2011 IEEE International Conference on Privacy, Security, Risk, and Trust, and IEEE International Conference on Social Computing

Read What You Trust: An Open Wiki Model Enhanced by Social Context

Haifeng Zhao*, William Kallander*, Tometi Gbedema*, Henric Johnson†, Felix Wu*†

* University of California, Davis, California, USA
Email: {hfzhao, wkallander, tkgbedema, sfwu}@ucdavis.edu

† Blekinge Institute of Technology, Karlskrona, Sweden
Email: {henric.johnson}@bth.se

***Abstract*—Wiki systems, such as Wikipedia, provide a multitude of opportunities for large-scale online knowledge collaboration. Despite Wikipedia's successes with the open editing model, dissenting voices give rise to unreliable content due to conflicts amongst contributors. From our perspective, the conflict issue results from presenting the same knowledge to all readers, without regard for the importance of the underlying social context, which both reveals the bias of contributors and influences the knowledge perception of readers. Motivated by the insufficiency of the existing knowledge presentation model for Wiki systems, this paper presents TrustWiki, a new Wiki model which leverages social context, including social background and relationship information, to present readers with personalized and credible knowledge. Our experiment shows, with reliable social context information, TrustWiki can efficiently assign readers to their compatible editor community and present credible knowledge derived from that community. Although this new Wiki model focuses on reinforcing the neutrality policy of Wikipedia, it also casts light on the other content reliability problems in Wiki systems, such as vandalism and minority opinion suppression.**

I. Introduction

Wiki systems are widely-used online collaborative applications which allow multiple contributors from diverse backgrounds and dispersed geographic locations to collaborate in creating and editing manuals, books, and other public knowledge bases. Among the numerous Wiki ecosystems, Wikipedia is probably the most widely known. Its essential idea, that a useful encyclopedia of knowledge can be created by allowing anyone (even anonymous users) to create and edit articles, is predicated on the principles of openness and neutrality. This openness ideal has led to challenging administrative problems in policing neutrality as Wikipedia has grown to over 3.6 million articles (in English alone) and with millions of contributors (as of June 2011). In the face of such scale, the openness policy has invited conflicts, or the inclusion of bias, debate, and abuse inside Wikipedia articles covering controversial topics.

The organization behind Wikipedia regards the concept of maintaining a neutral point of view (NPOV) as one of its founding principles. However, opening unrestricted editing access to everyone makes this lofty goal overly optimistic due to the inevitable biases and idiosyncrasies of human editors. Instead of collaborating harmoniously, some individuals attempt to dominate "their" articles and nullify all previous edits with which they disagree, often forcing administrators to lock down the editing access of those articles in contention. However, this "lock" method is, by itself, also at odds with the NPOV policy since Wikipedia administrators may subjectively choose their own preferred article edits before locking such articles.

Wikipedia assumes that "the articles are agreed on by consensus"[1]. This assumption treats unreliable content as a consensus problem, and thereby posits that, while misleading information can and will be contributed, over time the quality will improve as editors reach consensus and the resulting article moves toward a "stable" version. This effect works well in reducing certain types of unreliable content, such as vandalism, as the majority of editors are honest and responsible. However, it remains vulnerable to disputes existing in hundreds of thousands of pages, especially those related to contentious historical, religious and political subjects. Even if an article appears to be "stable", it may still retain bias, as some earlier contributors may have preferred to throw in the towel rather than engage in endless editing wars. With regard to a controversial article, there could be hundreds of editors with considerably diverse evidence, statements and viewpoints to present. For the vulnerable reader who is unfamiliar with a topic presented in an article, how does s/he evaluate the credibility of information from a plethora of unknown or anonymous editors?

Take a locked page "Muammar Gaddafi" (the leader of Libya) for example. Figure 1 shows two historical updates which resemble one another but have obviously different sentiments and evidence: The lefthand update positively expects the US government to restore diplomatic ties with Libya and implies that the country is not supporting terrorism. However, the righthand update suggests that the US government would only restore diplomatic ties contingent upon the cessation of Libya's weapons of mass destruction programs. Meanwhile,

[1] http://en.wikipedia.org/wiki/Wikipedia The version of 04:54, 29 May 2011

978-0-7695-4578-3/11 $26.00 © 2011 IEEE
DOI

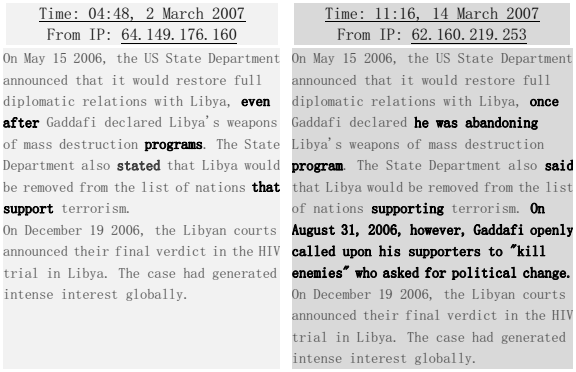

Fig. 1. Two Historical Updates on the Page of Muammar al-Gaddafi

the right update also puts forth evidence that denounces Libya's autocracy as further supporting terrorism. Both editors are anonymous. Which update should be more trusted? While Wikipedia authorizes administrators to evaluate opinions and evidence on some controversial topics[2], there are no guarantees that these "experts" will completely avoid personal bias and present universally fair viewpoints to readers.

There are some previous research efforts studying the conflict problem in Wikipedia[22], [11], [12]. However, their efforts focus on analyzing and visualizing conflict patterns rather than redressing the problem. In this paper, our primary focus is instead the subtler problem of resolving and fixing conflicted viewpoints in the article content, which remains a problem lacking an effective solution.

In our view, the ongoing presence of such dilemmas in Wikipedia reveals that the conflicts are not a consensus problem (as Wikipedia has been assuming), but rather they are due to the presentation of knowledge uniformly to all readers, irrespective of their social context. In other words, the content quality problems originating from conflicting viewpoints in Wikipedia are due to a failure to account for social backgrounds and interactions between contributors. The importance, trustworthiness, and compatibility of the same pieces of information vary significantly to users with different social contexts. The diversity of social contexts stretches across many variables, such as social relationships, cultural factors, ethnicity, education, and other human factors. In reality, when facing the dilemma of conflicting information, people tend to believe friends or authorities with whom common values and/or similar social backgrounds are shared. However, these social context factors are overlooked in the knowledge presentation layer of traditional Wiki systems where content is presented uniformly to all readers.

We assert that if more editor/contributor social context information behind contributed content such as editor background, friendships, and interactions were considered when presenting knowledge to different readers, it would guide the Wiki system to better decide what pieces of knowledge are important and meaningful to different readers. Moreover, if the knowledge selection process is customizable by readers, they will better be able to judge the credibility of knowledge content and thereby improve the quality of what is presented inside the system. As a result, unreliable content subjected to conflicts or vandalism could be mitigated, since the importance of these contributions would be reduced when being presented to socially-incompatible readers.

The goal of this paper is to solve conflicts in Wiki systems as discussed above. Our approach balances reader preference for socially compatible knowledge with opposing viewpoints by tuning the knowledge presentation layer. Assuming that the social information of Wiki editors and readers are available, we formally define the problem as follows:

Problem. In an open editing mode of a large Wiki system, how can we leverage social context information among readers and editors to reveal the underlying debates, provide personalized content, and present trustworthy knowledge to different readers?

We therefore introduce a new Wiki model, TrustWiki, which integrates data mining techniques to reveal the opposing groups of a controversial topic and present credible knowledge to readers consistent with their social context. The uniqueness of this paper rests on the customization of the presentation layer of online knowledge representation systems by incorporating social context to prioritize compatible information for each reader. In addition this approach offers a solution for redressing other misbehavior in Wiki systems, such as vandalism and minority opinion suppression.

The rest of this paper is laid out as follows. First, Section II introduces the solution, TrustWiki, and its key components. Next, Section III demonstrates the effectiveness of TrustWiki by three case studies, two of which are taken from controversial Wikipedia articles, with another taken from a contentious debate surrounding a New York Times article. Section IV further discusses the implementation of TrustWiki and its influence on the other Wiki problems. After reviewing related work in Section V, we summarize our conclusions and discuss next steps in Section VI.

II. TrustWiki Core Model

To reveal the conflicts and present credible knowledge, TrustWiki requires three components: first, the construction of an appropriate social context is fundamental to uncover the hidden factors behind contributed knowledge; second, an efficient algorithm is necessary to identify editor communities which have diverged from each other; and finally, a readable and customizable knowledge presentation layer reflecting reader-compatible or desired "credible" viewpoints. In this section, the three components will be explained in detail.

A. Social Similarity Network Construction

The elements of social context include relationships, interactions, gender, ethnicity, education, profession, and the communities in which we live. In TrustWiki, we separate these

[2]http://en.wikipedia.org/wiki/Wikipedia:Administrators

elements into two segments and represent each segment with a separate network. The first segment of social context is interpersonal social information, such as affinity, relationship and other forms of interaction, which we refer to as the trust network. We avoid using the term "friendship network", because we consider friendship as an non-directional relationship, but trust as a directional relationship. The second segment is individual social information, such as culture and background, where edges are imparted between two individuals who share background features in common, such as identical ethnicity, or common education. We refer to this latter network as the background network.

Thus, TrustWiki constructs a trust network using interpersonal social information and a background network using individual biographical features. Let us leave the source of social context information to Section IV-A. Assuming that we have already acquired the two kinds of social context information, we now focus on the functions and maintenance of the two networks separately.

Suppose we have a trust network $G_t(V, E)$ with users represented by vertices and weighted edges simulating the relationship intensity and interaction among users. In our model, this forms a fully connected **directional** graph. In $G_t(V, E)$, edge weights are determined by trust intensity and are influenced by user interactions. There are many possible methods to evaluate the trust intensity from u_i to u_j. Without loss of generality, we propose a simple measure herein to calculate and update trust values. Every user u_i keeps track of a trust value $trust(u_i, u_j) \in [0, 1]$ to another user u_j. By default, the trust value $trust(u_i, u_j)$ is initialized as α (e.g., 0.5 in our experiments) if u_i has never provided feedback to u_j's editing. If u_i's feedback to u_j's editing is γ, the trust value from u_i to u_j is dynamically refreshed as:

$$trust(u_i, u_j) = \frac{trust(u_i, u_j) + \gamma}{2}$$
$$\gamma = \begin{cases} 1.0 & \textit{positive feedback} \\ 0.0 & \textit{negative feedback} \end{cases} \quad (1)$$

That is, the more positive feedback u_i gives to u_j, the higher $trust(u_i, u_j)$ will be, implying the accrual of preference of u_i for u_j's editing.

Aside from the trust network, the background network also plays an important role in knowledge acceptance. Social background, such as gender, ethnicity, nationality, profession, and other similar factors potentially produce viewpoint divergence among people. We assume that, even without direct friendship, two people with similar social backgrounds are more likely to agree on a controversial issue than two people with different social backgrounds. To be noted here, most social background values are categorical values, rather than numeric values. The key characteristic of a categorical attribute is that the values are not inherently ordered. For example, the category of "Geography" may contain values like "California", "Massachusetts", and so on. To calculate the background similarity of two users, a categorical similarity measure is required. Anderberg[2] introduced a categorical similarity measure, which assigns a higher similarity to rare matches, and lower similarity to rare mismatches. The only limitation of the Anderberg measure is that each object may only possess a single value on an attribute. TrustWiki adapts the Anderberg measure to accept multiple-value attributes. If N_m denotes the count of all attribute values having appeared on the m^{th} category, and $p_m(A_k)$ denotes the probability of the attribute value A_k in the m^{th} category, $p_m(A_k)$ can therefore be computed by dividing the frequency of A_k (denoted by $f(A_k)$) with N_m as in Equation 2:

$$p_m(A_k) = \frac{f(A_k)}{N_m} \quad (2)$$

For two instances X and Y, TrustWiki computes their similarity as Equation 3 shows:

$$S(X,Y) = \frac{\sum_{m=1}^{d} \sum_{A_k \in U_m} \frac{\frac{1}{p_m(A_k)} \frac{1}{f_{A_k}+1}}{|W_m|}}{\sum_{m=1}^{d} \sum_{A_k \in U_m} \frac{\frac{1}{p_m(A_k)} \frac{1}{f_{A_k}+1}}{|W_m|} + \sum_{m=1}^{d} \sum_{A_k \in V_m} \frac{\frac{1}{p_m(A_k)} \frac{1}{f_{A_k}+1}}{|W_m|}} \quad (3)$$

where

$$U_m = X_m \cap Y_m$$
$$V_m = X_m \cup Y_m - X_m \cap Y_m \quad (4)$$
$$W_m = X_m \cup X_Y$$

and X_m, Y_m are respectively the values sets of X and Y on the m^{th} category.

Similar with respect to the Anderberg measure, this adapted measure likewise assigns higher similarity to rare matches, and lower similarity to rare mismatches. The range of $S(X, Y)$ is $[0, 1]$.

To take advantage of both trust network and background network, we use a measure named "social similarity" in TrustWiki, which linearly combines both similarity measures as described in Equation 5:

$$social\_sim(u_i, u_j) = \theta_1 \times trust(u_i, u_j) + \theta_2 \times S(u_i, u_j) \quad (5)$$

with $\theta_1 + \theta_2 = 1$.

Social similarity can be therefore seen as the weighted combination of both trust network and background network similarities. If we consider social similarity as a directional relationship, we build a new graph $G_{social}(V, E)$ which unifies both interpersonal and individual social context information.

B. Discovery of Credible and Compatible Editors

We turn our attention now to evaluating the importance and compatibility of knowledge information to different readers based on their social context.

In our approach, we analyze editors rather than articles due to current limitations of Natural Language Understanding (NLU) technologies. In an ideal world, knowledge in Wikis would be represented semantically, such that fine-grained knowledge facts could be extracted and evaluated for inclusion in a reader's view of the knowledge (Similar to Ontowiki[3]

but without the limitation of only being able to handle simple cases). However, NLU is not mature enough to accomplish this task with sufficient accuracy. Therefore, instead of analyzing the content directly, TrustWiki analyzes contributors as a proxy for subjective attitudes in dividing up the content.

TrustWiki separates contributors into groups and presents (to a reader) the information provided by an editor group which is most compatible and credible to the reader based on previous interactions and background similarity. More specifically, for each article, it first clusters editors based on social similarity to magnify the consensus within each group. Then, it evaluates which editor cluster is most similar to the reader and presents the textual content based on the aggregated contributions of members from this closest cluster.

As described in Section II-A, social context information is embedded in $G_{social}(V, E)$, therefore we want to find a partition algorithm on the graph. This partition can be formalized by the mincut problem. Given a weighted graph described by matrix A and a clustering $C = \{C_1, C_2, ..., C_k\}$, we can calculate the weighted cut by:

$$WCut(C) = \sum_{k=1}^{K} \sum_{k' \neq k} Cut(C_k, C_{k'}) \tag{6}$$

where

$$Cut(C_k, C_{k'}) = \sum_{i \in k} \sum_{j \in k'} A_{ij} \tag{7}$$

Usually, Equation 7 is normalized to counteract outliers as follows:

$$Cut(C_k, C_{k'}) = \sum_{i \in k} \sum_{j \in k'} \frac{A_{ij}}{vol(C_k)} \tag{8}$$

where $vol(C_k)$ is the combined weight of all edges in C_k.

So the goal of the mincut problem is to find a clustering $C*$ such that:

$$WCut(C*) = \min_{C} WCut(C) \tag{9}$$

An approximation to this mincut problem is spectral clustering[23], which recursively partitions the data set by removing edges and evaluating the mincut until k clusters are identified. The general process of spectral clustering contains two steps: first, compute the unnormalized Laplacian matrix L from the weighted adjacency matrix W; and second, apply the k-means algorithm to cluster the first k eigenvectors of L into $C = \{C_1, C_2, ..., C_k\}$. TrustWiki adopts the normalized spectral clustering introduced by Shi and Malik[21] to compute eigenvectors. The approximated clustering result is subjected to the centroids of the k-means algorithm.

As to one property of spectral clustering, if the cluster number is two, then the k-means algorithm is not necessary. The solution can be given by the second smallest eigenvector f of L. In order to obtain a partition of the graph we need to re-transform the real-valued solution vector f of the relaxed problem into a discrete indicator vector. The simplest way to do this is to use the sign of f as the indicator function:

$$\begin{cases} v_i \in C & \text{if } f_i \geq 0 \\ v_i \in \overline{C} & \text{if } f_i < 0 \end{cases} \tag{10}$$

TrustWiki accepts this (sign) property and sets the default clustering number as two, as it regards controversial articles as polarizing editors into either positive or negative camps. Note too that the number of clusters may also vary or be customized according to specific scenarios.

The editor clustering process employed by TrustWiki equipped with spectral clustering is summarized in the following steps:

1) From $G_{social}(V, E)$ which contains all the users of a Wiki, a subgraph $G' < V', E' >$ is extracted with nodes V', which are editors specifically related to the article.
2) Perform spectral clustering algorithm on the subgraph G'.
3) Assign the closest cluster to a reader as the most credible editor group with regard to the article. The reader's distance to a cluster depends on the average of his/her social similarity with everyone in the cluster.

After the clustering process, TrustWiki recognizes the editor group most compatible and credible to the reader. Even though the closest cluster may contain some untrustworthy editors (due to noise in social similarity values), the reader still has a high probability of acquiring his/her most compatible information. Moreover, if the reader does not like the provided content, his/her optional feedback will influence the decision of TrustWiki when choosing an editor cluster for him/her in future interactions.

C. Format of Knowledge Presentation

We discuss the content and format of the knowledge information that is to be provided to the reader in this section. As we mentioned in Section I, the knowledge presentation of an article provided by traditional Wiki systems like Wikipedia is the latest historical text version, which is easy to implement with current technology. However, if TrustWiki decides to maximize the knowledge provided by the most trusted cluster of contributors, it must first realize that the pieces of knowledge contributed by these trusted editors are inconsecutive, hardly distinguishable, or even incomplete natural language information spread over several (or all) historical article versions. It is quite difficult to assemble and organize these pieces of knowledge in a way as coherent, readable and smooth as a Wikipedia's latest version article version, since the finest granularity of knowledge representation in current popular Wiki systems is the free-text of an article. Bias is easily hidden in the narrative, and the technologies for re-writing the prose of an article to remove subjective bias are in their infancy.

One possible solution involves utilizing semantic web technologies in knowledge contribution and presentation. Semantic web technologies yield a much finer-grained knowledge store that can generate desirable content and remove biases, by associating knowledge facts with their contributors, and then

taking advantage of the social context between all editors and readers in order to tailor the presentation of the accumulated knowledge to an individual reader. However, an appropriate and general ontology for storing all the knowledge and social context necessary to produce natural language articles with the same readability as Wikipedia's latest-article version is unavailable.

To utilize the power of social context as well as semantic web technologies but make the problem tractable, TrustWiki makes a compromise by presenting knowledge in an adapted article version together with an auxiliary "Summary Board".

The adapted article version, with similar readability of a Wikipedia latest article version, contains the most accepted historical versions of paragraphs by the trusted editor cluster. Specifically, TrustWiki breaks down the original Wiki article into paragraphs, and for each controversial paragraph, TrustWiki essentially allows editors to provide votes on all historical versions through their modifications and contributions. It then combines three factors to decide which historical version of a paragraph is most popular with the trusted editor cluster. These three factors include the number of positive and negative votes ($FB \in \{-1, 1\}$), the social similarity SIM between a reader and each voter, and the time delta of each paragraph version. Then TrustWiki chooses a paragraph version v_i with the highest score $\widetilde{S}$ and presents it to the reader. A simple formula is provided in Equation 11 to combine the three factors described above, which also downweights older historical modifications, and normalizes the range of the trust score to lie between [-1, 1]:

$$\widetilde{S} = \frac{\sum_{e_i \in E} FB(e_i) \times SIM(r, e_i)}{|E|} \times (1 - \frac{\arctan(\Delta)}{\frac{\pi}{2}}) \quad (11)$$

where r is the current reader, E is the set of all *trusted* editors who provided feedback to the historical version and Δ is its time difference.

Intuitively, the score $\widetilde{S}$ for each paragraph version assumes that the higher the number of trusted editors who provided positive feedback, and the more recent a historical version is, the more reliable the paragraph is to the reader r. In this way, a readable article with credible paragraphs is produced.

The advantage of the adaptive article version is that it reads as smooth as a traditional Wiki, but presents more credible and compatible knowledge. Nevertheless, the weakness of the adaptive article version is that it might only capture partial facts and evidence if the most popular paragraphs do not fully include all credible knowledge content. To mitigate this, TrustWiki incorporates an auxiliary page, the "Summary Board", which looks like Wikipedia's "discussion" page. This "Summary Board" displays the combined facts, evidence, and arguments contributed by the nearest editor cluster (with respect to the reader). The difference between Wikipedia's "Discussion" page and TrustWiki's "Summary Board" is that the "Discussion" page as it currently stands is merely a fusion of editors' opinions on subtopics, which cannot guarantee content reliability. Conversely, the TrustWiki "Summary Board" is built on an underlying ontology model and generates a list of facts and evidence provided by trusted editors. At present, the simplest ontology for the "Summary Board" specifies that fact instances adhere to the form of $subject + predicate + object$, and are ordered by contribution time and social similarity. While this ontology allows for easy Natural Language Generation, a more general ontology that allows for more sophisticated language constructs is desirable for merging the adaptive article version and "Summary Board" into one unified form for knowledge presentation.

The most popular historical trusted article and the "Summary Board" are both forms of personalized knowledge presentation. Personalization of knowledge presentation has been widely used to filter information on the web to accommodate individual preferences by Google and Yahoo in search engine result rankings. We extend this approach to Wiki systems by allowing readers of such content to manipulate social parameters (when calculating editor clusters) which then adjusts the prioritization of the content being presented. Whereas traditional Wiki systems present predetermined knowledge to all readers, TrustWiki provides readers with the capability to flexibly customize the knowledge presentation. While we acknowledge that providing this functionality increases the risk of reinforcing reader biases, we observe the better utility of customizable tools for filtering online content in general and point out that TrustWiki allows the reader to customize their score threshold, such that no content is filtered. The additional level of control allows the user to set preferences according to their sensibilities, much like the many ubiquitous personalization mechanisms found elsewhere in online systems.

The need for manipulating social parameters derives from two concerns. One concern is that the social similarity may be inaccurate, depending on how little information Wiki users choose to provide. This, in turn, potentially causes TrustWiki to display opposing views to the reader in the presentation layer erroneously. Another concern is that the compatible viewpoints tailored to the user may only provide partial details, which do not reflect the whole story on a controversial article. These concerns can be counteracted by providing a customizable knowledge presentation interface to readers. First, with minor modification, TrustWiki allows a user to view the second, third, or n^{th} compatible historical versions of each paragraph within its interface, based on a readers selectable preference. Moreover, TrustWiki can alternatively compute $\widetilde{S}$ scores based on the "untrusted" editor cluster and provide the reader an article of the "opposing" viewpoints. Since advanced readers can manipulate the knowledge presentation by adjusting social parameters, such as trust value threshold and weights of background categories used in the model, a reader could obtain a clear idea about what kinds of people agree/disagree a certain historical version, why they agree/disagree, and better judge how credible they are.

III. Experiments and Evaluation

The ideal approach to verify TrustWiki's performance would be to experiment with real Wikipedia data, as it enjoys

significant amounts of content, contributors, and readers. Unfortunately, as a practical matter this is not achievable, since Wikipedia currently contains little social context, no reader information, and semantic content is locked away in natural language. To address these, one would have to implement TrustWiki outside of Wikipedia, accumulate significant numbers of contributors, readers, and content, and then evaluate the model. Needless to say, this would be a prolonged process, which is not necessary for evaluating how well the method differentiates viewpoints in text based on social similarity. We therefore adopt a case study approach in our evaluation, choosing two cases randomly from among Wikipedia's list of controversial articles[3] (specifically, "Net Neutrality" and "Smoking"), and another from the heated discussions surrounding a controversial article featured in the New York Times on the historical role of Africans in slavery.

Due to the lack of a standard measure to quantify the quality of an article (which is subjective to different readers), the experiments focus instead on the evaluation of editor clustering as a proxy for viewpoint differentiation. We believe correcting readers from the wrong cluster to the right one is less important since the system can quickly remedy that from user feedback. So the goal of our experiment is to determine whether consensus is obvious within a cluster, while divergence is significant across clusters. To complete the evaluation, three kinds of information are needed:

- The social background information of Wiki users, including nationality, education, profession and recognized communities.
- The feedback (e.g., positive/negative) among users on each other's modification(s).
- The attitudes (e.g. positive/negative) of each editor denoting his/her stance on the topic, which is used as ground truth to evaluate the clustering result.

We will first introduce the approaches we used to collect these two kinds of information and then present the performance of TrustWiki on the three experiments.

Since the data sources of the experiments are not identical, we collect the above information in different ways for Wikipedia articles (Net Neutrality and Smoking) and the New York Times article.

For the articles in Wikipedia (e.g. Net Neutrality and Smoking in our experiments), we take the top n non-anonymous/identifiable content contributors whose background information can be collected in their user pages [4], so that a background network can be generated for these editors. As to the feedback interaction and the ground truth attitude labeling, unfortunately, Wikipedia does not allow editors to indicate their attitude toward either an article or another editor's comment, so obtaining the interaction among editors and the ground truth of their attitude toward an article involved human tagging of individual revisions. To mimic the feedback interaction among editors, we adopt a trust model introduced by Adler[1] which simulates "positive"/"negative" feedback among editors based upon their modification actions on previous contributions. As such, in our trust model, if user u_i appends or reverts to an edit of user u_j's, we determine that u_i gives a positive feedback to the edit of u_j. Conversely, if u_i deletes or corrects an edit of u_j's, we determine that u_i gives a negative feedback to the edit of u_j. The trust value from u_i to u_j can be calculated by Equation 1 based on the number of positive feedbacks and the number of negative feedbacks.

To determine the attitude of an editor u_i toward an article, we go through u_i's historical edits on an article and discern whether the individual modification (with respect to the previous version exclusively) makes the topic more or less favorable and apply a label to that edit consistent with that determination. Once all edits are marked, we sum them to decide into which group to associate the contributor. As such, it is important to point out that "positive"/"negative" labels are indicative of edits that reflect the human-discernable attitudes of the editor toward the article topic, not necessarily a violation of the NPOV guidelines governing Wikipedia contributions. Further, some edits are not describable in this way; examples include the addition of inter-wiki links, minor spelling/capitalization/punctuation corrections, and neutral word substitutions. Edits corresponding to these events are labeled as "Neutral". For those edits which are discernable, we mark a change as "positive" if it casts the article about the topic in a more favorable light, and "negative" if it makes the article more critical of the topic. Examples in the Positive case might be the insertion of facts supporting Net Neutrality in the corresponding article, or the removal of content condemning Smoking in the article on that topic. Similarly, Negative labeling examples might be the removal of supportive remarks about Net Neutrality for the article on that topic, or the addition of new content regarding Smoking cessation programs in the article on Smoking. If u_i's "Positive" edits are obviously more than "Negative" edits (e.g., more than two), we label u_i as positive toward the article, and vice versa. Therefore, we can coarsely obtain the feedback among editors and their attitudes toward the same article. The jobs of feedback determination and attitude labeling for the two use cases were performed by two different people separately to avoid the same personal bias on revisions.

For the case study on the New York Times article, we selected an online debate surrounding an article entitled "Ending the Slavery Blame-Game" which was published in the New York Times on April 23, 2010 by Dr. Henry Louis Gates, Jr.[8]. His article highlighted the significance of the historical role played by Africans in the slave trade and advocated in-depth discussion on that role before offering reparations to decedents of victims. As one might assume, this essay generated strong reactions from many scholars and individual commentators, with fierce support on both sides of the issue.

We collected the opinions expressed from 23 non-anonymous respondents with disparate ethnic, professional, and geographical/biographical backgrounds, and used this information to build our background network. The correspond-

[3]http://en.wikipedia.org/wiki/Wikipedia:List_of_controversial_issues
[4]http://en.wikipedia.org/wiki/Wikipedia:User_pages

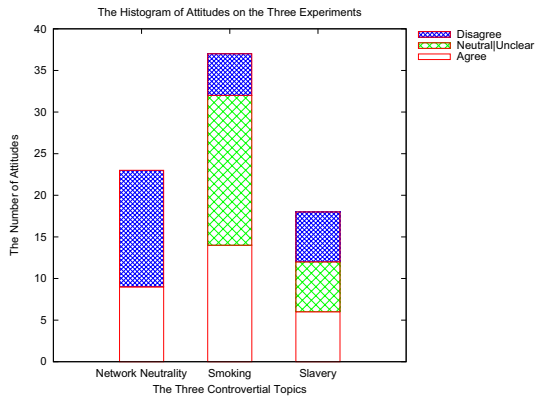

Fig. 2. The Number of Editors and Their Attitudes in The Three Use Cases.

ing trust network was then constructed from their commentaries. If a contributor quotes another contributor's comment/publication positively, we regard it as a positive feedback. By contrast, if a contributor rebuts another contributor's comment/publication, we regard it as a negative feedback. With support from the experts in the Geography Department of the University of California Davis, we determined that 8 contributors (including Dr. Gates) maintained that the African role must be examined before any discussion of reparation payment to descendants of slaves. The other 15 people, however, refuted the argument and called Dr. Gates' attention to the fact the enslavement of the Africans mostly occurred in America, along with other salient facts.

Figure 2 shows the number of content contributors corresponding to the three kinds of attitudes on each experiment.

Observing two distinct camps, we clustered the commentators into two groups. In order to get a clear idea of the effectiveness of our components, namely the trust network, background network and the combined social similarity network, we performed spectral clustering separately on the three different network types.

Since the Jaccard Index[9] has been commonly used to assess the similarity between different partitions of the same dataset, we evaluated clustering accuracies using the Jaccard Index calculated between our ground-truthed labeled groups and the clustering results for each network type. The level of agreement between a set of class labels P and a clustering result Q is determined by the number of pairs of points assigned to the same cluster in both partitions as in Equation 12:

$$J(P,Q) = \frac{a}{a+b+c} \tag{12}$$

where a denotes the number of pairs of points with the same label in P and assigned to the same cluster in Q, b denotes the number of pairs with the same label but in different clusters and c denotes the number of pairs in the same cluster but with different class labels. The Jaccard Index produces a result in the range $[0, 1]$, where a value of 1.0 indicates that C and K are identical.

We performed spectral clustering on the background, trust, and the combined similarity networks separately to investigate their effectiveness for viewpoint clustering. Figure 3 shows the performance of the TrustWiki clustering algorithm on the three experiments. Three important conclusions can be derived from the Jaccard Index result.

First, we find that the Jaccard Index of the trust network in the three experiments shows that sufficient editor interaction is a key factor for discovering editor communities. In the last experiment (pertaining to slavery), the Jaccard Index from trust network is much higher than the others because the editors wrote comments explicitly agreeing with or opposing one another and their words made apparent their attitudes towards the topic. On the contrary, the real-content-driven trust model[1] used in the Wikipedia experiments is too limited to deduce an editor's attitude toward the corresponding topic for a significant portion of the contributions. Further, contributor attitudes toward comments from the other editors were much harder to discern in Wikipedia. Therefore, a sufficient trust system is necessary in order to discover editor communities and obtain their driving factors, but that interpersonal network is currently hidden in Wikipedia.

Second, we show that the Jaccard Index between the background networks are low, which we attribute to the fact that all background categories are treated equal in our experiments. In reality, different background categories play variable roles of influence on editors' opinions given a particular topic. Taking the experiment pertaining to slavery as an example, if we only use the ethnicity category (black/non-black) to construct the background network, the Jaccard Index produced by this background network is 0.4486486, unsurprisingly reflecting more consistency than the 0.3684211 index achieved when considering all background categories equally. The weight of each category when calculating background similarity should therefore be trained based on sufficient training samples with statistical learning methods, such as regression and classification.

Finally, we note that the social similarity networks do not consistently outperform the component (trust and background) networks. As a consequence, the combination of trust and background networks produce better results only when both component networks are built on accurate social and biographical information.

Based on an editor clustering produced by an accurate similarity network, credible paragraphs can be selected and presented as an readable article. All the experiment data, including the background information, feedback information and labeling can be obtained on our web site[5].

Due to the limited information provided in Wikipedia, we mocked up an example of the "Summary Board" for the experiment pertaining to the slavery topic. In Figure 4, the left column contains the knowledge provided by the "for" cluster and the right column contains the knowledge provided by the "against" cluster.

[5]http://dsl.cs.ucdavis.edu/ zhf/homepage/TrustWiki

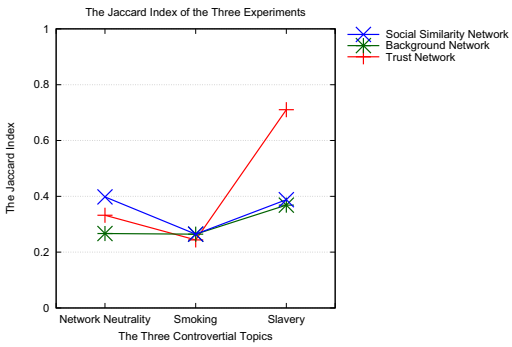

Fig. 3. The Evaluation of TrustWiki Based on Experiments

For	Against
JaeWhan is a contributor of an Asian American Intellectualism and Activism website. He thinks: •The key to moving forward is letting go of anger. •Gates steps back and looks at the facts. Gates demonstrates distancing from anger. •Most cases of racist behavior exists culpability on both sides of the color line. *Chauncey De Vega is an African American writer. In his opinion:* •We need a critical conversation about the role of African tribes and nations in the Black Holocaust. •Chauncey De Vega does not believe he is from Africa. •Black Americans has much to celebrate here. •Black Americans has created against unbelievable odds. •Black Americans does not need inspiration from the mythic past. ...	*Eric Foner is a professor of Columbia University. He thinks:* •The great growth of slavery in this country occurred after the closing of the Atlantic slave trade in 1808. •Africans did nothing to the slave trade within the United States. *Kwabena Akurang-Parry is a professor of History. He considers:* •The viewpoint that "Africans" enslaved "Africans" is obfuscating. Scholar questions the humanity of "all" Africans. •Western media characterizes local crisis in one African state as "African" problem. •Gates' essay supports by works of Linda Heywood and John Thorton. • Works of Linda Heywood and John Thorton supports by extant Eurocentric records. ...

Fig. 4. Example of Summary Board generated from knowledge facts stored in semantic form

The experimental results demonstrate that the TrustWiki model is effective for presenting knowledge on controversial subjects if provided with appropriate background information and a sufficient trust network. Based on the customizable knowledge presentation model, a reader is empowered to not only receive the most credible and compatible information by default, but also perceive knowledge from another angle by customizing the social context parameters.

IV. Further Discussion

A. Implementation of TrustWiki

We have introduced the model of TrustWiki and showed its performance with case studies. In this section, we will discuss its implementation and applicability to other problems in open Wiki systems.

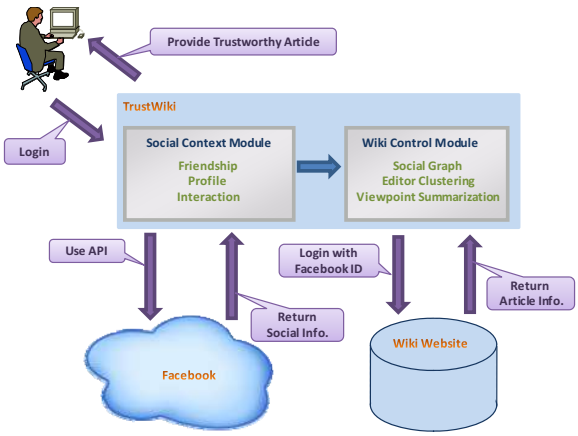

Fig. 5. The Architecture of TrustWiki

In our previous work, we presented a prototype Facebook application, SocialWiki[24], which mainly focuses on handling vandalism in Wikipedia. Even though TrustWiki serves a different purpose and adopts a completely different approach, this work can still utilize the general architecture of SocialWiki as Figure 5 shows. Similar to SocialWiki, TrustWiki can be built as an application in Facebook to lower the barrier of adoption for accumulating users over a prolonged period of time. With the Facebook API, TrustWiki can collect basic user information easily, such as profiles and friendship relationships. The only requirement for this to work is for the application to request permissions from users to access their profiles, and to further allow authentication in TrustWiki using Facebook credentials.

As described in Section II-A, the user social similarity network model of TrustWiki consists of distinct trust and background networks. Using the Facebook API as above, the trust network can be initialized with the friendship information extracted from a user's Facebook friendship list. As a consequence, when a new user joins TrustWiki, s/he brings along additional friendship information to TrustWiki. Although statically initialized from Facebook, the trust network is dynamically maintained by user interaction inside TrustWiki, independent of changes to the outside friendship status in Facebook. This means that sharing a friendship relationship in Facebook is neither a sufficient nor necessary condition for existence of a trust relationship in TrustWiki after initialization. The edges of the trust network are directional and their weights are initialized as α (e.g., 0.5) if the pairwise users are not friends in Facebook, or a higher value β if the pairwise users are linked. Weights of edges are updated according to users' explicit feedback.

Similarly, using the user's Facebook profile, the background network is constructed through matching fields between users or explicitly requiring additional user inputs in TrustWiki. Then, by combining trust and background networks, integrated social similarity can be calculated by Equation 5.

As we discussed in Section II-C, the compatible and credible knowledge presentation layer of TrustWiki is customizable, allowing user inputs to customize the social similarity parameters, such as weights and background categories used in clustering, as well as the parameters used in paragraph selection.

B. Applicability of TrustWiki to other Wiki Problems

The effectiveness of TrustWiki is not limited to conflict problems. Two other concerns, vandalism and minority opinion suppression, also benefit from the TrustWiki model. A 2007 peer-reviewed study[19] that measured the actual number of page views with damaged content concluded that 42% of damage is repaired almost immediately, i.e., before it can confuse, offend, or mislead readers. Nonetheless, given Wikipedia's popularity, this leaves hundreds of millions of damaged views. With TrustWiki, vandals are clustered and thereby only influence readers socially close to them. Individual vandals may be mingled with trusted editors, but their identities can be easily revealed since TrustWiki maintains a social context state. That is, their trust values with respect to other users can be decreased quickly by negative feedback events from readers. As a result, the influence of vandalism is diminished in TrustWiki.

Minority opinion suppression in Wikipedia is another problem which is paid little attention. With an open editing model, the Wiki articles mostly reflect the viewpoints of the majority of editors, though the truth may instead stand with the minority. Previous research[16], [10], [15] has manifested the influence power of minority group in social influence. However, in current Wiki systems, minority opinion is easily removed when these opinions are not tolerated by the majority. Without a fair channel to spread their viewpoints, it is unrealistic to share these opinions, truthful or not. TrustWiki provides a chance for the minority to be heard by allowing the user to adjust parameters that influence other readers socially close to them. When open-minded readers, who are not close to the minority, adjust their social parameters towards the minority editor cluster, it increases the likelihood for others who are similar to such open minded readers to also view these contributions, thus spreading the influence of the minority.

V. Related Work

A. Consensus Problem

The consensus problem has a long history in computer science and forms the foundation of the field of distributed computing[13]. In networks of agents (or dynamic systems), "consensus" involves reaching an agreement regarding a certain quantity of interest that depends on the state of all agents. Formal study of consensus problems in groups of experts originated in management science and statistics in the 1960s (See DeGroot[5] and references therein). The theoretical framework for posing and solving consensus problems for networked dynamic systems was introduced by Olfati-Saber and Murray[20], [17] which builds on the earlier work of Fax and Murray[6], [7]. The reason that the conflict problem in Wikipedia is not strictly a consensus problem is that consensus problems assume each individual cannot obtain any benefit without achieving agreement, which is not the case in Wikipedia.

B. Conflict and Coordination in Wikipedia

There are already some research efforts[22], [11], [12] which have delved into the conflict and coordination problems in Wikipedia. As we mentioned in Section I, their efforts focus on analyzing and visualizing the conflict patterns rather than fixing the problem for readers. Viégas[22] introduced a new visualization tool to reveal collaboration patterns within the wiki context. Kittur's earlier work[11] examined the research on conflict problem and described the development of tools to characterize conflict and coordination costs in Wikipedia. His later work[12] introduced a mapping technique that takes advantage of socially-annotated hierarchical categories while dealing with the inconsistencies and noise inherent in the distributed way that they are generated.

The key difference between the previous research and our work is that we focus on resolving the conflict problem for the reader and thereby reinforce the NPOV policy, rather than focusing on understanding the conflict patterns themselves.

C. Trust in Knowledge Presentation and Perception

The importance of a speaker's credibility has been studied in epistemology from ancient times. In Aristotle's systematization of rhetoric[4], a public speaker's character and credibility to the audience in a discourse is mentioned as ethos, which can influence an audience to consider the speaker to be believable. Recent rhetoric scholars[14] regard ethos as "source credibility" and suggested three dimensions for the credibility construct: expertness, trustworthiness and intention toward the receiver.

The concern of content accuracy in Wikipedia arises because "ethos" of editors are not disclosed. Several trust and reputation models have been proposed to discover reliable editors. Adler et al.[1] proposed a content-driven reputation system for Wikipedia authors where authors gain reputation when their edits are preserved by subsequent authors, and lose reputation when their edits are reverted. Thus, author reputation in their view is based on content evolution only and user-to-user comments or ratings are not used. In our previous research, SocialWiki[24], we described our prototype Wiki system which leverages the power of social networks to automatically manage reputation and trust for Wiki users in terms of the content they contribute and the ratings they receive. We extend this work to also address bias in contributions by leveraging social context as a proxy for user attitudes on controversial topics, and then use this information to generate personalized content for each reader.

VI. Conclusion and Future Work

An open Wiki system often involves the concern of content reliability, which is due to to several factors. First, the knowledge storage in the current popular Wiki systems are

text based, making the knowledge representation too coarse grained. Second, these same systems lack a way for automated systems to query and track contributor interactions, and provide no facetted interface for contributors to enter biographical background, even if they so wish. Finally, these systems present only that knowledge that appears in the last historical version, which may have been vandalized or subtly altered to further an agenda.

Here we evaluate a system that attempts to address these shortcomings. Our ultimate approach is to preserve contributed knowledge as facts in a semantic knowledge store, while taking the intermediate step of breaking up knowledge stored in natural language articles at the paragraph level to increase knowledge granularity. In concert with this, we track social context among contributors, inferring affinity relationships between contributors based on their modifications to articles and their biographical similarity with others. And finally, we combine these factors with a customizable knowledge interface to allow readers to better adjust their trust level with respect to groups of contributors, and thus drive the knowledge presentation layer towards more reliable content from the reader's perspective. Leveraging the social context, our model, TrustWiki, has great potential to reveal editors' social factors, provide credible knowledge, balance conflicting viewpoints, miticate minority opinion suppression and prohibit vandalism.

In our future work, we are interested in the implementation of TrustWiki model as a Facebook application to provide more accurate social context for readers and contributors, as well as provide a means of authenticated feedback on revisions. In concert with this, we would like to integrate the WYSIWYM[18] methodology in order to allow contributors to input knowledge facts in a fine-grained form directly into the knowledge base while simultaneously being able to read the natural language text that will be generated by those facts, given the ontology of our system. Further, Facebook integration allows for direct attribution of these contributed semantic facts to specific contributors, which greatly improves the social context awareness in our system. Meanwhile, we will study different methods to construct social graphs for clustering and test their effectiveness in clustering. We hope we can collect enough users to evaluate our solution, but large scale simulations will also be carried out first to obtain a rough idea of the practical feasibility at scale. We are also interested in applying TrustWiki model to other types of knowledge presentation applications, such as Q&A systems (e.g., Yahoo!Answers) and other collaboratively created knowledge repositories.

VII. Acknowledgment

We thank the anonymous reviewers for their insightful comments. This work was supported by the National Science Foundation FIND (Future Internet Design) program under Grant No. 0832202, GENI, MURI under ARO (Army Research Office), and was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-09-2-0053.

References

[1] B. T. Adler and L. de Alfaro, *A content-driven reputation system for the Wikipedia*, WWW '07: Proceedings of the 16th international conference on World Wide Web, 2007, pp. 261–270.

[2] M. R. Anderberg, *Cluster analysis for applications*, Academic Press, New York :, 1973.

[3] S. Auer, S. Dietzold, and T. Riechert, *OntoWiki - a tool for social, semantic collaboration*, Proceedings of The 5th International Semantic Web Conference (ISWC 2006), Springer, 2006.

[4] A. C. Braet, *Ethos, pathos and logos in aristotle's rhetoric: A re-examination*, HUMANITIES, SOCIAL SCIENCES AND LAW **6** (1992), 307–320.

[5] M. H. Degroot, *Reaching a consensus*, Journal of the American Statistical Association **69** (1974), 118–121.

[6] J. A. Fax, *Optimal and cooperative control of vehicle formations*, PhD thesis in Control and Dynamical Systems, California Institute of Technology, Pasadena, CA, 2001.

[7] J. A. Fax and R. M. Murray, *Information flow and cooperative control of vehicle formations*, IEEE Transactions on Automatic Control **49** (2004), 1465C1476.

[8] H. L. Gates, *Ending the slavery blame-game*, Op-Ed, April 23,2010.

[9] P. Jaccard, *The distribution of the flora in the alpine zone*, New Phytologist **11** (1912), 37–50.

[10] M. Kearns, S. Judd, J. Tan, and J. Wortman, *Behavioral experiments on biased voting in networks*, Proceedings of the National Academy of Sciences, January 2009.

[11] A. Kittur, E. H. Chi, and B. Suh, *What's in wikipedia?: mapping topics and conflict using socially annotated category structure*, Proceedings of the 27th international conference on Human factors in computing systems, CHI '09, 2009.

[12] A. Kittur and R. E. Kraut, *Beyond wikipedia: coordination and conflict in online production groups*, Proceedings of the 2010 ACM conference on Computer supported cooperative work, CSCW '10, 2010.

[13] N. A. Lynch, *Distributed algorithms*, Proceedings of the National Academy of Sciences, Morgan Kaufmann Publishers, Inc., 1997.

[14] J. C. McCroskey and T. J. Young, *Ethos and credibility: The construct and its measurement after three decades*, Central States Speech Journal **32** (1981), 24–34.

[15] G. Mugny and J. A. Perez, *The social psychology of minority influence*, Cambridge University Press, Feb 2010.

[16] C. J. Nemeth and J. L. Kwan, *Minority influence, divergent thinking and the detection of correct solutions*, Journal of Applied Social Psychology **17** (1987), 78–799.

[17] R. Olfati-Saber and R. M. Murray, *Consensus problems in networks of agents with switching topology and Time-Delays*, IEEE Transactions on Automatic Control **49** (2004), 1520–1533.

[18] R. Power, D. Scott, and R. Evans, *What you see is what you meant: direct knowledge editings with natural language feedback*, 13th European Conference on Artificial Intelligence (ECAI 1998), John Wiley and Sons, 1998.

[19] R. Priedhorsky, J. Chen, T. Shyong, K. Panciera, L. Terveen, and J. Riedl, *Creating, destroying, and restoring value in wikipedia*, Proceedings of the 2007 international ACM conference on Supporting group work, GROUP '07, 2007.

[20] R. O. Saber and R. M. Murray, *Consensus protocols for networks of dynamic agents*, Proceedings of the American Control Conference, 2003.

[21] J. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (2000), 888–905.

[22] F. B. Viégas, M. Wattenberg, and K. Dave, *Studying cooperation and conflict between authors with history flow visualizations*, Proceedings of the SIGCHI conference on Human factors in computing systems, CHI '04, 2004.

[23] U. von Luxburg, *A tutorial on spectral clustering*, Statistics and Computing **17** (2007), 395–416.

[24] H. Zhao, S. Ye, P. Bhattacharyya, J. Rowe, K. Gribble, and S. F. Wu, *Socialwiki: Bring order to wiki systems with social context*, Proceedings of The International Conference on Social Informatics (SocInfo), 2010.