Evaluating Authoritative Sources using Social Networks: an insight from Wikipedia

Nikolaos Th. Korfiatis

(nk.inf@cbs.dk)
Department of Informatics
Copenhagen Business School (CBS)
DK – 2000 , Frederiskberg
Copenhagen, Denmark

Marios Poulos, George Bokos

(mpoulos@ionio.gr, gbokos@ionio.gr)
Department of Archives and Library Sciences
Ionian University
Palaia Anaktora, GR-49100
Corfu, Greece

Abstract

Purpose – The purpose of this paper is to present an approach to evaluating contributions in collaborative authoring environments and in particular wikis using social network measures.

Design / methodology / approach – A social network model for wikipedia has been constructed and metrics of importance such as centrality have been defined. Data have been gathered from articles belonging to the same topic using a web crawler in order to evaluate the outcome of the social network measures in the articles.

Originality / Value - This work tries to develop a network approach to the evaluation of wiki contributions and approaches the problem of quality of wikipedia content from a social network point of view.

Practical Implications – We believe that the approach presented here could be used to improve the authoritativeness of content found in Wikipedia and similar sources.

Keywords Wikipedia, Social Networks

Paper type Research Paper

1. Introduction and Background

Since the invention of writing as a method of encoding human knowledge, the preservation and dissemination of information and knowledge has become a matter of great importance to humanity. People, as intelligent entities, produce and consume information which is preserved in and accessed from various sources such as books, articles and encyclopaedias. Since the organization of information is characterized by a high level of complexity, reference works to assist in the retrieval of relevant information resources are crucial for dissemination and further development of human knowledge in a particular subject. Encyclopaedias and dictionaries represent the major instances of such reference works since their principal scope is to assist, through associative trailing, the retrieval of the relevant resources through a particular domain (collection of relevant lemmas).

On the other hand, visions of the World Wide Web such as the "Memex" envisioned by Vanevar Bush (Bush, 1945) and the original intuition behind the design of WWW by Tim Berners Lee (Berners Lee et al., 1994) tend to represent the WWW as a huge encyclopaedia where lemmas are associated by using a hypermedia model. Nonetheless, an encyclopaedia and any other kind of reference work is often subject to evaluation as to the level of quality characterizing it. Since the production of such works is subject to a very small number of individuals, the development process is characterized by high complexity. Efforts in the WWW such as WIKIPEDIA try to distribute that kind of complexity to several contributing authorities by allowing the synchronous editing and publication of lemmas through its publication model. However, WIKIPEDIA since its early beginning has been subject to criticism (Fasoldt 2004 ; Orlowski, 2005; Lipczynska 2005) as to what level the information contained can be trusted and referenced in research works. In that case, models of credibility which are used extensively on search engine research and information retrieval can be used in order to evaluate the trustworthiness of the topic covered by WIKIPEDIA.

On the web several successful approaches to credibility such us the PageRank (Brin and Page, 1998; Brin et al., 1998) use methods from graph theory to model credibility using the connections of the resource for evaluation. Several graph theoretic models of credibility and text retrieval (Faloutsos, 1985) rely strongly on the consideration of the in-degree of the node (the sum of the incoming arcs of a node in a directed graph) so as to extract importance and trustworthiness. This is also implied by the publication workflow and the resulting context on which those models are based. For instance, on the World

Wide Web and similar hypermedia systems such models of credibility evaluate a webpage using the in-degree extracted by the hypertextual context. However, there are publication models supporting social activities (e.g. collaborative authoring) which derive much of their credibility by their productions (e.g. authorship) where the hyperlink context doesn't depict that kind of activity. In that case, the in-degree cannot provide input to evaluate the importance of that entity and therefore a holistic approach is needed. That kind of alternative evaluation has to consider the outputs of the entity (productions) as it happens with several informal social communication models (Festinger, 1950). In a graph theoretic interpretation, this can be modelled as the outer-degree of the node which conceptually represents the entity evaluated.

2. The WIKI publication Model

The web has given rise to new forms of collaboration and interaction facilitating the manipulation of shared artefacts and information spaces (Cadiz, Gupta, & Grudin, 2000). In the current state of the art, the web ecosystem consists of resources (web pages/ files) linked though hypertext connectors, thus forming a system of links denoting references to those 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 resources requested but also to allow editing and annotation of these resources in a simple way (Berners-Lee & Fischetti, 1999). Of the foremost approaches to this goal, the concept of wiki (Leuf & Cunningham, 2001) has given a response to this challenge. WikiWiki (Hawaian word for quick) 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 in cases such as formation of collaborative document editing (e.g. in communities of Practice) or formation of shared knowledge repositories such us dictionaries etc. One of the best-known implementations and example of the usefulness of the wiki system to support collaborative document editing is the wiki-based encyclopaedia WIKIPEDIA and its related projects[1].

Traditional encyclopaedias such us Britannica are often characterized with a high level of credibility by domain experts, taking into account the background process which has resulted (domain authorities contribute to the final outcome). On the other hand, since it uses the WikiWiki system, WIKIPEDIA allows the editing and creation of encyclopaedic articles by anyone who wishes to contribute. Its primary target is to provide free editing access and gather knowledge representing the consensus of the term presented and thus not to evaluate the contributing authorities. However, as the content increases along with the contributing sources (see figure 1.2), a critical issue has emerged regarding the credibility of WIKIPEDIA as an authoritative reference source (Andrew et al., 2005; Lih, 2004). 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 accepted or not accepted by him/her. WIKIPEDIA has internal mechanisms of managing those cases such as a permission ranking system, where a contributor is accredited by the level of participation in the shaping of the article,

[1] http://www.wikimedia.org

as well as a discussion tab on most of the articles or notifications and warnings regarding the content. Nevertheless, the research question to deal with is how to provide a clue to the credibility for an article based on the contributing authorities and their acceptance by the community of their fellow contributors.

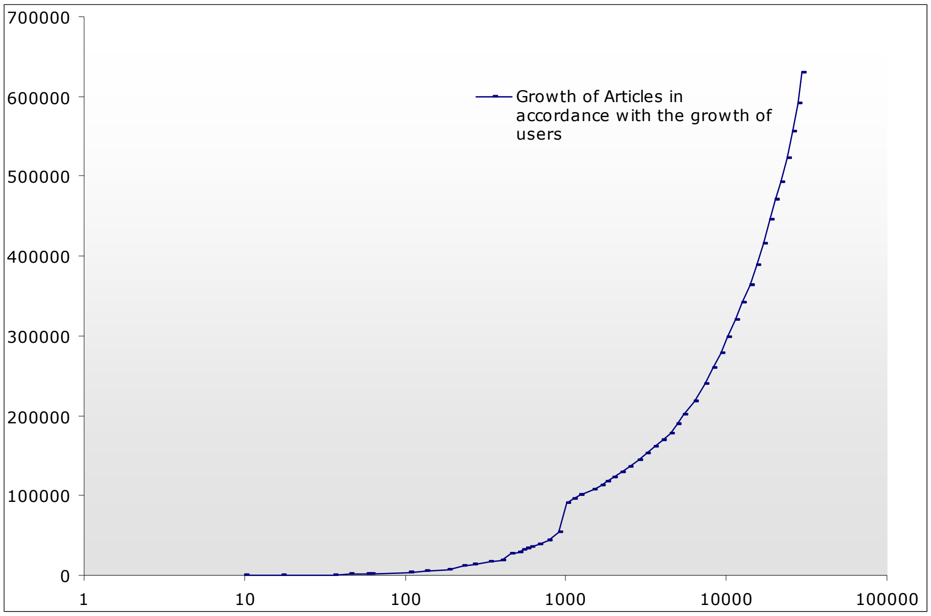

Figure 1: Growth of Articles in accordance with the growth of users in the English WIKIPEDIA (Statistics obtained from http://en.wikipedia.org/wikistats/EN/TablesWikipediaEN.htm). Values on X axis represent the articles on logarithmic scale. Values on Y axis represent the number of contributors.

In this paper we present an initial attempt to model the problem towards providing an authoritative ranking mechanism based on social interaction data collected through the wiki. Social interaction is approached from the side of social communication facilitated by the WIKIPEDIA platform (e.g. edits on edits) (Cobley, 1996). 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 which 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.

3. A network approach on the Wiki publication model

Social networks and social network analysis in particular (Wasserman, Faust, & Iacobucci, 1994; Scott, 2000), 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 actions which characterize most of the activity described by the network. Social network analysis focuses on a more rationalistic approach to research on organizations and social groups (Borgatti & Foster, 2003) since it aims to expand interdependent relations between the members of the group. WikiWiki applications facilitate a case where social relationships are established over a domain of social actions such as 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. This enables combining the contributions in an effective and democratic way allowing 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 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 certain 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 which 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 characterizing the final outcome and the dynamics of the process. In this paper, we will focus on the interaction ties between multiple contributors working on 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 which characterize such a network.

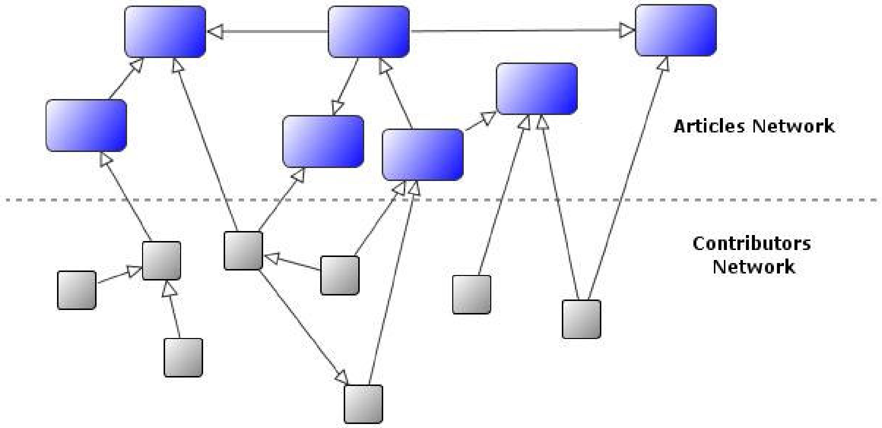

Figure 2: Network layers in the wiki publication model. Contributors are linked together by working on common projects (articles) in the WIKIPEDIA namespace.

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 to an article or topic in the WIKIPEDIA. Structural and compositional properties of the publication model can be found in the following use-cases:

1. When a contributor edits content 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 previous contributor's content that is visible after.
2. Every contributor who has a single contribution, or more, to the article establishes a relational tie with the other content contributors. 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 references 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 a relationship with another contributor if they work on the same article. In the resultant signed 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 resultant 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 that more interrelations will be found 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 and thus his background is depicted to have knowledge

of the domain. Let us consider, for instance, a contributor who 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 on Anatomy. However, the author is more affiliated to the articles regarding the history of the Spanish colonies than the medicine. Therefore his contribution in medicine may be considered as less authoritative than those in the other domain as his knowledge of Anatomy is not as extensive as of the other.

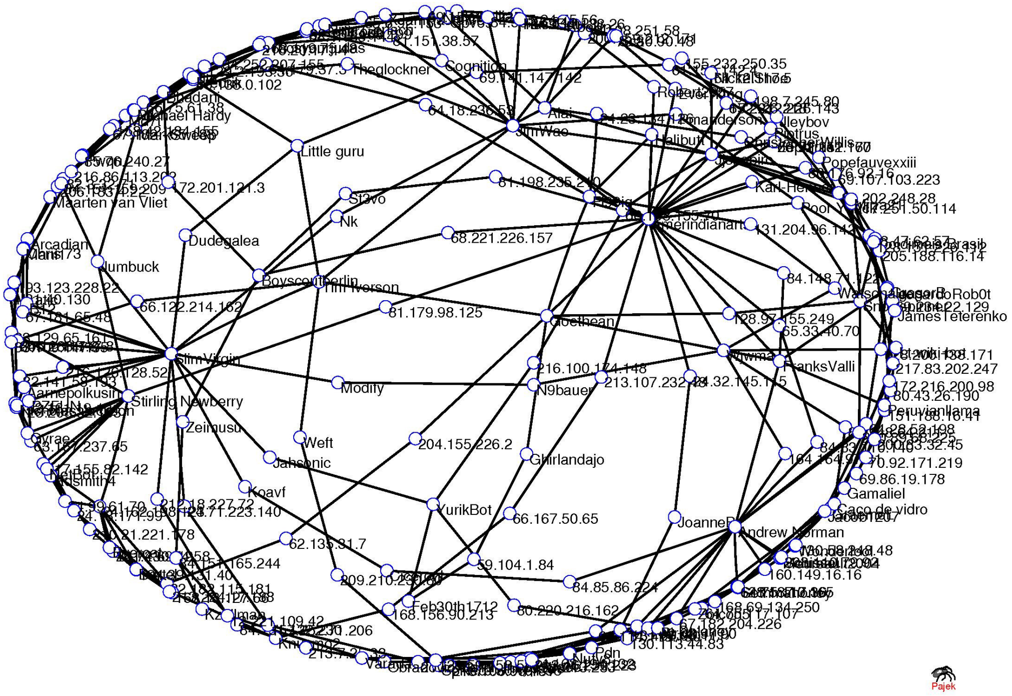

Figure 3: Visualization of the Social Network of the Contributors for the Article "Immanuel Kant". Nodes in the core denote high degree centrality.

In social network analysis, there are a variety of measures which 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.

4. Network measures in the Wikipedia Contributions

As previously mentioned, contributors contribute to one or more articles belonging to the same or different domains. Based on this, we can evaluate the activity of the contributors and thus extract metrics of their presence in the domain of the articles. The development of those metrics is based on the following assumptions:

- The more decentralized the editing of an article, then the better this article represents a consensus about it.
- The contributors whose content has been most accepted (seen from the result of the diff operation in the wiki) are attributed a level of authority regarding the article.
- This level of authority remains only in the domain of the article. However, domains which belong in the same topic can retain the level of authority for a contributor.

The graph that we model is a signed directed network with arcs signed as a factor depicting the level of acceptance of the content submitted by contributor ***A*** and accepted by contributor ***B***. In order to model the authoritativeness of contributors, we selected the Centrality index (Freeman, 1979; Sabidussi, 1966) of the resultant 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 accepted in social network analysis as there are numerous studies showing the usefulness of such a metric for measuring activity in social networks (Everett & Borgatti, 1999; Freeman, 1979). In sociometric studies, the usage of centrality is targeted to unfold the person/individual who is the most prominent in a network, thus ranking the actors according to their positions 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 at an 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.

4.1 Contributor Degree Centrality

In classical social network models, the inner degree (the amount of edges coming into a node) represents the 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 with the biggest inner degree has the biggest amount of objection/rejection in the contributor community and thus receives a kind of negative evaluation from his/her fellow contributors. On the other hand, the outer-degree of the vertex represents edits/participation in several parts of the article and thus gives a clue to the activity of the person in relation to the article and the domain. Mathematically we can represent such formalism as follows: Considering a graph representing the network of contributors for an article contributed in the wiki, then the Contributor Degree Centrality - a contextualized expression of actor degree centrality - is a degree index of the adjacent connections between the contributor and others who 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} represents the relational tie between the contributors and their contribution over the domain of the article. This also is characterized 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 remaining vertices (g) minus the self (g-1). Therefore the contributor degree centrality can be calculated as:

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

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

5. An Insight from Wikipedia

As previously mentioned, WIKIPEDIA follows a hypermedia model to categorize the articles (lemmas) in an associative taxonomy. In that particular taxonomic classification, we define the following structures:

- ***Domain:*** A collection of articles which tackle a common subject (e.g. philosophy).
- ***Category:*** A collection of domains which have a common categorical and etymological root. For example, the domains philosophy and economics have a kind of connection in the category of social sciences.

In order to provide a qualitative analysis of the metrics deployed in the evaluation of the articles, we picked ten articles with a similar number of contributors to 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[2] for each of the articles. The resulting networks contained an average of 259 contributors per article and the average article inter-relations 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 was then analyzed, using a python script in order to calculate the individual contributor degree centrality along with the article degree centralization.

[2] www.pywikipedia.org

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: The Wikipedia articles from which the empirical data was gathered.

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 inter-connections of the actors in the domain. However, it is enough to let us discuss some qualitative interpretations such us:

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

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

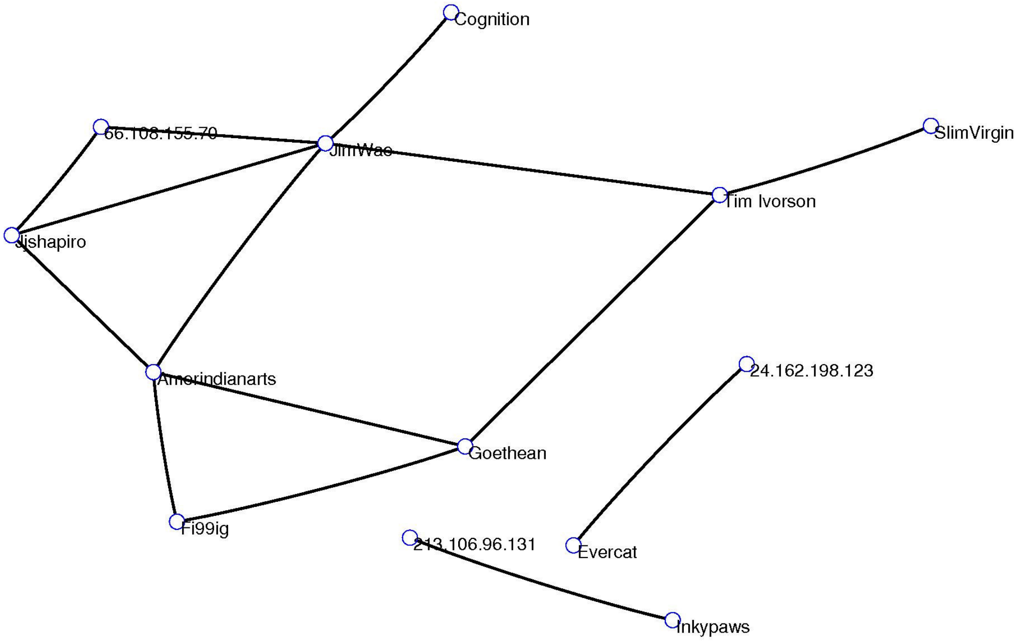

Figure 4: A decomposition of the network to the contributors with the highest degree centrality for the article "Immanuel Kant".

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

Cluster (Outerdegree)	**Freq**	**Freq%**	**CumFreq**	**CumFreq%**	**Representative**
1	1	0.4329	1	0.4329	65.6.92.153
2	199	86.1472	200	86.5801	82.3.32.71
4	14	6.0606	214	92.6407	80.202.248.28
6	6	2.5974	220	95.2381	Snowspinner
8	3	1.2987	223	96.5368	Tim Ivorson
10	1	0.4329	224	96.9697	StirlingNewberry
12	2	0.8658	226	97.8355	24.162.198.123
16	2	0.8658	228	98.7013	JimWae
18	1	0.4329	229	99.1342	Jjshapiro
20	1	0.4329	230	99.5671	SlimVirgin
31	1	0.4329 s	231	100	Amerindianart

Table 2: Contributor Degree Centrality for the Article “Immanuel Kant”.

On the other hand, contributors with lesser authority tend to have their content erased/objected by contributors with higher authority. As can be observed from Figure 3.2, there exist a number of contributors subject to objections regarding their submissions and therefore are situated on the periphery; whereas contributors with accepted contributions (authorities) tend to be in the centre.

6. Discussion and Further Research

The question of the reliability regarding WIKIPEDIA content is a challenging one. As long as the size of WIKIPEDIA grows, the problem becomes more demanding, especially for topics with controversial views such as politics or historical views. Our study represents an early attempt at getting to that problem and thus working towards a more sophisticated solution to address it. However, there are a number of open issues which can extend the merit of this report:

The in-degree can be calculated using a more sophisticated factor, representing how much of the text contributed by one actor has been edited by another. In our case, we represent the editing or the objection by using a scale from 0 to 1, thus aggregating the factors using simple sums. 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 relevant cases. Expressions of credibility using imprecise criteria (Sicilia and Garcia, 2004; Sicilia and Garcia 2005) can also contribute to further advance that direction.

The organization of topics and the definition of inter-connections is also a matter of research since there are related domains with contributing authorities. For instance, in the category of the social sciences, a contributor who edits the article of Adam Smith and has an acceptance factor can be retained on both the domains of *"Economics"* and *"Philosophy"* as an article about Adam Smith is represented in both. In that case, network modelling using two layer networks (document reference, authority reference) can enhance the trust of the contributions (Hess et al., 2006).

Furthermore, the measures developed and presented in this report do not actually measure the “subjective” quality of an article since such a task is a cognitive process characterized by a high level of complexity. Those measures can contribute in the direction of providing an indicator of “consensus” related to an article and thus assert it does not provide controversial views or expressions of a small group of persons (especially in articles with political content). Thus a level of neutrality expressed in the writing of this article is asserted. Finally, specific attention should be given to the diffusion of different affiliations related to one actor. For example, a contributor may have many affiliations to 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 as Renaissance can be attributed as a non-expert one. Therefore a classification of the competencies of each contributor may need to be promoted to strengthen their credibility and association with the subject or the article contributed.

References

Andrew, L., Jakob, V., Cathy, M., Samuel, K., & Reinhold, H. (Eds.). (2005). *Proceedings of wikimania 2005 - the first international wikimedia conference.*

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the web: The original design and ultimate destiny of the world wide web by its inventor.* Harper, San Francisco.

Berners-Lee, T., Masinter, L., & McCahill, M. (1994). Uniform Resource Locators (URL). RFC 1738.

Borgatti, S. P., & Foster, P. C. (2003). The network paradigm in organizational research: A review and typology. *Journal of Management, 29*(6), 991-1013.

Bush, V. (1945). As we may think. *The Atlantic Monthly. July Issue.*

Cadiz, J. J., Gupta, A., & Grudin, J. (2000). Using web annotations for asynchronous collaboration around documents. *CSCW '00: Proceedings of the 2000 ACM conference on computer supported cooperative work,* 309-318.

Cobley, P. (1996). *The communication theory reader.* Routledge.

Everett, M. G., & Borgatti, S. P. (1999). The Centrality of Groups and Classes. *Journal of Mathematical Sociology, 23*(3), 181-201.

Fasoldt, A. (2004). *Librarian: don't use Wikipedia as a source,* Syracuse Post Standard, 25 August

Faloutsos, C. (1985). Access methods for text. *ACM Computing Surveys, 17*(1), 49-74.

Freeman, L. C. (1979). Centrality in social networks: Conceptual clarification. *Social Networks, 1*(3), 215-239.

Festinger, L. (1950). Informal Social Communication. *Psychological Review, 57 (5), 271-282.*

Hess, C., Stein, K., & Schlieder, C. (2006). Trust-enhanced visibility for personalized document recommendations. *Proceedings of the 21st ACM symposium on applied computing,* Dijon, France.

Leuf, B., & Cunningham, W. (2001). *The wiki way: Collaboration and sharing on the internet.* Addison-Wesley Professional.

Lih, A. (2004). Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. *Proceedings of the international symposium on online journalism.*

Lipczynska S. (2005). Power to the people: The case for wikipedia. *Reference Reviews, 19*(2). Emerald Publishing.

Orlowski, A. (2005). *Wikipedia science 31% more cronky than britannica's*, The Register.

Sabidussi, G. (1966). The centrality index of a graph. *Psychometrika, 31*, 581-603.

Scott, J. (2000). *Social network analysis: A handbook* (2nd Ed.). London; Thousands Oaks, Calif.: SAGE Publications.

Sicilia, M. A., & Garcia, E. (2005). Filtering information with imprecise social criteria: A FOAF-based backlink model. *Proceedings of the fourth conference of the European society for fuzzy logic and technology (EUSLAT)*

Sicilia, M.A., García, E. (2004) Fuzzy Group Models for Adaptation in Cooperative Information Retrieval Contexts. *Lecture Notes in Computer Science 2932* Springer, 324-334.

Wasserman, S., Faust, K., & Iacobucci, D. (1994). *Social network analysis: Methods and applications (structural analysis in the social sciences).* Cambridge University Press.