EVALUATING WIKI CONTRIBUTIONS USING SOCIAL NETWORKS: A CASE STUDY ON WIKIPEDIA

Nikolaos Korfiatis and Ambjörn Naeve
Knowledge Management Research Group (KMR)
School of Computer Science and Communication (NADA)
Royal Institute of Technology (KTH), Stockholm
SE-100 44 Stockholm, Sweden
n-korf@nada.kth.se , amb@nada.kth.se

In this paper we present an approach to the problem of evaluating contribution in shared access repositories such us wikis based on the activity of the contributors as denoted by social network measures. An approach to the concept of Wiki is given using models and techniques of Social Network Analysis targeting the patterns of social ties between contributing authorities. A case study in the English Language is provided as a proof of concept using the metrics of degree centrality and group degree centralization.

1 Introduction

The web has given rise to new forms of collaboration and interaction facilitating the manipulation of shared artefacts and information spaces [1]. In the current state of the art the web ecosystem consists of resources (web pages/ files) linked though hypertext thus forming a system of links denoting references to these resources as well as providing views to the requesting authorities. However one of the initial design goals of the web was not only to facilitate views of the web resources requested but also to allow editing and annotation of these resources in a simple way [2]. On the foremost approaches of this goal the concept of wiki [3] has given a response to this challenge. WikiWiki applications facilitate a way of collaborative editing supported by a revision mechanism which allows the monitoring of changes and contributions to the sections edited. The use of WikiWiki applications is common on cases such us formation of collaborative document editing (eg. In communities of Practice) or formation of shared knowledge repositories (for instance the Portland Access Repository[a]). One of the most well known implementations and example of usefulness of wiki to support collaborative document editing is the wiki based encyclopaedia WIKIPEDIA[b] and it's related projects[c] .

Traditional encyclopaedias are often characterized of a high level of credibility by domain experts taking into account the background process that has resulted to the encyclopaedia (Domain Authorities contribute to the final outcome). From the other hand WIKIPEDIA since it uses the WikiWiki system allows the editing and creation of encyclopaedic articles by anyone who wishes to contribute as its primary target is to provide free editing access and gather knowledge regarding the term presented and thus not evaluate the contributing authorities. However as the content increases along with the

[a] http://c2.com/wiki
[b] http://www.wikipedia.org
[c] http://www.wikimedia.org

contributing sources (see figure 1) a critical issue has emerged regarding the credibility of WIKIPEDIA as an authoritative source that can be used as a reference [4, 5]. The question is extended not only to the outcome (article) but also to the process of shaping the article, in which a contributor would allow another authority to submit, change or delete a contribution that is accepted or not accepted by him/her. WIKIPEDIA has internal mechanisms of managing those cases such us a permission ranking system, where a contributor is accredited by the level of participation in the shaping of the article, as well as a discussion tab on most of the articles or notifications and warnings regarding the content. Nevertheless the research question deals on how to provide a clue of credibility for an article based on the contributing authorities and their acceptance on the community of their fellow contributors.

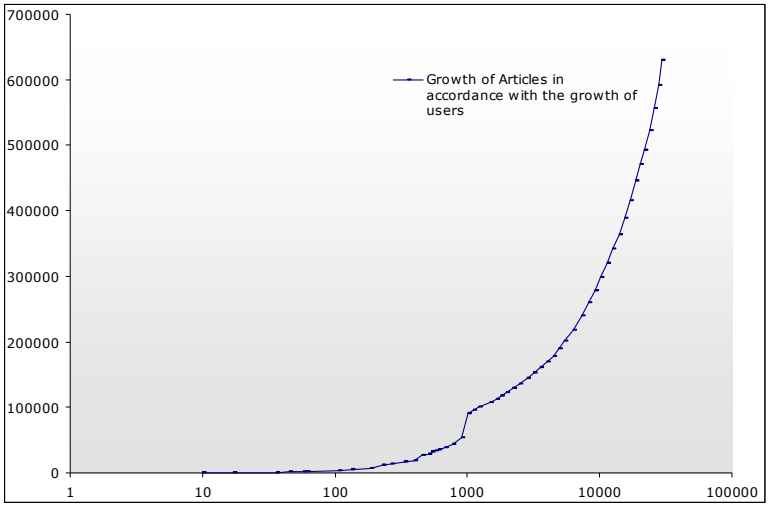

Figure 1: Growth of Articles in accordance with the growth of users in the English WIKIPEDIA (Statistics obtained from http://en.wikipedia.org/wikistats/EN/TablesWikipediaEN.htm). Values on X axis represent the articles are on logarithmic scale. Values on Y axis represent the number of contributors.

In this paper we present a first attempt to model the problem towards providing an authoritative ranking mechanism based on social interaction data collected through the wiki. We then model the credibility of each contributor using the metric of centrality thus producing an overall centrality measure for the article depicting the social activity / process that has taken place through the shaping of the article. We argue that this factor can be used as a metric of credibility representing the article and the contributing authorities.

2 A network approach on the Wiki publication model

Social Networks and Social Network Analysis in particular [7] [9], is a research paradigm which tries to unravel patterns of social relationships across various individuals in a social context. Following the patterns and measuring structural and compositional values in the networks we ought to understand the basic structure and properties of the network and explain its behaviour thus uncover those actors that their actions characterize the most of the activity described by the network. Social Network Analysis focuses on a more rationalistic approach on research on organizations and social groups [6] since it tries to expand interdependent relations. WikiWiki applications facilitate a case where social relationships are established over a domain of social actions such us acceptance, objection or rejection

of a contribution. Moreover as in the case of WIKIPEDIA the wiki facilitates a collaborative document editing effort relying on the contribution of multiple authors in a concurrent system that would be able to combine the contributions in an effective and democratic way which would allow all the ground knowledge about the article/lemma to be present in the most recent revision of the article. By democratic we also refer to the ability of anyone who uses the wiki to contribute content or to make modifications to content contributed by someone else. In that sense as the wiki-fication continues the final document (or the most recent revision) is the outcome of a community process involving a certain amount of social interactions embedded in the content modification used as a mean of expressing them.

From a social research point of view what makes such a case interesting is the negotiation process that takes place when writing and structuring the article. For example a user makes a contribution which is erased and this user tries to establish its contribution back (to make it visible and accepted by the others). In both cases (article and negotiation) there are interaction ties which characterize the final outcome and the dynamics of the process. In this paper we will focus on the interaction ties between multiple contributors working in the same article or domain of articles in the WIKIPEDIA namespace. However to do this kind of study we first need to define the structural and compositional variables that characterize such a network.

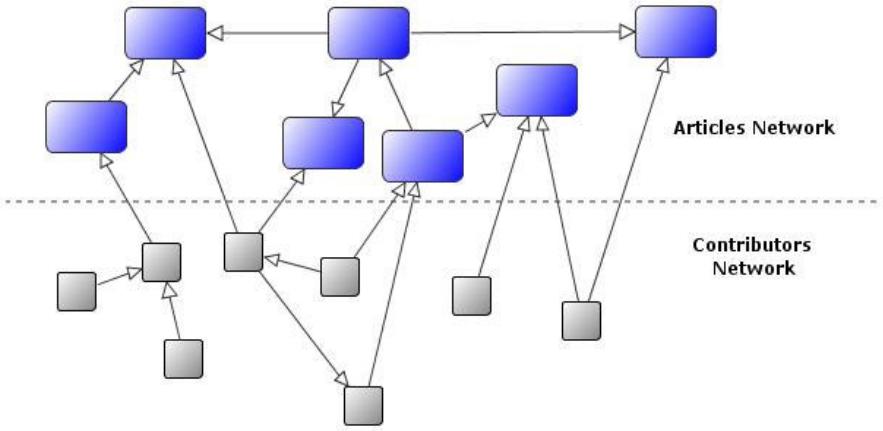

Figure 2: Network layers in the wiki publication model. Contributors are linked together by working in common projects (articles) in the WIKIPEDIA namespace.

In the wiki publication model we can see both the necessary structural and compositional variables important for the construction of a social network of contributors over an article or topic in the WIKIPEDIA. Structural and Compositional properties of the publication model can be found in the following use-cases:

- When a contributor edits content that has been submitted by someone else then it establishes a tie with him/her. This is depicted by an acceptance factor which represents the percentage of the content of the previous contributor that is visible after.
- Every contributor that has a single or more contribution to the article establishes a relational tie with the other content contributors of the article. Evidence of participation in common projects strengthens this tie.

We can also link actors through two different layers of networks (figure 2):

- *The Articles Network:* Every article in the WIKIPEDIA contains reference to other articles as well as external references. A set of links used for classification purposes is also available in most of the active articles of the encyclopaedia. Every article represents a vertex in the article network and the internal connections between the articles the edges of the network.
- *The Contributors Network:* WIKIPEDIA is a collaborative writing effort which means that an article has multiple contributors. We assume that a contributor establishes relationship with another contributor if they work on the same article. In the resulted weighted network each contributor is represented by a vertex and their social ties (positive or negative) are represented by an edge denoting the sequence of their social interaction.

The resulted graph is a two-mode network where we have two set of entities: articles and contributors. Contributors can be either connected (belong to the same article) or interconnected (common contributions on two or more articles in the same domain). In an article domain of high credibility it is expected to find more interrelations since the contributors may contribute content to more than one article thus depicting their common interest. In that case the more affiliated a contributor becomes with a domain then the most interested with the article he is thus his background is depicted to have knowledge of the domain. Consider for instance a contributor that has made a lot of contributions to the domain regarding the history of Spanish colonies in Latin America. The author has also done some contributions in the article of Anatomy. However the author is more affiliated with the articles regarding the history of the Spanish colonies than the medicine therefore his contribution in medicine may be considered as less authoritative than the contribution in the other domain since its knowledge of the domain is not as extensive as it is in the other.

In social network analysis there is a variety of measures that can assess this kind of social activity in a sociometric study. As we have already defined our graph we can use some common social network metrics to extract this kind of information from WIKIPEDIA data.

3 Network measures in the Wiki Contributions

As aforementioned contributors make contribution to one or more articles that belong to the same or different domain. Based on this activity we can evaluate the activity of the contributors thus extract metrics of their presence in the domain of the articles. To do this we select to use the Centrality index [8],[10] of the resulted graph and in particular a measure of centrality dealing with the degree (total of incoming, outgoing edges) of the vertex/contributor in the examined article.

The concept of centrality is well celebrated in social network analysis since there are numerous studies showing the usefulness of such a metric for measuring activity in social networks [11] [12]. In sociometric studies the usage of centrality is targeting to unfold the person/individual that is the most prominent in a network thus rank the actors according to their position in the network and is interpreted as the prominence of actors embedded in a social structure. In our study we use the degree centrality index which is the simplest definition of centrality and is based on the incoming and outgoing adjacent connections to other contributors in an article graph. To measure the centrality in individual level we define the Contributor Degree Centrality and to an article level the Article degree centralization which represents the collective of Contributor Degree Centrality.

Contributor Degree Centrality

In classical social network models the inner degree (the amount of edges coming into a node) is representing the amount of choices the actor has over a set of other actors. However in our wiki network model the amount of incoming edges represents edits to the text therefore the metric of inner degree is the opposite, meaning that the person that has the biggest inner degree has the biggest amount of objection/rejection in the contributor community thus it receives a kind of negative evaluation from his/her fellow contributors. From the other hand the outdegre of the vertex represents edits /participation in several parts of the article thus gives us a clue of the activity of the person in relation with the article and the domain.

Mathematically we can represent such formalism as follows: Considering a graph g representing the network of contributors for an article contributed in the wiki then the Contributor Degree Centrality - a contextualized expression of actor degree centrality [12] [13] - is a degree index of the adjacent connections between the contributor and the other contributors that edit the article. From graph theory the outer degree of a vertex is the cumulative value of its adjacent connections.

$$C_D(n_i) = d(n_i) = \sum_j x_{ij}$$

The adjacent x_{ij} represent the relational tie between the contributors i and j over the domain of the article they contribute and is characterized also by the visibility of the contribution in the final article and can be either *1* or *0*. To provide the centrality we divide the degree with the highest obtained degree from the graph which in graph theory is proved to be the number of vertices (g) minus one. Therefore the contributor degree centrality can be calculated as:

$$C'_D(n_i) = \frac{d(n_i)}{g-1}$$

Article Degree Centralization

We define an Article's degree centrality C_{DM} as the variability of the individual contributor centrality indices. The $C_D(n^*)$ represents the largest observed contributor degree centrality

$$C_{DM} = \frac{\sum_i^g [C_D(n^*) - C_D(n_i)]}{(g-1)(g-2)}$$

Again we divide the variability with the highest variability observed in the graph which is proved to be $(g-1)(g-2)$.

Having defined the metrics we apply them and explain their qualitative values in a case study on the English language WIKIPEDIA.

4 A case Study on WIKIPEDIA

In order to provide a qualitative analysis of the metrics deployed on the evaluation of the articles we picked ten articles with a similar amount of contributors of the domain "Philosophy" from the English language WIKIPEDIA. Table 1 shows the list of articles used in the case study as well as the values of their article degree centralization.

The data was collected using the Python WIKIPEDIA robot framework[d] for each one of the articles. The resulted networks contained an average of 259 contributors per article and the average article interrelations per contributor were approximately 2. We used the diff function of the wiki to assess the tie between the pair of contributors as modelled from Section 2. The data then was analyzed using a python script in order to calculate the individual contributor degree centrality along with the article degree centralization.

Article Name	**Number of Contributors**	**Article Degree Centrality (max 1)**
Adam Smith	276	0.039114
Aristotle	274	0.0232
Immanuel Kant	231	0.20484
Johann Wolfgang von Goethe	242	0.016682
John Locke	292	0.008581
Karl Marx	232	0.006601
Ludwig Wittgenstein	220	0.006328
Philosophy	280	0.00254
Plato	284	0.001207
Socrates	289	0.000405

Table 1: Articles Used in the Case study. The article domain is Philosophy.

As can be observed from the table the article degree centralization is relatively low because of the small collections of articles used in the case study and the interconnections of the actors in the domain. However it is enough to extract some qualitative measures which are the follows:

- The dispersion of the actor indices denotes how dependent this article is from individual contributors. For instance if an article has a very low degree centralization then it means that the social process to shape this article was highly distributed thus resulting an article that has been submitted by multiple authorities. In our case the articles represent a low degree of centralization which means that contributions on these articles have been done from individuals that have interests in other domains as well.

Contributor	**Actor Degree Centrality**	**Interrelations (Contributions to other articles in the same domain)**
Paul August	0.15217	9
Goethean	0.14348	5
Andrew Norman	0.13913	2
Everyking	0.13043	7
SlimVirgin	0.12609	2
Amerindianarts	0.1	1
Noisy	0.09565	3

[d] http://pywikipedia.sourceforge.net

Heah	0.09565	2
YurikBot	0.0913	4
Hadal	0.08696	3

Table 2: Contributor Degree Centrality for the Article "Immanuel Kant"

- The range of the group degree centralization reflects the heterogeneity of the authoring sources of the article. In our case the article "Immanuel Cant" has significantly higher degree of centralization which means that has been contributed by authorities that are most concentrated in the domain of the article thus have contributed on other articles.

Contributors with higher interrelation over the same domain represent higher authorities based on the assumptions that their interest spans the domain that the article belongs therefore they have conducted background research regarding the material that they have contributed.

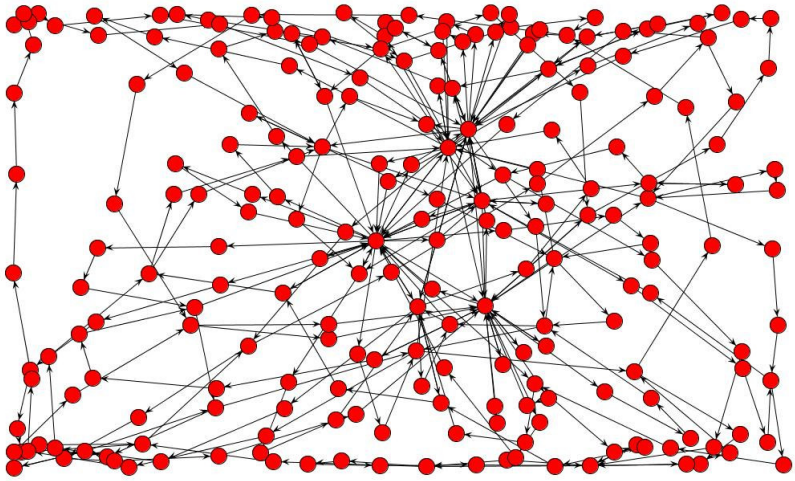

Figure 3: Network visualization of the Article referring to Ludwig Wittgenstein. Contributors with the most incoming edges considered to have a large amount on internal edits thus their actor (outer) degree centrality is minimal (positioned in the centre of the graph)

From the other hand contributors with less authority tend to have their content erased/objected by contributors with higher authority. As can be observed from Figure 4 there exist a number of contributors that are subject to objections regarding their submission therefore are situated in the centre whereas contributors with accepted contributions (authorities) tend to be on the periphery.

5 Discussion and Further Research

The question of the reliability regarding WIKIPEDIA content is a challenging one. As long as the reference of WIKIPEDIA grows the problem becomes more demanding especially on topics with controversial views such us politics or historical views. Our paper represents an early attempt on getting on that problem thus working towards a more sophisticated solution to address the problem. In our paper there is a number of open issues that need to be solved:

- The inner degree can be calculated using a more sophisticated factor representing how

much of the text contributed by one actor has been edited by another actor. In our case we represent the editing or the objection by using 0 or 1 whereas content cannot be subject to boolean choice. A fuzzy operator could provide a solution for aggregating the results obtained by doing a fuzzy diff between the current version of the article and the version submitted. In that case the social tie needs also to be expressed in terms of fuzziness along with the relevance cases.

- The organization of topics and the definition of interconnections is also a matter of research since there are related domains with contributing authorities. For instance the domain Philosophy is also dependent on the domain Psychology and it's contributing authorities.

Finally one specific attention should be given to the diffusion of different affiliations related to one actor. For example a contributor may have many affiliations to however unrelated subjects. This for instance may imply that the contributor may have knowledge of both fields but in cases such as special topics e.g cardiology, the contribution in subjects such us renaissance can be attributed as a non-expert contribution. Therefore a classification of the competencies of each contributor may need to be promoted so it can strengthen their credibility and association with the subject or the article that may contribute.

References

1. J. J. Cadiz, A. Gupta, and J. Grudin. Using web annotations for asynchronous collaboration around documents. In *CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 309-318, New York, NY, USA, 2000. ACM Press.
2. T. Berners-Lee and M. Fischetti. *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor*. Harper San Francisco, September 1999.
3. B. Leuf and W. Cunningham. *The Wiki Way: Collaboration and Sharing on the Internet*. Addison-Wesley Professional, April 2001.
4. Lih, A. (2004). Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative media as a news resource. Symposium. Proceedings of the International Symposium on Online Journalism 2004.
5. L. Andrew, V. Jakob, M. Cathy, K. Samuel, and H. Reinhold, editors. *Proceedings of Wikimania 2005 - The First International Wikimedia Conference*, 2005.
6. S. P. Borgatti and P. C. Foster. The network paradigm in organizational research: A review and typology. *Journal of Management*, 29(6):991-1013, December 2003.
7. S. Wasserman, K. Faust, and D. Iacobucci. *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, November 1994.
8. G. Sabidussi. The centrality index of a graph. *Psychometrika*, 31:581-603, 1966.
9. J. P. Scott. *Social Network Analysis: A Handbook*. SAGE Publications, January 2000.
10. L. C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215-239, 1979.
11. M. G. Everett and S. P. Borgatti. The centrality of groups and classes. *Journal of Mathematical Sociology*, 23(3):181-201, 1999.
12. N. E. Friedkin. Theoretical foundations for centrality measures. *The American Journal of Sociology*, 96(6):1478-1504, 1991.
13. M. Shaw. *Communication Networks*, volume 1, pages 111-147. Academic Press., New York, 1964.