Semantic Convergence of Wikipedia Articles

Christopher Thomas and Amit P. Sheth
Kno.e.sis Center,
Wright State University,
Dayton OH, USA
{thomas.258, amit.sheth}@wright.edu

Abstract

Social networking, distributed problem solving and human computation have gained high visibility. Wikipedia is a well established service that incorporates aspects of these three fields of research. For this reason it is a good object of study for determining quality of solutions in a social setting that is open, completely distributed, bottom up and not peer reviewed by certified experts. In particular, this paper aims at identifying semantic convergence of Wikipedia articles; the notion that the content of an article stays stable regardless of continuing edits. This could lead to an automatic recommendation of good article *tags but also add to the usability of Wikipedia as a Web Service and to its reliability for information extraction. The methods used and the results obtained in this research can be generalized to other communities that iteratively produce textual content.*

1 Introduction

Wikipedia has come a long way since it was introduced in 2001. Whereas in the early days, citing a Wikipedia article in a scientific paper would arouse suspicion, this practice has become commonplace for references to definitions, biographies, historical events and more recently even current events. Information on Wikipedia should be generally accepted, hence the rule not to publish original research. Many articles are well thought out, went through hundreds and thousands of rounds in iterative community review processes.

Change is fundamental to Wikipedia articles as it is arguably to good quality of information in general. New insights make adjustments necessary. Additionally, with an ever growing community, the number of changes also increases[4]. In fact, the distribution of authors and edits seems to follow a power law distribution[10]. Articles that are considered interesting are edited much more often than others and few authors contribute significantly whereas many edit very few articles. The question we want to answer is that in light of these constant content changes, can we give a prediction of maturity and quality of an article?

Here, we are attempting to find out whether there is a pattern in the life cycle of Wikipedia articles that we can exploit to make predictions about the usability and reliability of newer articles that a) have not been brought to the attention of the community to be labeled as *good articles*, b) do not meet some of the criteria for being considered *good articles*, but meet those related to validity and completeness of content and c) are in an early stage of their lifecycle and have not yet stabilized enough, but the content that is available is factually correct.

In a recent book, James Surowiecki [8] analyzed how large numbers of people are able to solve difficult problems, as long as they are independent, given a good infrastructure and their answers are aggregated in an intelligent manner. Wikipedia can be seen as such a social problem solving environment. In fact, Wikipedia allows authors full independence and the ability to change (almost) every article, regardless of the author's credentials. The underlying assumption is that the community will correct the mistakes of its single members. A recent Nature article found that the amount of factual errors in Wikipedia is not significantly different from those in the Encyclopedia Britannica [3].

We treat Wikipedia as a well developed example of a social problem solving or question answering Web Service. We want to stress that results gained from this analysis can be applied to similar services that use community control and iterative solution development as methodologies.

Questions about quality and completeness of free unstructured text cannot easily be answered computationally, even with a reference model. However, if we adopt some assumptions about the Wikipedia community process that give an indication about the maturing of an article, we can use the article revision history to make claims about the quality of an article.

The remainder of this paper is structured as follows: Chapter 2 addresses the motivation behind this research in more depth, Chapter 6 explores the related work on the analysis of Wikipedia and its use in text mining, clustering, accumulation of background knowledge, etc. Chapter 3 gives insight into the procedures used for the convergence evaluation, Chapter 4 describes the experiment and analyzes the results and chapter 5 applies the analysis to a possible predicition of the extent of future changes to an article. Finally chapter 7 concludes and hints at future directions.

2 Motivation

Recently, many knowledge mining attempts have focused on the Wikipedia data set. Not only its content, but also its structure make it an ideal candidate for fact extraction. Especially the so-called info boxes allow straightforward extraction of triples or triple-like statements. Projects such as DBPedia[1] and YAGO[7] have taken advantage of this structural property. High precision and recall in the extraction made it possible to extract roughly 93 million triples. However, most of the fact extraction algorithms are blind towards the quality of the article. One way would be to only extract from articles that have gone through a review process and have been deemed *good articles* by the Wikipedia community. As of May 16th, 2007, there are 1,393 *good articles* out of a total of 1,783,476, that is 1 in 1280 articles. Furthermore, the criteria for good articles are strict to assure good quality, but some of these criteria are purely based on formal design aspects and do not consider the trustworthiness of the content. One of the criteria for good articles is that it needs to be stable in its current form. This means, no major edits, no reverting back and forth because of vandalism, etc. This work mostly addresses this criterion. We developed means to assess stability of Wikipedia articles in order to automatically judge their maturity. Our hypothesis is, analogous to the one of the Wikipedia community, that good articles need to be semantically stable.

Ideally, we want to extract only true knowledge. Since so far, there is no way to assess truth computationally, unless logically derivable or explicitly stated, we have to resort to more feasible criteria, such as justification. We can say that the information contained in a Wikipedia article is justified, if, after going through the community process of discussion, repeated editing, etc, it has reached a stable state. A desirable computational solution would be one that can assess the reliability of a Wikipedia article computationally by taking advantage of the iterative nature of the evolution of articles. If a stable state is a criterion for being a good article, then it is likely that many stable articles are close to being good. In addition, we want to show that articles can be close to a stable state and assign a stability value that can be cast as a reliability measure.

With such a measure we will be able to:

- assign confidence values to articles
- assign confidence values to extracted facts
- predict (within a margin of error), the time it will take the article to reach a mature state

Such a stability value would need to be seen as a relative value with respect to the article's revision history. In a stable article the *semantic distance* between revision *n* and revision $n+k$, $k \geq 1$, should not exceed a threshold *t*. Naturally, a *true* measure of semantic distance would entail the ability to find a numeric representation of the exact meaning of a document; a function from the document to a point in a vector space would need to be found such that only documents with the exact same meaning have the same point as their representation. The restricted version of this requirement is shown in formula 1.

$$\forall d_i, d_j \, [f(d_i) = f(d_j) \iff d_i = d_j] \qquad (1)$$

In this case only the exact same documents map to the same function values. A better situation would occur if we had some sort of semantic oracle *SO*, that could tell us whether two documents have the same semantic content. This is shown in formula 2

$$\forall d_i, d_j \, [f(d_i) = f(d_j) \iff SO(d_i, d_j) = true] \qquad (2)$$

Given the lack of semantic oracles and the brittleness of many other techniques that take the actual content of documents into account, we decided to represent documents as *TF-IDF* vectors and measure the cosine distance between them as described in more detail in section 3.

In the end, such a distance measure by itself is of little practical use. It needs to be applicable to give an estimate of the future of the article. Hence the question is, given an article revision history of an arbitrary article, what is the likelihood that this article will change significantly in future revisions.

3 Methods

Talking about classifying a document as semantically stable requires the definition of a few terms.

Hypothesis 1 *A document can be seen as being mature, if, despite ongoing changes, it is semantically stable.*

Hypothesis 2 *A document is semantically stable, if, after the k^{th} revision, no significant changes have been made until the current n^{th} revision, with (n-k) > t being above a stability threshold t.*

There are different ways to computationally measure such a semantic stability, all of which can only metaphorically measure the actual stability of the document's meaning. However, statistical methods that transform a document into a vector space have been proven successful in related applications such as clustering, following storylines, etc.

Our approach is based on such a vector space model for computation of semantic distance. A matrix is built for each Wikipedia topic with the rows representing the different revisions in order of the date they were entered and the columns representing words in the articles. We chose a global lexicon for all matrices and a *TF-IDF* representation of the term occurrence that is also based on the global word count rather than a word count for each topic. Stop words as well as very rare occurrences of a word were removed. The vector representation of a revision step is defined as follows: Let $w_1, w_2, ..., w_m$ be the words in the lexicon. The vector for a revision document is the sparse representation of the *TF-IDF* value of the words in the document.

$$\vec{r_i} = \{\textit{TF-IDF}(w_{1,rev_i}), ..., \textit{TF-IDF}(w_{m,rev_i})\} \quad (3)$$

To be able to align the revision histories of different articles, all revisions of one week (see formula 4) were combined to one vector by taking the median of all the revisions of that week (see formula 5).

$$rc_t = \{\vec{r_i} | \text{timestamp of } \vec{r_i} \text{ is after } t \text{ and before } t\text{+one week}\} \quad (4)$$

$$\vec{R_i} = median\,(rc_t) \quad (5)$$

The rationale behind this was that some articles get edited more often than others, but not necessarily producing more stable results. See e.g. [9]. So the intention was to set an arbitrary time frame for one revision-milestone. With this technique, revert wars and complete deletions and following restorations play only a minor role, unless the article is permanently altered as a result of these actions, in which case the revision milestone will reflect this change. It also allows us to align articles more easily. Certainly, there are many facets to the alignment of revision histories of different articles. Instead of taking an arbitrary time frame, an arbitrary number of edits could have been chosen or the milestones themselves could have been determined by following each article's revision history individually and finding points of drastic change versus points of stability. In particular, a page about a current event may mature faster than a historical page. On the other hand, these pages are in a constant flux and their maturity may not be determined by the proposed method which relies on the assumption that mature pages will not experience major edits of their content.

Let $\vec{R_i}$ be the vector-space representation of the i-th revison milestone. The distance between the revision vectors is then determined using a cosine distance measure as defined in formula 6.

$$\cos(X, Y) = \frac{\vec{X} \cdot \vec{Y}}{|X||\dot{Y}|} \quad (6)$$

We considered two different measures to be meaningful in our context:

1. The pairwise distance between revision milestones towards the final edit: $\cos(\vec{R_i}, \vec{R_{i+1}})$
2. The distance between every revision milestone and the final edit: $\cos(\vec{R_i}, \vec{R_n}, 0 \leq i < n)$

One criticism of these measures is that the number of edits that contribute to the revision milestones is not taken into account. We propose the following formula to assess a value that considers the degree of user involvement.

$$Q_i = \cos(\vec{R_i}, \vec{R_{i+1}}) \cdot (\ln(edits(R_i, R_{i+1})) + 1) \quad (7)$$

Since previous research observed a power-law growth of the editing Wikipedia population, the natural logarithm of the number of edits per milestone is used to reflect a normalized interest in the article.

Two distinct data sets were used for this experiment. One with the revision histories of all 1393 articles that had been labeled *good articles* by the Wikipedia community as the reference data set, henceforth referred to as labeled dataset (L). The other data set consisted of 968 random articles with the requirements that there were no stubs, each article had already undergone at least 50 revision milestones and is not in the set of labeled articles. This set will be referred to as the unlabeled dataset (UL).

4 Experiment and Evaluation

We chose the articles that are labeled *good articles* by the Wikipedia community as the data set for testing the hypothesis that Wikipedia articles converge. We will show how it could be corroborated by this test set. The *good articles* are then compared to random Wikipedia articles above a certain length, stubs were not considered, because we assume that these can by definition not yet be classified as *good articles* or do simply not contain enough knowledge to be taken into account. Furthermore, we restricted ourselves to articles that already had an edit history of at least 50 milestones. The figures in this section give different views of the two measures mentioned in the previous section: Pairwise distance between the revisions and absolute distance to the final revision. Figure 1 shows the conversion rate for the labeled data set. We can see that quite early the average of the articles reaches a much more stable state. After about 20 weeks of editing, changes do not affect the complete document any more, but parts are altered. However, until about

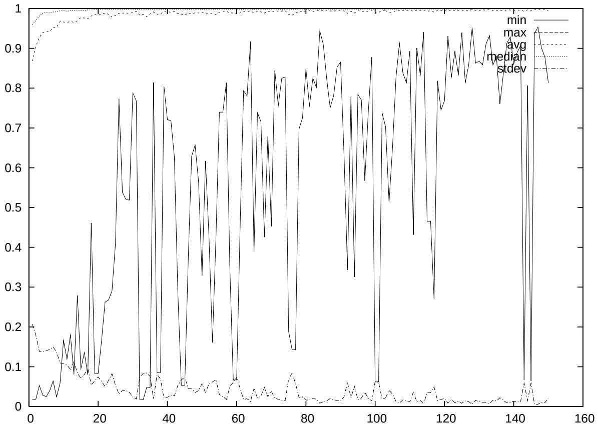

Figure 1. Convergence of the labeled dataset with more than 150 editing milestones.

week 120, this trend is not uniform over the data set, as can be seen in the fluctuating value of the standard deviation. Many articles experience major changes at some point.

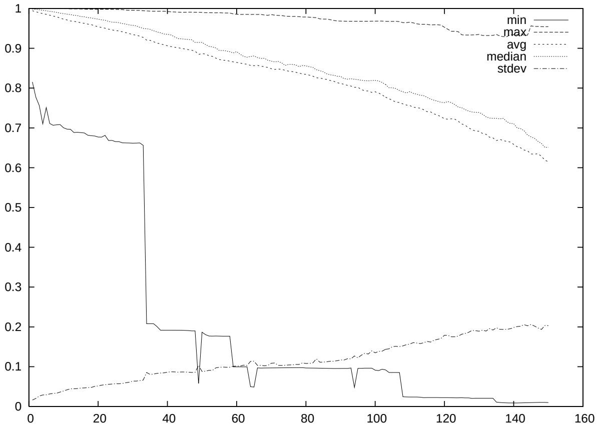

Figure 2. Distance from the final edit of the labeled dataset with more than 150 editing milestones.

Figure 2 shows how the revisions in the labeled dataset work towards the final revision. The leftmost data point represents the final revision, the further away, the less developed the revision is and the lower the cosine distance value. Interesting to see is that the minimum of the cosine distance is quite low until roughly 30 weeks before the final edit. This value indicates that since this point none of the articles have undergone complete revisions.

The following figures depict comparisons of the two data sets.

Figure 3 is the analog to figure 1. It is interesting to see

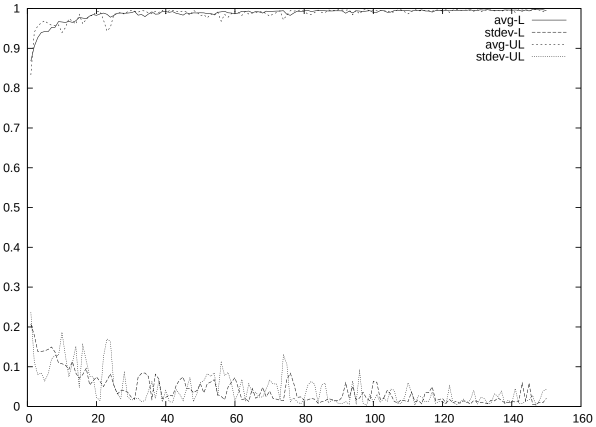

Figure 3. Average and Std-Deviation for both labeled and unlabeled data sets, pairwise comparison.

that the differences between the labeled (L) and the unlabeled (UL) data sets are not significant. This could raise two types of questions. The first is about the validity of the approach. Does a cosine similarity measure between vector space representations of revisions accurately reflect the semantics of change? The other is, accepting the rationale behind the approach, can we generalize the findings in the L-set to the UL-set and deem articles as sufficiently reliable before the community agrees on it?

Figure 4 is similar to figure 2, but aligned at the first edit milestone. It compares the convergence with respect to the so-far final edit of both data sets directly. The dimensions chosen were average and standard deviation. It reflects that articles from both data sets seem to monotonically and linearly strive towards the so-far state. The higher standard deviation of the UL set indicates that this trend is more erratic for the unlabeled documents. However, as in the pairwise comparison, the difference is not significant enough to make a clear distinction between articles in both sets just by looking at their revision history.

Another interesting aspect much of the related work focuses on is the user involvement. Figure 5 uses formula 7 to assess the quality of the articles wrt. the number of edits that contributed to each milestone. This value is more illustrative than meaningful, because it can not empirically be determined how many edits are enough. It shows, however, that user involvement and stability are comparable in both data sets. No significant differences could be found. For the further discussion, it is assumed that enough edits were performed to achieve a revision milestone. User involvement will not be part of the maturity assessment and stability prediction formulae.

An unexpected observation is that many articles in the

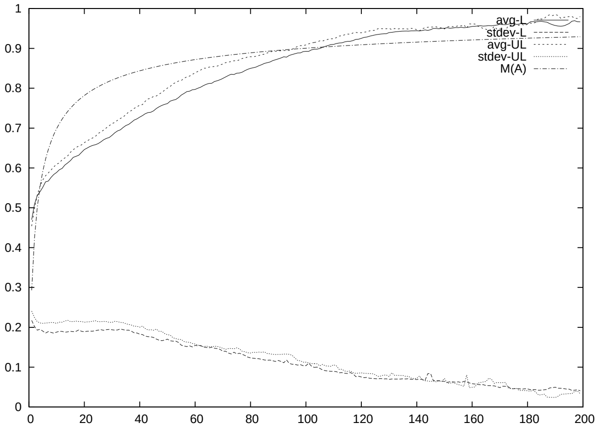

Figure 4. Average and Std-Deviation for both labeled and unlabeled data sets, comparison current to final edit, aligned at first edit milestone.

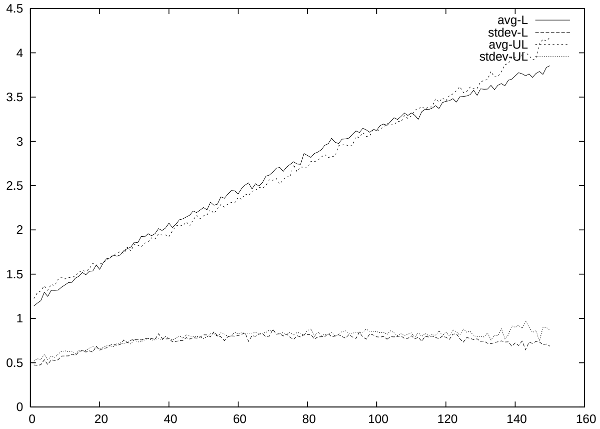

Figure 5. Average and Std-Deviation for both labeled and unlabeled data sets, pairwise comparison, considering edits per milestone.

UL set behave very similar to the ones in the L set. If the premises of our approach are right, it is an indication for their maturity and it gives support for the use of Wikipedia for knowledge extraction.

These well behaved charts tell us that there is a pattern to the maturity of Wikipedia articles. On an average, all articles that have already been edited multiple times tend to stabilize despite growing numbers of edits. The exact pattern will not be as simple as stating that after 30 weeks of editing we can trust an article. A stability prediction function must also take the number of edits that have been made between the revision milestones into account. It is not difficult to have a stable article, if it does not get edited or only edited by a single author. The next section will discuss our approach to maturity prediction.

5 Predicting maturity

Assuming that the community quality control works, our evaluation has shown that articles in general tend to move towards a stable state. It is harder to predict, where in its history towards maturity an arbitrarily chosen article is at a given point in time.

There are at least two dimensions that have to be taken into account for the prediction of the quality of the current edit.

1. the slope of the current stabilization in quality, measured wrt. the distance of the revision milestones

2. the maturity of the article, measured wrt. previous stages of comparable articles that have already matured

We can define the probability of change as the probability that the cosine similarity between two revision milestones is below a given threshold: $\cos(\vec{R_i}, \vec{R_{i+1}}) < 1 - \epsilon$. On an average, this is the number of unchanged articles per revision milestone divided by the total number of articles. The ”p(change)” lines in figure 6 show the graph of this function. Again, this figure basically corroborates the hypothesis of the previous sections. To summarize, all evidence points towards articles getting more stable over the course of their revision history. This becomes even clearer when considering useful patterns that indicate stability. Intuitively, the longer an article is stable, the more likely it should be that it stays stable in the future. Figure 6 shows this for both data sets. The analysis considers the probability of a change in general after *n* revision cycles and after 3 cycles of stability (see the p(change) after 3 lines). The chart confirms the intuitive notion that stability fosters more stability. Articles that have been stable are less likely to change again. In figure 7, we compare the change probabilities in the labeled and unlabeled data sets. In the beginning, More changes happen on average in the labeled set. This could be explained by a higher interest in these topics to begin with. Between 50 and 90 revision milestones, the situation changes and the unlabeled set undergoes more significant changes. This could potentially be caused by the community taking interest in the neglected topics after a core of knowledge has been satisfactorily built. The extent of the difference, however, is very limited. The chart shows more commonalities than differences between the sets. Overall, our analysis shows that there is little difference between the labeled and the unlabeled data sets. Hence, the insights gained from the analysis of the labeled data set can be applied to the articles of the unlabeled set and also to other articles that have already undergone a substantial number of revisions. The maturity of an article can only be estimated with respect to a reference revision time line. In our case, this is the curve given by the labeled data set for the distance between an arbitrary and the final edit. It turns out that this curve can be well approximated by the following formula, where M(a) stands for a maturity measure of the article a and i is the $i^t h$ revision number:

$$M(a) = 1 - \frac{1}{\sqrt{i}} \tag{8}$$

Summarizing, we can say that

1. On average, an article that has shown at least one stable revision milestone is less likely to change significantly. The probability that it will remain stable increases with the number or successive stable revision milestones.
2. formula 8 as shown in figure 4 approximates the maturity curve of the analyzed labeled and unlabeled articles with little error. In addition to the stability estimate, it can give an estimate of the maturity of the article.

While this is statistically correct, it does not hold for a single article. More sophisticated prediction techniques would need to be used to have better accuracy on an individual level. Since single nodes in social networks behave chaotically, we could deploy techniques for prediction of chaotic time series[2].

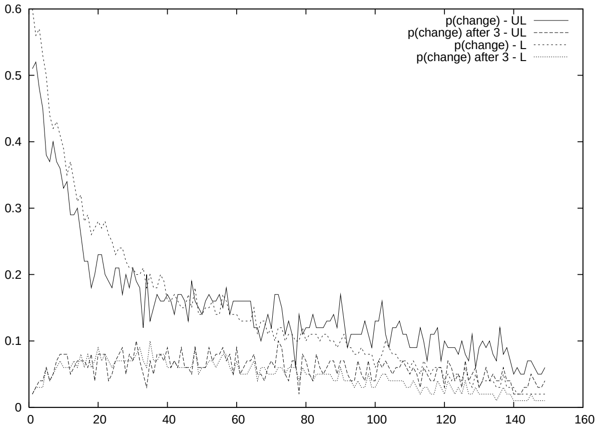

Figure 6. Average change probability for labeled (L) and unlabeled (UL) data set.

6 Related Work

To the best of our knowledge, nobody has taken a vector space representation of Wikipedia revision history into account to assess the quality of articles. In [9], the revision history is used to visualize types of changes made to the article, but without proposing measures for article maturity. In [6], the authors assess seven information quality dimensions. The analysis shows a clear distinction between the reference set of *good articles* and the set of randomly chosen articles. This diverging assessment can be explained by the fact that the authors did not restrict their random data set to articles with a significant edit history. In fact, the average number of edits of the random set was 8, whereas in our case it was at least 50. Voss [10] analyzed user and edit patterns quantitatively, showing that the number of editors and the number of edits follow power law distributions.

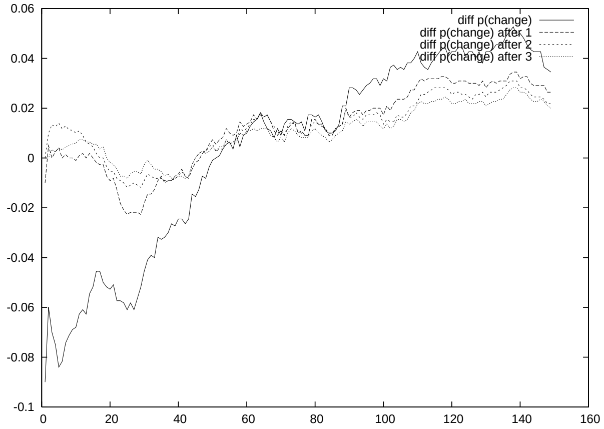

Figure 7. Difference in change probability between both data sets (trend averaged over 10 data points). A positive number indicates that the unlabeled set had a higher probability of change, negative means the labeled set had a higher change probability.

7 Conclusion

This work assumes that the community editing and review process works and will produce justifiable outcome. Based on this assumption we conducted experiments to show the conversion of articles that have been deemed *good articles* by the Wikipedia users. We could show that according to our measurement standards, these articles converge to a stable state. Using the supposed convergence rate as a basis, we compared articles that lack this label with the good articles. We showed that many of these articles exhibit the same behavior as the good ones. Assuming that the methodology for determining stability is correct, we can put high confidence in the correctness of articles that have reached a stable state, regardless of being labeled as good.

Summarizing, we can conclude that

- Articles that have been labeled good tend to have a significant edit history

- There is no statistically significant difference between articles labeled as good and others, given that both have already experienced enough edits

- We were able to predict the current stability and maturity of an article with little error based on its edit history and the insight gained from comparable edit histories.

In future work, we want to explore more sophisticated methods for time series prediction, such as the ones mentioned for chaotic behavior in [2], to make the predictions more accurate. Also, incorporating some of the measures tested in [6] might prove beneficial, especially for articles that have not gone through extensive revision cycles. Lastly, we want to address the above mentioned limitations of the statistical methods by combining the vector space method with an analysis of facts extracted over the revision history using the InfoBox methods, but also relationship extraction from full text, as described in [5]. This will allow us to get closer to a real measure of semantic distance between revisions.

References

[1] S. Auer and J. Lehmann. What have Innsbruck and Leipzig in common? Extracting Semantics from Wiki Content. In E. Franconi, M. Kifer, and W. May, editors, *Proceedings of the European Semantic Web Conference, ESWC2007*, volume 4519 of *Lecture Notes in Computer Science*. Springer-Verlag, July 2007.

[2] M. Casdagli. Nonlinear prediction of chaotic time series. *physicaD*, 35:335–356, 1989.

[3] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, Dec 2005. 10.1038/438900a.

[4] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. symposium. In *Proceedings of the International Symposium on Online Journalism*, 2004.

[5] C. Ramakrishnan, K. Kochut, and A. P. Sheth. A framework for schema-driven relationship discovery from unstructured text. In *International Semantic Web Conference*, pages 583–596, 2006.

[6] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *IQ*, 2005.

[7] F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge - unifying WordNet and Wikipedia. In *Proceedings of the 16th nternational World Wide Web Conference (WWW'07)*, New York, NY, USA, 2007. ACM Press.

[8] J. Surowiecki. *The Wisdom of Crowds*. Anchor, August 2005.

[9] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, New York, NY, USA, 2004. ACM Press.

[10] J. Voss. Measuring wikipedia. In *Proceedings International Conference of the International Society for Scientometrics and Informetrics*, Stockholm, Sweden, 2005.