. That is, even if the recommendation system finds connections between a user and other people in his or her social network, these results may not in fact be appropriate for a particular task, and exactly how the results were automatically determined is not clearly shown. Ultimately this effects the system's usability; [17] showed that users like and feel more confident about the recommendations or search results that they perceive as transparent.

We introduce a novel system that effectively classifies a user's friends in terms of particular shared interests. Since a typical user

- *Angus Graeme Forbes is with the Media Arts & Technology Department at the University of California, Santa Barbara, E-mail: angus.forbes@mat.ucsb.edu.*
- *Saiph Savage is with the Department of Computer Science at the University of California, Santa Barbara, E-mail: saiph@cs.ucsb.com.*
- *Tobias Höllerer is with the Department of Computer Science at the University of California, Santa Barbara, E-mail: holl@cs.ucsb.edu.*

of a social network may have hundreds or even thousands of friends, our system drastically reduces the time it would take to manually select a target audience for a directed social query, and enables users to execute such queries dynamically, based on the latest information about their friends. Furthermore, we present an interactive visualization that represents how our system infers that particular friends are good candidates for the social query. Moreover, the visualization allows a user to interactively explore the relations between these friends and their common interests, allowing the user to rapidly verify that our classification scheme is accurate and/or appropriate for a particular task, and to select friends for exclusion when it is not. A short video demonstrating the interactive verification visualization can be found at `http://www.mat.ucsb.edu/a.forbes/DSQ.mp4`.

Previous work, such as [8, 9, 16, 19], is also concerned with the automatic modeling and classification of users via online data. [11] explored how the communications patterns of a user could be utilized to detect the person's gender, and [15] had some success at modeling biographical attributes in online conversational patterns. Earlier work on automatic classification of online users analyzed blog posts, emails, and search queries. More recently, [13] has classified users via the more highly networked microblogs found in Twitter, where the only available data are short textual updates. In this work, we take an approach similar to modeling users from short available textual descriptions. However, we expand the range of data that we use to classify users by introducing external data that is related to the initial brief snippets of text. That is, we create a knowledge base by retrieving new terms from various crowd-sourced repositories, such as Wikipedia and Google Merchant Center. Regarding our efforts on system transparency [17], our work takes insights from early work on explanations in recommender systems [7] and more recent research on revealing the inner workings of recommendations and topic modeling using interactive visualization [4, 5, 6, 20].

2 Shared-Interest Modeling

We classify each of the user's friends in terms of their potential to fulfill the user's information needs. Although our model could be extended to other social networks, we use the Facebook social network in this study as it is one of the most popular social media websites, and because recent studies have shown that many Facebook users already utilize their status-message updates to ask questions to their friend community [10]. Specifically, we determine the friends' interests as indicated by their Facebook *likes*. These interests are then turned into a larger knowledge base by using these interests as search phrases to gather related words from crowd-sourced databases. We then model the friends, the *likes*, and the words in terms of topic vectors, where each component of the vector represents the probability of an affiliation to a particular shared interest. The use of a probabilistic topic vector was chosen as it presents simple methods, via vector operations, for the retrieval and ranking of a user's friends given a social query (which itself can be described in terms of a topic vector). When a user types a social query, such as "Should I replace my Blackberry with an iPhone, or just upgrade my Blackberry?", the system determines how strongly related the social query is to each of the discovered shared interests. The friends who present a similar relationship to the shared interests are returned and recommended to the user as appropriate people to direct the social query to.

Given that the textual information describing a *like* can be very sparse, and thus difficult for a machine to interpret, our system retrieves additional contextual information that can provide a broader semantic description that exposes potential meanings of the *like*. The additional contextual information that is obtained depends on the type of the *like*. *Like*s on Facebook are designated as products, bands, or one of various other categories. If the *like* corresponds to a specific product, we use Google's Shopping API to query the crowd-sourced Google Merchant Center database in order to gather a fuller textual description of what that product is. Otherwise, our system attempts to retrieve information from Wikipedia, retrieving articles that correspond to that *like*. We use DBpedia to programmatically gather this information from Wikipedia. If no further information can be gathered, we use only the information from Facebook itself.

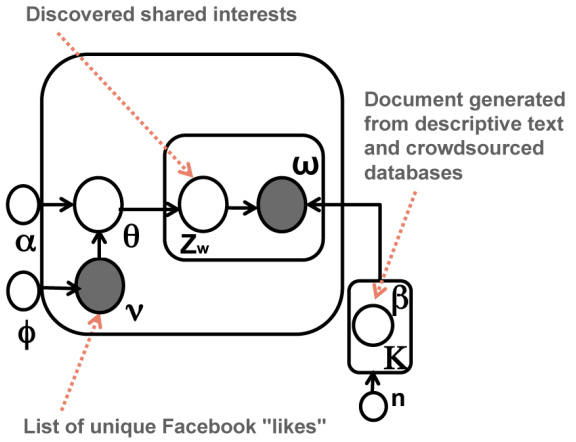

Fig. 2. An illustration of shared-interest modeling, based on labeled LDA [14]. In our model, both the label set v and the topic prior α influence the topic mixture θ. The labels we consider are the titles of the Facebook *likes*, and the documents are generated from related text gathered from Wikipedia and via the Google Shopping API, as well as from Facebook itself.

A generative process is used for discovering the interests shared by the user's friends. This process infers the set of overall interests shared by the user's friends, associates friends to one or more of these shared interests, and then, given a user's social query, finds which interests the query is related to and then which friends this query is most relevant to. The generative process first detects the K number of unique categories that the friends' *like*s belong to. This sets the initial number of shared interests that will be considered. For each shared interest, a unique *like* and its associated data is drawn with a Dirichlect distribution α. A multinomial mixture distribution θ^d over all K shared interests is drawn for each friend with a Dirichlet prior α_ϕ. Now, because we have information about the *like*s that the friend explicitly linked himself to, θ^d is restricted to be defined only to the shared interests that correspond to that friend's *like*s. This methodology is analogous to defining a document's topic mixture in Labeled LDA [14]. After this step, each friend is represented as a mixture over shared interests, capturing the friends' tastes and knowledge. A user's social query is also modeled as a mixture of shared interests, except that because the social query does not have any explicit labels, θ^d is not restricted. The friends who present a shared interest mixture similar (using the L_1 norm similarity metric) to that of the directed social query are presented to the user via the interactive visualization.

We were motivated to use a topic modeling approach as, in general, topic models are highly effective when the query topics are very broad [21]. On the other hand, topic modeling methods most often make use of approximation algorithms, based on either sampling approaches or optimization approaches, that may introduce modeling errors. In particular, the unsupervised nature of topic modeling makes it difficult to preform precise evaluations [18].

3 Transparent Interactive Verification

Our prototype interactive visualization has two primary functions. First, it transparently exposes exactly *how* our system correlates a subset of the user's friends to particular shared interests. That is, it presents a visual representation of why these friends were selected in response the user's directed social query. Second, it allows a user to verify whether or not the resulting list of friends is in fact appropriate for a particular directed query.

On the right side of the visualization, a list of the "topics" (i.e., shared interests) that define the friends of a particular user (as inferred

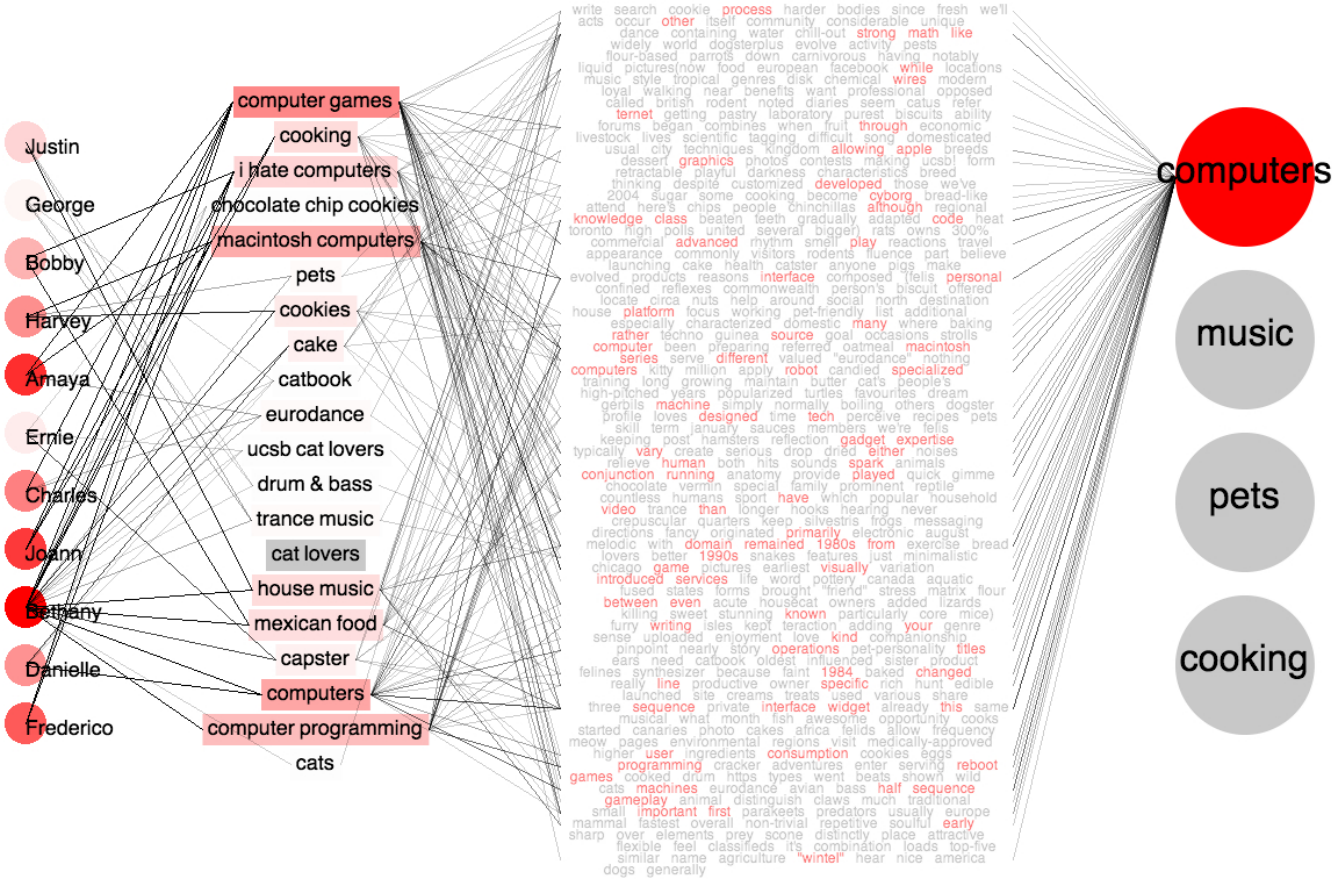

Fig. 3. Screenhot of the interactive verification visualization. Here (in the first column on the left) we see the friends that are correlated to the shared interest "computers," as well as the likes and keywords (the second and third column from the left, respectively) that were involved in the matching (the rightmost column) of these friends to the shared interest. The strength of the correlation is indicated from light red to dark red.

by our system) is shown in a column. On the left side, we display all of the users' friends that are highly correlated to these topics – that is, who have the potential of providing the user with feedback for their social query. By selecting any of the topics (via a mouse over event) the friends that are correlated to that topic are instantly highlighted. We use a simple gradation from light red to dark red to indicate the strength of the correlation. Users that are not correlated remain gray, the default color of all elements when nothing is selected. Similarly, when selecting any of the users, topics related to that user are also highlighted in the same manner. More interestingly, we display the main elements that contribute to our algorithm's determination of how users and topics are correlated. The *likes* that are highly correlated to the users are positioned in a second column next to the users, and a sample of the words that contributed most to establish a correlation between the likes and one or more topics are also displayed in a third column. Highlighting either any of the topics or any of the users immediately draws thin lines indicating a connection between the users, *like*s, words, and topics and also highlights them in a gradation of red to indicate the strength of their correlation. The *like*s and words can also be selected to show which topics and users they are correlated to. Figures 1 and 2 show screenshots of the interface when the user has selected a *like* and a topic, respectively.

The chief advantage to exposing this extra information is that the user can verify whether or not the model has accurately classified users in terms of the selected topics. Because the most relevant information that led to the classification is visible, a user can quickly determine if it is appropriate or not. Moreover, it provides more insight than the classification alone, as a user can "zoom in" to see what words are correlated to the topic, and thus gains a fuller concept of the semantic range of the topic. We ran a series of in-depth cognitive walkthroughs with a small number of subjects to solicit feedback about the basic design as well as to identify how effective this type of transparency might be for identifying appropriate users for direct queries. All of our subjects were able to navigate the represented information within a few seconds to a few minutes of experimentation and only minimal instruction. The highlighted path between users and topics were clearly interpreted, although one subject thought that the visual representation of connectivity through lines seemed "cluttered" and another wondered aloud "where do the words come from?" However, for the most part, the subjects had positive responses to the visualization, and noted that it aided them in verifying friend lists for directed social queries.

As a test case, we included two users who were purposefully correlated with topics incorrectly– that is, the connections between the friend's *likes* and the topics were scrambled. We wanted to see if users would notice that some correlations were problematic. For instance, we changed one friend to be highly correlated with the topic "Computers," but who had likes whose keywords that had more to do with "Cooking." Without exception, users independently noticed that something was awry without any prompting from the authors. Our motivation in providing this example was to show that our visualization of the underlying user modeling data was sufficient to allow a subject to confirm or deny a classification. We also included a second artificially-constructed friend that had the name of one of the user's real friends, but was highly correlated with *all* of the topics. Again, each of our subjects pointed out that it was strange that a single friend would be so uniformly correlated with all of the topics. One asked, "is this friend a bot?" Another asked whether or not "our program had a made a mistake" with this friend. While further evaluation is needed to identify the amount of error that a user would detect, the simple visual cues in

these instances were sufficient to indicate issues with the model and to cause users to investigate their cause.

While this visualization is displays only a limited number of users and a set number of topics, it is clear that it could be extended to include a larger number of results. Some design choices would need to be addressed however, including a dynamic determination of the thresholds for how strongly correlated a *like*, word, or topic would need to be to be included in the visualization. Providing user control over these thresholding parameters might be a good first step, but is not a substitute for experimentally validated heuristics for suitable defaults.

4 Conclusions and Future Work

This paper introduces a novel system for modeling and visualizing social network users in terms of their interests. By leveraging crowdsourced databases we were able to generate a rich description of each user's interests which we then used to create a shared-interest topic model that effectively correlated a user's friends to shared interests. During our testing we found that, due to the nature of using unstructured data from multiple sources, our model occasionally incorrectly matched friends to particular social queries. That is, in a small percentage of cases, regardless of which similarity metric we used, some mitigation of modeling errors was necessary. In creating our visual application, we had the realization that the verification could be presented interactively. We developed an interactive verification prototype and conducted cognitive walkthroughs to gain insight into the effectiveness of presenting the underlying data that generated the topic model. Our qualitative analysis indicated that users appreciated the transparency and found that it was a useful tool for determining appropriate friends for directed social queries. More generally, this work hints at the potential that pairing machine learning algorithms with interactive visualization strategies may have in aiding social decisions. Future research will investigate how well our system scales to a larger number of users and a larger of number of shared interests, and will examine how the generation of knowledge bases from different sources might impact the generation of the shared-interest model. Finally, we plan to explore the application of our model and our visualization to other social data sets. Source code for the visualization code using a sample data set is available via a git repository at `https://github.com/angusforbes/DirectedSocialQueries/`.

5 Acknowledgments

This work was partially supported by CONACYT-UCMEXUS, by NSF grant IIS-1058132, and by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of NSF, ARL, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] F. Adu-Oppong, C. Gardiner, A. Kapadia, and P. Tsang. Social circles: Tackling privacy in social networks. In *Symposium on Usable Privacy and Security (SOUPS)*, 2008.

[2] S. Amershi, J. Fogarty, and D. Weld. Regroup: Interactive machine learning for on-demand group creation in social networks. In *CHI '12: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 21–31, 2012.

[3] M. Bernstein, A. Marcus, D. Karger, and R. Miller. Enhancing directed content sharing on the web. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 971–980. ACM, 2010.

[4] J. Chuang, C. D. Manning, and J. Heer. Termite: Visualization techniques for assessing textual topic models. In *Advanced Visual Interfaces*, 2012.

[5] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *CHI '12: Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, pages 443–452, 2012.

[6] L. Gou, F. You, J. Guo, L. Wu, and X. Zhang. Sfviz: interest-based friends exploration and recommendation in social networks. In *Proceedings of the 2011 Visual Information Communication-International Symposium*, page 15. ACM, 2011.

[7] J. Herlocker, J. Konstan, and J. Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 241–250. ACM, 2000.

[8] R. Jones, R. Kumar, B. Pang, and A. Tomkins. I know what you did last summer: query logs and user privacy. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 909–914. ACM, 2007.

[9] S. Kim and E. Hovy. Crystal: Analyzing predictive opinions on the web. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1056–1064, 2007.

[10] M. Morris, J. Teevan, and K. Panovich. What do people ask their social networks, and why?: a survey study of status message q&a behavior. In *Proceedings of the 28th international conference on Human factors in computing systems*, pages 1739–1748. ACM, 2010.

[11] J. Otterbacher. Inferring gender of movie reviewers: exploiting writing style, content and metadata. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 369–378. ACM, 2010.

[12] M. Pennacchiotti and S. Gurumurthy. Investigating topic models for social media user recommendation. In *Proceedings of the 20th international conference companion on World wide web*, pages 101–102. ACM, 2011.

[13] M. Pennacchiotti and A. Popescu. Democrats, republicans and starbucks afficionados: user classification in twitter. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 430–438. ACM, 2011.

[14] D. Ramage, D. Hall, R. Nallapati, and C. Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.

[15] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, pages 37–44. ACM, 2010.

[16] S. Savage, M. Baranski, N. E. Chavez, and T. Hollerer. I'm feeling loco: A location based context aware recommendation system. In *Advances in Location-Based Services: 8th International Symposium on Location-Based Services, Vienna 2011*, Lecture Notes in Geoinformation and Cartography. Springer, 2011.

[17] R. Sinha and K. Swearingen. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 830–831. ACM, 2002.

[18] C. Wang, J. Paisley, and D. Blei. Online variational inference for the hierarchical dirichlet process. In *Artificial Intelligence and Statistics*, 2011.

[19] I. Weber and C. Castillo. The demographics of web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 523–530. ACM, 2010.

[20] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.

[21] X. Yi and J. Allan. A comparative study of utilizing topic models for information retrieval. *Advances in Information Retrieval*, pages 29–41, 2009.