Utilizing Crowdsourced Databases for Social Media Question Asking

Saiph Savage[1] **Angus Forbes**[2] **Rodrigo Savage**[3] **Norma Elva Chávez**[3] **Tobias Höllerer**[1]

[1]University of California, Santa Barbara
{saiph@cs, holl@cs}@.ucsb.edu

[2]University of Arizona
angus.forbes@sista.arizona.edu

[3]Universidad Nacional Autónoma de México
{rodrigosavage@comunidad, norma@fi-b}.unam.mx

ABSTRACT

This paper discusses an interview-based study conducted to explore why people *like* particular pages on Facebook. We analyze how the different relationships that people have with their Facebook *like*s could affect the design of intelligent interactive systems. Specifically, we examine how the *like*s of a user's friends can be used to gather information from external crowd-sourced knowledge bases in order to identify relevant candidates for user-posed questions. Finally, we consider how different forms of interaction affect the visual representation of Q&A systems.

Author Keywords

Interactive verification, user modeling, social network group creation, access control lists, friendsourcing, social transparency

ACM Classification Keywords

H.5.2. [Information Interfaces and Presentation]: User Interfaces

INTRODUCTION

Many online social networks encourage their users to share biographical information, interests, and political views, among other personal traits. This information can be gathered by an autonomous system that, through statistical machine learning techniques, can infer the expertise of users in regards to particular topics. This data can then be used to identify users to which community-posed questions should be directed.

A problem with statistical approaches that utilize social media data is that the small snippets of text found in a user's profile can be very sparse. And it can therefore become difficult for the system to relate information from a user's profile to his or her knowledge about different topics. For example, a user might state that he or she is interested in "Michael Jackson." But without any additional knowledge, it is very difficult for the system to identify that it should consider that user when a question related to the topic of "Pop Music", or "80's Music" arises. To leverage this problem, approaches [3] [4] have been proposed where external crowdsourced knowledge bases are utilized for aggregating information, thus allowing the system to obtain a better understanding of the meaning behind each tag. But it is unclear what type of crowdsourced database should be used to enhance the system's understanding of the tag.

In this paper, we conduct an interview-based study that analyzes how users engage and interact with their stated interests in their online profile. We discuss how this knowledge can be used for deciding which are the best crowd-sourced knowledge bases to accurately model users in order to identify relevant candidates in a Q&A social media based system. Specifically, we explore how these findings could be used in our on-going Q&A research prototype, which models the Facebook friends of a user in terms of their shared interests. Given a user-posed question or query, the prototype identifies which friends are good candidates for answering that question, and help the user fulfill his or her information need. Our prototype uses the Facebook *like*s of the user's friends to infer shared interests. Given that the information associated with a Facebook *like* can be very sparse, we utilize external crowdsource knowledge bases to enhance the Facebook data. We also provide different interactive visualizations that help users further decide how appropriate the list of recommended friends is.

Participant	Age	Gender	Occupation	Hobbies
Ana	29	F	Youth Pastor	Running
Bernardo	30	M	Teacher	Anime
Carlos	23	M	PhD Student	Cooking
Dolores	19	F	Student	Drawing
Elena	20	F	Master Student	Running
Federico	26	M	Journalist	Reading

Table 1. Participant Behavior

Our current study was undertaken with the aim of understanding what leads people to *like* certain Facebook pages and to study the interactions people have with their Facebook *likes* further on. Our study also begins to raise some interesting questions. What problems does the act of introducing extra data into the machine learning algorithm solve? What are some of the potential issues with introducing extra terms from databases that are commercially-oriented? What are some of the potential issues with introducing extra terms from databases that are not specifically geared to social media tasks? What qualities of such databases are most effective for increasing the accuracy of user model? Through working with crowd knowledge and crowd behavior we hope to find appropriate ways to think about these questions. In the following sections we first discuss our preliminary study that analyzes how people engage with their stated interests on Facebook, and we relate how our findings from this study can be used to enhance Q&A systems. We introduce our own Q&A system that is utilizing some of the findings from this study. We provide motivation for our Q&A system, its machine learning component, and interactive visualizations.

RESEARCH STUDY

To explore current practices for *liking* Facebook pages, we conducted semi-structured interviews with six participants (three men, three women.) We wanted to learn from the experiences of individuals the variety of reasons why people *like* particular pages.

Participants were primarily recruited from mailing lists for college students. The recruitment email asked for participants who had over 30 Facebook likes, and who had liked a page in the last two weeks. This allowed us to recruit participants who frequently liked different Facebook pages. The study had a duration of approximately 40 minutes per participant, and we compensated them with snacks. All interviews were conducted by one of the co-authors of this paper; all interviews were conducted face-to-face and notes were taken during the interviews. The interviews consisted of a series of open-ended questions about why participants liked the pages that they liked on Facebook, and the interactions that they had with the page after they liked it. Table 1 shows the demographics of the six interview participants and the diversity of personal activities among them (all names have been changed for privacy reasons.)

Presented February 24th, 2013 CSCW Workshop on Social Media Question Asking San Antonio, Texas

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

FINDINGS

Most participants (4 out of 6) mentioned they primarily liked Facebook pages to obtain periodic updates on the content generated from the page, especially for brands and/or products. Users said they often liked these branded pages in order to obtain information on the latest new products or announcements on discounts and sales.

One of the participants (Ana) said that many times she was forced to like certain things on Facebook in order to enroll for race events. Ana said that these pages which she indicated she liked did not actually reflect at all her own preferences. Another of the participants (Elena) said that many times she liked pages on Facebook because it forced her to "be surrounded with school related activity even while relaxing in social media." Elena said she liked Facebook pages that were related to her career, and that provided periodic updates related to her field of study. She mentioned these pages functioned more as a sort of online forum where other Facebook users pose questions and experts respond. Elena said she had never participated in the discussion, as she felt she did not have enough knowledge to respond to the questions posed, but she still found the discussions interesting, and enjoyed reading and learning from them.

Another participant (Dolores) said she liked Facebook pages in order to obtain periodic updates from the page. But she mentioned that in recent months, her interactions with Facebook Likes had changed. She said that before she would hide her Facebook likes from her Friends and engage in active controversial discussions on various of her Facebook Like pages. She said that because her interests were "so out-there", she was sure she would never find one of her friends on that same page, and it allowed her to freely and openly express herself without worry. But after Facebook implemented the live-status bar that constantly tells you what each of your friends is doing, many times, regardless of the privacy surrounding that content, she said that she panicked about one of her controversial conversations appearing in the feed and the online image she maintains with her Facebook friends being compromised. Dolores said that after the insertion of this live feed bar she decided to no longer engage in these controversial discussions. Dolores's interactions with her online identity intrigued us, and we further asked her whether she would engage in conversations with a Facebook friend who by mere chance was also participating in these controversial online discussions. Dolores replied that she would only respond if she could respond anonymously.

Another user who also had interesting online identity patterns was Carlos. Carlos stated that many times he liked certain Facebook pages because it allowed him to present his lifestyle to his friends and boost his online image. He liked pages that were related to clothes he bought, where he went to school, or places he had visited. He said he liked to do this in a subtle form. He would like a certain restaurant, which would tell his audience that he belonged to a certain social class. He would like certain school groups, so his friends would know he was studying abroad.

Only two of the participants (Federico and Bernardo) said they shared with their friends actual content from the Facebook page. Bernardo mentioned he enjoyed sharing content from certain Facebook pages because he believed it allowed him to create philosophical reflections with his friends. He liked sharing animal rights pictures, as well as pictures related to politics, that encouraged reflection among his friends. Bernardo also said, that because he was a high school teacher, and many of his former students were among his Facebook friends, he felt it was his moral obligation to provide material for philosophical reflections. Bernardo also said he enjoyed sharing material that was humorous and could bring smiles and laughs to his friends. Federico stated he engaged with Facebook pages more as a personal activity. He liked pages that had visual elements in its feed. He said he enjoyed sitting at home and looking at the content of these Facebook pages. He said he shared content from the Facebook page to maintain contact with his friends. He said he had many different types of friends, and the content he shared helped him keep connected with these friends.

Reason for Liking Page	Mode of Interaction
Obtain Updates on a Product	Read news feed
Participate in a real life event	Just like the page
Obtain knowledge on a subject	Read news feed
Discuss topics of interest	Participate in discussions
Bolster their online persona	Just like the page
Create philosophical reflections	Share material
Stay in touch with friends	Share material

Table 2. Participant Info

DISCUSSION

From the interviews we notice seven different behaviors for engaging with a Facebook *like*. We show a summary of each of the detected behaviors in Table 2. In the following, we discuss how each of these different behaviors could be used for determining the type of crowd-sourced knowledge base to use to enhance the knowledge of a social media based Q&A system.

Obtain product updates: This scenario involves being interested in a particular product or brand and receiving constant updates about that product. It therefore makes sense to utilize a crowd-sourced database, such as Google Merchant, that could provide an initial definition of the product, and then utilize the discussions on the Facebook page to obtain a better measure of the user's current expertise on the matter. The main difficulty with this approach is the fact that it can become difficult to quantify how much of the data that the Facebook *like* page posted is absorbed by the user. In terms of the visualization, it might make sense to show that the user is interested in updates related to that product. As this might invite other users to invite him/her into Q&A sessions that involve discussions on the latest features of the product, or discussions involving what new features should be added to the product.

Participate in a real life event: The fact that many companies or events require users to *like* certain pages, even when the user is not at all interested in the data related to the page, taints studies such as ours that use the stated interests of users for recommending who to direct particular content to. We consider that in this scenario, an alternative design consideration is to have options where users can state they are liking a certain page or item for the mere reason of getting something else from it. This way their online identity would not be compromised. In that case, what would be important is to understand what is the data item that the user

really is interested in, and this depends on the nature of the event and the reasons the user has for attending the event. The user might be interesting in attending a race, for example, because they wish to live a healthy life. In this case the knowledge base to use could be more related to medical terms that provided information about the benefits of such activity. Likewise, the user might be going to the race to promote an identity, e.g. they belong to a certain social class that can participate in races. In that scenario, the crowdsourced knowledge base to use would need to be created from people in the user's social class that could provide information as to what it socially means to be linked to such data. Additionally, it might be interesting for other users to visualize the data items that their friends have accepted to like in order to get something in return for it. This visualization allows others to better understand the boundaries of how their friends manage and engage with their online identities.

A possible technique that could be used to detect when a user is liking something that is more related to an event than the actual Like itself is by analyzing the time of day and GPS coordinates of when the user liked the data item, and analyzing whether it matches an event the user has stated that they will be attending. This scenario considers whether the user is liking the required Facebook page while at the event itself. In such case, the item that the system would link to the user's interests would be the event and not the Facebook page itself. Information about what the event implies could be obtained from the description of the event itself and through polling of information from the user's social circle about what it means to attend such an event.

Obtain knowledge on a subject: This type of stated interest involves knowledge and an interest for matters related to that subject. Therefore the best crowd-sourced database that could be used in this scenario is one where a definition of the subject is provided, as well as an ontology of related fields. The ontology provided from Wikipedia could be a good option. This would allow the system to obtain a better understanding of the type of profile of the person. We believe the majority of *like*s that fall under this category are labeled as "interest" in Facebook. In which case a simple analysis of the type of the *like* could be used to infer the type of knowledge base to use. In terms of the visualization, we believe it would be interesting to provide users with information about what data their friends follow as interests, but do not consider themselves experts in; or conversely, which interests they do consider themselves experts in.

Bolster their online persona: We believe that for this type of engagement, the best way to model the user is to obtain the social definitions of what it means to link to certain data from people in the same social class of the user. It can be extremely difficult to detect when a user is liking certain items for the purpose of promoting an online image, as in this case there is no clear category that is exclusive to this behavior. In this case, a system which could ask the crowd whether a group of Likes appeared to be tailored for promoting a certain online image could be adequate.

Create collective awareness: Given that the engagement with this like is to share the content of the Facebook page to create awareness, it becomes adequate to obtain from Wikipedia a definition of the like (a definition of the cause) along with public opinion polls, open-record laws or sunshine laws on topics related to the cause. This data would help understand the nature of the cause, as well as to obtain a spectrum of where the user's friends lay. It would be interesting to provide users with a visualization that showed them, based on the what their friend had shared, what aspect of a particular cause their friends support and how much. Additionally it would be interesting to provide to users' opinions related to the cause that are different from those held by their friends. We believe the majority of *likes* that are related to this behavior fall under the category of *Cause*–a Facebook-defined category– and this label could be used to identify when there is a case for such a data relationship.

Stay in touch with friends: This type of behavior is stating that users will engage with certain material even when it does not personally interest them, for the mere reason that they wish to entertain their audience. In terms of interface design, we believe in such cases, it could be presented to users a visualization of friends that enjoy sharing content to others, and the type of content they enjoy sharing. Such a data visualization could help users identify mavens – people that are able to find others who have knowledge about a subject, even though they themselves don't have that knowledge.

Another interesting design consideration we obtained from our interview-based study is the fact that there may be many questions for which users would enjoy participating in an anonymous form, even when discussing with friends. We believe that exploring functionality where users could respond and help their friends in their information needs, without exposing themselves or compromising their online image, is needed. It is important to understand the different forms of self expression that users have with their stated identities, and analyze how this affects the design of Q&A systems. In the following section we present our on-going research project that actively explores aspects of the findings we have described here.

DIRECTED SOCIAL QUERIES SYSTEM

Our system models users of a social network in terms of their shared-interests. Our machine-learning algorithm is based on the popular Latent Dirichlet Analysis and uses an influx of additional information from external sources based on similarities to the information more readily available within the social network. We describe the task of matching questions asked by a particular user (via their status bar) to appropriate friends within their social network. We explore a different aspect of research first described in [5]; here we discuss the use of crowdsourced databases as a methodological approach of gathering sufficient data for discovering appropriate friend lists for question asking tasks. Our machine learning algorithm lends itself to different types of visualizations: [1] examined the use of a novel visualization system to aid the user in verifying automatically generated friend lists, and [5] introduced a novel interactive system to allow users to explore aspects of their friends (such as geography and shared interests) that would enable them to refine the results of the task of matching friends to questions. Here we present more detail about what external crowdsourced databases should be used to learn hidden traits and characteristics of a user's friends, based on an interview based study that provides a preliminary analysis of the meaning behind stating an interest on a social network. We provide a discussion that analyzes how the different relationships that exist for why users' state particular interests can be used to identify the type of crowdsourced databases that should be utilized, and also how this can affect the visualization of information in the system.

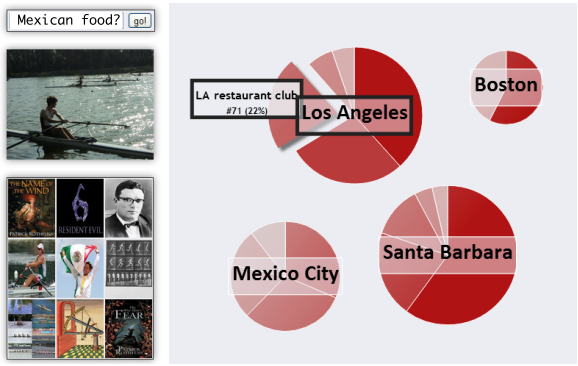

Figure 2. Screenshot of the interactive visualization. Here we see clusters of friends that our model has correlated to the directed query. A user can click on the wedges in the pie chart to determine which group these friends belongs to, and choose to include or exclude that friend based on that information.

Our system includes a learning algorithm that explicitly attempts to augment this *internal* data (profile information of users) with related data from *external* sources (crowd-sourced knowledge bases). In particular, we use large, publicly accessible, crowd-sourced databases to increase the amount of information we have to characterize each of the user's friends as well as to define the question directed toward the user's friends. As described in [5]– an earlier description of this ongoing research– we were able to get promising results from our learning algorithm. In that paper however we did not focus on what we have come to see as a surprising aspect of the system– that the introduction of potentially arbitrary, unrelated, or even seemingly contradictory data successfully improved our results.

Our original approach was to find synonyms of each of the words. We used a list of synonyms that came from course material used in teaching English as

a second language [2]. Although this increased the number of words available to our learning algorithm, it did not dramatically increase its effectiveness (effectiveness of the system was measured using precision and recall for matching friends to question in a "ground-truth" dataset.) We decided that it might be useful to add larger amounts of data, even though the data would include a number of elements that were less clearly related, or even completely unrelated. Additionally, because a number of the *likes* indicated that user liked a particular band or product, we thought that by using the Google Shopping API to get related terms and products would be useful. This did indeed increase the effectiveness of our method, even though we introduced ambiguity into the system. Since we gathered the data simply by keyword, a search for the related term "cat" using the API would return, for example, information about dogs, pets, sports teams, zoos, as well as about specific types of cats, cat toys, products that had cats on them, and a variety of other items. We also used DBpedia to gather short Wikipedia blurbs of the keyword. Retrieving these Wikipedia information introduces an even wider range of words that had ostensibly nothing to do with the original terms in the *like*s. As we note, it can be extremely difficult to identify what is the best external source that should be utilized to complement the information associated to a Facebook *like*.

This was the motivation for the study we presented previously. In the next section we present the system's interactive visualizations. The visualizations of our system were inspired by the fact that to construct an optimal system capable of accurately and precisely identifying which friend's best match a social query or question the system needs to be well trained to identify optimal threshold values for each task (threshold values for finding the friends that are most relevant for a social query, as well as identifying the *likes* that are the most relevant to the question.) These findings were what motivated the second part of our work, in which we created interfaces that exposes the information that our system uses to determine classifications. By bringing the user into the loop when making classification decisions we have the best of both worlds, as a human user is more tolerant to errors (since the user can easily correct the system), while the machine learning algorithm is vastly more efficient at making an initial classification of a large amount of crowd generated data.

INTERACTIVE VERIFICATION VISUALIZATION

We designed a prototype interactive visualization with two primary functions: first, it transparently exposes how exactly our system correlates a subset of a user's friends to a particular social query, and how it identifies the ones that are the experts in the field; second, it allows a user to verify whether or not the resulting list of friends and their classification (whether or not they are experts) is appropriate for a particular social query. On the left side of the visualization, we display all users that were found to be highly correlated to a user's particular social query. In the left side of the visualization we present the information that was used to determine that these friends had the potential of responding to the user's question. The user can select any of those friends, and further explore the information from the crowd that the system utilized to determine that a friend was indeed a good candidate to respond to a query, and that they are potential experts in the field. When selecting any of their friends, the topics the system considers the friend is interested in are highlighted. 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 shows screenshots of the interface when the user has selected a *like* and a topic, respectively.

SOCIAL AWARENESS VISUALIZATION

Although the interactive verification process described above is effective at weeding out problematic correlations, it does not contextualize the friend lists within the full spectrum of available social network data, and this social information could play a role in the user's decision to include that friend in their social question. We extended our prototype application to address potential real-world scenarios for question asking scenarios to leverage this available data. Specifically, we organized the highly-correlated users hierarchically in terms of their geographical location and their inclusion in particular Facebook groups. Through visualizing friends in this way, a user can quickly determine which of the users are appropriate for a particular social query. The data used for this social awareness visualization is not exhaustive, but was chosen to show the potential that this social data and visualization techniques have in aiding social decision making. Figure 2 presents a screenshot of the prototype application in which the user is in the process of determining which friends to ask about the best place to eat dinner. In the top left corner, the user has entered "Where is the best place to get Mexican food?" This resulted in our model returning a list of correlated friends, which are visualized within the context of their geography and most related Facebook groups. Each of the piecharts represents a particular geographical location, and each wedge within the piecharts represents a particular group.

CONCLUSIONS AND FUTURE WORK

This paper discussed a preliminary study on how users engage with their Facebook *likes* and analyzed the implications that the different forms of interaction have in deciding which crowd-sourced database to use to augment the social media data. We further discussed how these different forms of interaction affect the design of Q&A systems that use the stated interests (Facebook *likes*) of users to identify candidates for user posed questions, as well as to present users with transparent information on how the particular list of friends was selected. We are currently investigating how the act of utilizing databases that contain some unrelated information serve to increase the effectiveness of modeling shared-interests. We hope to find ways of characterizing the elements of these database which improve the results so that the we can refine our system.

ACKNOWLEDGMENTS: This work was partially supported by CONACYT-UCMEXUS and by NSF grant IIS-1058132.
Special thanks to our users.

REFERENCES

1. A.G. Forbes, S. Savage, and T. Höllerer. Visualizing and verifying directed social queries. IEEE Workshop on Interactive Visual Text Analytics. Seattle, WA. 2012.
2. English as a second laguage. About.com. Retrieved April 11, 2012.
3. M. Michelson and S.A. Macskassy. Discovering users' topics of interest on twitter: a first look. In Proceedings of ACM Workshop on Analytics for Noisy Unstructured Text Data (AND). New York, NY. 2010.
4. R. Savage, T. Nava, N.E. Chavez, S. Savage, Search on the cloud file system, In Proceedings of Parallel and Distributed Computing and Systems (PDCS). Dallas, TX. 2011.
5. S. Savage, A.G. Forbes, R. Savage, T. Höllerer, N. E. Chavez, Transparent user modeling for directed social queries. In Adjunct Proceedings of ACM User Interface Software and Technology (UIST). Cambridge, MA. 2012.