

Wiki Trust Metrics based on Phrasal Analysis

Mark Kramer
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730
+1-781-271-3629
mkramer@mitre.org

Andy Gregorowicz
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730
+1-781-271-8228
andy@mitre.org

Bala Iyer
The MITRE Corporation
202 Burlington Rd.
Bedford, MA 01730
+1-781-271-4620
biyer@mitre.org

ABSTRACT

Wiki users receive very little guidance on the trustworthiness of the information they find. It is difficult for them to determine how long the text in a page has existed, or who originally authored the text. It is also difficult to assess the reliability of authors contributing to a wiki page. In this paper, we create a set of trust indicators and metrics derived from phrasal analysis of the article revision history. These metrics include author attribution, author reputation, expertise ratings, article evolution, and text trustworthiness. We also propose a new technique for collecting and maintaining explicit article ratings across multiple revisions.

Categories and Subject Descriptors

K.4.3 [Computers and Society - Organizational Impacts]: Computer-supported collaborative work; H.3.1 [Information Storage and Retrieval - Content Analysis and Indexing]

General Terms

Algorithms, Measurement, Experimentation

Keywords

Wiki, collaboration, shingling, reputation, authorship, attribution

1. INTRODUCTION

Wikis are becoming an increasingly popular vehicle for creating, storing, and sharing information. Wikis pose new challenges for both authors and readers. Authors are in the unfamiliar situation where their contributions are immediately subject to scrutiny and revision from the entire readership. Readers may find it difficult to ascertain the trustworthiness of the information they find. The resulting openness and fluidity has led educators and professionals to create policies against citing Wikipedia, which do not exist for conventional encyclopedias. Our present goal is to strike a small blow against such fears and policies by providing tools and methodologies to help wiki readers evaluate what they read.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WikiSym 2008, September 8-10, 2008, Porto, Portugal.

Copyright 2008 ACM 978-1-60558-128-3/08/09...\$5.00.

In this paper, we propose a new, multi-faceted approach to trustworthiness. We incorporate both machine-derived and human-derived quality metrics, since each provides a different perspective on the trustworthiness of authors and articles. Human users rate the quality of the text, evaluating aspects such as organization, writing quality, and completeness. Automatic analysis is used to track statistical information derived from the revision history, such as author contributions, text stability, and author reputation. Human judgment is ultimately behind both types of information, since it is human decisions that drive everything in a wiki. However, we know of no data mining technique that will tell us that an article has logical inconsistencies, may be politically slanted, or might benefit from reorganization. Using human and machine analysis together provides a rich set of information to help a reader ascertain the quality of the information in a wiki.

The basis for our approach is phrasal analysis of the article revision history. The revision history contains a wealth of information about contributors and the progression of articles that can be leveraged to help assess article quality. In addition, it provides guidance on how to interpret human ratings information in the context of a document undergoing revision. We begin in Section 3 by decomposing each revision into a set of phrases, and mapping the phrases to revisions. We then present our trust and reputation applications in Section 4. The techniques discussed here can be used individually or collectively.

2. Related Work

Adler and de Alfaro [1] colorize the text of Wikipedia articles by computing a trust value for each word of an article, derived from the computed reputation of the authors of the article. Reputations are determined by a scoring system where an author distributes a number of points to previous authors, depending on whether the current author retains, modifies, or deletes existing text. WikiDashboard [15] displays a list of authors for any Wikipedia article ranked by the number of edits, as well as a timeline of editing activity for any article or author. History Flow [13] is a visualization that plots author contributions across revisions as colored lines, giving insight into author contributions over time. Flagged Revisions (http://en.wikipedia.org/wiki/Wikipedia:Flagged_revisions) is an extension to MediaWiki that allows revisions to be tagged. Tags are arbitrary, but it has been proposed to tag articles based on the level of review they have undergone.

A number of authors have studied the reliability of Wikipedia. Wilkinson and Huberman demonstrate a strong overall correlation between number of edits, number of distinct editors, and article

quality [14]. Chesney compares evaluations of subject matter experts to those of non-experts, and found that 13% of articles contain mistakes [4]. While these and other similar studies help gauge the average or overall reliability of information in wikis, and Wikipedia in particular, they do not help address the trustworthiness of individual articles or authors.

The current work uses a metric of text similarity based on shingling (also called chunking or n-grams) [2] [3]. These techniques operate by passing a sliding window over the document to create word sequences (phrases). Semantic approaches have also been utilized to assess similarity [8]. This is achieved by marrying shingling approaches with WordNet to determine semantic matches to shingles. Other approaches compute the number of character operations necessary to transform one text string into another [10]. Still others use Term Frequency Inverse Document Frequency (TF-IDF) [11], and suffix trees [5].

3. DOCUMENT MODEL

3.1 Phrasal Decomposition

In this work, we model an article as a set of phrases. There are several possible ways to decompose an article into phrases. A syntactic approach would involve parsing the article into units of grammar, and a semantic approach would attempt to decompose the text into units of meaning. In this work, we adopt a simple technique known as shingling. Shingling involves moving a length- N sliding window over the text, one word at a time, to create overlapping phrases consisting of every continuous sequence of N words in the document D . For example, if an article contains the 6-word sequence A B C D E F and the shingle length is 3, the phrases are {ABC, BCD, CDE, DEF}. Shingling is language-independent and robust to complex (and ungrammatical) uses of language. Despite this choice, our techniques are independent of the method of phrasal decomposition, and more sophisticated approaches of phrase generation could be used.

In the scope of a specific article, we give credit to the author who first contributes a unique phrase. Each time the same phrase recurs in a subsequent revision of the same article, the original author is credited. This assumes that the probability of two authors independently creating the same phrase is small. We focus on phrases, as opposed to words or sentences, because (a) it would not make sense to assign authorship credit for each use of a single word, and (b) sentences are relatively fragile, in the sense that a change as tiny as inserting a comma results in a new sentence, which takes authorship credit away from the original author [13].

Under the phrase-set model, two revisions are similar if they contain a large number of common phrases, regardless of the order of the phrases. This model will favor authors who contribute novel phrases, as opposed to editors who rearrange existing phrases. However, even an editor who only rearranges existing sentences is creating some new phrases, by our definition of a phrase.

In implementing the shingling approach, text can be preprocessed to remove stop words and/or wiki markup. Words can also be stemmed. We have found that the trust applications are essentially independent of these decisions, so in this paper, we process the

original wikitext. The same authors rank highest in terms of reputation rankings, regardless of whether we remove stop words.

One adjustable parameter is the choice of the shingle width (N). Other authors have used 4- to 7-word shingles for detecting duplicate documents [16]. For our purposes, N is the smallest number of consecutive words for which an author should receive authorship credit. N should be large enough so the probability for two authors to independently creating the same phrase is small. For the purposes of this paper, we have chosen $N=6$, intuitively and without rigorous study¹.

3.2 Implementation

Our work is based on MediaWiki, although the same approach could be applied to many other wiki frameworks. We added two new tables to MediaWiki database, *Phrase* and *Revision_Phrase*. The *Phrase* table contains the phrase ID (primary key), page ID, the text of the phrase, and a CRC-32 hash of the phrase. The *Revision_Phrase* table contains the phrase ID, revision ID, and offset into the original wikitext. The size of the tables is relatively small, because many phrases are repeated from revision to revision. For example, one page we examined had 714 unique phrases over 19 revisions. Of those phrases, 371 were used in 10 or more revisions.

The decomposition of Wikitext into phrases and storage occurs when a new revision is saved, via the MediaWiki ArticleSave hook. A new entry into the *Phrase* table is created for each new phrase. Existing phrases are detected by comparing the CRC and page ID. Consecutive revisions from the same author are reduced to a single revision from that author². There is also a one-time cost in populating the tables for any existing articles.

The result of this processing is a complete accounting of all phrases associated with a page, and a many-to-many mapping of these phrases to revisions. The author of each phrase, date/time, and other information in the Mediawiki database can be easily determined by dereferencing the revision ID.

4. TRUST APPLICATIONS

4.1 Author Attribution

“Says who?” is a very basic trust question. In conventional writing, authorship is usually clearly attributed. In a wiki, only a complex analysis of the revision history can trace authorship. An average user cannot determine whether the current revision contains any significant contribution from one or more authors found deep in the revision history.

Phrasal analysis can be used to calculate percentage authorship for any revision. Start by considering a single word W in a specific position in a certain revision. Except at the beginning and end of the article, W is part of N shingles S_n , $1 \leq n \leq N$, where N is the shingle length. Denote the set of phrases containing word W in document D by $S_N(W,D)$. There is a one-to-many relationship

¹ This choice gives credit to Lincoln for the phrase “four score and seven years ago” but not to Martin Luther King for saying “I Have a Dream”. Oh well, nothing is perfect.

² Some authors do frequent saves to avoid losing work during long editing sessions.

between a phrase and the revisions in which it occurs. These revisions are not necessarily consecutive, because deleted text can be restored. Denote the set of revisions containing any phrase containing W is $R(S_N(W,D))$, and the earliest revision in this set is R^* . The author of R^* is $A(R^*)$. Word W is credited to the author of the earliest phrase in the document containing W :

$$A(W) = A(R^*)$$

An example of the author attribution process is shown in Fig. 1. In this example, the word “seven” in the text “Four score and seven years ago our fathers...” is contained in four 4-word shingles. The earliest occurrence of the first shingle (“four score and seven”) appears in the fifth revision, attributed to Author C; the earliest occurrence of the third shingle “and seven years ago” appears in the fourth revision, attributed to Author B, etc. Of all the earliest occurrences of the four shingles, the overall earliest is “seven years ago our” which appears in revision 2, whose author is A. Therefore the word “seven” is attributed to author A.

four	score	and	seven				<i>Rev. 5 (Author C)</i>
	score	and	seven	years			<i>Rev. 5 (Author C)</i>
		and	seven	years	ago		<i>Rev. 4 (Author B)</i>
			seven	years	ago	our	<i>Rev. 2 (Author A)</i>

Figure 1. Length-4 shingles and their first occurrences result in attribution of the word “seven” to author A.

To facilitate the authorship calculation, we record the current revision ID whenever a new phrase is created. Finding authorship for each word in an article is implemented using a queue of N phrases containing the target word, finding the first occurrence of any phrase in the queue, then moving to the next word by inserting one phrase into the front of the queue, and dropping the last phrase from the back.

Once authorship has been determined at the word level, aggregate authorship can be rolled up to the sentence, paragraph, section, article, category, or even the wiki level by considering the percentage of words contributed by each author. For example, if a certain author accounts for 2 out of 10 words in a particular sentence, that author is 20% responsible for that sentence.

In our demonstration system, we present the contributing authors in descending order of contribution on a related page. The author name is presented as a link to the corresponding user page, where more information about the author can be found, including our version of the author’s reputation.

4.2 Author Reputation

Using phrasal analysis, it is possible to calculate an author’s reputation along three dimensions: quality, quantity, and value. Quality measures the fitness of the contribution, quantity is how many contributions have been made, and value is a measure of usefulness from the consumer’s point of view. These three reputation dimensions are independent. It is quite different for an author to make a few high-quality contributions, as opposed to a large number of low-quality contributions. Therefore, we calculate three dimension of reputation separately, rather than conflating them into a single overall score.

4.2.1 Quantity Reputation Dimension

Quantity is measured in terms of the number of phrases attributed to an author (essentially equal to the number of words, due to shingling). The count can be limited to the current contribution (counting only phrases that are “alive”) or extended over a specified period of history (considering all phrases contributed over a specified time period, whether currently “living” or “dead”). Given the phrase-revision incidence data in the database, it is straightforward to tally the quantity metric.

4.2.2 Value Reputation Dimension

We measure value in terms of an author’s total readership, measured in effective number of page views. Page views are weighted by the author contribution to the viewed page. More specifically, we count the number of page views corresponding to each revision of each article, and combine this data with the authorship percentage for that revision, to determine the total effective page views for each author for that revision. The total readership is then the sum of the effective page views across all revisions. Although MediaWiki does not currently record page view data per revision, we added a column to the *Revision* table to record the number of page views each time a new revision is created.

Like quantity, the readership can be scoped over a specified period of time to avoid summing over the entire history of the wiki. An author contributing to a highly-viewed page (such the Main Page) can score higher on the value scale than someone with more overall contributions.

4.2.3 Quality Reputation Dimension

Compared to quantity and value, it is less obvious how to define quality. It has been previously observed by Alder and de Alfaro that the lifetime of a text fragment (counted in terms of the number of subsequent edits that the text survives) is a good proxy for the quality of the text fragment [1]. Each time a person edits an article, he or she implicitly approves or disapproves of its contents, and on average, high quality contributions will last longer than low-quality contributions.

Similar to Adler and de Alfaro, we use text survival as a proxy for text quality. However, we propose a much simpler and direct way of calculating reputation. Adler and de Alfaro’s scoring system applies a large penalty if an author’s text is edited by a high-reputation author and a smaller penalty the text is edited by a low-reputation author. Scoring rules involve a number of factors: the difference of reputations between the editor and the editee, the amount of text affected by the edit, and the number of revisions between the original contribution and the edit action. Additional scoring rules account for text rearrangement. The parameters in these rules must be optimized to fit specific wikis. Alder and de Alfaro’s method also appears capable of producing the same reputation score for an author with many low-quality edits, as opposed to one with a few high-quality edits, something we have tried to avoid.

By contrast, our approach is based on a standard data analysis technique, survival analysis. Survival analysis quantifies the expected lifetime of phrases contributed by a given author. The quality metrics considered here are the phrase life expectancy (PLE) and the edit survival rate (ESR). These metrics have the

advantage of being easily interpreted, since they relate directly to the longevity of phrases, and they are pure quality measures, independent of the number of contributions or their value.

To calculate either metric, we first create a Kaplan-Meier Survival Curve [9] for each author. The Kaplan-Meier Curve shows what fraction of the phrases contributed by an author survives after a given number of edits. Importantly, the KM accounts for right-censored data, in this case, phrases that have survived to the current revision. If a phrase is still “alive” in the current revision, we do not know its ultimate lifetime, but we do know it has survived longer than its current “age” (counted in terms of number of revisions, discounted for consecutive edits by the same author). Other phrases are “dead”, i.e., not included in the current revisions, and so have a known lifetime. The KM curve accounts for both “living” and “dead” phrases in determining lifetime. An example is shown in Fig. 2.

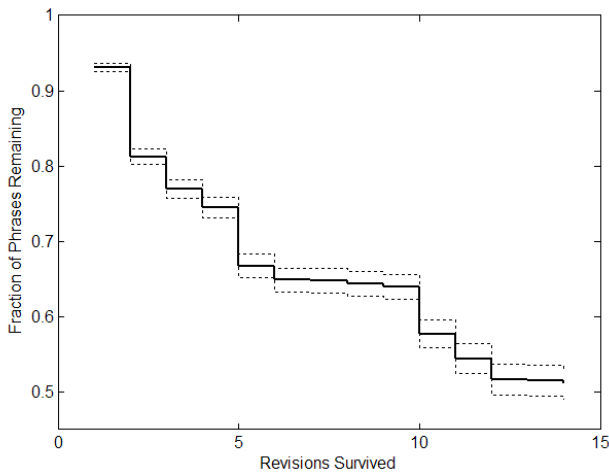


Figure 2. Kaplan-Meier Curve for an author in MITREpedia, with 10-edit survival rate of approximately 60%. Upper and lower confidence limits are shown as dashed lines.

The first step in applying survival analysis is to determine how many revisions the phrase has appeared in, and whether the phrase is currently alive or dead. Since phrases can be resurrected, there is some ambiguity about being dead, but if they do not appear in the current revision, we consider them dead for this purpose. We then calculate the KM curve using the well-known KM algorithm. This calculation can be run off-line, periodically, since reputations change slowly. Confidence limits on the KM curve are calculated using Greenwood’s formula [7]. Typically, the confidence limits tighten as we collect more data on an author, and the author’s reputation becomes more established.

The PLE at a given age τ is determined by integrating the KM curve from τ to the maximum lifespan, and normalizing by the fraction of the phrase population still alive $t = \tau$. Since most authors have some fraction of living phrases, the maximum lifetime is unknown, and we cannot simply integrate the KM curve to infinity. In our corporate wiki, MITREpedia, the KM curves of many authors do not extend below 50% survival. To calculate the PLE, it would be necessary to fit the data with a hazard model, such as exponential decay or Weibull, to extend the curve to 0% survival. Because of the uncertainties involved, we have not implemented the PLE.

The ESR, on the other hand, is a model-free descriptive statistic. Survival rates are often used in medical contexts to express prognoses. For example, a certain type of cancer might have a 10-year survival rate of 60%. Taking this to the wiki domain, we can talk about the K-edit survival rate, defined as the percentage of phrases from an author remaining alive after K edits. Recall that in the database, sequential edits by the same author are reduced to one edit. From experience with our corporate wiki, we have decided that K=10 provides good differentiation between authors. If K is too small, the survival rates for most authors will be high, and the spread between authors will be small. If K is too large, all the survival rates will be low, and again, the spread between authors will be small. K should be chosen so the spread between authors is as large as possible. In our corporate wiki, K=10 provides excellent dispersion, with a range of ESR between 10% and 90% with an average about 60%. We have found that users have little difficulty interpreting the ESR -- it is simply the percentage of an author’s work that has survived 10 subsequent edits.

An additional factor about ESR that is intuitively correct is that novice authors (or previous authors with new identities) do not get reputations right away. They must wait until their contributions undergo the test of time -- until articles they contribute to accumulate at least 10 subsequent edits -- before they get a reputation score.

To help users interpret the three reputation metrics, we divide the authors into quintiles, and award five stars to the top quintile, four stars to the next quintile, and so on. The user can sort the authors by quantity, quality, or readership, search for a given author, or follow a link to detailed data on user pages (Fig. 3).

4.3 Subject-Matter Expertise Ratings

The Wiki Expert Finder is an application we created to help readers find most trustworthy authors in a given subject area. To begin, the user inputs a set of keywords describing the topics of interest. The keywords are mapped to one or more categories or articles using normal search capabilities. The articles may be ranked according to relevance. For each article, we know the authorship percentages, and therefore, we can determine the top contributors associated with the keywords. In addition, for each author thus identified, we have reputation information relating to the quality, quantity, and value of their contributions. Especially in the context of an enterprise wiki, where authors have known real identities and published contact information, the Wiki Expert Finder can direct the reader to one or more knowledgeable individuals.

Author	Author Ratings		
	Quantity (word count)	Quality (edit survival %)	Readership (effective page views)
Carl (contributions)	★★★★★ (14054)	★★★★★ (94.84)	★★★★★ (991)
Thclark (contributions)	★★★★★ (1652)	★★★★★ (92.56)	★★★★★ (211)
Costello (contributions)	★★★★★ (2406)	★★★★★ (88.91)	★★★★★ (1590)
Pforst (contributions)	★★★★★ (12792)	★★★★★ (85.31)	★★★★★ (67575)
Roliver (contributions)	★★★★★ (1595)	★★★★★ (81.64)	★★★★★ (256)
Sring (contributions)	★★★★★ (4336)	★★★★★ (81.23)	★★★★★ (264)
Jprnce (contributions)	★★★★★ (1838)	★★★★★ (80.96)	★★★★★ (1107)
Cosans (contributions)	★★★★★ (1167)	★★★★★ (79.29)	★★★★★ (608)

Figure 3. User reputation page in MITREpedia

4.4 Imputed Text Quality

We can obtain an indicator of text quality through the combination of author quality ratings and author contribution percentages. Since author quality is measured in terms of edit survival rate (ESR), text quality is also measured in ESR. The resulting metric of text quality is a rating that varies between 0 and 1. Again, the assumption is that high-quality text will tend to survive subsequent edits to a greater extent than low-quality text, so a rating of 1 suggests the entire text will survive 10 subsequent edits, while a rating of 0 indicates no part of the text will do so. Articles primarily authored by high-reputation authors will have the highest imputed quality ratings.

When we determine authorship percentages, we also know the revision number when the each phrase was first introduced, and thus we can determine the current age of each word in the text. The future life expectancy of the text depends on its current age. For example, if you were born in 1970, your life expectancy at birth may have been 70 years, but given that you are still living in 2008, your remaining life expectancy might be 42 years, giving an expected total life expectancy of 80 years. We considered measuring text reliability in terms of remaining life span, rather than ESR. However, as we previously observed, calculation of life expectancy requires fitting hazard models to the survival data, so the data can be extrapolated to 0% survival. If we had used this approach, we would calculate the remaining life expectancy of the text, based on its current age, which in theory is a better estimate of future survival. However, we felt the marginal difference in the results did not justify the added work.

Our method of imputing text quality can be applied at the level of articles, sections, paragraphs, or even sentences. As discussed in Section 4.1, we can calculate author contributions to any portion of text, from a single word on up. Therefore, we can determine the imputed text quality for a sentence, paragraph, or section.

Edit wars, if they occur, tend to lower the reputations of the participants. Therefore, text associated with an edit war will have low imputed text quality. Since vandals are unlikely to be high-reputation authors, bad text in vandalized articles also will have low imputed text quality.

For ease of interpretation, we can transform the text quality rating into a five-star rating, by taking a representative sample of imputed text lifetimes across the wiki, sorting and dividing into five quintiles. To display text quality, we can colorize the text, using background shading or font color to indicate levels of reliability, in a manner similar to [1].

In passing, it is worth mentioning the technical challenge involved with superimposing background shading or text color to visualize quality. The analysis performed in this work is being done on the wikitext of an article. MediaWiki transforms the wikitext into HTML for presentation in a browser. When attempting to color sections of text, we know specific locations in the wikitext that should receive coloring; however, it is difficult to ensure this coloring will correctly pass through the transformation into HTML. Text coloring in MediaWiki is typically achieved using HTML span tags. The Wikipedia template provided for this purpose (Template:TextColors) is not reliable because wikitext may contain structure elements, such as tables, which are transformed into HTML table, <tr> and <td> tags. Often, coloring must traverse table boundaries. Therefore, blindly inserting span tags at arbitrary locations in the Wikitext can create invalid HTML. To properly provide colorization to an article, one needs to create an engine that understands Wikitext structure to appropriate break and restart span tags. This work is outside the scope of our research, and probably better addressed with changes in the MediaWiki rendering engine.

4.5 Aging Human Ratings

The imputed text quality is a somewhat indirect way to assess the writing quality and accuracy, since it depends only on author reputation, and not on any specifics of the text. In real life, reputation is not destiny. We do not assume that because George Clooney has an excellent reputation as a movie actor, all movies starring George Clooney are excellent. Instead, we rely on a combination of reputation and film-specific reviews. In addition, our author reputation only captures the opinions of “wiki insiders” (editors), and leaves out the majority of wiki users who are readers only. For these reasons, human ratings are desirable.

Several of our MITRE colleagues have created the Wiki Quality Index (WQI), which allows readers to rate wiki articles in terms of overall quality, objectivity, sources, relevance, consistency, and other aspects [Williamson, Schuler, and LaHousse, personal communication, 2008]. A five-star rating system is used, and any reader can express his or her opinion on any article. Summary statistics are presented to users at the top of the article, with details under a separate quality tab (Fig. 4)

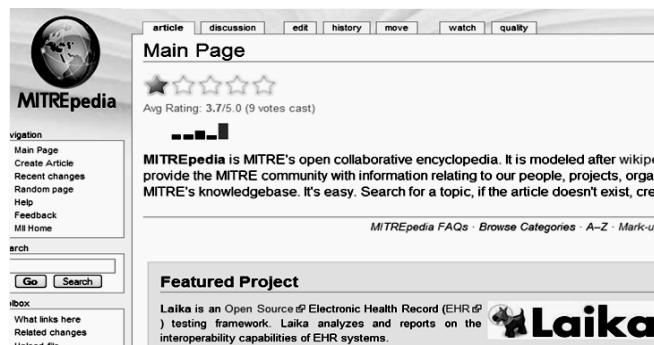


Figure 4. MITREpedia main page, showing average user rating and quality tab.

The difficulty with human ratings in wikis is that wiki articles evolve, and a rating corresponds to a particular (past) revision. This leads to the following dilemma: If a user awards revision A two stars, does this rating still apply to revision B, C, or D? If one

were to discard all existing ratings at each revision, it would be difficult to build up a meaningful number of ratings for each revision. Worse, anyone who didn't like the current ratings could effectively void them by making a minor edit. On the other hand, keeping all ratings forever is also a poor strategy, because a vandalized article could inherit a five-star rating from the previous revision.

Our approach to this problem is to measure the amount of change to the article, and retain old ratings in proportion to the similarity to the current revision. For example, if an article changes by 20% from revision A to revision B, then 80% of the rating of revision A will apply to revision B. On the other hand, if the article changes by 80% from revision A to revision B, only 20% of the rating will apply to revision B.

Clearly, we are making some assumptions here. It is quite possible that an article is vastly improved by the addition of a very small amount of text (i.e., if we plug a major hole), and conversely, the quality could remain unchanged despite large changes. However, there is some intuitive merit to the idea that Moby Dick would still be a classic of American literature if one paragraph were removed, but it would suffer greatly if the character of Captain Ahab were deleted. Overall, we believe revision similarity is a valid means of decaying user ratings over time.

Shingling can be used to measure the similarity between two documents [3, 16]. Each revision is modeled as a set of phrases. The similarity between sets is measured by the Jaccard coefficient, the ratio of intersection to the union of the sets:

$$J(A,B) = |A \cap B| / |A \cup B|$$

where A and B represent the set of phrases in the two revisions we are comparing, and the vertical bars represents the size of the set, in terms of number of elements. The Jaccard coefficient varies from 0, if there are no common phrases, to 1, if the sets are identical.

When a user rates an article, the rating (1 to 5 stars) is stored, along with the user name and revision ID. One vote per user per article is allowed; it is assumed that if a user revisits an article and changes his rating, the old rating is no longer relevant. If there are different aspects (such as completeness, neutrality, writing quality, etc.), then the votes in each aspect are maintained separately. The rating of the current revision is calculated as follows:

1. The similarity of each revision to the current revision is calculated, using the Jaccard coefficient. Let $J_{ik} = J_{ki}$ be the similarity of the i-th revision to the current version, k.
2. Each star-rating category is considered separately. Let V_{ij} be the number of votes for revision i in ratings category j, where (for example), $j = 1, \dots, 5$. The effective number of votes for the current revision in category j is calculated as:

$$V_{kj} = \sum_{i=1}^k J_{ik} V_{ij}$$

3. (Optional) The average rating for revision k is calculated as:

$$\bar{V}_k = \left(\sum_{j=1}^5 j * V_{kj} \right) / \sum_{j=1}^5 V_{kj}$$

The result is a pro-rated user rating that accounts for article evolution.

4.6 Visualizing Article Evolution

Modeling articles as sets of phrases gives us the ability to calculate the distance between any two revisions of an article. The Jaccard distance between two revisions A and B is defined as the number of phrases not shared between A and B, divided by the size of the union of the two sets:

$$J_\delta(A,B) = 1 - J(A,B) = (|A \cup B| - |A \cap B|) / |A \cup B|$$

This is a scaled distance, bounded between 0 and 1. It is also useful to consider the unscaled distance, equivalent to the city block distance:

$$\delta(A,B) = |A \cup B| - |A \cap B| = |A| + |B| - 2 * |A \cap B|$$

Given R revisions of an article, the distance matrix will be a symmetric RxR matrix with zeros on the diagonal (since the distance between a revision and itself is zero). Using the city block distance, we can visualize the evolution of the article over the course of many revisions by plotting the revisions as points in a 2- or 3-dimensional space. To do this, we utilize a dimensionality reduction technique called multidimensional scaling (MDS). The path resulting from connecting sequential revisions is termed the *evolutionary path* of the article.

The proposed technique allows for different distance metrics. One could use edit distance [10] [6], or semantic distance, if it could be measured.

4.6.1 Shape of the Evolutionary Path

Before detailing MDS, we present a few examples to show the shape of the evolutionary path in several simple scenarios.

Article Growth. Suppose there are three revisions, A, B, and C, where each subsequent revision is a superset of the previous revision. Suppose revision A has 20 phrases, revision B has 30 phrases, 20 retained from revision A plus 10 more. In revision C, another 10 new phrases are added, retaining all previous phrases. We can express this in the following phrase origin table:

Revision	Phrase origin		
	A	B	C
A	20	0	0
B	20	10	0
C	20	10	10

Using city-block distance, the corresponding distance matrix is:

$$D = \begin{bmatrix} 0 & 10 & 20 \\ 10 & 0 & 10 \\ 20 & 10 & 0 \end{bmatrix}$$

Since the distance between A and C is the sum of distances between A and B, and B and C, the geometrical representation of the revisions is three points on a line.

Reversion. Suppose that revision B removes all text from A, but revision C reverts the change, returning the article to its original state. In this case, the phrase origin table is:

	Phrase origin		
Revision	A	B	C
A	20	0	0
B	0	20	0
C	20	0	0

In this case, the distance between A and C is zero, so the evolutionary path reverses direction and returns to the previous point. Often, a reversal of the path is a sign of vandalism, since vandalism is usually quickly reverted.

Full Replacement. Suppose that successive revisions fully replace the text, as illustrated in the following table and distance matrix:

	Phrase origin		
Revision	A	B	C
A	20	0	0
B	0	20	0
C	0	0	20

$$D = \begin{bmatrix} 0 & 20 & 20 \\ 20 & 0 & 20 \\ 20 & 20 & 0 \end{bmatrix}$$

Since the distance between points A, B, and C are the same, the evolutionary path consists of two sides of an equilateral triangle, with subtended angle of 60°.

Simultaneous Addition and Removal. Consider the evolution of an article where each revision removes some phrases and adds an equal number of novel phrases. For illustration purposes, assume the phrases are removed in equal proportions with their origins. For example, if revisions A and B originated 20 phrases and 10 phrases respectively, and if 3 phrases are removed in revision C, then 2 will come from A and 1 will come from B. Here is an example of three consecutive revisions with 10% turnover at each revision, showing the number of phrases for each author at each revision, starting with 100 phrases:

	Phrase origin		
Revision	A	B	C
A	100	0	0
B	90	10	0
C	81	9	10

$$D = \begin{bmatrix} 0 & 20 & 38 \\ 20 & 0 & 20 \\ 38 & 20 & 0 \end{bmatrix}$$

The evolutionary path consists of two sides of a triangle of length 20, with hypotenuse of 38. By the law of cosines, the subtended angle is 143.6°.

If the phrase turnover at each edit is greater, then the subtended angle decreases. Here is the same example with 50% turnover at each revision:

	Phrase origin		
Revision	A	B	C
A	100	0	0
B	50	50	0
C	25	25	50

$$D = \begin{bmatrix} 0 & 100 & 150 \\ 100 & 0 & 100 \\ 150 & 100 & 0 \end{bmatrix}$$

The evolutionary path consists of two sides of a triangle, with subtended angle of 97.5°. The general rule of thumb is that the greater the degree of replacement, the more acute the change of direction in the evolutionary path, approaching the limit of 60° for full replacement. Major revisions that change the direction of the article literally change the direction of the evolutionary path.

4.6.2 Multidimensional Scaling

Whenever there are more than three revisions, it may be impossible to geometrically place the revisions in a way that satisfies all mutual distances. For example, it is impossible to place more than three points in a plane equidistant to one another. Multidimensional scaling (MDS) is a method that places points in a low-dimensional space in a way that best preserves the point-to-point distances [12]. MDS minimizes the sum of squared differences between the distances in the reduced-dimensional space and the distances in the original space.

By applying MDS, we can visualize the evolutionary path of any article in one, two, or three dimensions. Using this visualization, it is easy to spot cases of vandalism and reversion (indicated by out-and-back excursions, usually large and sudden), article growth (movement along a smooth curve), and addition-deletion (zig-zag movement). At a glance, a user can determine if an article has been very contentious (a “hairball”), attracted vandals, or shown consistent growth. The user can easily differentiate minor edits (short segments) from major edits (long segments). In addition, since each segment on the evolutionary path is associated with a particular author, the graphical interface can allow the user to select any segment interactively, to see the author, date, and other pertinent information about the edit. We also provide a link from each line segment to the side-by-side comparison of the article, before and after a particular revision.

4.6.3 Examples

Figures 5-8 shows evolutionary paths for four articles from Wikipedia: Evolution, Chocolate, Abortion, and Islam. We have chosen these articles and the time ranges (approx. 2001-2003) to facilitate comparison of the evolutionary path to IBM history flow. The corresponding history flow diagrams can be found in

[13] and exist on the web (http://www.research.ibm.com/visual/projects/history_flow/gallery.htm).

Evolution (Fig. 5) evolved steadily and in a relatively orderly fashion during the time period under study, indicated by the relatively even path length distribution (top). Adding a time dimension (bottom), a helical structure emerges.

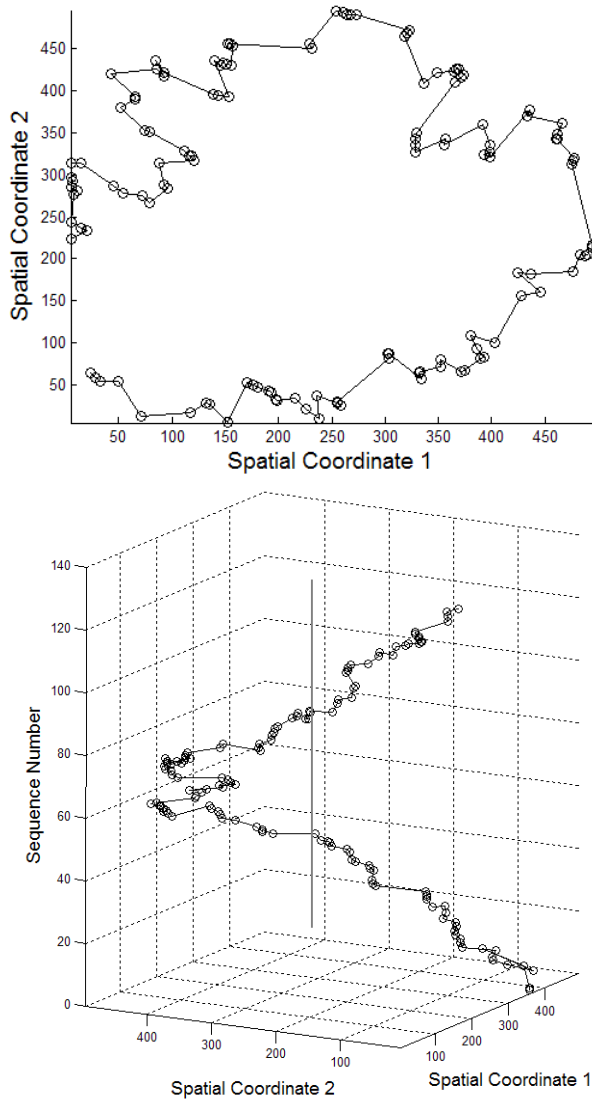


Fig. 5. Evolution, Dec. 2001 - July 2003, shown in two spatial dimensions (above), and with the addition of sequence number dimension (below).

Viégas et al. [13] use Chocolate to illustrate an edit war, which appears as a zig-zag path in history flow. Fig. 6 (top) shows the evolutionary path for Chocolate, in two dimensions. The edit war is not visible, because it involves traveling back and forth between the same two points. With an added dimension representing revision order, the edit war becomes visible (middle, bottom).

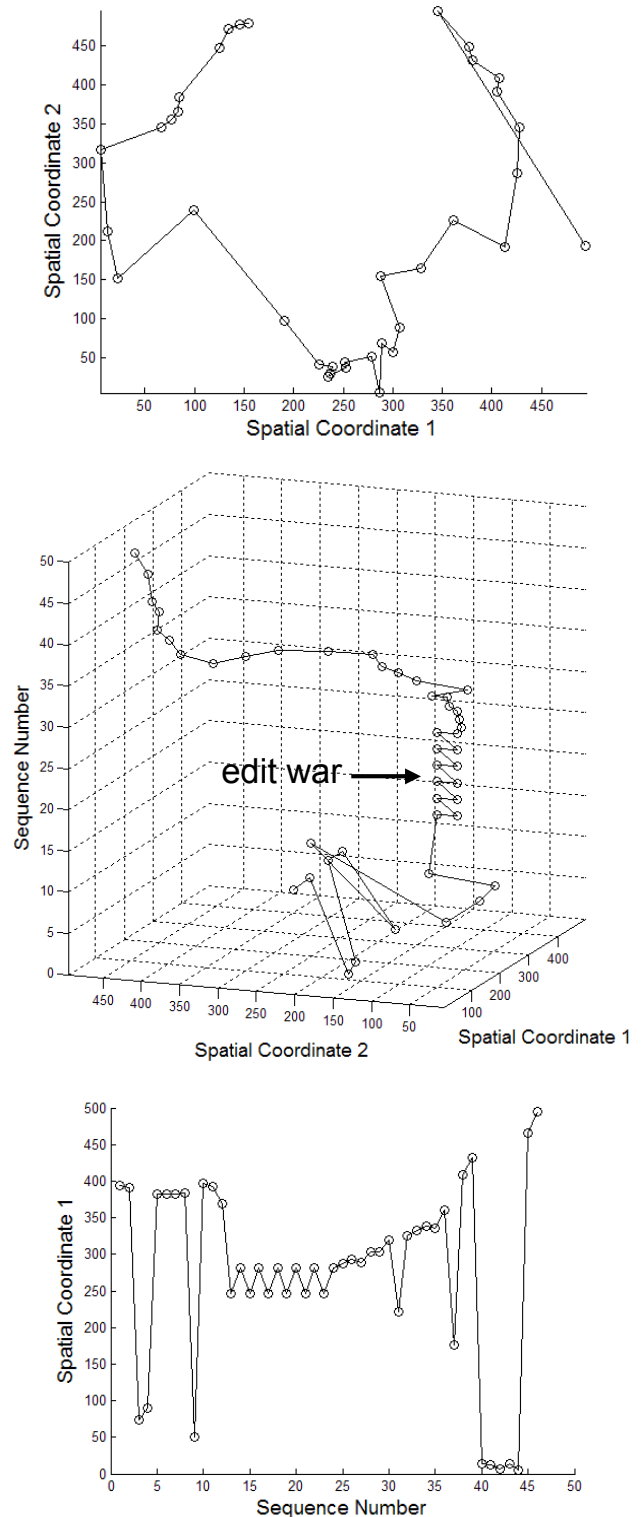


Fig. 6. Chocolate, Dec. 2001 - Aug. 2003, shown in two spatial dimensions (above), and in 2D and 1D with the addition of a sequence number dimension (middle, bottom).

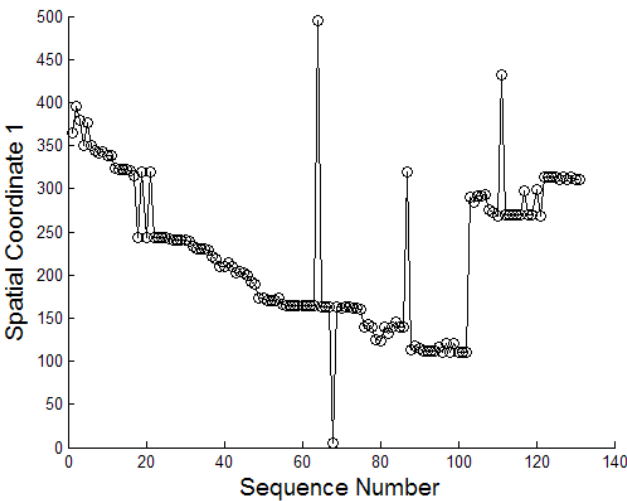
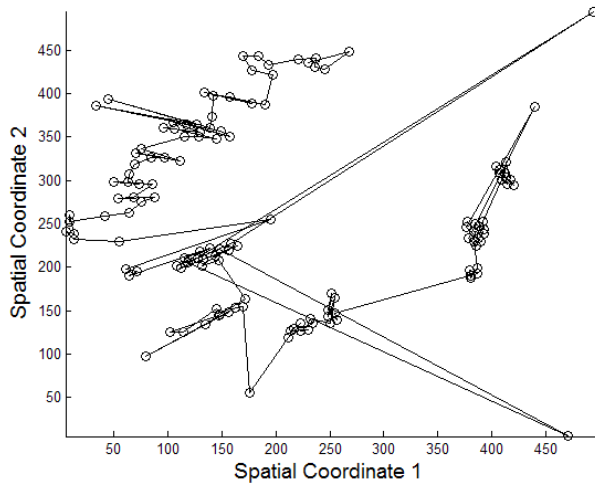


Fig. 7. Abortion, Dec. 2001 - June 2003, shown in two spatial dimensions (above), and in one dimension with the addition of sequence numbers on the x-axis (middle, bottom).

Abortion (Fig. 7) and Islam (Fig. 8) represent controversial topics, which can be seen from the relative disorganization and sudden changes. When projected into one dimension, and presented with a time coordinate, vandalism is evidenced by spikes.

5. DISCUSSION

In this paper, we have shown that modeling wiki articles in terms of phrase sets facilitates a variety of trust and reputation applications. We have discussed six of these applications: author attribution, author reputation, wiki expert finder, text trustworthiness, revision-sensitive handling of human ratings, and visualizing article evolution.

Unlike previous attempts at wiki reputations, our reputation is developed along three dimensions: quality, quantity, and value.

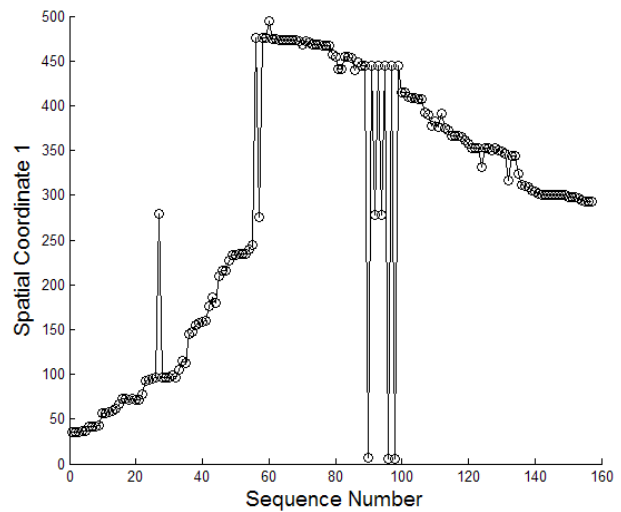
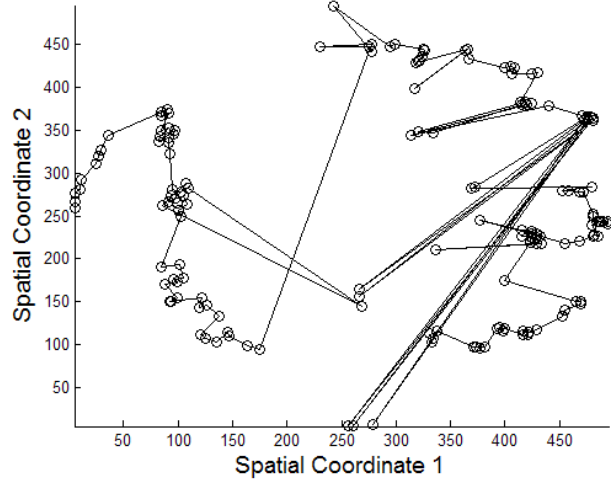


Fig. 8. Islam, Nov. 2001 - June 2003, shown in two spatial dimensions (above), and in one dimension with the addition of sequence numbers on the x-axis (middle, bottom).

Each reputation dimension has a clear interpretation. The quality dimension is a quantitative measure of the survival rate of contributions by the author in question, value is an estimate of the readership of an author in terms of effective page views, and quantity is the contribution volume. This enables the reader to not only compare authors, but also understand the meaning of the reputation metrics.

Reputation information is not only useful to readers, but it also can provide feedback to authors or information stewards, perhaps helping them to determine where to focus their efforts. Reputations can also be used for granting or removal of privileges. However, when reputations are used for gatekeeping, we must be sure reputations have not been manipulated. The topic of creating manipulation-proof and collusion-resistant reputations remains for future study.

One of the more interesting trust-related applications proposed here involves plotting the trajectory of an article, using the

distance between revisions as input. The user can immediately see which revisions constitute large changes, and associate these changes with specific authors. Since sudden changes in direction are associated with removal of text, the reader can also see from a zigzag trajectory, which articles have been controversial or have undergone many changes of direction.

Another point we have emphasized is that automatic analysis of history cannot totally substitute for human rating. Only humans can determine if the article is complete, consistent, neutral, and has adequate citations, among other quality aspects. Since readers typically far outnumber authors, capturing and publishing their feedback can potentially enhance trustworthiness. One reason this has not been done in the past is that the article is undergoing change, and that ratings apply to particular historical revisions. We propose a solution to this problem based on weighting the ratings based on the similarity between the current article and the article when it was rated. This enables article ratings to be carried forward across revisions with appropriate discounting, creating valid ratings for the current version.

6. REFERENCES

- [1] Adler, B.T. and de Alfaro, L. 2007. A Content-Driven Reputation System for the Wikipedia. Proceedings of the 16th International World Wide Web Conference, 261-270, Banff, Canada. ACM Press, 2007.
- [2] Broder, A., Glassman, S., Manasse, M. and Zweig, G. 1997. Syntactic Clustering of the Web. In Proceedings of the 6th International Web Conference.
- [3] Brin, S., Davis, J. and García-Molina, H. 1995. Copy detection mechanisms for digital documents. In Proceedings of the 1995 ACM SIGMOD international Conference on Management of Data (San Jose, California, United States, May 22 - 25, 1995). M. Carey and D. Schneider, Eds. SIGMOD '95. ACM, New York, NY, 398-409.
- [4] Chesney, T. 1006. An empirical examination of Wikipedias credibility. *First Monday*, 11(11), 2006.
- [5] Chim, H. and Deng, X. 2007. A new suffix tree similarity measure for document clustering. In Proceedings of the 16th international Conference on World Wide Web (Banff, Alberta, Canada, May 08 - 12, 2007). WWW '07. ACM, New York, NY, 121-130.
- [6] Cormode, G., and Muthukrishnan, S. 2007. The string edit distance matching problem with moves. *ACM Trans. on Algorithms*. 3, 1 (Feb. 2007).
- [7] Greenwood, M. 1926. The "errors of sampling" of the survivorship tables. Appendix 1. In: Reports on public health and medical subjects. No. 33. London, UK: Her Majesty's Stationery Office.
- [8] Hao, T., Lu, Z., Wang, S., Zou, T., Gu, S., and Wenyin, L. 2008. Categorizing and ranking search engine's results by semantic similarity. In Proceedings of the 2nd international Conference on Ubiquitous information Management and Communication (Suwon, Korea, January 31 - February 01, 2008). ICUIMC '08. ACM, New York, NY, 284-288.
- [9] Kaplan, E.L. and Meier, P. 1958. Nonparametric estimation from incomplete observations, *J Am Stat Assoc* 53(282), 457-481.
- [10] Levenshtein, V.I. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, Vol. 10, p.707.
- [11] Salton, G. and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.* 24, 5 (Aug. 1988), 513-523.
- [12] Schiffman, S. S., Reynolds, M. L., and Young, F. W. 1981. *Introduction to Multidimensional Scaling*. Academic Press, New York.
- [13] Viégas, F.B., Wattenberg, M. and Dave, K. 2004. Studying cooperation and conflict between authors with history flow visualization. Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 575-582 (April, 2004), Vienna, Austria.
- [14] Wilkinson, D.M. and Huberman, B.A. 2007. Assessing the value of cooperation in Wikipedia. *WikiSym '07*. October 2007, Montreal, Canada. ACM. arXiv:cs/070214v1.
- [15] Chi, E., Chi, H. Suh, B. and Kittur, A. 2008. Providing social transparency through visualizations in Wikipedia. *Social Data Analysis Workshop*, April, 2008, Florence, Italy.
- [16] Fetterly, D., Manasse, M., Nojork, M. 2005. Detecting phrase-level duplication on the World Wide Web. *SIGIR '05*, Salvador, Brazil (Aug. 2005).