Mining Revision History to Assess Trustworthiness of Article Fragments

Honglei Zeng, Maher A. Alhossaini, Richard Fikes and Deborah L. McGuinness
Knowledge Systems, AI Lab, Department of Computer Science, Stanford University
{hlzeng, maherhs, fikes, dlm}@ksl.stanford.edu

Abstract—Wikis are a type of collaborative repository system that enables users to create and edit shared content on the web. The popularity and proliferation of Wikis have created a new set of challenges for trust research because the content in a Wiki can be contributed by a wide variety of users and can change rapidly. Nevertheless, most Wikis lack explicit trust management to help users decide how much they should trust an article or a fragment of an article. In this paper, we investigate the dynamic nature of revisions as we explore ways of utilizing revision history to develop an article fragment trust model. We use our model to compute trustworthiness of articles and article fragments. We also augment Wikis with a trust view layer with which users can visually identify text fragments of an article and view trust values computed by our model.

Index Terms—Mining Revision History, Trust Computation, Trust Visualization, Wiki, Wikipedia

I. INTRODUCTION

“Wikis” are collaborative web sites that allow users to create and revise shared content. The most popular and well known Wiki is Wikipedia1, an online encyclopedia with more than 3.5 million articles in 200 languages (approximately one million articles in English). Wikipedia is also one of the most visited websites according to the Alexa web traffic ranking2.

The popularity and proliferation of Wikis have created a new set of challenges for trust research. Most Wikis are open to the general public and have liberal editing policies that allow users to write and revise their content, thus leaving many ways for creating erroneous information. Additionally, while user rating schemas have worked fairly well in traditional reputation systems, they may not work well in wikis, e.g., an article rated trustworthy may not still be trustworthy if revised. Wikis normally lack explicit trust mechanisms to help users decide how much they should trust an article or a fragment of an article. Trust management has become an important requirement for the continued growth and acceptance of collaboratively developed and maintained information resources.

In our previous work [8], [18], we have developed two methods for computing trustworthiness of articles in collaborative information repositories such as Wikis. Although these methods showed promising results, they addressed document size granularity and did not address issues related to fragments. We consider a fragment of an article to be text in the article that was contributed by one author in one revision. Articles in Wikis typically contain many fragments that should not be trusted equally. Thus, any trust management scheme aimed at practical use will require fragment-size granularity. In addition, fragment granularity can be used to develop more accurate measures of trust.

In this paper, we extend our previous work by developing a revision history-based fragment trust model to compute the trustworthiness of articles and article fragments. We also provide a trust view layer augmentation to wikis. This trust view allows users to visually identify text fragments of an article and view computed trust values.

The rest of our paper is structured as follows. We review our previous work in computing trustworthiness of Wikipedia articles in section 2. In section 3, we discuss fragment formulation and a fragment identification algorithm. In section 4, we develop a Bayesian network for our revision history-based fragment trust model. We describe the experimental results in section 5. In section 6, we show an application of a trust view layer in Wikipedia in which trust information is visually presented to end users. Finally, we discuss related work in section 7 and conclude our paper with a discussion of future work in section 8.

II. OVERVIEW OF OUR PREVIOUS WORK

In the work reported in [8], we described an initial investigation into trust computation and a vision of trust visualization for collaborative information repositories. We developed a method for determining article trust based on the citation relationship among Wikipedia articles. For example, a frequently cited article is usually trustworthy. Nevertheless, our experiments showed that the citation-based trust measure is not an accurate trust measure in isolation; it should work as a component in a solution that utilizes multiple trust computational measures.

Motivated by this observation, we explored in [18] another approach based on the hypothesis that revision history may be used to compute a measure of trustworthiness of articles. For example, one may assume that an article may become more trustworthy if it revised by a trustworthy author. Based on our hypothesis, we developed a revision history-based article trust model for computing and monitoring the trustworthiness of articles. Given the rich and accessible revision information in Wikipedia3, our revision history-based model performed

---

1www.wikipedia.com
3Wikipedia authors have made approximately 41 million revisions, an average of 12 versions per article, over the last four years.
far better than the citation-based model in assessing the trustworthiness of Wikipedia articles.

Figure 1 shows three consecutive versions (i.e., the 23\textsuperscript{rd}, 24\textsuperscript{th} and 25\textsuperscript{th}) of the Wikipedia article U.S. National Forest\footnote{en.wikipedia.org/w/index.php?title=U.S._National_Forest}, contributed by three authors, Willmcw, Jengod and Trivial, respectively. This article had a total of 43 revisions between March 2002 and January 2006. The highlighted content in one version are the modifications to its previous version. For instance, author Willmcw inserted a new sentence in V23 that reads, “The national forest system was created by an act of Congress in 1892.”, which in a later version was revised by another author Trivial to “The national forest system was created by the Land Revision Act of 1891.” It seems that from this example the article U.S. National Forest is more trustworthy with more revisions.

The abstract view of the article trust model that we developed in [18] is shown in Figure 2. This model is represented as a chain of successive article versions, starting with the original version and ending with the last revised version. The chain represents the changes in the article that result from the revisions applied by the authors. We used the article trust model to compute trustworthiness of a revised article based on the trustworthiness of the previous version, the author of the last revision, and the amount of text involved in the last revision.

III. FRAGMENT FORMULATION AND IDENTIFICATION

We consider a fragment of an article to be a collection of text in the article that was contributed by one author in one revision. If an author revised an article several times, each of his or her revisions would form different fragments. Text in a fragment is typically continually located but not necessary in this work for computational reasons that we will discuss in later sections. Because most Wikis do not store individual fragments, fragment formulation is open to many interpretations.

In this paper, a revision on an article refers to the action of editing the article by an author. We consider a revision to be a combination of zero or more delete operations each of which removes some content from an article, and zero or more insert operations, each of which adds some content to an article. When an article is revised, a new version of the article is created to archive the revised content. Thus, the $i$\textsuperscript{th} version of an article is the article after $i$ revisions. We define the original article to be the 0\textsuperscript{th} version of the article. A revision history of an article is a sequence of its versions ordered by their creation time.

A. Constitutes of a fragment

A fragment can be formulated at a word level, a sentence level or a paragraph level.

At a word level, a fragment consists of a collection of words. If an author modifies one or more words in a fragment, a new fragment that comprises just the modified words is created in the revised article. All other words remain in the original fragment. The contribution of that author is thereby limited to the words he or she modified.
At a sentence level, a fragment consists of one or more sentences. If an author modifies one or more words in a fragment, the entire sentences that contain these words are considered modified. Thus, these sentences are removed from the original fragment and the newly inserted sentences form a new fragment.

At a paragraph level, a fragment consists of one or more paragraphs. Even if one word or one sentence is modified in a fragment, the entire paragraph that contains that word or that sentence is considered modified. The ownership of the paragraph changes from the author of the original fragment to the author who revised it. Consequently, this paragraph is removed from the fragment and newly inserted paragraphs form a new fragment.

Ideally, fragment formulation should depend on the context of the revision. For example, if an author is to correct an obvious spelling error in a fragment, then word-level fragment is more appropriate. However, if an author removes a sentence that contains a critical assumption from a fragment, then paragraph-level fragment may be more desirable in that case.

In this paper, we chose the sentence-level fragment formulation as we assume that the semantics of a revision may be interpreted at a sentence level for the most part. Additionally, our decision was based on practical implementation considerations: article fragments at a word level may be too factored to model, while comparing articles on a paragraph basis may not be very helpful in most of cases.

Automated identification of the semantics of revisions is beyond what current technologies can offer, although Natural Language Processing (NLP) techniques may help to a certain degree. For example, a diff program with an embedded NLP parser may produce fragment formulation at a phrase or a clause level. We also propose to extend Wikipedia editing markup language such that an author may explicitly encode fragment identification information during a revision.

### B. Sentence Boundaries Identification

The Wikipedia databases store articles in raw text rather than in individual sentences. Therefore, in order to identify fragments at a sentence level, we need to divide raw text into sentences. The problem of identifying sentence boundaries is efficiently solved with NLP techniques; e.g., Reynar and Ratnaparkhi present a maximum entropy approach in [12]. Nevertheless, our model can tolerate small inaccuracies in dividing sentences; thus, we used a list of sentence-final punctuation marks (e.g., “.” and “?”) to divide text and a set of manually decided rules to handle exceptional cases; for instance, “.” is also used in e-mail addresses.

### C. Fragment Identification

Assume the $i^{th}$ version of an article has $k$ fragments, namely $f_i^0, f_i^1, ..., f_i^{k-1}$. If an author revises this version, he or she first performs deletion operations on all $k$ fragments and then performs an insertion operation to insert new text. Therefore, there are $k + 1$ fragments in the $(i+1)^{th}$ version: $f_{i+1}^0, f_{i+1}^1, ..., f_{i+1}^{k-1}, f_{i+1}^k$, where $f_{i+1}^j$ is the fragment $f_i^j$ in the $i^{th}$ version minus the text that the author removed from it, and $f_{i+1}^k$ is the new fragment comprised of the newly inserted text (note that a fragment could be null). Since the first version (i.e., the $0^{th}$ version) of an article has exactly one fragment, it is easy to verify by induction that the $i^{th}$ version of an article has $i+1$ fragments. Without confusion, a fragment could be empty in two cases: the inserted text is empty (i.e., no insertion was made) or the text in a fragment is subsequently deleted by other authors in later revisions.

While the Wikipedia databases do not store fragment information, it archives complete revisions of articles. Therefore, one approach to fragment identification (FI) is to compare successive article versions, e.g., using diff and identify changes. Assume function diff($x, y$) returns a list of sentences in text $x$ but not in text $y$. Our implementation of diff was based on the well known longest common subsequence algorithm (e.g., [9]); and the differences between article versions were counted at a sentence level.

It is easier to tag every sentence with the number of a fragment to which the sentence belongs than to keep track of the sentences in a certain fragment. Given a revision history $V$ (i.e., $V[0], V[1], ..., V[n]$), the following fragment identification algorithm returns $tag[i][j]$ which is the tag (the fragment number) of the $j^{th}$ sentence of article version $V[i]$.

We consider repeated sentences as different sentences.

Algorithm 1: Fragment Identification

1. $tag[i][j] = -1$ for each $i, j$ // Init
2. for each sentence $c$ in $V[0]$
3.   $tag[0][c] = 0$
4. for $i = 1$ to $n$
5.   for sentence $c$ in $V[i-1]$ and $V[i]$
6.     $tag[i][c] = tag[i-1][c]$
7. for sentence $c$ in diff($V[i], V[i-1]$)
8.   $tag[i][c] = i$
9. return $f$;

### IV. COMPUTING TRUST FROM REVISION HISTORY

We developed a Bayesian network (BN) to model the trustworthiness of article fragments over revisions. The BN is defined by a pair $(B_s, B_o)$ where $B_s$ is the graph structure of the network defined in Figure 3 and $B_o$ is the set of the network’s conditional density functions.

Because both the article trust model and the fragment trust model are based on revision information, the BN we developed here was constructed in a similar setting as in [18]. For example, we used beta distributions for approximating the conditional density functions in BN in both works. Additionally, we define trust values in a continuous range $[0, 1]$, where 0 is the total absence of trust and 1 is the total presence of trust. For example, a trust value of 0.6 of a fragment means 60% content of the fragment is trustworthy. On the other hand, an author with a trust value 0.6 means 60% of the content he or she contributes is considered trustworthy.

The differences between the two models are nevertheless significant. In the fragment trust model, we decomposed the revision chain of the article trust model (in Figure 2) into $n+1$ distinct fragment chains, thus making the computation of fragment trust feasible. In addition, any revision in the article
trust model is applied to the whole article because the article is considered as an indivisible unit. While in the fragment trust model, we know precisely what revision is applied to what fragment. This finer-grained model can overcome some limitations of our previous approach and lead to more accurate trust metrics. For example, we assigned a new trust, called “review trust”, to unchanged fragments on the basis that fragments reviewed but unrevised by independent authors are likely to be trustworthy.

A. Notation

We use the following notation throughout this paper. Consider an article with \( n + 1 \) versions, i.e., \( n \) revisions were made on the original article:

- \( V_i \) refers to the \( i \)-th version of an article which resulted from a revision applied to the previous version \( V_{i-1} \). \( V_0 \) is the original article and \( V_n \) (and \( V_{end} \)) is the final version.
- \( A_i \) refers to the author who revised \( V_{i-1} \). \( A_0 \) is the creator of the original article.
- \( V_i \) consists of \( i + 1 \) fragments, namely \( f_i^0, f_i^1, \ldots, f_i^n \), as depicted in the \( i \)-th column in Figure 3.
- \( D_i^j \) is the text that the author \( A_i \) deleted from \( f_i^j \). Thus, \( f_{i+1}^j = f_i^j - D_i^j \), where the minus symbol is interpreted as a deletion operation. \( D_i^j \) is undefined because \( f_{i+1}^j \) is the last fragment in the \( i \)-th chain.
- If \( x \) is a piece of text, we use \( |x| \) to denote the size (i.e., the number of sentences) of \( x \). We can see that \( |f_{i+1}^j| = |f_i^j| - |D_i^j| \) and \( |V_{end}| = \sum_{j=0}^{n} |f_i^j| \).

B. BN structure \( B_o \)

The BN graph structure \( B_o \) is fully defined in Figure 3. We use \( t_{f_i^j} \) to represent the trust value of a fragment \( f_i^j \), and similarly \( t_{V_i} \) for \( V_i \) and \( t_{A_i} \) for \( A_i \). The nodes in Figure 3 are now viewed as trust variables, e.g., \( t_{f_i^j} \) rather than \( f_i^j \).

Chain \( j \) represents the revision process of fragment \( f_i^j \) which is originally inserted by \( A_j \) during the \( j \)-th revision. Every successive fragment in the same chain, i.e. \( f_i^j \) (where \( j + 1 \leq i \leq n \)), is the remaining text of \( f_i^j \) after being revised, by delete operation only, by the authors \( A_{j+1}, A_{j+2}, \ldots, A_i \). In particular, \( f_i^j \) is the remaining text of \( f_i^j \) in the final version of the article. The length of Chain \( j \) is \( n - j + 1 \) because there are \( n - j \) deletions (which could be null) on fragment \( f_i^j \).

Every pair of nodes \( t_{f_i^j} \) and \( t_{f_i^{j+1}} \) (where \( 0 \leq j \leq i \leq n-1 \)) are connected by an arc from \( t_{f_i^j} \) to \( t_{f_i^{j+1}} \); additionally, there are arcs from \( t_{A_i} \) to every \( t_{f_i^j} \) (where \( 0 \leq j \leq i \leq n \)) and arcs from \( t_{f_i^j} \) (where \( 0 \leq j \leq n \)) to \( t_{V_{end}} \).

In BN, nodes are independent of their non-descendants given their parents. Thus, \( B_o \) is fully characterized by three types of conditional probability density functions \( f(t_{f_i^j}|t_{A_i})^5, f(t_{f_i^j}|t_{f_i^{j+1}}, t_{A_{j+1}}) \) and \( f(t_{V_{end}}|t_{f_0}, t_{f_1}, \ldots, t_{f_n}) \), which will be defined in the following subsections.

C. Beta distributions

When deciding the conditional density functions in our BN, we are faced with the problem of how to handle continuous trust variables. A common approach is to assume normality and model conditional density functions with Gaussian distributions. In this work, we chose beta distributions because trust variables are defined in the range \([0, 1]\) where beta distributions are also defined. Nevertheless, we do not view revisions as independent events, even though beta distribution is normally associated with binomial process. The beta distributions are a family of distributions with two parameters \( \alpha \) and \( \beta \):

\[
\text{beta}(p|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} p^{\alpha-1}(1-p)^{\beta-1},
\]

where \( B(x, y) = \int_0^1 t^{x-1}(1-t)^{y-1} \) dt.

\( t_{f_i^j} \) is only dependent on \( t_{A_i} \) because \( f_i^j \) is completely owned by \( A_j \) per our fragment formulation.

![Fig. 3. Revision History-based Fragment Trust Model.](image-url)
Our choice of beta distributions adds some extra
revisions. Our choice of beta distributions adds some extra
to make a simplifying assumption that the variance \( \sigma \) should be neither too small nor too large when the mean \( \mu \) is close to 0.5. On the other hand, if \( \mu \) is close to 0 or 1, \( \sigma \) does not make much difference in the choices of \( \alpha \) and \( \beta \) because \( \sigma \) is bounded by a very small value \( \mu (1 - \mu) \). Given this assumption and the fact that the mean of \( \text{beta}(\alpha, \beta) \) is \( \mu = \frac{\alpha}{\alpha + \beta} \), we have
\[
\alpha = 10, \beta = \frac{10 - 10\mu}{\mu}; \text{ or }
\alpha = \frac{10\mu}{1 - \mu}, \beta = 10.
\]
(1)

While the actual \( \alpha \) and \( \beta \) values are automatically derived from the \( \mu \) value based on Equation 1 (assuming that either \( \alpha \) or \( \beta \) is 10), it is important to note that the posterior trust distributions are not necessarily beta distributions (thereby their variances are not determined by Equation 1) even though the conditional trust distributions defined below are. Our experiments suggest that a posterior trust distribution typically has a smaller variance as a document goes through more revisions. Our choice of beta distributions adds some extra degree of freedom that we can exploit other than the mean values.

**D.** \( f(t_{f_i} | t_{A_j}) \)

The trustworthiness of a new fragment is only dependent on its author. Since we chose beta distributions to model conditional density functions, we have
\[
f(t_{f_i} | t_{A_j} = a_0) = \text{beta}(p|\alpha_i^j, \beta_i^j)
\]
(2)

For example, if an author is 0.8 trustworthy, then the trustworthiness of the fragment he or she authors could be in a range of 0.7–0.9 with a high probability. The uncertainty in the trustworthiness is caused by the fact that we do not know the exact context in which the author writes the fragment.

**E.** \( f(t_{f_{j+1}} | t_{f_i}, t_{A_{i+1}}) \)

Similarly, we assume
\[
f(t_{f_{j+1}} | t_{f_i} = t, t_{A_{i+1}} = a_{i+1}) = \text{beta}(p|\alpha_{i+1}^j, \beta_{i+1}^j)
\]
(3)

Given Equation 1, we only need to compute \( \mu_{i+1}^j \) to fully determine \( \text{beta}(p|\alpha_{i+1}^j, \beta_{i+1}^j) \). Since \( f_{i+1}^j = f_i^j - D_i^j \), there are two possible scenarios:

1. If \( |D_i^j| = 0 \) fragment \( f_i^j \) is reviewed but unchanged by author \( A_{i+1} \). In this case, we could add a fraction of the trustworthiness of the author, called “review trust”, to the unchanged fragment:\n\[
\mu_{i+1}^j = min(t + a_{i+1} \times \delta, 1)
\]
(4)

We chose \( \delta = 0.01 \) in the Wikipedia experiments.

2. If \( |D_i^j| > 0 \) fragment \( f_i^j \) could be more trustworthy if the deletion made by \( A_{i+1} \) has removed inaccuracies in \( f_i^j \). Since \( t_{f_i} = t \), we consider the \( |D_i^j| \) portion of \( f_i^j \) trustworthy and \( (1 - |D_i^j|) \) portion untrustworthy. Additionally, since \( t_{A_{i+1}} = a_{i+1} \), we assume \( A_{i+1} \) would (incorrectly) remove \((1 - a_{i+1})|D_i^j| \) amount of trustworthy content from \( f_i^j \). In our approach, the trustworthy content of \( f_i^j \) is the trustworthiness of \( f_i^j \) minus the trustworthy content that \( A_{i+1} \) incorrectly removed from \( f_i^j \):
\[
\mu_{i+1}^j = \frac{|D_i^j| - (1 - a_{i+1})|D_i^j|}{|D_i^j|}
\]
(5)

In Equation 5, \( \mu_{i+1}^j \), the mean of \( \text{beta}(p|\alpha_{i+1}^j, \beta_{i+1}^j) \), is the size of the trustworthy portion in \( f_i^j \) divided by the total size of \( f_i^j \). There are two exceptional cases when Equation 5 does not apply:

1. \( (1 - a_{i+1})|D_i^j| > |D_i^j| \) when \( A_{i+1} \) (incorrectly) removes all the trustworthy content in \( f_i^j \). Thus, \( u_{i+1} = 0.6 \)
2. \( a_{i+1}|D_i^j| > (1 - t)|f_i^j| \) when \( A_{i+1} \) (correctly) removes all the untrustworthy content in \( f_i^j \). Thus, \( u_{i+1} = 1.6 \)

**F.** \( f(t_{V_{end}} | t_{f_0}, t_{f_n}, ..., t_{f_k}) \)

We represent the trust of the final version as a deterministic function of the weighted trust of its fragments:
\[
t_{V_{end}} = \frac{\sum_{j=0}^{n} t_{f_j} |f_i^j|}{\sum_{j=0}^{n} |f_i^j|}
\]
(6)

Our BN is complete with both \( B_a \) and \( B_o \) fully defined.

**G. Inferences in the BN**

Our goal in the fragment trust model is to seek the posterior trust distribution of the final version (i.e., \( f(t_{V_{end}}) \)) and the posterior trust distributions of the fragments (i.e., \( t_{f_i} \)), given the revision history. Note that the posterior distributions are not necessarily beta distributions even though the conditional distributions are. The inference computations in our DBN were performed using the freely available software BUGS (Bayesian inference Using Gibbs Sampling [13]).

Our BN contains \( \Theta(n^2) \) nodes, where \( n \) is the number of revisions in the final article version. Wikipedia articles often comprise several hundred revisions, thus making trust computation very expensive. We have taken many measures to
reduce the computational cost. For example, we do not assume that fragment text is continually located and we remove empty chains from the BN. Additionally, each chain in our BN satisfies the Markov property:

$$f(t_{j,i+1} | t_{j,i}, \ldots t_{j}, t_{\Delta i+1}, \ldots, t_{A_i}) = f(t_{j,i+1} | t_{j,i}, t_{\Delta i+1})$$

(7)

While the one time computation of $$t_{j,i}$$ is expensive, once $$t_{j,i}$$ is known, the trustworthiness of fragments in future revisions can be computed incrementally according to Equation 7, thus making the computational cost less of a concern.

V. Experiments

The data set we used to evaluate our approach consists of articles crawled from the English Wikipedia in February 2006. In total we collected revision histories of 1068 articles (i.e., 107, 165 article versions). In order to compare this work to our previous work, we conducted the experiments in a similar setting as in [18].

The articles we collected were classified by Wikipedia into three groups:

featured articles which are considered highly trustworthy in the Wikipedia community because they have been thoroughly reviewed for style, prose, completeness, accuracy and neutrality. Only 0.1% of Wikipedia articles are featured articles;

clean-up articles which are considered untrustworthy because they need major revision to meet Wikipedia’s quality standards. 1.3% of Wikipedia articles are clean-up articles;

normal articles which are the remaining articles that have a great variety of trustworthiness.

Although the exact trust values of Wikipedia articles are not explicitly asserted and are subjective, we can use featured articles and clean-up articles as references to evaluate our model. For example, we expect a featured article to have a high trust value in our model.

Since Wikipedia articles are constantly under revision, even a featured article could become untrustworthy if it has been changed. We found an average of 7.3 mass deletions, a form of vandalism that removes more than 90% content of an article, in our set of featured articles. Similarly, a clean-up article could become trustworthy upon major revision. Therefore, in the data set, we only consider the revision history of a featured article (or a clean-up article) up to the point where it was initially approved by Wikipedia authors, rather than using its complete revision history which includes the most recent article versions whose trustworthiness may be unknown. We use complete revision histories for normal articles.

We synthesized the trust values of the Wikipedia authors based on the background knowledge from Wikipedia because these values are also unavailable. Wikipedia currently supports four levels of authorship: administrators (including stewards and developers), registered authors, anonymous authors, and blocked authors, with the authors at each level having decreasing editing powers and trustworthiness. We use the following beta distributions to approximate the trustworthiness of the authors : beta(p[190, 10]) for administrators, beta(p[23, 10]) for registered authors, beta(p[15, 10]) for anonymous authors, and beta(p[10, 190]) for blocked authors. The means of the distributions are 0.95, 0.7, 0.6, and 0.05 respectively. The exact values are not important for this evaluation since it is the relative differential between the trust values that reflects our trustworthiness assessment of authors’ Wikipedia editing status.

Evaluation of article trust In the rest of this section, we will focus on the means of the posterior trust distributions, which we refer as $$\mu$$.

Table I summarizes the results of $$\mu$$ for featured articles, clean-up articles and normal articles in the fragment trust model.

<table>
<thead>
<tr>
<th></th>
<th>featured articles</th>
<th>clean-up articles</th>
<th>normal articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average of $$\mu$$</td>
<td>0.920</td>
<td>0.717</td>
<td>0.817</td>
</tr>
<tr>
<td>Variance of $$\mu$$</td>
<td>0.002</td>
<td>0.010</td>
<td>0.011</td>
</tr>
</tbody>
</table>

We compared the accuracies of the classification rules developed from the fragment trust model and the article trust model.
We reused the aforementioned 50 featured articles and 50 clean-up articles into two training sets. Each training set contains 100 pairs, \((x_i, y_i)\), where \(x_i\) is the trust value of an article in a trust model, and \(y_i\) is the class of the article. We assume \(y_i = 1\) for featured articles and \(y_i = 2\) for clean-up articles. Since we just use one feature (trust value) to predict a class that an article belongs to, the learned classification rule for featured articles in the fragment trust model is:

\[
x > 0.842
\]

where \(x\) is the trust value of an article. The classification rule for featured articles in the article trust model is:

\[
x > 0.823
\]

A test size of 200 articles (100 featured articles and 100 clean-up articles, none of which is from the Geography category) were then prepared. Table II shows the accuracies of both models in classifying the test data.

<table>
<thead>
<tr>
<th>Classification accuracy</th>
<th>featured articles</th>
<th>clean-up articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fragment trust model</td>
<td>91%</td>
<td>84%</td>
</tr>
<tr>
<td>Article trust model</td>
<td>82%</td>
<td>84%</td>
</tr>
</tbody>
</table>

The fragment trust model substantially improves the classification accuracy of featured articles, though the same accuracy is observed in classifying clean-up articles. We see little improvement in clean-up articles because any author could mark an article as a clean-up article (with a reasonable reason) and hence the untrustworthiness of clean-up articles is less certain. However, it takes a lengthy and rigorous process to approve featured articles; thus, we are highly confident about their trustworthiness.

It is clear that featured articles have far more revisions than clean-up articles and normal articles (726 vs. 56 vs. 27 in the training data set). While it appears that the number of revisions of an article could predict the class of the article, it is not particularly useful in assessing the trustworthiness of a version of an article. For example, a featured article could become untrustworthy if it has been changed despite the fact that the number of revisions is monotonically increasing. While the number of revisions is one of the most important factors in determining the trustworthiness of an article, other factors, such as the trustworthiness of authors, the amount of the revisions, and the order of the revisions are also important, and have been taken into considerations in our fragment trust model.

**Evaluation of fragment trust**  Since the trust value of an article is determined by the trust values of the fragments in that article, strong results produced by our model in assessing the trustworthiness of articles, as shown in the previous experiments, also indicate a good performance of our model at a fragment level. Nevertheless, we are interested in a more direct evaluation of fragment trust. One possible approach is to survey a group of people and manually decide the trustworthiness of article fragments, and then compare the results with the values computed from the fragment trust model. However, manual determination of trust is time-consuming and possibly impractical.

In this work, we developed an evaluation method based on the “survival rates” of sentences. Consider, for example, an article that has 501 versions, \(V_0, V_1, \ldots, V_{500}\). One interesting question is how many sentences in an intermediate version, for example \(V_{100}\), will still be in the final version \(V_{500}\)? Our intuition is that the more trustworthy a sentence is, the more likely it will not be deleted in the subsequent 400 revisions and will “survive” in the final version of the article.

Figure 6 shows the survival rates of sentences with different levels of trust values, where survival rate is considered to be the probability that a sentence in an early version will remain in the final version. We computed survival rates for sentences in four intermediate versions, namely \(V_{0.2n}, V_{0.4n}, V_{0.6n}\) and \(V_{0.8n}\). A data point in Figure 6 is the median of the survival rates of 36 feature articles in our data set that have 400 to 800 versions. The sample size is limited by the number of featured articles with a similar and a sufficient number of versions.

We chose median over mean value because we found large variances in survival rates, even though similar patterns were observed with mean survival rates. While the median survival rate of trustworthy sentences over the entire data set appears
to grow linearly, it is not the case for individual articles. Additionally, there are very few sentences with trust value less than 0.6 in feature articles, e.g., a total of 109 sentences vs. 10587 sentences with trust value over 0.8 in $V_{0.8n}$. The median survival rate for sentences with trust value less than 0.6 is 0 and hence it is not depicted in Figure 6.

The survival rates of sentences in early versions (i.e., $V_{0.2n}$ and $V_{0.4n}$) are usually low, while they are gradually increasing in late versions ($V_{0.6n}$ and $V_{0.8n}$). Moreover, the survival rates of more trustworthy sentences are normally higher than less trustworthy sentences. We used a fraction of the revision history (e.g., $V_{0.8n+1}, V_{0.8n+2}, \ldots, V_{n}$) to evaluate the fragment trust of previous versions (e.g., $V_{0.8n}$); thus, this experiment indicates that our model may be accurate in estimating the trustworthiness of article fragments. Nevertheless, we assumed in this evaluation that a completely trustworthy sentence would have a 100% survival rate. This assumption does not consider different points of view, i.e., that there may not be universal agreement on trustworthiness.

In summary, even though several simplifying assumptions were made in this work, our approach showed strong results and an improvement over our previous work in assessing the trustworthiness of articles and article fragments. Our fragment trust model may be refined and improved in several directions; for example, we expect better performance when the simplifying assumptions in our model are removed.

VI. TRUST VIEW: AN APPLICATION OF THE FRAGMENT TRUST MODEL

In this section, we augment Wikipedia with a trust management component and develop a trust view associated with Wikipedia articles. Figure 7 shows a snapshot of the 25th version of the Wikipedia article U.S. National Forest. A new “trust view” tab is added to the conventional tabs of Wikipedia, e.g., “article”, “edit”, and “history”. When a user clicks the trust view tab, the fragments of the Wikipedia article that the user is viewing are rendered in different colors based on their trustworthiness.

Fragments of higher trustworthiness are displayed in a more trustworthy color than fragments of lower trustworthiness so that users can gain insight on relative trust from just glancing at the trust tab presentation (and color legend) of an article. For example, in Figure 7, the fragment inserted by the author Trivial in the 25th revision (which is colored in green) is considered less trustworthy than the first two paragraphs of the article (in blue) because the latter has already been revised by several independent authors. Moreover, the second fragment in the third paragraph (in orange) is considered untrustworthy because it was written by a blocked author. Nevertheless, issues such as more intuitive mapping from color to trustworthiness are still under investigation.

The benefits of revision trust are far beyond the trust view in Wikipedia. We can build many applications to utilize the available trust information. For example, users may also have the option of viewing the most trustworthy version of an article, in addition to the most recent one. Moreover, revision trust may improve Wikipedia’s quality control process; for example, our model may provide an automated method to monitor changes in trustworthiness and thereby provide timely notifications of vandalism and other forms of malicious content modifications.

VII. RELATED WORK

Wikipedia has taken many measures to address the trust challenges. For example, it recently tightened the privileges of anonymous authors to create new articles. More recently, a new feature called “article validation” is under consideration of anonymous authors to create new articles. More recently, a new feature called “article validation” is under consideration which may allow users to rate an article in an open voting form or a more restricted form. Additionally, the quality of Wikipedia articles among many factors including the number of revisions. Viegás et al. [15] presented a tool that visualized revision history flow and through which they revealed several interesting patterns in Wikipedia. For example, they found that half of mass deletions were reverted within 2 minutes!

There have been extensive studies on web logs mining for different purposes, e.g., Srikant and Yang used access logs to improve website organization [14]; Yang et al. studied web logs to improve GDSF caching policies [17]; Livshits and Zimmermann mined software revision history to identify common error patterns [7]. While article revision history could be considered as a special case of web logs, we have very different goals and approaches in this work. For example, most current web logs mining applications are built on association rules.

Theories and techniques to compute trust in social networks have also been widely studied, e.g., Kamvar et al. [5] presented a reputation system to minimize the impact of malicious peers in p2p networks; Guha [2] discussed the propagation of trust and distrust in social networks like ePinions.com. These approaches are typically based on the transitivity property of trust, i.e., if A trusts B and B trusts C, then A would trusts C (to a certain degree). As aforementioned, simple rating based reputation system may not work well in Wikis.

9 However, manual inspection showed that this author’s insertion here was accurate. Our model may be improved by developing author trust models for modeling complicated author behaviors. For example, a blocked author could make trustworthy contributions in some cases.

10 meta.wikimedia.org/wiki/Article_validation
Wang and Vassileva[16] showed a Bayesian network trust model for building reputation based on recommendations in p2p network. Studies of trust formulation using beta distributions can be found in [4] and [11]. For more background knowledge on Bayesian networks, one may refer to [10] for a thorough discussion. Although we used beta functions to model conditional distributions of continuous variables in our BN, we are also interested in nonparametric density estimations to model trust [3] and are investigating methods to learn distributions from article revision history.

VIII. Conclusions

As Wikis and other collaborative information repository systems grow in popularity and use, issues concerning the trustworthiness of information become increasingly important. In this paper, we have developed a revision history-based fragment trust model to compute and monitor the trustworthiness of articles and article fragments. Our Bay-esian network provided a flexible method to model revision history of article fragments and combine different factors that affect trust. The experimental results with Wikipedia articles showed the veracity of the fragment trust model and a significant improvement over our previous work on the article trust model. In addition, we augmented Wikipedia with a trust view tab that helps users to view and filter information of varying levels of trustworthiness.

Fragment trust is just one component of the complex trust framework that may be built from revision history. In this work, we made a simplifying assumption about the trustworthiness of Wikipedia authors based on their editing privileges. We also observed in the experiments that a blocked author could make trustworthy contributions in some cases. Thus, one direction of future work is to develop author trust models, and in particular, author trust models for specific domains. For example, it may be possible to derive an author’s trustworthiness in the “Wine” domain with over 500 revisions of the Wikipedia article Wine or one’s trustworthiness in “Brandy” with 150 revisions of the article Brandy. Rich revision information may help to infer domain-specific trust in a finer granularity than other approaches.

We treat trust as a global belief about articles and authors. Nevertheless, trust assessments may differ from person to person and even among experts. Another future direction to extend our work is to develop personalized trust models based on revision history; e.g., when a trustworthy author makes no change to an article, he or she expresses a degree of personal belief to that article.

References


