Social SQL: Tools for exploring social databases

Marc Smith, Vladimir Barash
Microsoft Corportation
Marc.Smith,t-vladba@microsoft.com

Abstract

Social media are constructed from the collective contributions of potentially millions of individuals and present an increasingly common and large scale form of database. As these databases grow in social and technical importance exploration of their structure is needed. The social nature of much of this data is an added opportunity and challenge, providing contributions to social science questions about large scale social behavior while raising technical thresholds caused by the scale and complexity of the data. The detailed records of the interactions of users of social media form a foundation for higher level representations of the behavior of users and communities in these systems. Social networks are a key structure in social media databases. In the following several visualization tools are used to illustrate social networks and other structures within these data sets. These images highlight behavioral motifs that can be understood through social science theories about roles and social structure. The result is a deeper understanding of the dynamics that drive the creation of user generated content in social media. This process suggests the need for an extension for the structured query language (SQL) that explicitly supports social queries.

1 Introduction

Social life increasingly takes place through computer mediated interaction systems, and these systems are growing in terms of affordances and social importance. Social networking and Web 2.0 services are the most recent examples in a long line of social media. Ever since email and email lists started to accumulate, social media have grown dramatically in volume and function. Many people's experience of internet communication is the product of multiple complex channels. It is now commonplace for many people to use tools like email, email lists, newsgroups, discussion boards, web forums, blog comments, wiki document and talk pages, instant message conversations, SMS messages, Social Networking Services, photo, audio and video sharing services and several other mechanisms for communication and relationship management. These channels allow for the exchange of a rich collection of digital objects among select or global populations.

When people gather and interact in computer mediated spaces they often leave traces behind which record who does what with whom when. Many computer mediated spaces are linked to databases that record these traces for logging and backup purposes. These databases can be processed to reveal patterns of association and patterns of individual differences present in the data. These patterns tell a story about an ecosystem and its inhabitants, a story about variation and the emergence of stable types of social spaces and the roles participants

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

play within them. A great opportunity exists in the river of real data coming from these systems that enables a focus on empirical studies of large scale naturally occurring data sets composed of large interacting populations. Letting this data tell its story is a goal shared by many who want to understand what happens when millions of people interact through computation.

As these databases grow in social and technical importance exploration of their structure is needed. The emergent structures that result from millions of people using these systems are opaque; the systems themselves rarely provide tools to gain a meta-overview of the system itself. The confluence of increasing computer power and the widespread adoption of Internet social media applications offer a particular opportunity within the larger space of network science. Network structures are becoming the focus of a diverse range of disciplines from physics, biology, and information science to sociology and other social sciences. Commonalities across diverse disciplines are emerging as network structures are found within complex processes from protein biology to international finance.

Despite the variety of social media databases in operation, they often share common core structures in the forms of social networks, conversation and document structures, hierarchies and user profiles. Many of these structures have corresponding time stamps. With effort these data sets can be transformed into visualizations that illustrate higher level structures. In social media data sets these higher level structures include attributes like social roles, cliques, communities, and their historical changes.

An integrated view of social media remains elusive, if only because of the distributed way in which such data is stored across multiple systems, services and geographies. Today, only fragmentary images that feature one or a few systems are available. Metaphorically, studies of social life on the Internet remain in a state similar to meteorology prior to satellite photography. Work to build local maps of social databases is occurring in a number of disciplines.

The challenge of mapping even individual social databases is heightened by the absence of standard “social queries”. Most social media database systems lack tools for dealing with higher level social structures in their datasets. A ”Social SQL” is called for that provides higher level forms of interaction with social databases.

Geographic data is a good analogy for what a Social SQL could be. In GeoSQL, a higher level set of relationships for database objects related to location and connection is provided. GeoSQL offers custom support for the geographic properties of roads and bounded regions so that developers avoid calculating these attributes from scratch. This extension to SQL provides for ”topological relationships between two geographic objects with seven spatial predicates. These predicates are: “EQUAL, MEET, INSIDE, CONTAIN, CROSS, OVERLAP and DISJOINT.” [12] A similar set of operations could be applied to social media datasets. Social queries often want to know if a person is a member of a group or if a group is a sub-group of a larger organization. Other queries want to discover if there is an intersection between people or groups in terms of the common relationships or interest they share. It is common to want to know if two slightly different names refer to the same physical person.

The comparison has limits, however. Geographic space is an objective reality that is not in widespread dispute (even if the boundaries sometimes are). In contrast, social media lie in an amorphous data space whose internal structure remains poorly understood. While many of these geographic operators could make sense in a social media space, there is an absence of a unified underlying terrain. Tracking the shifting memberships of well defined organizations is a challenge as people enter and exit and move from one part of the organization to another each day. Building such a map for more informal groupings is even more challenging. Cities and other geographic entities only slowly move from one jurisdiction to another while informal computer mediated groups may resemble more closely flocks of birds that merge and divide while moving from one perch to another [4].

Social media repositories are populated by multiple entities which are interconnected in complex ways. Mapping social media databases requires the production of higher level representations of the potentially large volumes of individual records of transactions and interactions. Each entity must be aggregated in terms of other entities and across time. For example, individual rates of activity over time and the interaction patterns between many individuals at some point in time are two common elements of many social database visualizations. These

higher level elements are a form of accounting system for social interaction. The production of social accounting metadata is a prerequisite for all the efforts to produce social database visualizations.

Maps of computer-mediated social interactions reveal the range of social variation taking place within these systems, the variety of social roles being performed, and the different groupings, clusters and communities into which people aggregate. Traces of conflicts or group efforts to collaboratively construct artifacts are topics of great interest to social scientists who can use these computer-mediated social spaces to further the study of these and other group processes. Social databases have many advantages over traditional data collection methods that involve direct observation and manual coding of events. Data collection is often costly, necessarily limited in scope and time period, and error prone. In contrast, social databases are high fidelity records of some elements of social interactions. Attributes like the time of an event, the identifiers of participants and other objects are all likely to be recorded very accurately.

The attractions of social databases are often simultaneously their greatest challenge in terms of building maps. Before any form of Social SQL can be developed the first step is to define the terrain that it would map. Doing that takes tools for surveying social media data sets. The following is a review of several tools that reveal social patterns in these data sets.

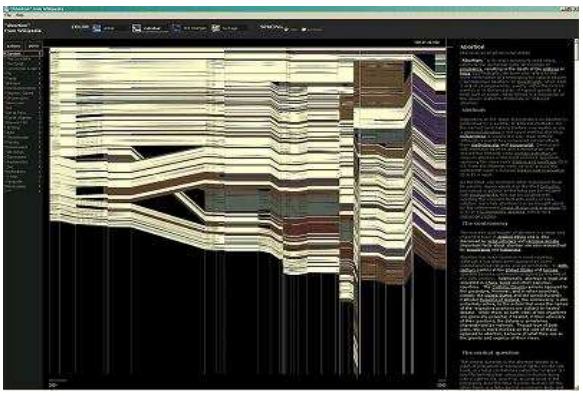
2 Related Work

Visualization has frequently been used in the sciences to illustrate findings of prior quantitative analysis or succinctly summarize those findings to non technical audiences. Visualization can also be used to reveal new relationships, develop hypotheses, refine classification systems, or otherwise discover new insights about how the online social world operates. Many researchers have adopted visualization as an integral strategy of discovery and investigation in research.

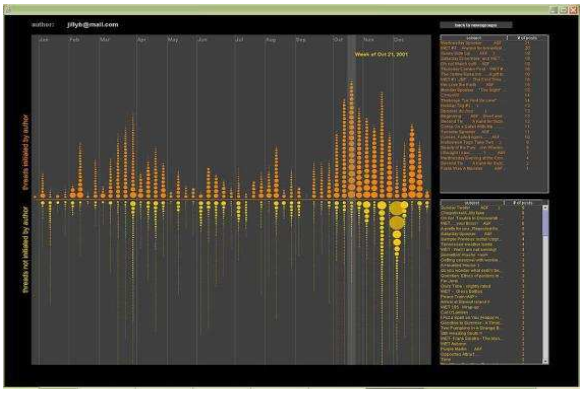
Key data structures in social media are time series, hierarchy, and directed graph. For each of these structures, researchers make use of several visualization strategies are to investigate social media spaces at various scales and levels of detail.

2.1 Time Series

Viegas et al. [10] created “HistoryFlow,” a visualization for examining activity over time on Wikipedia pages. Each version of a Wikipedia page is represented by a colored vertical strip, where different colors represent different editors and the length of the strip represents the volume of contribution in that particular version. Pieces of text that remain the same from version to version are connected by regions of the color corresponding to the respective editor, which allows the viewer to see parts of the page that persist over time. Insertions and deletions manifest themselves as gaps between the connected regions. History Flow analyzes data from online communities to give insight into the evolution of digital artifacts produced by these communities. We present a sample History Flow diagram of the “Abortion” article on Wkipedia below. Notice the shift down and then back up towards the right edge of the graph, which suggests a large (and controversial) section was added to the article, then deleted as status quo was restored.



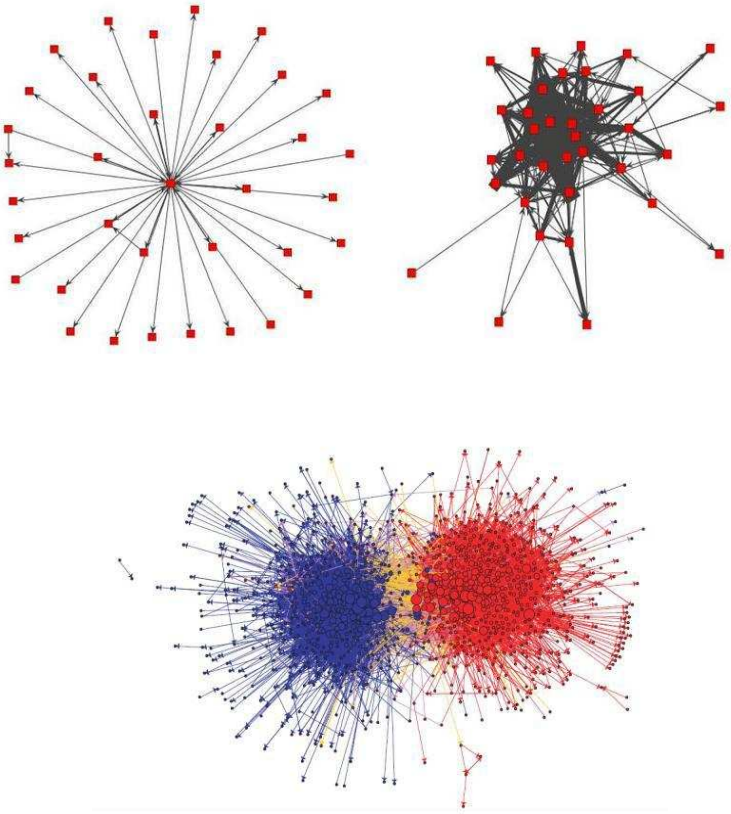
Viegas and Smith [9] take a higher-level approach to visualizing individual activity over time. They develop Author Lines, histograms of user activity in online spaces. These histograms are broken into two halves to represent two contrasting types of activity (e.g. starting newsgroup threads vs. replying to newsgroup threads). The dividing plane is a temporal axis, usually broken down by weeks. For each week, the author line contains zero or more circles of different size. Individual circles represent individual threads started / replied to, and the size of the circle represents the size of the post. Author Lines allow for clear identification of certain iconic roles, such as "Answer person" (his author line would have no activity in the upper half-plane) or "Discussion Person" (his author line would have a small number of large circles). The image below shows the author line for a user with answer person tendencies.



2.2 Directed Graph

A different set of visualizations looks at the social network of interactions in online communities. Welser et al [13] pursue the discovery of iconic roles in Usenet, employing both Author Lines and network diagrams. They construct simple directed networks where nodes represent posters and edges represent replies to other posters. Focusing on individual users, they show only "1.5" degree networks that show only some user, the alters he or she replied to, and the reply edges between those alters. The exemplary networks for an answer person (above, left) and discussion person (above, right) underscore the contrast between these two roles, which is not always visible from Author Lines alone.

At the community level, Adamic and Glance [2] study the network structure of political blogs. In the image below, individual circles are blogs, and edges are URL links between them. Red dots are conservative blogs, blue dots are liberal blogs. The visualization shows not only the clear blog divide along partisan lines, but also the interactions "across the aisle:" orange edges are links from liberal blogs to conservative ones, and purple edges are links from conservative blogs to liberal ones. Given this visualization, a researcher can pinpoint "crossover" blogs that tie the two halves of the online political community together and study them in more detail.



2.3 Hierarchy

Another series of visualizations looks at the community hierarchy. Fiore and Smith [3] use a Tree Map, which collocates all Usenet newsgroups in a rectangle. The highest-level newsgroup labels (alt, soc, comp) partition the rectangle into regions of size proportional to the number of messages that fit under the label. Lower level labels (e.g. soc.culture) partition the region allotted to their parent higher level label in a recursive fashion. The tree map provides a birds eye view of even extremely large communities, such as Usenet. Further, the regions are color coded by the change in number of messages in the respective region since the last time period. Green regions indicate labels with growing numbers of child messages, red regions labels with falling numbers of child messages. The tree map below lays out the Usenet newsgroups with the “comp” label.

3 A Social SQL?

These examples of visualizations of key aspects of social media spaces share common data requirements. A step towards defining and implementing a Social SQL is to enumerate the many facets of social databases. Researchers need to make routine workflow operations on social databases. These frequently repeated operations could be standardized so that researchers can build social accounting metadata in a consistent and simple manner.

Managing social graphs is a database chore that often overwhelms non-specialists. Better tools for storing, indexing, searching, and extracting data from social databases would address the need for managing multiple views of large networks that are changing rapidly.

3.1 Social Queries

Social queries range from standard database operations to those that require the creation of complex and specialized logic to generate more complex data like social networks. It is often a fairly straight forward process to extract a time series of behaviors from social databases. But more complex events over time, like the patterns of connections that develop among participants in social databases, are far more challenging to extract. Social network queries often want to limit or extend the network sub-graph which results from the set of nodes returned by the query. Below, we propose an outline of essential queries, written in natural language, to be supported by a SQL extension for social databases. The highest level of the outline contains three fundamental queries (the third query is an ORDER BY operation). Lower levels of the outline include more specific queries that extract the data necessary for generating visualizations above.

- Extract activity time series for a single person
  - Extract activity time series for all people in a community
  - Extract statistical breakdown of activity for all people in a community by user-defined patterns
    - Extract all activity time series in a community that fit a particular pattern (role)
- Extract the directed social network of all people in a community (user provides definition of relationship between two people)
  - Extract the directed social network of all entities in a community (user provides definition of entity and of relationship between two entities)
  - Extract the 1.5 degree network (ego, alters, ties between them) of all entities in a community
  - Given a node in a social database, find all the other nodes that have a similar network structure (e.g. all the nodes with overlapping connection networks)
- Improve the relevance of search results based on the social network attributes of the author of the result documents

3.2 A meta move to ecological models

As social databases are explored and better tools for studying them become available the effect should be to shift our focus from the detail of events or even of roles to a broader focus on the ecosystem of social media spaces. Once roles are well defined it becomes clear that multiple roles exist and play different and sometimes complementary functions within social databases. Ecologies of interactions become the next unit of analysis as tools lift our focus to the ways whole populations vary in structure and performance over time.

4 Discussion

The mining of directed graphs, particularly social networks, is a topic of growing interest [1]. Visualizing these graphs along with other key structures like hierarchies and time series, enables researchers to observe these patterns and gain deeper insights into the dynamics of social databases. As humans gain these insights, more quantitative approaches can pick up on insights gleaned from the observation of rich visualizations. These quantitative measures can evaluate empirical observations about user and community behavior and provide some measure of the goodness of answers. This approach differs from the typical database research effort that is likely to use algorithms to build and test synthetic data and only later test results on real world systems. Instead, this approach seeks to explore naturally occurring social databases to learn about their basic structure and explore opportunities for enhancement and augmentation.

Naturally occurring social data is the key driver and attraction for many people working around social databases. The goal is to let the data tell its story. Initial work derived from the peculiarities of specific datasets drawn from newsgroups, web boards, email lists, wikis and similar repositories will, over time, compose a picture of social databases in general. Evaluating maps of social databases will become possible once enough of these stories are told and generic structures become visible. Patterns discovered in one social database silo may or may not be corroborated in another silo. Only the widespread collection of data across time and systems will allow for the creation of a systematic taxonomy of social databases. These findings may even apply to other forms of datasets that have similar structures even if not socially constructed. For example, complex networks are present in many large biological datasets. These are a set of tools that help analyze any large collection of data.

5 Conclusion

Picturing the complex data structures that are created when humans interact in and through computational media is a challenging but potentially richly rewarding method for discovery. Information visualization techniques have been increasingly applied to the data generated by social media on the Internet resulting in insights that may have been far more difficult to grasp with either qualitative methods based on reading message content or quantitative statistical methods alone. Finding ideal images for various forms of complex data remains a challenge. Nonetheless, several examples of discoveries about the nature and dynamics of social structures point to the value for research based on graphical representations. Data structures like hierarchies, time series, and directed network graphs are common in most forms of computational social spaces. All of these efforts rest on a common set of queries that, like geographic extensions to databases, should ultimately be supported as a special domain of the structured query language (SQL).

Acknowledgements

We would like to thank Derek Hansen and Eric Gleave for assistance with the paper, and Lada Adamic, Andrew Fiore, Fernanda Viegas, and Ted Welser for screenshots of online community visualizations.

References

[1] Adamic, L. and E. Adar, “How to search a social network,” Social Networks, vol. 27, no. 3, pp. 187-203, 2005.

[2] Adamic, L. and N. Glance, “The political blogosphere and the 2004 u.s. election: Divided they blog,” Working Paper, 2005.

[3] Fiore, A. and M. Smith. “Treemap visualizations of Newsgroups.” Proceedings of CHI, 2002

[4] Rosen, D., Woelfel, J., Krikorian, D., and Barnett, G. (2003). Procedures for analyses of online communities. Journal of Computer-Mediated Communication, 8 (4). Retrieved July 12, 2005 from http://jcmc.indiana.edu/vol8/issue4/rosen.html

[5] Shneiderman, B. (2004). Treemaps for Space-Constrained Visualization of Hierarchies. Retrieved July 12, 2005 from http://www.cs.umd.edu/hcil/treemap-history/

[6] Tufte, E. R. (1995). Envisioning Information (5th printing, August 1995 ed.). Cheshire, Conn.: Graphics Press.

[7] Tufte, E. R. (1997). Visual Explanations: Images and Quantities, Evidence and Narrative. Cheshire, Conn.: Graphics Press.

[8] Turner, Tammara, Marc Smith, Danyel Fisher and Howard T. Welser. 2005. “Picturing Usenet: Mapping Computer-Mediated Collective Action.” Journal of Computer Mediated-Communication. 10 (4).

[9] Vigas, F. B., and Smith, M. A. (2004). Newsgroup Crowds and Authorlines: Visualizing the activity of individuals in conversational cyberspaces. Proceedings of the 37th Hawai’i International Conference on System Sciences. Los Alamitos: IEEE Press.

[10] Viegas, F. B., Wattenberg, M. and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In Proceedings of SIGCHI,pages 575–582, Vienna, Austria, 2004. ACM Press

[11] Wellman, B. 2001. Computer Networks as Social Networks. Science, 293(5537), 2031-2034.

[12] Wang, F., Sha, J., Chen, H. and S. Yang. GeoSQL: A Spatal Query Language of Object-oriented GIS. Retrieved from: http://citeseer.ist.psu.edu./475924.html on 6/16/08.

[13] Welser, Howard T., Eric Gleave, Danyel Fisher, and Marc Smith. 2007. “Visualizing the Signatures of Social Roles in Online Discussion Groups.” The Journal of Social Structure.8(2).

.