**Social Network Analysis and Mining (SNAM), http://www.springerlink.com/content/1869-5450/**
(This is a preprint of an article accepted for publication in SNAM. DOI: 10.1007/s13278-011-0037-5)

# Mining Latent Relations in Peer-Production Environments: A Case Study with Wikipedia Article Similarity and Controversy

**Chenliang Li · Anwitaman Datta · Aixin Sun**

Received: date / Accepted: date

**Abstract** As people participate actively in social networking and peer production sites, there are additional, implicit relations that emerge from various user activities. Mining such latent relations, or wisdom of crowds, is in itself an important area of ongoing research, with both general as well as domain specific custom-made techniques. In this paper, we propose a new similarity measure, which we call expert-based similarity to discover semantic relations among Wikipedia articles from the co-editorship perspective. Also, different kinds of relations among entities may reveal diverse information. Both to explore and expose such a premise, we carry out a case study leveraging on multiple relations among Wikipedia articles. Specifically, we use expert-based similarity as well as other standard similarity measures, to discern the influence and impact of several factors which are hypothysed to generate controversies in Wikipedia articles. In the context of Wikipedia specific research, our case study helps better differentiate the degree of impact of some of the possible causes of controversies.

**Keywords** Social Networks, Similarity Measure, Wikipedia, Controversy

## 1 Introduction

From infotainment sites to citizen reporters, blogs, Q&A sites such as Yahoo! Answers to Wiki-based encyclopedic corpus like Wikipedia, in recent years social media has become an integral part of our daily life. Compared to traditional websites, social media sites enable users more interactions with information items as well as with other users, like sharing photos, tagging webpages, submitting and commenting on news stories, as well as making friends online. These interactions interplay with each other, and make social media sites complex networks as well as peer-production environments. Latent relations encoded within these interactions may provide us with a new avenue to discover new knowledge that may complement our understanding about the system of interest based on the declared explicit relations. Also, harnessing information from mining different relations based on the different perspectives would help us discover knowledge that can't be deduced by considering various aspects in isolation.

Wikipedia, as a multilingual, web-based, free-content encyclopedia, has more than 3.6M articles in English, attracting around 400M visitors monthly as of March 2011[1]. Moreover, Gartner, the Wall Street Journal and Business Week have identified the Wikipedia paradigm of 'peer-production' of knowledge repository as an up-and-coming technology to support collaboration within and between corporations. Enterprise wiki has been increasingly adopted in companies and organizations.[2]

Wikipedia's open strategy that allows anyone to create and edit articles, leads to its unsurprising success. The open access property makes knowledge creation in Wikipedia a dynamic process evolving over time by contributions and collaboration among different people. It is arguably an outcome and ongoing venue for the most massive collaboration online. Furthermore, all the actions of all the contributors are logged meticulously, and are also openly available for analysis. The edit related metadata information can be used to help us understand the collaboration dynamics of Wikipedia. Specifically, these can be analyzed to obtain a deeper and clearer insight on the characteristics of contributors as well as articles.

In this paper, we investigate similarity measures over Wikipedia articles based on different perspectives of the collaborative knowledge building system enabled by Wikipedia. For example, by analyzing the logs of edit history, we observe that individual contributors only edit a relatively small number of articles. This shows that people have only focused expertise and/or interest areas with respect to the areas covered by the entire Wikipedia. Based on this observation, we propose a similarity measure, *expert-based similarity*, to evaluate the relevance of articles among each other. In contrast,

Corresponding author: Chenliang Li, lich0020@ntu.edu.sg.

Chenliang Li, Anwitaman Datta, Aixin Sun
School of Computer Engineering
Nanyang Technological University, Singapore
E-mail: {lich0020,anwitaman,axsun}@ntu.edu.sg

[1] `http://en.wikipedia.org/wiki/Wikipedia:About`
[2] `http://en.wikipedia.org/wiki/Enterprise_wiki`

existing state-of-the-art approaches that have been adopted widely in Information Retrieval (IR) areas to quantify the relevance focus on other perspectives such as content-based (Manning et al, 2008) and structure-based similarities (Jeh and Widom, 2002; Zhao et al, 2009).

We conduct a case study to illustrate that knowledge from different perspectives can be obtained by mining the various relations among Wikipedia articles. We show that by combining knowledge obtained from such different perspectives (essentially based on the different alternate measures), we better discern the origin of some controversies in Wikipedia, which can not be deduced by considering any single particular aspect in isolation.

To the best of our knowledge, this paper is the first attempt to explore different relations encoded in Wikipedia by studying the edit related metadata. Section 2 reviews the similarity measures that quantify the relevance of Wikipedia articles from different perspectives and the related studies on exploration of different aspects in social media networks. Then, we discuss recent works about controversy in Wikipedia, since we show later that by harnessing knowledge from different perspectives, it is possible to identify the dominant cause of controversy in Wikipedia.

In Section 3 we propose a new measure, *expert-based similarity* to evaluate relevance relationship between articles based on the observation that Wikipedia contributors often edit a small number of articles each. We additionally discuss the relevance relationships between Wikipedia articles based on content, hyperlinks, each of which focuses on one specific perspective.

In Section 4 we conduct extensive experiments to evaluate different similarity measures. Specifically we evaluated the similarities based on expertise, and existing metrics such as cosine similarity, P-Rank and SimRank.

In Section 5 we cluster articles based on the various similarity measures, and study the distribution of controversial articles in the resulting clusters. We carry out a detailed analysis on the distribution of the controversial articles in the resulting clusters. From our analysis, at least in the considered data sets, we determine that out of several plausible hypothesis for controversies (Brandes et al, 2009; Brandes and Lerner, 2007; Kittur et al, 2007; Vuong et al, 2008), one reason dominates. Namely, the specific controversial topics contained in articles are the principal source of controversy in Wikipedia, instead of the aggressive contributors or controversial concept in general. Finally, we conclude this work in Section 6.

## 2 Related Work

Firstly, we summarize in Section 2.1 existing similarity measures used in this work. Recently, a number of researchers have explored the explicit/implicit relations for knowledge discovery and IR tasks in social media that we elaborate in Section 2.2. Because of its massive scale of collaboration as well as usage, and open access to all edit related history, Wikipedia analysis has become a research subtopic in its own right in recent years. Among various aspects of Wikipedia being studied, the works focusing on the coordination and conflict between users are reviewed in Section 2.3.

### 2.1 Similarity Measures

In this paper, we analyze the edit history and derive the inherent information about users and articles from it. An expert-based similarity measure is proposed in Section 3 based on the observation that people expose their expertise with the articles they have edited. Effectiveness of expert-based similarity in grouping relevant articles is validated and compared with cosine similarity of article content, as well as similarity based on hyperlinks, namely SimRank (Jeh and Widom, 2002) and P-Rank (Zhao et al, 2009).

SimRank (Jeh and Widom, 2002) is based on the intuition that similar objects are related to similar objects, which in turn are mutually similar. More precisely, objects $a$ and $b$ are likely similar if they are related to objects $c$ and $d$ respectively, and objects $c$ and $d$ are themselves similar. Formally, given two objects $a$ and $b$ in a graph, let $I(a)$ and $I(b)$ be their corresponding sets of in-link neighbors. The SimRank between $a$ and $b$ is computed recursively by Equation 1. In this equation, $I_i(a)$ is the $i^{th}$ in-link neighbor of $a$; $C$ is a damping factor with value between 0 and 1. The equation is initialized by setting $s(a, b) = 1$ if $a = b$ and $s(a, b) = 0$ otherwise.

$$s(a,b) = \frac{C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) \tag{1}$$

P-Rank (Zhao et al, 2009) extends SimRank by considering out-neighbors between two objects, see Equation 2. In this equation, $I(a)$ and $O(a)$ denote the set of in-link neighbors and the set of out-link neighbors of object $a$ respectively. Similar to that in SimRank, $C$ is the damping factor and $\lambda$ is a variable for the weight of similarities derived from in-link neighbors and out-link neighbors respectively.

$$\begin{aligned} s(a,b) = & \frac{\lambda \times C}{|I(a)||I(b)|} \sum_{i=1}^{|I(a)|} \sum_{j=1}^{|I(b)|} s(I_i(a), I_j(b)) \\ & + \frac{(1-\lambda) \times C}{|O(a)||O(b)|} \sum_{i=1}^{|O(a)|} \sum_{j=1}^{|O(b)|} s(O_i(a), O_j(b)) \end{aligned} \tag{2}$$

To evaluate P-Rank and SimRank, a clustering evaluation measure named *compactness* was used in (Zhao et al, 2009). The results from their extensive experiments showed that P-Rank consistently outperformed SimRank.

### 2.2 Different Perspectives in Social Media

By considering different relations mined from multi-dimensional social data, Lin et al. (2009) conducted community discovery study. For these different relations, they built a hypergraph by encoding the relations into a tensor. The communities were discovered by applying factorization of the tensor. Similarly, Rendle et al. (2009) explored directly the ternary relationship in tagging systems, i.e., tag, user and item, by applying a tensor factorization model in their work about tag prediction. Hamouda and Wannas (2011) built a personalized tag recommendation algorithm by considering the tag, user and content similarity together. Bross et al. (2011) combined different information in their weblog ranking algorithm. Zhou et al. (2009) proposed an approach to identify the expert for a given question by considering different relations contained in CQA systems. In detail, they not only measured the relevance of the questions, expertise of the users, but also built a user-to-user interaction graph and measured the authorities of users. Then, they incorporated all these relations into a Naïve Bayes model to quantify the overall expertise regarding to the question of concern. Bhattacharyya et al. (2010) studied the effect of homophily in online social networks from the perspectives of the user proximity and the topological distance.

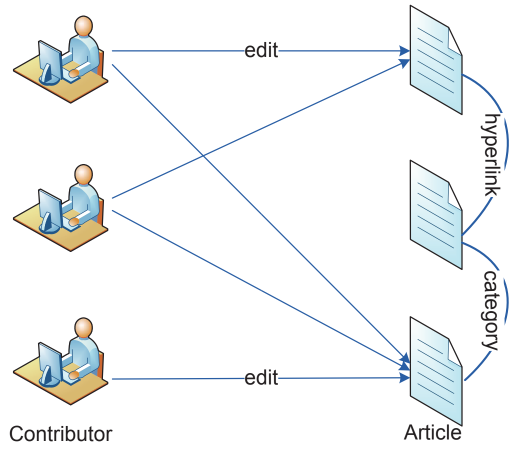



**Fig. 1** Explicit relations among articles and contributors in Wikipedia

### 2.3 Controversy in Wikipedia

Brandes and Lerner developed a visualization tool which reveals the dominant authors that are most involved in a controversy and who plays what role in the article building process (Brandes and Lerner, 2007). Subsequently Brandes et al. (2009) offered an edit network derived from the edit history to illustrate the collaborative work of contributors in Wikipedia. They analyzed the interaction of the contributors in an article to characterize the role each individual user plays during article writing. Potthast et al. discussed the characteristics of vandalism and develop a number of features related to identify vandalism edits (Potthast et al, 2008). By training a classifier over these features, their experimental results showed high precision and recall could be achieved.

Vuong et al. (2008) proposed several models to measure the controversy in an article by the amount of disputes occurring in articles and the degree of controversy in each dispute. The models are designed based on the premise that an article is more controversial if more disputes are from the less controversial contributors while a contributor is more controversial if s/he invites more disputes in less controversial articles. Such a model implicitly assumes that *the source of controversial articles is inherently the nature of the individual contributors*, rather than the subject matter of the articles. With similar assumptions, Kittur et al. investigated a set of page metrics including revisions, the length of content, number of contributors etc. as the features of Wikipedia articles to train a Support Vector Machine (SVM) classifier (Kittur et al, 2007). The experiments showed that the learnt classifier was able to rank the controversial articles consistent with their actual degree of controversy. Also, they demonstrated the use of visualization in making sense of disputes between users. Similarly Le et al. analyzed edit history of individual articles to cluster the contributors with concurring opinions, and with antagonistic relation with those with conflicting opinions (Le et al, 2008).

This work explores different plausible sources of controversy, and based on our study we refute the previously made assumption that contributors are the principal source of controversy. In contrast, we establish that the origin of controversies in Wikipedia is inherent to the subject matter of the articles themselves, rather than the contributors. The detailed analysis is presented in Section 5.

## 3 Article Similarity in Wikipedia

Wikipedia is the largest online encyclopedia and collaborative knowledge building system. Compared with the traditional encyclopedia that are edited by a limited number of experts, it allows everyone in the world to contribute and share their knowledge. Thus, vandalism and incorrect description are spotted and corrected in a very short time by the wisdom of crowds in online communities. Also, traditional encyclopedias are sequential, i.e. ordered along alphabetical, topical or historical lines (West et al, 2009a), while Wikipedia articles have rich semantic structures that every article is connected to other articles by the hyperlinks annotated by the contributors on the places where the concepts are mentioned in the original article. Based on the Wikipedia policy, the hyperlinks in an article should be created to relevant topics of the article and technical terms mentioned that are likely to enhance reader's understanding. Moreover, articles are also organized by Wikipedia categorization system, which an article may be assigned to at least one category based on the concepts it covered. Figure 1 depicts the above explicit relations existing in Wikipedia.

Besides the explicit relations mentioned above in Wikipedia, we can also derive latent (implicit) relations hidden from the explicit ones, e.g. expert-based similarity which will be explained. In the following we first propose a implicit relations regarding to the relevance of Wikipedia articles, expert-based similarity, in detail; then discuss relevance relationships based on content and hyperlinks.

### 3.1 Expert-based Similarity

A Wikipedia article is an artifact which evolves from the contributors' contributions, which induce interactions among these contributors. In the process, contributors also manifest their expertise and interest by making contributions to related articles. By analyzing the distribution of the number of the revisions and the articles edited by contributors, we find

**Table 1** Symbols and semantics

| Symbol | Semantic |
|---|---|
| $a, c$ | the instance of a variable: $a$ for article, $c$ for contributor |
| $V_a$ | a collection of articles |
| $V_c$ | a collection of contributors |
| $r_{c,a}$ | the number of revisions of article $a$ made by contributor $c$ |

that most contributors only edit a small number of articles, and make limited number of revisions. Furthermore, contributors often have very biased contribution focus. That is, if a contributor has contributed to a large number of revisions, then it is likely a large portion of these revisions went to very few article (Zhang et al, 2010). From this observation we infer that individuals have focused interest and familiarity with topics where they frequently contribute.[3] Specifically, in our dataset, on an average each contributor edits only 6 articles. This provides the intuition that an article normally contains a topic which attracts a limited scope of readers and contributors. The scope is article dependant.

A two-way implicit selection process can be identified here: *contributors choose articles relevant to their focused expertise and each article has only limited audience who is expert or interested in the related subject matter*. Based on these observations, we hypothesize that two articles are similar to each other (up to a certain degree) if they have been edited by the same contributors. Thus, we use the commonality of contributors of two articles to determine similarity among these articles based on a metric which we call the *expert-based similarity*. Such a measure complements other similarity measures like content similarity (measured using metrics like cosine similarity) as well as network neighborhood analysis based on graph induced by hyper-links. Table 1 lists the main symbols that we use in the following.

Formally, the *expert-based similarity* is an implicit relation between Wikipedia articles based on the explicit editorship relation, i.e. $r_{c,a}$ (see Fig 1 & Table 1). Let $C_{c_i,a_j}$ represents the contribution score of contributor $c_i$ in article $a_j$, the *expert-based similarity* for articles $a_u$ and $a_v$ is calculated using the standard metric of cosine similarity defined below, where $1 \le u, v \le |V_a|$

$$s_e(a_u, a_v) = \frac{\sum_{c_i \in V_c} C_{c_i,a_u} C_{c_i,a_v}}{\sqrt{\sum_{c_i \in V_c} C^2_{c_i,a_u}} \sqrt{\sum_{c_i \in V_c} C^2_{c_i,a_v}}} \tag{3}$$

The contribution score $C_{c_i,a_j}$ is chosen based on the context and the application applied to. It can be the number of words added by the contributor, the number of revisions made by the contributor, or other quantities calculated by other measures or algorithms. Throughout our study, we define the contribution score to be the number of revisions a contributor has made to the article multiplied by the significance score of the contributor across the collection. The formula to calculate the contribution score, $C_{c_i,a_j}$, is as follow:

$$C_{c_i,a_j} = r_{c_i,a_j} \times f_{c_i} \tag{4}$$

where $f_{c_i}$ refers to the significance of contributor $c_i$. Here, we simply define the significance score of contributor $c_i$ as follows:

$$f_{c_i} = log(\frac{|V_a|}{\sum_{a_j \in V_a} I(r_{c_i,a_j} > 0)}) \tag{5}$$

where the boolean function $I(x)$ takes value 1 when $x$ is true, otherwise 0, i.e., whether contributor $c_i$ has edited article $a_j$. Equation 5 is analogous to inverse document frequency ($IDF$) in the widely adopted $TF \times IDF$ weighting scheme. A contributor has a lower significance score if s/he has edited more articles. The boolean function is used here since, for the considered data-set, contributors only edit few articles very relevant to their expertise. Although number of revisions $r_{c_i,a_j}$ is not a very fine grained discriminative feature compared to say taking the number of words contributed, it holds enough discriminative ability to measure the similarity of two articles in that contributors on average make 13 revisions. By using the number of revisions made by a contributor to an article, the computation cost of calculating the expert-based similarity between two articles is substantially decreased since it ignores the details associated with each revision.

### 3.2 Article Relevance Aspects

As being discussed previously, the relevance relationship between Wikipedia articles can be evaluated by similarity measure defined based on the perspective of the co-editorship, (e.g., *expert-based* similarity). Other than from the perspective of contributors, similarity explicitly evaluated based on the textual content of the articles also help us understand the relevance relation from the perspective of textual content. Furthermore, articles are connected by hyperlinks in their content. Recall that, the link structure in Wikipedia is quite different from the hyperlinks in traditional web pages since most of the internal links in Wikipedia point to semantically related content (Kamps and Koolen, 2008, 2009).

In this work, we evaluate the relevance relationship between Wikipedia articles based on content, hyperlinks, and co-editorship, and the similarity metrics used are summarized in Table 2. Among them, content based cosine similarity, which is widely used in IR and related areas, is an explicit relation as discussed previously, i.e., the cosine similarity of bag-of-word models of two articles $a_i$ and $a_j$. To capture the relevance information from the perspective of the link structure, we use SimRank/P-Rank similarities which are detailed in Section 2.1.

## 4 Article Similarity Evaluation

We conduct three sets of experiments to evaluate expert-based similarity against alternative similarity measures - co-

[3] Note for clarification: From our analysis, we noticed that same contributor may actually contribute to many articles spread across different unrelated categories (and in that sense, the focus is not limited) - e.g., on articles related to American football, Scientology and Biology, but within each specific category, the contributions are few and rather focused. The rest of the discussion is pertaining to contributions within one such category, namely "Religious Objects", with which we carry out our case-study.

**Table 2** Similarity aspects and metrics

| Relevance aspect | Similarity metric | Relation type |
|---|---|---|
| Content | Cosine similarity | Explicit |
| Hyperlink | P-Rank, and SimRank similarities | Implicit |
| Co-editorship | Expert-based similarity | Implicit |

sine similarity using article content, and SimRank and P-Rank based measures using link structure among articles, as well as to validate its generalization to the larger scale.

In the first set of experiments, we adopt the methodology presented in (Zhao et al, 2009) where *compactness* was used to evaluate the clusters produced by $K$-Medoids clustering algorithm using SimRank and P-Rank respectively. As the results of $K$-Medoids maybe impacted by the $K$ chosen, we also apply DBScan (Ester et al, 1996) which does not need the number of expected clusters as input.

In the second set of experiments, we utilize the category labels from Wikipedia as partial ground truth to evaluate the clusters produced by $K$-Medoids and Agglomerative Hierarchical Clustering (AHC) (Tan et al, 2005) algorithm using various similarity measures. With these two clustering algorithms, the number of expected output clusters can be set to be the same as that in the ground truth. The experimental results are evaluated using purity and entropy (Manning et al, 2008).

While the performance of expert-based similarity is evaluated based on a specific Wikipedia category above, we validate the generalization of expert-based similarity to larger scale of whole Wikipedia in the third set of experiments.

Before we report the experimental setting and results, we detail the dataset used in the experiments.

## 4.1 Dataset

In related existing studies (Vuong et al, 2008), *Religious Objects* was identified to have many controversial articles. Since we attempt to study controversy in Wikipedia as a case study, we used a similar set of articles in our experiments.

We extracted articles under *Religious Objects* category from the English Wikipedia dump generated on 03 January 2008. The dump consists of edit history of articles from August 2001 to January 2008. Since our dataset is extracted from a more recently generated dump than (Vuong et al, 2008), the number of articles in our dataset is slightly different. In our dataset, there are a total 18,973 articles, 69,481 registered contributors and 891,231 revisions after we filter the revisions made by anonymous contributors.

## 4.2 Evaluation with Compactness

**Evaluation metric.** *Compactness* metric was used in (Zhao et al, 2009) to measure the quality of clustering results by considering *intra-cluster* distances and *inter-cluster* distances of the clusters. Formally, *compactness* is defined as:

$$C_{\mathcal{F}} = \frac{\Sigma_{i=1}^{K} \Sigma_{x \varepsilon C_i} d(x, m_i)}{\Sigma_{1 \leq i < j \leq K} d(m_i, m_j)} \quad (6)$$

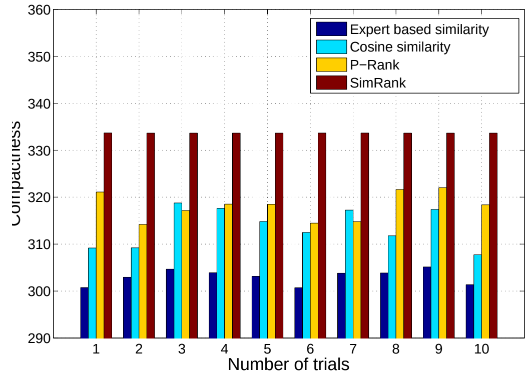



**Fig. 2** Compactness vs. Number of Trials for expert-based similarity, cosine similarity, P-Rank and SimRank

where $\mathcal{F}$ denotes the specific similarity measure adopted, and $d(x_1, x_2)$ denotes the distance between two data points $x_1$ and $x_2$ with regards to the measure $\mathcal{F}$. In this equation, $K$ is the number of clusters generated by a clustering algorithm; $C_i$ is the $i$-th cluster; $m_i$ and $m_j$ are the centers of the clusters $i$ and $j$ respectively. Note that the numerator and denominator represent the *intra-cluster* and *inter-cluster* distances respectively. A smaller compactness value means that the clustering results reflect the inherent relationships of the data better.

**Similarity Computation.** For a fair comparison between the four similarities, we used the articles with more than 5 distinct contributors, leading to 15,018 articles in our experiments. This automatically also ensured that each article had more than five revisions.

- The pair-wise expert-based similarity was then computed for these 15,018 articles according to Section 3.1.
- For content based similarity, the last five revisions for each article were aggregated to represent it since only the last revision alone may not fully represent an article due to the fast evolving nature of Wikipedia.
- For P-Rank and SimRank based similarity computation, only hyperlinks that existed in at least three revisions among the latest five were considered. We set the damping factor $C$=0.8 for both SimRank and P-Rank, and the relative weight $\lambda$ is set to 0.5 for P-Rank. The calculations were run until the scores converged.

**Experimental Results by K-Medoids**. $K$-Medoids is a center-based partitional clustering algorithm which is similar to the widely used $K$-Means (MacQueen, 1967) algorithm. Different from $K$-Means, $K$-Medoids chooses the most representative points as centers of clusters (Tan et al, 2005). These representative points are called medoids. The medoids are chosen with respect to some measure, e.g., minimize the sum of the distance of a point from the medoid of the cluster.

**Table 3** Compactness for similarity measures w.r.t percentages of the noise

| Similarity | Percentage of Noise 45% | 60% | 75% | 90% |
|---|---|---|---|---|
| Expert Based | 5.093 | 0.603 | 0.105 | 0.035 |
| Cosine | 56.540 | 6.613 | 0.719 | 0.170 |
| P-Rank | 38.448 | 6.450 | 0.707 | 0.492 |
| SimRank | 587.104 | 44.985 | 7.402 | 2.090 |

Similar to that in (Zhao et al, 2009), we ran $K$-Medoids clustering algorithm over the dataset with the 4 similarity measures respectively with $K$ set to 10. In total 10 trials were performed to minimize the impact of randomly selected initial medoids. Figure 2 plots the compactness values of the four similarity measures w.r.t the 10 trials. We note that the expert-based similarity is consistently the best in the compactness measure. SimRank achieves the worst compact clustering results, while the cosine similarity outperforms P-Rank in 8 out of 10 trials. It partially confirms that the expert-based similarity measure is robust and achieves better performance in the Wikipedia context. In order to further evaluate the performance of the expert-based similarity, we run the same experiments using DBScan algorithm.

**Experimental Results by DBScan**. The DBScan algorithm clusters the data points by finding clusters of points which are density-reachable to each other within the cluster (Ester et al, 1996). In contrast to $K$-Means and $K$-Medoids algorithms which partition all data points into $k$ clusters, the DBScan algorithm groups the points within an area with the high density into a cluster and classifies the points of the low density areas as noise. A noise point doesn't belong to any cluster and can be considered as an outlier. Two parameters are required in DBScan algorithm: (1) $\varepsilon$ is the radius of the neighborhood of a point; (2) *minPts* specifies the minimum number of points required to form a cluster. The suggested way of setting the parameters is given in (Ester et al, 1996) which involves plotting a sorted $k$-dist graph. All points with a higher $k$-dist value are considered to be noise in the clustering whereas the remaining points are assigned to some clusters. Hence the percent of noise points can be specified and the parameters can be derived correspondingly. In our experiments, we picked four points which set the percentage of noise to be 45%, 60%, 75%, 90% by fixing the *minPts* to 4 respectively. We conduct clustering with the four similarity measures and summarize results in Table 3. We see that expert-based similarity again outperforms the rest while SimRank is still the worst in these experiments. The cosine similarity and P-Rank roughly tie for the four parameter settings.

**Manual Verification.** From the clustering produced by DBScan, we examine the clustering results manually to verify whether the expert-based similarity reflects the reality. Figures 3(a) and 3(b) show two clusters produced by DBScan with the percent of noise being 60%. As illustrated in Figure 3(a), the cluster 1 consists of six Pathis[4] and a worship place (Nizhal Thangal[4]) of Ayyavazhi, and a famous theertha of the temple (Muthiri kinaru) which is located half a kilometer west from the Swamithope pathi[4]. Ayyavazhi is a dharmic belief system that originated in South India in the 19th century[4]. Pathi is the name asserted to the primary centres of congregational worship for Ayyavazhi, having a relatively large structure like that of a temple. And there are 7 Pathis in Ayyavazhi religious system. Vakaippathi[4], the one missed in the cluster result isn't included in the dataset since it had just 5 contributors by 03 Jan 2008 as explained previously. Figure 3(b) shows another cluster, which consists of articles relevant to metropolitan community church. The clusters are quite intuitive and self-explanatory upon human inspection. Combined with all experiment results, we conclude that the expert-based similarity is an effective measure to find semantically relevant articles in Wikipedia. Although the expert-based similarity is proposed and evaluated in Wikipedia context, it is also applicable to other online social collaboration systems - and is expected to be useful to identify and recommend experts, for example in a forum or a Q&A system.

[4]`http://en.wikipedia.org/wiki/{Pathi,Nizhal_Thangal,Muthiri_kinaru,Ayyavazhi,Vakaippathi}`

### 4.3 Evaluation with Partial Ground-truth

The evaluation with compactness reported in the above section reflects the structural cohesiveness of the clusters produced by the clustering algorithm with the chosen similarity measure. However, the semantic cohesiveness of the articles is not captured by the compactness measure. In this section, we report on experiments evaluating the semantic cohesiveness of the clusters utilizing the category labels in Wikipedia which are manually assigned. We used 3033 articles (subset of the 15018 articles used in Section 4.2) from 39 categories in Wikipedia and evaluate the clustering results with the four similarity measures respectively. Since each article has a class label (i.e. category), we evaluate the clustering result by using the standard clustering validity indexes: Purity and Entropy. In this set of experiments, $K$-Mediods and AHC were used since the number of expected clusters can be specified for both clustering algorithms to be same as that in the ground truth.

**Dataset.** A category in Wikipedia contains a list of articles belonging to it and a list of subcategories within it, as well as its parent categories. However, as categories in Wikipedia can be created based on topic, location, time and others, the categories do not form a strict hierarchy tree but rather a graph. In this experiment, we extracted all descendent categories under "Religious objects" containing a total of 2708 categories. We then manually selected 39 categories each of which refers to a specific topic. For each of the selected 39 categories, the articles belonging to it and its sub-categories are considered to form one cluster. The articles that belong to more than one category are ignored. We finally have 3033 articles from these 39 categories which we consider to reflect the ground truth.

**Experimental Results by K-Medoids.** We perform the clustering using $K$-Medoids algorithm over the 3033 articles, with $K$ set to 39. The initial medoids are chosen randomly

| No. | Article |
|---|---|
| 1. | Muthiri Kinaru |
| 2. | Poo Pathi |
| 3. | thamaraikulam Pathi |
| 4. | Nizhal Thangal |
| 5. | Ambala Pathi |
| 6. | Mutta Pathi |
| 7. | Pancha pathi |
| 8. | Swamithope pathi |

(a) Cluster 1

| No. | Article |
|---|---|
| 1. | Metropolitan Community Church East London |
| 2. | Metropolitan Community Churches in London |
| 3. | Metropolitan Community Church |
| 4. | Metropolitan Community Church in South London |
| 5. | Metropolitan Community Church of New York |
| 6. | Metropolitan Community Church of Toronto |
| 7. | Metropolitan Community Church of Manchester |
| 8. | Metropolitan Community Church of Edinburgh |

(b) Cluster 2

**Fig. 3** Two example clusters

**Table 4** K-Medoids clustering results

| Similarity | Purity | Entropy |
|---|---|---|
| Cosine | 0.7592 | **0.9817** |
| Expert Based | 0.7312* | 1.2695* |
| P-Rank | **0.7689** | 1.0383 |
| SimRank | 0.6632* | 1.6279* |

**Table 5** Agglomerative clustering results

| Similarity | Purity | Entropy |
|---|---|---|
| Expert Based | 0.625 | 0.448 |
| Cosine | **0.725** | **0.199** |
| P-Rank | 0.513 | 0.576 |
| SimRank | 0.543 | 0.510 |

from the 39 categories respectively. 10 trials were performed with different sets of initial medoids and the average was taken as the result to minimize the effect of the initialization. For each trial, the same initial medoids were used for the four similarity measures. Table 4 shows the experimental results by using the four similarity measures. The bold values indicate the best performance obtained for purity and entropy measures. The symbol * indicates the change is significant according to the paired $t$-test at the level of $p < 0.05$, compared to the cosine similarity. From Table 4, we can observe that the cosine similarity achieves the best performance according to the entropy, while P-Rank outperforms others in the measure of purity. The changes between the cosine similarity and P-Rank are not significant according to both purity and entropy. The expert-based similarity achieves much better performance than SimRank and is much closer to the cosine similarity and P-Rank. As consistent with the compactness measure, SimRank achieves the worst performance in the experiments.

**Experimental Results by AHC.** We perform agglomerative hierarchical clustering using the Cluto package[5] with the complete link function. Table 5 lists the results of agglomerative clustering. The bold values indicate the best scores obtained for purity and entropy measures. From Table 5, it is observed that the cosine similarity achieves the best performance according to both the purity and entropy, while the P-Rank performs worst in the experiment. For the expert-based similarity, it achieves 21.8% increase in purity and 22.2% increase in entropy as compared to the P-Rank. While the SimRank only improves the performance by 5.8% and 11.5% as regards to the P-Rank.

From the experimental results from $K$-Medoids and AHC clustering, we can see that P-Rank and SimRank achieve the modest performance. We think the main reason is the deviation of the relevance carried by the hyperlinks as well as the missing links among articles. In detail, some articles are underlinked so that many relevant topics are not linked together, while some articles are overlinked which result in the existence of the links of less value. Some previous works have been conducted to address such problem (Adafre and de Rijke, 2005; West et al, 2009b). Thus, the structure based similarity measures, such as P-Rank and SimRank, suffer from the noisy links in the Wikipedia context. As shown in the experiment of $K$-Medoids, by picking an article from a category as a initial medoid can decrease the impact of the noisy links which may bring about the topic drift for some cluster to some extent. Comparing to the cosine similarity, the expert-based similarity performs much worse in the measure of entropy than in purity (see Tables 4 and 5). This indicates that a small number of articles from the different categories also share common authors to some extent which makes them clustered into the same group. We think the reason is that the ground truth is selected from a dataset of high cohesion. Thus there is no explicitly strict boundary between different categories. This is due to the ground truth we build as well as the nature of the dataset we study on. However, combined the experimental results together, we can conclude that the expert-based similarity is better than hyperlink based similarity in most cases and not too far from the content based similarity.

## 4.4 Evaluation with Linear Correlation

Since the previous evaluation is conducted based on articles from a specific category, *Religious Objects*, it is arguable that expert-based similarity may only work well in some specific categories. In order to validate that expert-based similarity generalizes well, we measure the linear correlation between expert-based similarity and the most widely adopted measure, cosine similarity using a corpus of Wikipedia articles spanning all across the Wikipedia. We sample 100 random articles as seeds from the whole Wikipedia. In detail, given an article, we first extract the articles within its neighborhood of 2-hops by following the out-going hyperlinks of Wikipedia. We call the original article as a seed. The similarity scores between the seed and its neighbors are calculated using cosine similarity and expert-based similarity respectively. Then we measure the Peasrson's linear corre-

[5] `http://glaros.dtc.umn.edu/gkhome/views/cluto/`

**Table 6** Correlation within five segments of cosine similarity

| Segment | Correlation |
|---|---|
| [0.0, 0.2) | 0.3033 |
| [0.2, 0.4) | 0.1518 |
| [0.4, 0.6) | 0.0239 |
| [0.6, 0.8) | 0.0155 |
| [0.8, 1.0] | 0.2562 |

lation between the values generated by the two measures. Each seed article has on average 6,458 neighbors within the network of 2-hops. The average correlation coefficient for these 100 seed articles is $0.4012 \pm 0.1913$. This indicates that the two measures are correlated with each other positively. Moreover, it also indicates that the two measures expose different information to some extent. We further split the neighbors of a seed into 5 segments based on the cosine similarity (i.e. each segment with a value range of 0.2). And we calculate Pearson linear correlation coefficient for each segment separately. The average correlation coefficient for each segment are reported in Table 6. From Table 6, we can see that expert-based similarity is much stronger correlated with cosine similarity at the two extreme ends than in the middle-range. It means that when articles are too similar/dissimilar, the two similarity measures concur, while for the other articles, each similarity measure can tell us something different based on different perspectives. We have repeated the same experiments with different samples 10 times, similar results are observed. Thus, we conclude that expert-based similarity generalizes well.

In summary, we compare the performance of the four similarity measures by using agglomerative hierarchical clustering and $K$-Medoids clustering on the ground truth we build from the Wikipedia's category system. We do not intend to demonstrate that a specific similarity measure is better. Instead, we evaluate the appropriateness of the similarity measures with regard to the different aspects and purposes in the context of Wikipedia. Moreover, our experiments validate that expert-based similarity is an effective metric to quantify relevance relationship between articles in Wikipedia.

### 4.5 Qualitative Comparison

Wikipedia keeps a history of all operations from contributors. Thus, the contribution matrix needed to compute expert-based similarity score is obtained without much overhead cost. P-Rank and SimRank are computationally expensive because of the iterative calculation process involved and the huge size of Wikipedia. Moreover, with any change in the structure which may not even concern directly the articles being compared, the whole computation needs to be redone. Wikipedia's open strategy would be misused by some editors. Vandalism, defined by Wikipedia, as "any addition, removal, or change of content made in a deliberate attempt to compromise the integrity of Wikipedia". Cosine similarity can't tolerate the effect of vandalism. On the contrary, the IDF scheme used in the expert-based similarity to attenuate the effect of editors that edit too many articles to be meaningful for relevance determination can make it more defensible against vandals. The expert-based similarity measure is thus more efficient and robust in comparison.

## 5 Source of Controversy

In the previous section, we evaluate the performance of relevance measures over Wikipedia articles from different perspectives. We see that focusing on different perspectives results in the varying measure qualities, respectively. Moreover, since a social network consists of multi-dimensions as well as explicit or implicit relations associated with, replying on only one dimension or one relation provides limited, or even misleading, knowledge rather than the underlying truth. Considering the interplay between different relations with different dimensions involved, it is hard to derive the cause and effect by just looking at one specific relation. Taking Wikipedia as an example, controversy during the knowledge building process may originate from the fight of aggressive contributors, or the controversial property of the topic covered in articles, etc,. It is obvious that considering all aspects that would bring about controversy is the only correct way to resolve it. Here, we demonstrate that by examining relations based on different perspectives, we can obtain a clear insight about the origin of controversy in Wikipedia.

The investigation of origin of controversies is important - because some conflicts in collaborative activities in general is inevitable - and because of the rich meta-information available openly in the case of Wikipedia, it provides an unique avenue to understand collaboration dynamics in online social collaborative environments in general. It is also specifically relevant for Wikipedia in that - while presence of articles which deal with controversial content is natural, controversies may also arise simply because of discord among collaborating contributors. The later reflects poorly upon the overall collaboration environment as well as quality of content. In this section, we illustrate by combining the relevance information from different perspective, the essential factor regarding the controversy during the collaborative knowledge building process can be identified clearly.

### 5.1 Methodology

In group theory (Johnson and Johnson, 2002), controversy is defined as "the conflict that arises when one person's ideas, information, conclusions, theories, and opinions are incompatible with those of another person, and the two seek to reach an agreement". Given that Wikipedia is a collaboration system of knowledge building, controversy in Wikipedia is manifested by high volume of delete operations of others' contributions. This happens because people argue with each other by adding their opinions or revising others' work. We ask and investigate in this paper, "What is the root cause of such behavior?". As said by John Stuart Mill (1982), "Since the general or prevailing opinion on any subjects is rarely or never the whole truth, it is only by the collision of adverse opinion that the remainder of the truth has any chance of being supplied". Thus controversy could happen in the articles inherently containing some specific concept about which the contributors hold adverse opinions. Besides the issues of topics, contributors expose their social relations, such as prejudice, aggression, intimacy, through the interaction of their edit behaviors. Among multiple types of social relations identified in the study of social psychology (Myers, 2009), prejudice is the most relevant one with regards to controversy in the context of Wikipedia. Gordon Allport (1979) defined prejudice in his book *The Nature of Prejudice* as "an antipathy based upon a faulty and inflexible generalization". Consequently, it is possible that a group of contributors hold negative tendencies towards a class (a category) of topics, which inevitably causes argument. Since contributors are the main force driving the development of Wikipedia, contributors with emulative or aggressive personalities could also cause conflicts. Based on the above discussion and speculations in existing Wikipedia specific studies (Brandes et al, 2009; Brandes and Lerner, 2007; Kittur et al, 2007; Vuong et al, 2008), we identify and investigate three plausible hypothesis.

- The article deals with *specific controversial concepts* which are championed by different groups of users. For example, the article on Michael Jackson deals with specific controversial topics like child abuse, drugs, etc. In this example, the controversial concepts are sections of the article where the other sections of the article are not controversial.
- Alternatively, the article may belong to *a category of topics* which is generally controversial in nature. For example, all articles on nuclear technology and related scientific concepts may get controversial. Users have different understanding or points of views, which they may be championing across all associated articles.
- It is also possible that some aggressive contributors fight against each other, more because of *personality and egoistic reasons* rather than to do anything with the content of the articles themselves, and thus inadvertently make the articles look controversial.

We first determine the relevance relationships of controversial articles based on the above assumed causes. If the controversy is from the topical category of an article, then the relevant articles with similar content should attract the attention of the community too and will be controversial as well. Thus, controversial articles should be grouped in regards to the topical categories they deal with. Alternatively, if the aggressive contributors are the source of controversy, then controversial articles must share a lot of such contributors. By measuring the relevance of articles from commonality of their contributors, the controversial articles should be much closer to each other. We can then cluster the controversial articles together by considering their contributors.

By clustering the controversial articles over different aspects, we expect to identify the common properties among them and confirm or discount the plausible source(s) of controversy. However, it is difficult to measure the relevance of controversial articles which contain some specific controversial topics (e.g., some specific sections in an article). Articles dealing with specific controversial topics should be less relevant to each other either in terms of content similarity or in terms of common contributors.

## 5.2 Controversial Articles

Using the method used to identify the controversial articles in (Vuong et al, 2008), we build up a ground truth of 68 controversial articles by looking for the dispute tags assigned to the articles in their whole lifespan. We denote this corpus of controversial articles as CA. There are 6 dispute indicative tags. Table 7 shows these tags and explains their meanings.

We first analyze the coordination and conflict of the contributors of these 68 articles. 5,203 contributors, excluding bots,[6] had contributed to these. For identifying the disputes between contributors, we compare two successive revisions by counting the words in the old revision that were deleted in the new revision. It is likely that a contributor makes several successive revisions. In that case we consider only the last revision. Two contributors are said to have disputes in an article if one contributor has deleted some words from another's contribution, or they both have deleted each other's words. This simplistic and albeit somewhat flawed, nevertheless useful model to determine disputes was proposed in (Adler and de Alfaro, 2007). Results over the ground truth set are as follows:

- 4,285 contributors have edited exactly one controversial article, which is 82.4% of all the contributors. They can be downright discounted from being disputative.
- 15,444 contributor pairs have edited at least 2 common controversial articles. There are 917 unique contributors in these 15,444 contributor pairs.
- Among these 15,444 pairs only 81 contributor pairs (0.58%) comprising 71 unique contributors (0.46%) have disputes in at least 2 controversial articles.

The above results indicate that most contributors, 99.54%, definitely don't stalk each other even if they had disagreement on some specific article and argued with each other, which would otherwise have led to more disputes in other articles. It also means that users don't intentionally form groups with any specific agenda and pursue such agenda across articles - even if such groups may form automatically in any given article, as has been witnessed previously (Le et al, 2008).

So far we observed the disputes by counting the words deleted by contributors. Next, we zoom in on these 15444 contributor pairs by measuring edit wars among them. An edit war occurs when individual contributors or groups of contributors repeatedly override each other's contributions, rather than try to resolve the disagreement by discussion[7]. In this paper, we define an edit war as a pair of contributors who have deleted each other's contributions at least 3 times

[6] The bots in Wikipedia are automated or semi-automated tools designed by contributors to carry out some edits, for example, adding some content and some links, reverting vandalism or removing some images, to a specific class of articles. Bots must be harmless and useful and be approved by Wikipedia.

[7] `http://en.wikipedia.org/wiki/Wikipedia:Edit_war`

within an article. The following results are obtained from the analysis of edit wars among contributors:

– 104 contributors are involved in 93 edit wars within 29 articles.
– 2 contributor pairs have edit wars within 2 controversial articles, and 1 contributor pair has edit wars within 3 controversial articles.

From the above results we note that while contributors have different opinions on some articles and argue with each other, even starting an edit war to repeatedly override each other's contributions, still most of these contributors don't fight against each other again in other articles. However, one cannot say that no contributors are aggressive or disputative. 81 contributor pairs comprising 71 unique contributors have disputes in more than 1 article. This indicates that an aggressive contributor may choose different contributors to argue with in different articles. Also, the results of this section are restricted to a small corpus of 68 documents. We explore further the role of contributors next.

## 5.3 Role of Contributors

As described in Section 3.1, the expert-based similarity measure considers the common contributors shared by two articles. If there are more common contributors with more revisions, there will be higher expert-based similarity scores between such articles. Thus, the expert-based similarity measure opens an avenue for us to identify recurring disputes among disputative contributors.

In (Vuong et al, 2008), a controversy model was built based on the assumption that the controversial contributors are the sources of disputes. If it is true, then controversial articles should have more shared controversial contributors with high contribution scores, leading to the very high expert-based similarity score for two controversial articles.

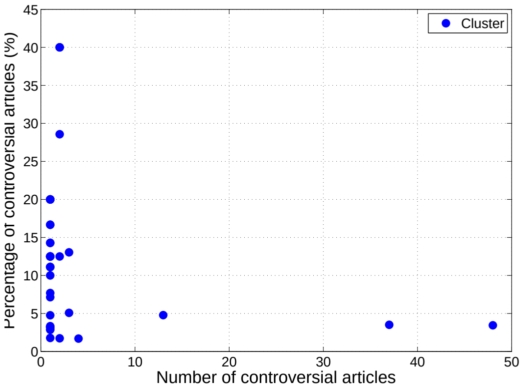
To determine whether contributors are the source of disputes, we first retrieve the top 30 relevant articles for each of the 68 CA articles by using the expert-based similarity and aggregate them, including the original 68 CA articles, as a sub-dataset. The sub-dataset has 1696 unique articles. This is because some of these articles are close to at least two CA articles. We then use DBScan clustering algorithm to group these articles.

If contributors are the cause of controversy, the controversial articles should have high expert-based similarity scores, so the CA articles should be clustered into the same cluster along with a large proportion of controversial articles in general. If this is true, then we can conclude that contributors are the cause of controversy. If not, we can likewise conclude that the disputes originate from the controversial nature of the topic that is specific to the article or the category to which the article belongs, rather than because of the contributors.

Since there is no obvious threshold point in the sorted $k$-dist graph, we set the percent of noise to 15%, 30%, 45%, 60% and 75% respectively. We then check how many clusters contain CA articles and the percentage of CA articles in

| Feature | Percentage of Noise | | | | |
|---|---|---|---|---|---|
| | 75% | 60% | 45% | 30% | 15% |
| Cluster | 24 | 37 | 29 | 16 | 9 |
| Cluster with CA | 7 | 14 | 14 | 8 | 5 |
| CA as Noise | 59 | 50 | 36 | 24 | 15 |

(a) Statistic about clustering result w.r.t the different parameter settings

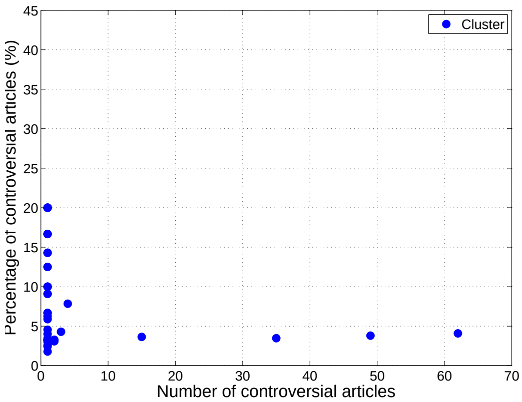



(b) Number of controversial articles vs. Percentage of controversial articles in their clusters

**Fig. 4** Distribution of CAs using expert-based similarity

| Feature | Percentage of Noise | | | | |
|---|---|---|---|---|---|
| | 75% | 60% | 45% | 30% | 15% |
| Cluster | 22 | 26 | 16 | 5 | 1 |
| Cluster with CA | 10 | 24 | 11 | 4 | 1 |
| CA as Noise | 52 | 39 | 23 | 16 | 6 |

(a) Statistic about clustering result w.r.t the different parameter settings




(b) Number of controversial articles vs. Percentage of controversial articles in their clusters

**Fig. 5** Distribution of CAs using cosine similarity

the clusters. Figure 4(a) shows the number of clusters produced, the number of clusters containing the CA articles and the number of CA articles classified as noise w.r.t the five settings for the percent of noise. As illustrated, most of CA articles are classified as noise and only a small proportion of generated clusters contain CA articles.

When the percent of noise is 30% or 15%, more than half of the CA articles are put into the clusters. It looks like controversial articles are closer to each other. But as the number of generated clusters decreases, each cluster becomes larger and contains a large number of articles. Thus, the percentage of the CA articles, instead absolute numbers in their cluster

**Table 7** Dispute Tags Used

| Tag | Meaning |
|---|---|
| `{{disputed}}` | The factual accuracy of this article is disputed. |
| `{{totallydisputed}}` | The neutrality and factual accuracy of this article are disputed. |
| `{{controversial}}` | This is a controversial topic, which may be under dispute. |
| `{{disputed-section}}` | Some section(s) has content whose accuracy or factual nature is in dispute. |
| `{{totallydisputed-section}}` | The neutrality and factual accuracy of some section are disputed. |
| `{{pov}}` | The neutrality of the article is disputed. |

can help us figure out the real distribution of CA articles. If some cluster contains almost only CA articles and a major proportion of CA articles, then we could say that the controversial articles have higher expert-based similarity scores between themselves. It would then indirectly confirm that the source of controversy is controversial contributors.

Figure 4(b) illustrates the percentages of the CA articles in their clusters w.r.t the absolute number of CA articles each cluster contains. We note that there are two kinds of clusters: 1. clusters with small number of CA articles (less than 10); 2. clusters with a large number of CA articles (more than 10). The relatively low percentage of CA articles in the latter type of clusters indicates that the size of cluster increases along with the number of CA articles it contains. Studying Figures 4(a) and 4(b), we conclude that the controversial articles aren't highly related to each other. In other words, their mutual expert-based similarity is not higher than that to other articles in the dataset. Based on the property of the expert-based similarity, we can say that these controversial articles don't have a large number of shared contributors, even when they have a large number of revisions. Thus, one can refute that the contributors are the source of disputes.

Thus it must be the article itself that contains some specific controversial subject matter or otherwise belong to a category involving general controversial concepts. For example, the article "Michael Jackson"[8] is an example that invites the disputes by its specific controversial matter contained in the article. People argued with each other about his changing appearance, the child sexual abuse, his marriages and even his death. However, we can't find such similar conflicts in other articles about dance musicians. On the other hand, there are some debates in the article "Nuclear power"[9] about its pollution, radioactive waste and safety issues. It is possible that similar disputes occur in articles related to nuclear power, for example, in nuclear reaction, nuclear power stations and nuclear entombment, etc. Next, we investigate disputes from the concept perspective.

### 5.4 Concept Perspective

Let's assume that the article deals with some general controversial concepts - which would recur in other articles dealing with the same concept. Then relevant articles under the same concept should attract the attention of the community too. Thus, these articles could be clustered together as they are semantically relevant in terms of their content. As discussed in Section 4, the semantic relevance can also be derived from the link structure of Wikipedia. But considering the mediocre performance of the P-Rank and SimRank similarity measures obtained in Section 4, we only use the cosine similarity here and repeat the experiments conducted above in Section 5.3. We build the sub-dataset by retrieving the top 30 relevant articles for each CA article by using the cosine similarity. There are 1,589 unique articles, including the original 68 articles. The results are illustrated in Figure 5(a) and Figure 5(b). Similar to our previous experiments exploring the role of contributors, we note that most of the clusters contain only one or two controversial articles. And the percentage of CA articles in clusters is very low (less than 20%), especially in the clusters containing a large number of CA articles. We can say that the controversial articles don't have high relevance with each other in their semantic content, which means that the controversial concept is not a principal source of controversy either.

Having eliminated the other possibilities, we infer that specific controversial topics contained in articles are the primary source of controversy in Wikipedia. This conclusion is specific to the "Religious objects" category, but our methodology can be emulated for any other data-set.

## 6 Conclusions

Over the last decades, an increasing amount of our daily life and business is being carried out online using digital technologies, with the advent of Web 2.0 and online social networking sites. The interactions among users and information items offer us a new avenue to discover information that complement with those from other approaches. Thus, rich information regarding different kinds of relations may be obtained by considering relations mined from different perspectives, which offer interpretations not always available when investigating individual facets in isolation. In this paper, we examine different relations from different perspectives in the context of Wikipedia by studying the multiple relations induced among contributors and articles based on edit history, link structure, contributors' expertise, etc.

Besides the link structure and content information associated with Wikipedia articles, we propose a new similarity measure, named *expert-based similarity*. Experiments show that the expert-based similarity is an effective and efficient similarity measure to measure the relevance of articles.

Moreover, as a case study, we studied the source of controversy from different aspects of Wikipedia, including the contributors' edit history, general controversial subject which an article belongs to, as well as topics specific to individual articles. By leveraging different dimensions, i.e., the content, the semantic link structure as well as the editorship of

[8] `http://en.wikipedia.org/wiki/Michael_Jackson`
[9] `http://en.wikipedia.org/wiki/Nuclear_power`

Wikipedia articles, we find that, while isolated edit-wars and group formations have been witnessed in previous studies, most contributors however don't continue such antagonisms across articles (i.e., there is no stalking effect) - and we conclude that disagreements among contributors is nothing personal. We similarly find that general topics and concepts do not cause controversies. Thus, we conclude that controversies arise from specific content typically confined to individual articles themselves.

This conclusion on 'origin of controversies' is valid specifically for the articles in "Religious Objects" category of Wikipedia, but our methodology to infer the conclusion is generic.

In the context of the bigger picture of social network mining and analysis, this paper is a tiny albeit important step in systematically demonstrating the possibility of using different aspects of relations to determine new information, which, without the use of a specific relation aspect may simply not be derivable, or worse still, may be misleading.

**Acknowledgements** The work is funded by A-STAR grant no: 072 134 0055.

## References

Adafre SF, de Rijke M (2005) Discovering missing links in wikipedia. In: Proc. of LinkKDD, ACM, pp 90–97

Adler BT, de Alfaro L (2007) A content-driven reputation system for the wikipedia. In: Proc. of WWW, pp 261–270

Allport GW (1979) The nature of prejudice. Basic books

Bhattacharyya P, Garg A, Wu S (2010) Analysis of user keyword similarity in online social networks. Social Network Analysis and Mining pp 1–16

Brandes U, Lerner J (2007) Visual analysis of controversy in user-generated encyclopedias. Information Visualization 7(1):34–48

Brandes U, Kenis P, Lerner J, van Raaij D (2009) Network analysis of collaboration structure in wikipedia. In: Proc. of WWW, pp 731–740

Bross J, Richly K, Kohnen M, Meinel C (2011) Identifying the top-dogs of the blogosphere. Social Network Analysis and Mining pp 1–15

Ester M, Kriegel HP, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proc. ACM KDD, pp 226–231

Hamouda S, Wanas N (2011) Put-tag: personalized user-centric tag recommendation for social bookmarking systems. Social Network Analysis and Mining pp 1–9

Jeh G, Widom J (2002) Simrank: A measure of structural-context similarity. In: Proc. of ACM KDD, pp 538–543

Johnson DW, Johnson FP (2002) Joining together : group theory and group skills, 8th edn. Allyn & Bacon

Kamps J, Koolen M (2008) The importance of link evidence in wikipedia. In: Proc. of ECIR, pp 270–282

Kamps J, Koolen M (2009) Is wikipedia link structure different? In: Proc. of WSDM, pp 232–241

Kittur A, Suh B, Pendleton BA, Chi EH (2007) He says, she says: Conflict and coordination in wikipedia. In: Proc. of SIGCHI Conf. on CHI, pp 453–462

Le MT, Dang HV, Lim EP, Datta A (2008) Wikinetviz: Visualizing friends and adversaries in implicit social networks. In: Proc. of Intelligence and Security Informatics (ISI), pp 52–57

Lin YR, Sun J, Castro P, Konuru R, Sundaram H, Kelliher A (2009) Metafac: community discovery via relational hypergraph factorization. In: Proc. of ACM KDD, Paris, France, pp 527–536

MacQueen JB (1967) Some methods for classification and analysis of multivariate observations. In: Proc. of Berkeley Symposium on Mathematical Statistics and Probability, pp 1:281–297

Manning CD, Raghavan P, Schtze H (2008) Introduction to Information Retrieval. Cambridge University Press, New York, NY, USA

Mill JS (1982) On Liberty. Penguin Classics

Myers D (2009) Social Psychology, 10th edn. McGraw-Hill

Potthast M, Stein B, Gerling R (2008) Automatic vandalism detection in wikipedia. In: Proc. of ECIR, pp 663–668

Rendle S, Balby Marinho L, Nanopoulos A, Lars ST (2009) Learning optimal ranking with tensor factorization for tag recommendation. In: Proc. of ACM KDD, Paris, France, pp 727–736

Tan PN, Steinbach M, Kumar V (2005) Introduction to Data Mining, (First Edition). Addison-Wesley Longman Publishing Co., Inc.

Vuong BQ, Lim EP, Sun A, Le MT, Lauw HW (2008) On ranking controversies in wikipedia: Models and evaluation. In: Proc. of WSDM, pp 171–182

West R, Precup D, Pineau J (2009a) Completing wikipedia's hyperlink structure through dimensionality reduction. In: Proc. of ACM CIKM, Hong Kong, China, pp 1097–1106

West R, Precup D, Pineau J (2009b) Completing wikipedia's hyperlink structure through dimensionality reduction. In: Proc. of CIKM, ACM, pp 1097–1106

Zhang Y, Sun A, Datta A, Chang K, Lim EP (2010) Do wikipedians follow domain experts?:a domain-specific study on wikipedia knowledge building. In: Proc. of JCDL, Gold Coast, Australia

Zhao P, Han J, Sun Y (2009) P-rank: A comprehensive structural similarity measure over information networks. In: Proc. of CIKM, pp 553–562

Zhou Y, Cong G, Cui B, Jensen CS, Yao J (2009) Routing questions to the right users in online communities. In: Proc. of IEEE ICDE, Washington, DC, USA, pp 700–711