Persons, Communities and Artifacts on The Web

Eleni Stroulia

Department of Computing Science,
University of Alberta, 221 Athabasca Hall,
T6G 2E8, Edmonton AB, Canada
stroulia@ualberta.ca

Abstract. In this paper, we review some recent trends in the evolution of social systems on the web today and we discuss how these systems may be brought to bear to improve existing services and to enable new innovative services for web users.

Keywords: Groupwork, Groupware, Social networks, Web 2.0, Web 3D

1 Background

The World Wide Web was conceived and born out of the desire to support information exchange, communication and collaboration. In its 30-year history (and it is flabbergasting to think about how short, in terms of time, this history is and how dense, in terms of events and innovations) it has more than fulfilled its promise and vision while at the same time undergoing three interesting transformations.

In the beginning, the objective of the web community was to enable document publishing and to advance large-scale information communication. The first beneficiaries of this platform were the academic and research community who had the knowledge and skills (a) to develop "web portals" even without any development tools and (b) to access the published information through the original crude client applications. Through this activity, the first broadly usable clients and web-development toolkits were developed and gave rise to portals supported by traditional and new content owners, such as mainstream print publishers (MIT's Tech newspaper in 1991, BBC's TV program in 1994, and the Clinton White House in 1994) and new content providers (Yahoo in 1994). In this stage, the web was ***a web of information*** broadcasted by few to many.

The second phase transition in the Web's history was brought by the advent of ecommerce sites (Amazon and eBay in 1995), which gave rise to ***the web of applications;*** the web became a ubiquitous platform through which to deliver innovative services. The number of providers increased dramatically as the community became ever more creative about the types of services that could migrate to the web. The number of consumers also exploded with the increased availability of user-friendly browsers, search engines (Alta Vista, the first multilingual engine, was launched in 1995) and email-service providers for individuals (Hotmail was launched in 1996). Still, the communication model was broadcasting by relatively few to many.

This changed with the advent of bulletin boards, originally associated with ecommerce web sites, and wikis and blogs, easy to use publication tools for individuals. These tools brought about ***the personal web***, a continuously available whiteboard, hosting everyman's opinions and personal expressions, across the world.

And as the tools for searching, tagging, visualizing and connecting personal posts, published through any of the multitude of available platforms, became increasingly available, ***the social web*** emerged. Today each one of us is linked to a multitude of others through our on-line presence: to the authors of the blogs to which we comment, to the other buyers of the products and services we have bought, to the members of our professional communities (linked-in and ning), to the people whose micro-blog postings we follow (twitter), to our on-line friends (facebook), to the members of our virtual-world communities (second life), and to the users of the on-line tools we use.

Clearly, the original web vision, of supporting collaboration, has been evolving throughout these phase transitions, and today, it appears that the potential for innovative modes of web-enabled collaboration has reached new heights. It is in fact at the core of the "smart planet" ***interconnectedness*** vision, which includes (a) data, (b) system and (c) people interconnectedness.

In our work, motivated primarily by the need to support collaborative software development, we have developed a family of systems for supporting, managing and analyzing different types of collaborative activities. In the rest of this paper section, we review this family of systems and we place it in the context of related work (Section 2). Next, we identify what we believe are some interesting questions in terms of which to understand and analyze social systems (Section 3) and we review our work on SociQL, a social query language designed to support the expression of such analyses (Section 4). Finally, we discuss some ways in which social systems can be brought to bear in service delivery (Section 5) and we conclude with some thoughts on what we expect to be the next important innovations to come (Section 6).

2 Collaboration in the Social Web

In the past several years, our team has developed four different web-based systems to support, manage and/or analyze four different types of collaborative work. Looking back through this work, we have attempted to place it within a coherent conceptual framework by categorizing each tool in terms of two dimensions: (a) the type (and flexibility) of collaborative practices they support and (b) the type of technology/platform they assume.

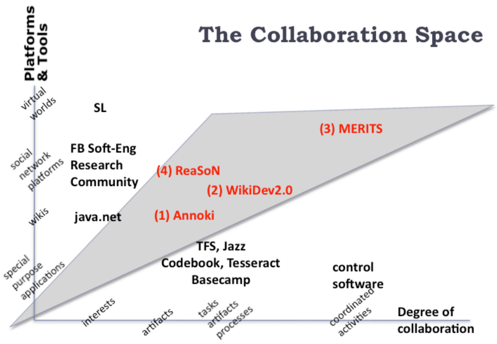

Figure 1: Collaboration Tools in a Two-Dimensional Space

As shown in the vertical axis of Figure 1, the adopted platforms range from task-specific to general-purpose, with the latter category including wikis (and blogs), social-network platforms and virtual worlds. In the horizontal axis of Figure 1, we have identified several interesting spots in the continuum of collaborative practices, from simply establishing communities with common interests, to groups of people sharing artifacts of interest, to teams that collaborate by exchanging artifacts and information according to established process, to very regimented workflow tools that enact well-defined processes to which people contribute well-structured artifacts.

2.1 Annoki [1,8]

Annoki was built on top of the popular MediaWiki, to support the collaboration of our research team. We chose MediaWiki as the platform because its default features fulfill many of our original requirements for our envisioned research-collaboration tool. First, it provides "user" pages, for the personal use of the wiki members, and "regular" wiki pages where content is collaboratively edited. Second, it has a "discussion" page for each "regular" page, thus enabling a distinction between "content" and "reviewer's comments" among the collaborators. Third, it supports concurrent editing of pages (with the multiple versions getting merged a-la SVN) and notifications of users when a page of interest changes; these two features enable a tighter, more synchronous coordination among collaborators. Finally, it supports templates, so that in addition to free-formatted pages, structured information can be collected.

To the MediaWiki features, the Annoki toolkit adds the following set of extensions.

Namespace-based access control: Each Annoki user has an associated namespace and all pages he creates belong in this namespace. Group namespaces can also be defined to organize wiki pages that "belong" to a group of users. Pages belonging to a "public" namespace are visible to all. In this manner, layers of protection can be supported for personal, project-specific, organization-related, and publicly accessible content.

Annotations and visual editing of template instances: To enable lightweight cross-referencing of pages, users can annotate pages with their own tags. In this manner, users can superimpose a personal layer of their own on the wiki resources. Users can also create and edit pages based on templates using a graphical editor, removing the burden of writing wiki page code.

Visualizations: Annoki is equipped with two types of rich, interactive, Ajax-based visualizations: WikiMap and wiEGO. WikiMap is a visualization of the whole wiki structure (users, pages, links among pages and authorship relations between users and pages). The set of wiEGOs are visualizations of the semantic structures implicit in a set of special template-based pages corresponding to concepts in Blooms taxonomy (i.e., tree, topic, persuasion, brainstorm, story, and decision maps, as well as flowcharts).

Collaboration and contribution analysis: extending the default differencing capability of MediaWiki, Annoki supports analysis of the page edit history at the level of sentences, and collects metrics of each user's contribution to each page and to the wiki as a whole in terms of sentences added, deleted, and edited.

2.2 WikiDev2.0 [3,4,6]

The WikiDev2.0 tool for collaborative software development was conceived as a lightweight platform through which to integrate information about various software artifacts produced in the variety of tools used by the software team (code, documentation, communication messages etc), to analyze this information in order to infer interesting relations among these artifacts, the team members and their activities, and to present views on this information that cut across the individual tool boundaries.

The ***code and communication clustering*** process of WikiDev2.0 consists of the following steps. The first step involves parsing of all the textual information associated with the input information feeds, to recognize mentions of team members (their names, nicknames, or IDs) and software artifacts (classes, methods and interfaces). The recognized references introduce the explicit relations between people, code and communication artifacts. A subsequent step calculates the implicit relations based on triangular inequality thus providing further insights about hidden dependencies among these artifacts. Using this information, one can see who works on what artifact currently, who has discussed a specific artifact that should be potentially consulted about changes to it, and how a member's own work might affect other people's work.

The ***syntactic-semantic text-analysis feature*** of WikiDev2.0 is meant to further enhance the ability of the tool to recognize relations among people, code and communication artifacts, implicit in the large amounts of textual information collected through the software process. The process consists of (a) a syntactic parsing stage for all textual content in WikiDev2.0, (b) an annotation stage, where the syntax trees of the parsed sentences are annotated with semantic information (such as team-members' names and code artifact names), and (c) a pattern-matching query stage that extracts subject-predicate-object triples from the annotated parse trees, corresponding to relations such as "who worked on what", "who has experience in what" etc.

We have developed ***a variety of visualizations*** in WikiDev2.0 to communicate the state of the project to team members and managers. Traditional line- and graph-based charts communicate the amount, type and frequency of team-members' activities. The UMLViewer resents an UML-like view of the code artifacts, annotated with information about their developers and evolution.

Finally, ***WikiDev2.03D*** is an extension built in the Open Wonderland virtual world, which visualizes the discovered clusters adopting a 3D city metaphor. This virtual-world view of the project can be visited and discussed by multiple interested parties at the same time, thus enabling a shared understanding of the software project.

2.3 MERITS [2,5]

The MERITS system is the third of our collaborative-work support tools and it focuses on activities that are more complex and involve, in addition to information exchange, interactions among people and between people and the real world. To that end, it combines the immersive, collaborative potential of virtual worlds with BPEL-based process specification to enable (a) instructors to specify educational scenarios, and (b) students to experience those scenarios in a realistic, interactive manner.

The MERITS framework offers two important features. The first involves a method and tool support for ***specifying complex collaborative processes,*** including tools for specifying the behavioral capabilities of the various roles in the process in terms of web services invoked by avatar actions in the virtual world, as well as developing behavioral scripts for real-object simulacra in the virtual world.

The second important feature is a ***comprehensive action-recording*** tool that produces a compact trace and a synchronized trace of all in-world actions, which can then be parsed to identify interesting action patterns.

2.4 ReaSoN [7]

Shifting focus from supporting to analyzing collaborative activities, we developed ReaSoN (again based on Annoki), a comprehensive set of tools for visualizing and exploring the social networks, implicit in academic research practices. In doing so, ReaSoN contributes to the understanding as well as fostering of the social networks underlying academic research.

In terms of ***visualizations***, ReaSoN offers specially structured pages to communicate information about individual and collections of *publications*, *authors*, the *communities* around conferences and journals, the *keywords* of publications, and the geographical distribution of people, keywords and communities.

More interestingly, ReaSoN also provides infrastructure for asking customized queries about researchers, their collaborators, their publications and their citations. The results of a query can be visualized in tabular form, explored in the WikiMap graph format (making visually explicit the network of authors and publications) or plotted on a map (visualizing the geographical relations among people organizations and their research activities).

3 Analyzing the Social Web

Developing and reflecting upon the four systems we described in the last section, we have come up with a set of research questions (and associated technical challenges) that cut across most (all?) web-based social systems today. We review these questions in the remainder of this section, organized in two different groupings of "analysis questions" around social systems and possible "services supported" by social systems.

Today, there exists a plethora of social-networking sites, each one supporting different means of "connecting" among members and catering to different demographics. Some sites enable bi-directional connections, like Facebook, where others enable directed connections, like Twitter. MySpace caters to a younger demographic than Facebook, which in turn is surpassed in popularity by Orkut in Brazil. In addition to these "superficial" differences, each of these social networks encourage different types of communications. Facebook appeals to people who want to keep in touch with family and friends where twitter seems to be the medium of choice for people to share and access information from a wide variety of channels. Facebook favors deeper connections and enables the organization of these connections in groups so that different personas can be projected to each of them. Twitter, on the other hand, encourages maximization of connections (followers) and enforces a single persona its users who cannot distinguish their followers in groups. Clearly these differences deserve deeper analysis; in the mean time, all of these networks share three important concepts, i.e., community, contribution and influence.

3.1 Recognizing Communities

Groups of collaborating people are not uniformly cohesive. Some members are more highly connected to each other than to the rest of the group. This is a corollary to the "homophily" phenomenon. Homophily is the tendency of individuals to associate and bond with similar others. Individuals in homophilic relationships share common characteristics (beliefs, values, behaviors, etc.) that make communication and relationship formation easier. In principle, graph algorithms for connected-components' recognition can be applied to recognize such "cliques". Alternatively, domain-specific notions of subcommunities can be defined.

Let us review the issue of "recognizing communities" in the context of our systems above. In ReaSoN, we have analyzed the communities of authors who have published in specific conferences over a period of time and the intersections of these communities with each other. In WikiDev2.0, we have analyzed the email communications of team members to recognize subgroups who have communicated most frequently with each other. We have also clustered communication artifacts around the code artifacts they relate to, and by implication the authors of these code and communication artifacts. In Annoki members belong in project-related Namespaces, which essentially define the communities of members and documents that are associated with a project; thus there seems to be no point in recognizing implicit subcommunities. In MERITS workflow-defined simulations, the activities of the various participants are understood in terms of the workflows they enact; however, in cases of more open-ended activities, special-purpose relations (like communication) could be defined in terms of which to recognize dense subcommunities.

3.2 Recognizing Contribution

As the collaborating community increases, the roles of individuals become blurred and unclear. In WikiDev2.0, for example, most teams consist of four to six developers (plus TAs and instructors). Contrasting this to the about 200 members of the Annoki installation for the software-engineering group at the University of Alberta, it becomes clear that the latter community is much more complex than the former. Recognizing the contribution of individuals in the latter context becomes challenging.

MediaWiki, as well as most wikis, offers a differencing capability, which summarizes the contribution of an individual to a specific version. Annoki provides a more sophisticated contribution analysis and visualization tool, which summarizes the contribution of an individual to a wiki page over its lifecycle. WikiDev2.0 implicitly recognizes contribution in terms of frequency of SVN commits, wiki-page edits, and email communications. ReaSoN offers a variety of bibliometrics-inspired statistics to measure the "importance" of each author, including their h-index and various pagerank calculations of the influence of their papers to other papers and their corresponding authors through citations. MERITS does not offer an explicit contribution-measurement solution since it is a framework, and contribution measures, in general, have to be aware of the nature of the collaboration activity. Instead, through its recorded activity logs one can define metrics of interest based on the participants' in-world activities and measure contribution in different ways. For example, one can imagine that it would be interesting to identify the persons who talked the most during a session or the person who made the most interactive gestures (like shaking hands for example) with others.

This discussion assumes that "importance" is semantically equal to "contribution" which is not necessarily the case. Domain-specific importance metrics can be based on different person attributes, but contribution appears to be a cross-domain importance metric.

3.3 Recognizing Influence

Related to the concept of contribution is the concept of influence. Within a collaborating community, people influence their collaborators through their contributions. Not all contributions however are equally likely to be consumed and to influence other people's contributions.

In ReaSoN, we measure influence through pagerank calculations over the implicit coauthorship and citation networks. Through these metrics, one can recognize authors with broad co-authorship networks, i.e., who have written papers with many other authors who have similarly written papers with many others etc., as well as authors with broad citation networks, i.e., authors whose papers have been cited by many authors whose papers have been cited by many other papers etc. In WikiDev2.0, the clustering process implicitly attempts to recognize the members' influence to code artifacts by collecting references of other materials, associated with team members, to these artifacts. It does not offer however any insight on how to compute any type of transitive closure of these relations.

4 SociQL

Aiming at understanding the various types of collaborative work exemplified by the above systems and at supporting a general conceptual framework in which to address the above research questions, we are now working to design asocial query language. SociQL is a query language, and an associated prototype implementation, that supports for the representation, querying and exploration of disparate social networks.

Unlike generic web query languages, SociQL is designed to support the examination of such sociological questions, incorporating social theory and integration of networks that form a single unified source of information. In sociology, object-centered sociality characterizes social relations between individuals by means of objects. In this setting, specific constitute evidence of social relations (Knorr-Cetina 1997). Essentially, while recognizing the social interaction between *individuals*, this theory exalts the role of *specific objects* as the reasons why social actors affiliate with each other.

For this reason, we define SociQL's data model around the concept of an **object**. For instance, in the context of ReaSoN, we have that a paper (an object) connects the researchers who authored it; similarly, a publication venue (an object) connects authors who publish their work in it. In the MediaWiki-based, Annoki and WikiDev, the wiki pages are the objects that connect the pages authors. In MERITS, the avatars are connected through the simulacra of the real-world objects they manipulate as well as through their communication objects, i.e., their text and voice utterances.

In our model, both objects and relationships are described by **properties** (actual data), such as the name of an author or the date in which a citation is made from a paper into another. We also distinguish the **context** in which properties are defined to describe the objects and relationships. For instance, the same query might return different email addresses for the same individual depending on the context in which the query is asked (professional or personal). In practice, each context will correspond to different social network system—thus, each context may have its unique data access methods and privacy restrictions, which complicates query processing to a great extent (as discussed below).

As it turns out, this problem is extremely hard to solve in practice. In order to correctly interlink the different communities, different social network sites describing the same object would have to refer to it with a globally consistent identifier. In practice, however, each site has its own local identifier, unique only in its particular context. This practice results in a proliferation of identifiers that make it harder to merge social networks.

5 Services with Social Support

As the number and types of social systems increase and so is their membership, the question becomes to identify the means through which they can be brought to bear in delivering novel and/or improved services.

An interesting new technology than can provide a catalyst for the deployment of social-network information to consumers is the combination of QR tags and the availability of tag readers on almost all new mobile devices. By scanning the QR tags annotating real products and business cards, mobile apps can inform the individuals' networks of their real-world consuming behavior and social interactions. In this manner, the social network itself can seamlessly expand through traditional real-world practices (like business-cards' exchanges) and the word-of-mouth advertisements for products and services can efficiently travel through it. Similarly, information about individuals' entertainment choices can be propagated through their networks as, increasingly, we are consuming entertainment, games, audio and video, through the Internet. As more information is shared by the network, collaborative filtering becomes more effective in advising the network members about what their connections buy, play, listen and watch. And to the extent that more network members choose to make similar choices they can negotiate better prices and improved quality for their "group buying".

6 Summary

In this paper, we discussed our recent work on four collaborative/social systems, and our more recent work on a social query language designed to express the types of questions that users (and applications) may want to answer in the context of such systems. Further, we reviewed the types of analyses that we believe are relevant in the context of social systems and the ways in which social systems can be deployed to improve current services to Internet users and to enable new innovative services. This is clearly an active and fascinating area with a huge number of open questions and substantial opportunities for the development of innovative intelligent services.

Acknowledgements

This work was supported by NSERC, iCORE, IBM, and was conducted in collaboration with B. Tansey, K. Bauer, M. Fokaefs, D. Chodos, D. Serrano and D. Barbosa.

References

1. O. Arazy, Stroulia E., Ruecker S., Arias C., Fiorentino C., Ganev V., and Yau T., Recognizing Contributions in Wikis: Authorship Categories, Algorithms, and Visualizations, Journal of the American Society for Information Science and Technology (JASIST), 61(6):1166-1179, Mar. 2010.
2. D. Chodos, P. Naeimi, E. Stroulia: A simulation-based training framework for health-science education on video and in a virtual world, Journal of Virtual Worlds Research 2(1), Apr. 2009.
3. M. Hasan, E. Stroulia, D. Barbosa, M. Alalfi: Analyzing Natural-Language Artifacts of the Software Process, ICSM 2010, ERA track, 26th IEEE International Conference on Software Maintenance, September 12-18, 2010, Timișoara, Romania.
4. M. Fokaefs, D. Serrano, B. Tansey and E. Stroulia: 2D and 3D Visualizations in WikiDev2.0, ICSM 2010, ERA track, 26th IEEE International Conference on Software Maintenance, September 12-18, 2010, Timișoara, Romania.
5. D. Chodos, E. Stroulia, P. Kuras, M. Carbonaro, and S. King: MERITS Training System - Using Virtual Worlds for Simulation-based Training, CSEDU 2010 (the International Conference on Computer Supported Education), April 6-9 2010, Valencia Spain.
6. K. Bauer, M. Fokaefs, B. Tansey and E. Stroulia: WikiDev 2.0: Discovering Clusters of Related Team Artifacts. CASCON 2009, Toronto, Canada, November 2-6, 2009.
7. V. Ganev, Z. Guo, D. Serrano, B. Tansey, D. Barbosa, E. Stroulia: An Environment for Building, Exploring and Querying Academic Social Networks. The International ACM Conference on Management of Emergent Digital EcoSystems (MEDES) 2009, October 27-30, Lyons, France.
8. B. Tansey, E. Stroulia: Annoki: A MediaWiki-based Collaboration Platform, In. Web2SE: First Workshop on Web 2.0 for Software Engineering, ICSE 2010.