Learning about team collaboration from Wikipedia edit history

Adam Wierzbicki
Polish-Japanese Institute of Information Technology
Warsaw, Poland
adamw@pjwstk.edu.pl

Piotr Turek
Polish-Japanese Institute of Information Technology
Warsaw, Poland
piotr.turek@pjwstk.edu.pl

Radosław Nielek
Polish-Japanese Institute of Information Technology
Warsaw, Poland
radoslaw.nielek@pjwstk.edu.pl

ABSTRACT

This work presents an evalation method of teams of authors in Wikipedia based on social network analysis. We have created an implicit social network based on the edit history of articles. This network consists of four dimensions: trust, distrust, acquaintance and knowledge. Trust and distrust are based on content modifications (copying and deleting respectively); acquaintance is based on the amount of discussion on articles' talk pages between a given pair of authors and knowledge is based on the categories in which an author typically contributes. As authors edit the Wikipedia, the social network grows and changes to take into account their collaboration patterns, creating a succinct summary of entire edit history.

Categories and Subject Descriptors

J.4 [**Social And Behavioral Sciences**]: Sociology; K.4.3 [**Computers And Society**]: Organizational Impacts—*Computer-supported collaborative work*

General Terms

Human Factors, Experimentation

Keywords

Wikipedia, Social networks, Collaboration

1. INTRODUCTION

Wikipedia is today not only one of the most prominent examples of Web-based collaboration sites, but is also used as a model for the new ways of working in which people communicate bypassing hierarchies and collaborate outside of artificial organizational boundaries. This kind of work called *swarm creativity* has been described by Gloor [2] as being part of a Collaborative Innovation Network (COIN), defined as a cyberteam of self-motivated people with collective vision, enabled by Web to collaborate in achieving a common goal by sharing ideas, information and work.

WikiSym '10 July 7-9, 2010, Gdańsk, Poland

To learn more about the social structure of such collaboration, studying Wikipedia is a good place to start. Wikipedia edit history is a rich source of information about the collaboration patterns of teams that emerge from a COIN. The purpose of this work is to create a model of the social structure of such teams and to evaluate this model through an attempt to distinguish between excellent teams and the less successful ones. Teams of authors in Wikipedia can be seen as model of collaborative work in the future economy [4], so it seems worthwhile to ask: what social factors determine the quality of a team in Wikipedia?

To model the relationships between authors we used the notion of multidimensional social network. A social network is a directed graph, where nodes represent people or organizations and weighted edges represent their relationships. Multidimensional social network consists of one set of nodes and several sets of edges to model various kinds of relationships, in this case: trust, distrust, acquaintance and knowledge.

2. PROCESSING THE EDIT HISTORY

For the purpose of this research we have selected the Polish edition of Wikipedia as it was sufficiently developed (over 650,000 articles) and known to us. The edit history contains every revision of every page since the inception of Polish Wikipedia (over 200GB of uncompressed text).

We needed to be able to recognize changes in text including finding the fragments which were pasted from somewhere else or moved around. We need this to preserve authorship information for each word in the text. We have used the Karp-Miller-Rosenberg algorithm [1] to build the structure called *dictionary of basic factors*. DBF allows us to check in constant time if two fragments of text are identical. If long enough fragment of text appears in next revision then we know that it wasn't deleted, though it could be pasted or moved from somewhere else. This way when someone pastes or moves content, he or she doesn't become its author.

The output of this stage of processing is the edit history with each word having assigned original author. It is easy to find whose text was modified and by whom.

3. THE SOCIAL NETWORK

Basing on preprocessed edit history we have created a social network. Contrary to most social networks, this one is based on real interaction history, not on information provided by the members, thus it is implicit, contains only behavioral and not declarative data. To cover only real, identifiable Wikipedia contributors, we have filtered out the users,

who did not log in (anonymous) and automated scripts that perform edits (bots). Anonymous contributors cannot be well identified as the only information available about them is the IP address or hostname.

Inspired by [3], we have defined four measures of relationships (network dimensions): trust, distrust, acquaintance and knowledge. Each of them represents one aspect of collaboration between a pair of contributors.

3.1 Trust

We define trust as a tendency of one author to believe in the credibility of another. In Wikipedia this may be operationalized by the copying and pasting of text. When one author copies or moves content within an article we assume that he or she has read it and considered it trustworthy. Surprisingly such actions are very common in the edit history, as text is constantly moved around and reorganized by various contributors.

3.2 Distrust

Distrust is considered as conceptually different from trust and has been separated from the trust network. Distrust values cannot just be subtracted from trust due to significant differences in scale of these measures. Distrust is operationalized through the deletion of text. We assume that when an author deletes some words and does not paste them elsewhere, he or she considers conveyed information as irrelevant, unreliable or not true.

3.3 Acquaintance

Acquaintance is the measure of how well two people know each other, usually in person. This information is not directly available from Wikipedia edit history, so we used as an approximation the activity on the articles' talk pages, where authors usually discuss the scope of an article and resolve disputes and other problems. This is the virtual equivalent of normal meetings to address specific issues and – to some extent – to socialize.

3.4 Knowledge

Knowledge of team members is a measure of their expertise in given fields. In Wikipedia this is easily modeled by authors' activity in categories. Most contributors are active in many categories, partly because of categories' granularity, most articles belong to many categories. On the other hand, those who make most edits in a narrow set of categories most probably have some level of expertise in them. Contrary to previous network dimensions, knowledge consists of a bipartite graph where links form ordered pairs of authors and categories in which they have contributed. It may be easily transformed to create standard author-to-author links used in typical social networks by connecting authors that share links to the same category.

4. TEAM QUALITY

A team is a subgraph of the entire social network. We have selected a team for each Wikipedia article as a set of contributors who wrote the final version of the text. We distinguished good teams as those associated with featured or good articles to evaluate our approach.

We have defined simple criteria based on trust, distrust and acquaintance focused on aggregating the intra-team link strengths. The simplest measure is just the sum of those strengths. It is strongly related to the team size, the more members of a team, the more opportunities to have strong links. To address this, we propose two averages: one is just the sum divided by the number of those links, thus it is average link strength. The second is the sum divided by number of possible links between nodes, it represents average link strength with nonexistant links counted as links of zero weight.

Criteria based on knowledge are different due to the nature of this dimension. As teams have been created from articles, for each team we have the relevant categories for it. The first measure is calculated as the average team expertise in each of these categories. Team expertise in the average member's expertise (weight of a link between contributor and particular category). The second measure is based on looking for an expert (the member with highest expertise level) in relevant categories and selecting the category for which the knowledge of chosen expert is lowest.

5. CONCLUSIONS

Our initial results indicate that acquaintance and trust have a significant positive impact on team quality. Average criteria values for those two dimensions for good teams are higher than for other teams, and the difference is statistically significant (also for the variance).

On the other hand, distrust behaves unexpectedly. It turns out that distrustful behaviour is also beneficial for team quality. This finding can be interpreted as follows: our operationalization of distrust is consistent also with critical behaviour, and if accompanied by a high activity on talk pages, can improve the overall team result. This is an important finding for Wikipedia. We are working on new distrust criterion that would put an emphasis on mutual distrust and plan to carry out experiments that would see whether this criterion has is negatively correlated with team quality (mutual distrust has been linked to edit wars).

The knowledge dimension has been based on Wikipedia categories. It has turned out that many of these categories are used for an improvement of content organization and navigation and do not represent domains of knowledge. Because of that the knowledge criteria have not been useful in evaluating team quality so far; we are working on a distinction of Wikipedia categories that represent knowledge domains to improve the knowledge dimension.

6. ACKNOWLEDGEMENTS

This work has been supported by the research grant 69/N-SINGAPUR/2007/0 of the Polish Ministry of Science.

7. REFERENCES

[1] M. Crochemore and W. Rytter. *Jewels of stringology.* World Scientific Publishing Co. Pte. Ltd. Singapore, 2002.

[2] P. A. Gloor. *Swarm Creativity : Competitive Advantage through Collaborative Innovation Networks.* Oxford University Press, USA, January 2006.

[3] A. Hupa, K. Rzadca, A. Wierzbicki, and A. Datta. Interdisciplinary Matchmaking: Choosing Collaborators by Skill, Acquaintance and Trust. 2009.

[4] D. Tapscott and A. D. Williams. *Wikinomics.* Hanser, München, 2007.