Computational Trust in Web Content Quality: A Comparative Evalutation on the Wikipedia Project

Pierpaolo Dondio and Stephen Barrett
Trinity College Dublin, School of Computer Science and Statistics, Dublin, Ireland
E-mail: {dondiop, stepehn.barrett}@cs.tcd.ie

Keywords: computational trust, Wikipedia, content-quality

Received: March 17, 2007

The problem of identifying useful and trustworthy information on the World Wide Web is becoming increasingly acute as new tools such as wikis and blogs simplify and democratize publication. It is not hard to predict that in the future the direct reliance on this material will expand and the problem of evaluating the trustworthiness of this kind of content become crucial. The Wikipedia project represents the most successful and discussed example of such online resources. In this paper we present a method to predict Wikipedia articles trustworthiness based on computational trust techniques and a deep domain-specific analysis. Our assumption is that a deeper understanding of what in general defines high-standard and expertise in domains related to Wikipedia – i.e. content quality in a collaborative environment – mapped onto Wikipedia elements would lead to a complete set of mechanisms to sustain trust in Wikipedia context. We present a series of experiment. The first is a study-case over a specific category of articles; the second is an evaluation over 8 000 articles representing 65% of the overall Wikipedia editing activity. We report encouraging results on the automated evaluation of Wikipedia content using our domain-specific expertise method. Finally, in order to appraise the value added by using domain-specific expertise, we compare our results with the ones obtained with a pre-processed cluster analysis, where complex expertise is mostly replaced by training and automatic classification of common features.

Povzetek: Ocenjena je stopnja zaupanja v strani v Wikipediji.

1 Introduction

In the famous 1996 article *Today's WWW, Tomorrow's MMM: The specter of multimedia mediocrity* [1] Cioleck predicted a seriously negative future for online content quality by describing the World Wide Web (WWW) as *"a nebulous, ever-changing multitude of computer sites that house continually changing chunks of multimedia information, the global sum of the uncoordinated activities of several hundreds of thousands of people"*. Thus, the WWW may come to be known as the MMM (MultiMedia Mediocrity). Despite this vision, it is not hard to predict that the potential and the growth of the Web as a source of information and knowledge will increase rapidly. The Wikipedia project, started in January 2001, represents one of the most successful and discussed example of such phenomenon, an example of *collective knowledge*, a concept that is often lauded as the next step toward truth in online media. Wikipedia is a global online encyclopaedia, entirely written collaboratively by an open community of users, it now supports one million registered user, delivers 900.000 articles in its English version alone, and it is one of the ten most visited web sites.

On one hand, recent exceptional cases have brought to the attention the question of Wikipedia trustworthiness. In an article published on the 29th of November in USA Today [2], Seigenthaler, a former administrative assistant to Robert Kennedy, wrote about his anguish after learning about a false Wikipedia entry that listed him as having been briefly suspected of involvement in the assassinations of both John Kennedy and Robert Kennedy. The 78-year-old Seigenthaler got Wikipedia founder Jimmy Wales to delete the defamatory information in October. Unfortunately, that was four months after the original posting. The news was further proof that Wikipedia has no accountability and no place in the world of serious information gathering [2].

On the other hand, Wikipedia is not only being negatively discussed. In December 2005, a detailed analysis carried out by the magazine Nature [3] compared the accuracy of Wikipedia against the Encyclopaedia Britannica. Nature identified a set of 42 articles, covering a broad range of scientific disciplines, and sent them to relevant experts for peer review. The results are encouraging: the investigation suggests that Britannica's advantage may not be great, at least when it comes to science entries. The difference in accuracy was not particularly great: the average science entry in Wikipedia contained around four inaccuracies; Britannica, about three. Reviewers also found many factual errors, omissions or misleading statements: 162 and 123 in Wikipedia and Britannica respectively.

This paper seeks to face the problem of the trustworthiness of Wikipedia by using a computational trust approach; our goal is to set up an automatic and transparent mechanism able to estimate the trustworthiness of Wikipedia articles. In the next section 2 we review related work on trust and content quality issues; in section 3 we argue that, due to the fast changing nature of articles, it is difficult to apply the trust approaches proposed in related work. In section 4 this discussion will lead us to introduce our domain-specific approach, that starts from an in-depth analysis of content quality and collaborative editing domains to give us a better understanding of what can support trust in these two Wikipedia related fields. In section 5 we map conclusions of the previous section onto elements extracted directly from Wikipedia in order to define a new set of sources of trust evidence. In section 6 we present our evaluation conducted trough three different experiments. The first is a study case over 250 articles from the single category "country of the world"; the second is an extension conducted over almost 8,000 Wikipedia. In the third experiment we perform a cluster analysis to isolate article of low and great quality and we compare the results obtained with this implicit approach to the previous one based explicitly on domain expertise. Section 7 will collect our conclusions and future work.

2 Related Works

There are many definitions of the human notion trust in a wide range of domains from sociology, psychology to political and business science, and these definitions may even change when the application domains change. For example, Romano's definition tries to encompass the previous work in all these domains: "trust is a subjective assessment of another's influence in terms of the extent of one's perceptions about the quality and significance of another's impact over one's outcomes in a given situation, such that one's expectation of, openness to, and inclination toward such influence provide a sense of control over the potential outcomes of the situation."[4].

However, the terms trust/trusted/trustworthy, which appear in the traditional computer security literature, are not grounded on social science and often correspond to an implicit element of trust. Blaze et al [5] first introduced "decentralized trust management" to separate trust management from applications. PolicyMaker [6] introduced the fundamental concepts of policy, credential, and trust relationship. Terzis et al. [7] have argued that the model of trust management [5,6] still relies on an implicit notion of trust because it only describes "a way of exploiting established trust relationships for distributed security policy management without determining how these relationships are formed".

Computational trust was first defined by S. Marsh [8], as a new technique able to make agents less vulnerable in their behaviour in a computing world that appears to be malicious rather than cooperative, and thus to allow interaction and cooperation where previously there could be none. A computed trust value in an entity may be seen as the digital representation of the trustworthiness or level of trust in the entity under consideration. The EU project SECURE [9] represents an example of a trust engine that uses evidence to compute trust values in entities and corresponds to evidence-based trust management systems. Evidence encompasses outcome observations, recommendations and reputation. Depending on the application domain, a few types of evidence may be more weighted in the computation than other types. When recommendations are used, a social network can be reconstructed. Golbeck [10] studied the problem of propagating trust value in social networks, by proposing an extension of the FOAF vocabulary [11] and algorithms to propagate trust values estimated by users rather than computed based on a clear count of pieces of evidence. Recently, even new types of evidence have been proposed to compute trust values. For example, Ziegler and Golbeck [12] studied interesting correlation between similarity and trust among social network users: there is indication that similarity may be evidence of trust. In SECURE, evidence is used to select which trust profile should be given to an entity. Thus similar evidence should lead to similar profile selection. However, once again, as for human set trust value, it is difficult to clearly estimate people similarity based on a clear count of pieces of evidence. However, the whole SECURE framework may not be generic enough to be used with abstract or complex new types of trust evidence. In fact, in this paper, we extracted a few types of evidence present in Wikipedia (detailed in the next sections) that did not fit well with the SECURE framework and we had to build our own computational engine.

We think that our approach to deeply study the domain of application and then extract the types of trust evidence from the domain is related to the approach done in expert systems where the knowledge engineer interacts with an expert in the domain to acquire the needed knowledge to build the expert system for the application domain. In this paper, we focus on trust computation for content quality and Bucher [14] clearly motivates our contribution in this paper because he argues that on the Internet "we no longer have an expert system to which we can assign management of information quality".

We finish this section by two last computational projects related to content quality in a decentralised publishing system. Huang and Fox in [15] propose a metadata-based approach to determine the origin and validity of information on the Web.

3 The problem of Wikipedia Articles Trustworthiness and our method

Wikipedia shows intrinsic characteristics that make the utilization of trust solutions challenging. The main feature of Wikipedia, appointed as one of its strongest attribute, is the speed at which it can be updated. The most visited and edited articles reach an average editing rate of 50 modifications per day, while articles related to recent news can reach the number of hundreds of modifications. This aspect affects the validity of several trust techniques.

Human-based trust tools like feedback and recommendation systems require time to work properly, suffering from a well know ramp-up problem [16]. This is a hypothesis that clashes with Wikipedia, where pages change rapidly and recommendations could dramatically lose meaning. Moreover, the growing numbers of articles and their increasing fragmentation require an increasing number of ratings to keep recommendations significant.

Past-evidence trust paradigm relies on the hypothesis that the trustor entity has enough past interactions with the trustee to collect significant evidence. In Wikipedia the fact that past versions of a page are not relevant for assessing present trustworthiness and the changing nature of articles makes it difficult to compute trust values based on past evidences. In general, user past-experience with a Web site is only at 14th position among the criteria used to assess the quality of a Web site with an incidence of 4.6% [17]. We conclude that a mechanism to evaluate articles trustworthiness relying exclusively on their present state is required.

Our method starts from the assumption that a deeper r understanding of the domains involved in Wikipedia, namely the content quality domain and the collaborative editing domain, will help us to identify trust evidence we required to set up an automatic trust computation. The procedure we followed can be summarized in a 4-stage process.

We begin by modelling the application under analysis (i.e. Wikipedia). The output of the *modelling Phase* should be a complete model showing the entities involved, their relationships, the properties and methods for interacting: here we will find out trust dynamics. It is also necessary to produce a valid theory of domain-compatible trust, which is a set of assertions about what behaviours should be considered trustworthy in that domain. This phase, referred as *theories analyser*, is concerned with the preparation of a theoretical trust model reasonable for that domain. To reach this goal a knowledge-based analysis is done to incorporate general theories of Trust, whose applicability in that domain must be studied, joined with peculiar domain-theories that are considered a good description of high-quality and trustworthy output in that domain.

The output is a domain compatible trust theory that acts like a sieve we apply to the application model in order to extract elements useful to support trust computations. This mapping between application model and domain-specific trust theory is referred as *trust identifier*. These elements, opportunely combined, will be the evidence used for the next phase, our trust computation. The more an entity (a Wikipedia page) shows properties linked to these proven domain-specific theories, the more is trustworthy. In this sense, our method is an evidence-based methodology where evidences are gathered using domain related theories.

In other words, after understanding what brings trust in those domains, we mapped these sources of evidence into Wikipedia elements that we previously isolated by defining a detailed model of the application. This resulting new set of pieces of evidence, extracted directly from Wikipedia, allow us to compute trust, since it relies on proven domains' expertise. In the next three paragraphs we will apply our method: *theories analyzer* phase (section 4), *modelling phase* and *trust identifier* (section 5) and our expertise-based *trust computation* in the evaluation section.

4 Wikipedia Domain Analysis

In this section we identify a trust theory derived from domain-specific expertise relevant to Wikipedia, the *theories analyzer* phase of our method. Wikipedia is a combination of two relevant areas involved in Wikipedia: the *content quality* domain and *collaborative editing* domains In this section, we analyse what can bring high quality in these two domains. The quality of online content is a critical problem faced by many institutions. Alexander [18] underlines how information quality is a slippery subject, but it proposes hallmark of what is consistently good information. He identified three basic requirements: objectivity, completeness and pluralism. The first requirement guarantees that the information is unbiased, the second assesses that the information should not be incomplete, the third stresses the importance of avoiding situations in which information is restricted to a particular viewpoint. University of Berkeley proposes a practical evaluation method [19] that stresses the importance of considering authorship, timeliness, accuracy, permanence and presentation. Authorship stresses the importance of collecting information on the authors of the information, accuracy deals with how the information can be considered good, reviewed, well referenced and if it is comparable to similar other Web content, in order to check if it is compliant to a standard. Timeliness considers how the information has changed during time: its date of creation, its currency and the rate of its update; permanence stresses how the information is transitory or stable.

In a study already cited [17], presentation resulted in the most important evaluation criterion with an incidence of 46%. The Persuasive Technology Lab has been running the Stanford Web Credibility Research since 1997 to identify which are the sources of credibility and expertise in Web content. Among the most well-known results are the ten guidelines for Web credibility [20], compiled to summarize what brings credibility and trust in a Web site. The guidelines confirm what we described so far and again they emphasize the importance of the non anonymity of the authors, the presence of references, the importance of the layout, the constant updating and they underline how typographical errors and broken links, no matter how small they could be, strongly decrease trust and represent evidence of lack of accuracy.

Beside content quality domain, Wikipedia cannot be understood if we do not take into consideration that it is done entirely in a collaborative way. Researches in collaborative working [21] help us to define a particular behaviour strongly involved in Wikipedia dynamics, the balance in the editing process. A collaborative environment is more effective when there is a kind of emerging leadership among the group; the leadership is able to give a direction to the editing process and avoid

fragmentation of the information provided. Anyway, this leadership should not be represented by one or two single users to avoid the risk of lack of pluralism and the loss of collaborative benefits like merging different expertises and points of view. We summarize our analysis with the prepositions shown in table 1: in the first column are theoretical propositions affecting trust, second column lists the domains from which each preposition was taken.

Preposition 1 covers the authorship problem. Preposition 2 derives from the accuracy issues. Preposition 3, 4 and 5 underline the importance that the article should have a sense of unity, even if written by more than one author. Preposition 7 underlines the fact that a good article is constantly controlled and reviewed by a reasonable high number of authors. Preposition 8 stresses the stability of the article: a stable text means that it is well accepted, it reached a consensus among the authors and its content is almost complete. Preposition 9 emphasizes the risk, especially for historical or political issues, that different authors may express personal opinions instead of facts, leading to a subjective article or controversial disputes among users. In order to have meaning, these prepositions need to be considered together with their interrelationships along with some conditions. For example, the length of an article needs to be evaluated in relation to the popularity and importance of its subjects, to understand if the article is too short, superficial or too detailed; the stability of an article has no meaning if the article is rarely edited, since it could be stable because it is not taken in consideration rather than because it is complete.

Table 1. A Trust domain-compatible theory. CQ is Content Quality domain and CE is Collaborative Editing domain.

	Propositions about Trustworthiness of articles (T). T increases if the article…	Domain of origin
1	was written by expert and identifiable authors	CQ
2	has similar features or it is complaint to a standard in its category	CQ
3	there is a clear leadership/direction in the group directing the editing process and acting like a reference	CE
4	there is no dictatorship effect, which means that most of the editing reflects one person's view.	CQ/CE
5	the fragmentation of the contributions is limited: there is more cohesion than dissonance among authors	CE
6	has good balance among its sections, the right degree of details, it contains images if needed, it has a varied sentence structure, rhythm and length	CQ
7	is constantly visited and reviewed by authors	CQ
8	is stable	CQ
9	use a neutral point of view	CQ
10	the article is well referenced	CQ

5 Mapping Theories onto Wikipedia

In this section we produce a model of Wikipedia and we map over the model the domain-specific trust we identified in the previous section. We first need a model of Wikipedia in order to extract elements useful for our purpose. Wikipedia has been designed so that any past modification, along with information about the editor, is accessible. This transparency, that by itself gives an implicit sense of trust, allows us to collect all the information and elements needed.

Our Wikipedia model is composed of two principal objects (Wiki Article and Wiki User) and a number of supporting objects, as depicted in fig. 1. Since each user has a personal page, user can be treated as an article with some editing methods like creating, modifying and deleting article or uploading images. An article contains the main text page (class wiki page) and the talk page, where users can add comments and judgments on the article. Wiki pages include properties such as its length, a count of the number of sections, images, external links, notes, and references. Each page has a history page associated, containing a complete list of all modifications. A modification contains information on User, date and time and article text version.

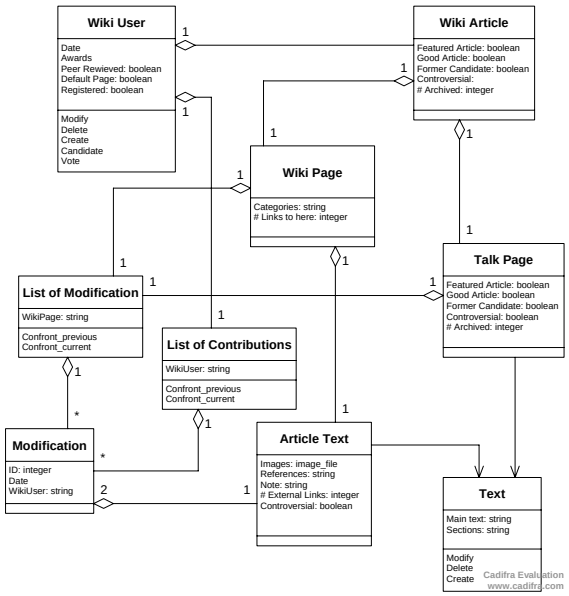

Figure 1 The Wikipedia UML model

The community of users can modify articles or adding discussion on article's topic (the talk page for that article).

We are now ready to map the proposition listed in table 1 onto elements of our Wikipedia model. We remind that the output of this phase will be a set of trust evidence to be used in our trust computation. In general, computing trust using a domain-specific analysis means to aggregate some elements of the application into formulae, that in general maybe be not intuitive and elaborated, in order to model more accurately as possible expert conclusions.

By mapping the conclusions achieved in section 4 - the ten propositions - over our model we identified about 50 sources of trust evidence classified in 6 macro-areas:

Quality of User, *User Distribution and Leadership*, *Stability*, *Controllability*, *Quality of editing* and *Importance of an article*. We now analyses as an example two of the six macro-areas.

5.1 User's Distribution/Leadership (p. 3,9)

Given an article *w* in the set *W* of all Wikipedia articles we define:

$$U(w)$$

as the ordered set of all users *u* that contributed to the article *w*. Thus, the set *U* is a property of a single article. Then we define a set of formulas that are properties of a single user *u*.

$$E(u,w): U \otimes W \rightarrow \Im .$$

Or only *E(u)*, the number of edits for user u for article *w*. We define:

$$T(w): w \rightarrow \Im ,$$

the total number of edits for article *w*. We then define

$$P(n): [0..1] \rightarrow \Im \quad P(n) = \sum_{U_a} E(u)$$

Table 2. Users Distribution factors

Trust Factors	Comments
Average of *E*	Average number of edits per user.
Standard Deviation of *E*	Standard deviation of edits
$\frac{P(n)}{T}$	% of edits produced by the most active users
$\frac{Pe(t)}{T}$	% of edits produced by users with more than n edit for that article
Number of discussions (talk edit)	It represents how much an article is discussed
Blocked (the article cannot be edited)	The article is blocked due to vandalism. Provided by Wikipedia
Controversial	Article's topic is controversial. Provided by Wikipedia

Where U_a is the set of n% most active users in *U(w) P(n)*, given a normalized percentage *n*, returns the number of edits done by the top n% most active users among the set *U(w)*. Similar to *P(n)* is

$$Pe(n): \Im \rightarrow \Im$$

$$Pe(n) = \sum_{U_n} E(u)$$

$$U_n = \{u \in U \mid E(u) > n\}$$

that, given a number of edits *n*, represent the number of edits done by users with more than n edits. The different between *P* and *Pe* is that *P* considers the most active users in relation to the set of users contributing to the article, while *Pe* considers the most active users in relation to an absolute number of edit n.

We explain the meaning of the functions defined: P(n)/T tells us how much of the article has been done by a subset of users. If we pose n=5 and we obtain:

$$\frac{P(5)}{T} = 0.45$$

this means that the 45% of the edits have been done by the top 5% most active users. If the value is low the article leadership is low, if it is high it means that a relatively small group of users is responsible for most of the editing. We introduced the function *Pe(n)/T* to evaluate leadership from a complementary point of view. *Pe(n)/T* is the percentage of edits done by users that did more than n edits for the article. If we pose n=3 and we obtain:

$$\frac{Pe(3)}{T} = 0.78$$

This means that 78% of the edits were done by users with more than 3 edits and only 22% by users that did 1,2 or 3 edits. Thus, 1-Pe(n)/T with n small (typically 3) indicates how much of the editing's process was done by occasional users, with a few edits. Thus, it can represent a measurement of the fragmentation of the editing process. The average and standard deviation of the function *E(u)* (total edits per user u) reinforces the leadership as well: average close to 1 means high fragmentation, high standard deviation means high leadership. The last three factors are a clue of how much an article is discussed and controversial.

5.2 Stability (propositions 8)

We define the function

$$N(t): t \rightarrow \Im$$

That gives the number of edits done at time t. Then we define:

$$Et(t) = \sum_{t}^{P} N(t)$$

that, given time t it gives the number of edits done from tine t to the present time *P*. We than define

$$Txt(t): t -> \Im$$

that gives the number of words that are different form the version at time *t* and the current one. We define *U* as the largest period of time for that article, i.e. its age. We define L as the number of words in the current version.

Table 3. Article's stability factors.

Trust Factors	Comments
$\frac{Et(t)}{U}$	Percentage of edits from time *t*
$\frac{Txt(t)}{U}$	Percentage of text different between version at time *t* and current version

We evaluate the stability of an article looking at the values of these two functions. If an article is stable it means that Et, from a certain point of time t, should decrease or be almost a constant that means that the number of editing is stable or decreasing: the article is not being to be modified. The meaning of Txt(t) is an estimation of how different was the version at time t compared to the current version. When t is close to the current time point, Txt goes to 0, and it is obviously 0 when t is the current time.

An article is stable if Txt, from a certain point of time t not very close to the current time is almost a constant value. This means that the text is almost the same in that period of time. As mentioned above, an article can be stable because it is rarely edited, but this may mean it is not taken in consideration rather than it is complete. To avoid this, the degree of activity of the article and its text

quality are used as a logic condition for stability: only active and articles with good text can be considered stable.

6 Evaluation

We developed a working prototype in C able to calculate our trust factors. A diagram of the prototype is depicted in figure 3. The system, using the *factors updater* module, is continuously fed by the Wikipedia DB and it stores the results in the factor DB. The Wikipedia database is completely available for download. When we want to estimate the trustworthiness of an article, the Data Retrieval module query the Wikipedia DB (it could retrieve information directly from the web site as well), and it collects the needed data: article page, talk page, modification list, user's list, article category and old versions. Then, the *factors calculator* module calculates each of the trust factors, merging them into the defined macro-areas. Using the values contained in the *Factors DB* about pages of the same or comparable category, it computes a ranking of the page for each macro-area. Finally, the *trust evaluator module* is in charge for estimating a numeric trust value and a natural language explanation of the value. The output is achieved by merging the partial trust value of each macro-area using constraints taken from the *Logic Conditions* module. This contains logic conditions that control the meaning of each trust factor in relationship to the others:

- *IF leadership is high AND dictatorship is high THEN warning*
- *IF length is high AND importance is low THEN warning*
- *IF stability is high AND (length is short OR edit is low OR importance is low) THEN warning*

By looking at the page rank in each macro-area and considering the warnings coming from the logic condition module, explanations like the following can be provided:

"*The article has a strong editing leadership. The very high standard deviation of the edits suggests that it could be an article written mainly by few people. The quality of editing is good but its length is the highest in its category and the topic has average importance. The number of discussions is below average.*"

We present now three different experiments we conducted. Two experiments were performed using our trust factors identified using domain-specific expertise.

The third and last experiment was preformed with a radical different approach. We performed a cluster analysis to isolate featured and standard articles. The experiment was performed on the same set of data used for the second article. The radical difference of the approaches is that in the first two experiments we exploit explicit rules and factors deducted from expertise, while the last approach is obviously implicit. The comparison of the results will show the added value, if any, of using domain-specific expertise in the Wikipedia context.

In all our three experiment, in order to test our predictions we should know if the quality of an article is actually good. Wikipedia gives its best articles some awards that guarantee that these articles represent the highest standard of the encyclopaedia. There are two levels of awards. The first is the *featured article* status, which means that it has been identified as one of the best articles produced by the Wikipedia community, particularly well written and complete. Only 0.1% of the articles are featured articles. The second level is the *good article status*: articles contain excellent content but are unlikely in their current state to become featured; they may be too short, or about too an extensive or specific topic, or on a topic about which not much is known. We focused on featured articles: they should represent the trustworthiest ones, and the evaluation phase will succeed if our trust computation indicates these articles among the most trustworthy.

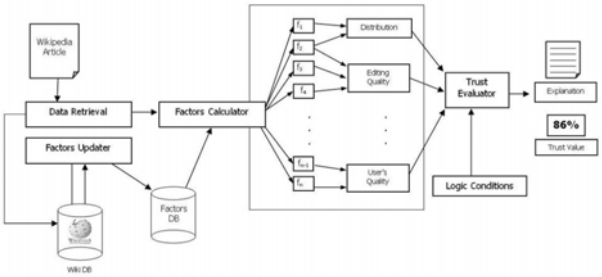

Figure 2 Trust Calculator for Wikipedia

6.1 Study-Case 1: A category of articles

We consider a subset of Wikipedia pages, the articles that describe geographical countries. We analyzed 250 countries. We decided to use these articles because they are among the more visited and edited pages; their topic is multidisciplinary, inter-cultural, they interest almost the whole community of wikipedians and they tend to have a standard that lets us meaningfully compare them to each other. For each page, we calculated a trust value in [0.1], where 1 is defined to be the most trustworthiness. The experiment was done on the 30th January 2006. On this date, there were 8 featured articles: *Australia*, *Belgium*, *Cambodia*, *Bhutan*, *Hong Kong*, *India*, *Nepal* and *South Africa*.

Table 5. User distribution ranking.

R	Article	T. V.	R	Article	T. V.
1	Portugal	1	9	Pakistan	0.921
2	Cuba	0.994	10	Trinidad and T.	0.914
3	Australia	0.974			
4	Cambodia	0.971			
5	India	0.961	16	Hong Kong	0.907
6	Canada	0.960	36	Bhutan	0.776
7	Belarus	0.953	49	Nepal	0.682
8	Belgium	0.924	55	S. Africa	0.633

Table 6. Dictatorship effect ranking.

R	Article	T. V.	R	Article	T. V.
1	Portugal	1	9	Venezuela	0.455
2	Cuba	0.787	10	Trinidad and T.	0.433
3	India	0.676			

4	Chile	0.606			
5	Australia	0.576	20	Hong Kong	0.339
6	Belarus	0.526	29	Bhutan	0.285
7	Cambodia	0.506	47	Nepal	0.24
8	Canada	0.480	70	S. Africa	0.199

Table 5 and 6 estimate *users distribution* and the *dictatorships effect*. The article *Portugal* seems to have a high possibility of suffering from the "dictatorship effect", while the same trust value decreases rapidly for the other articles. Our hypothesis is proven by reading a discussion on the *Portugal Talk page* on Wikipedia, where users complained about an author that did 35% of the edits, writing that *"Wikipedia is not a personal web page"*.

The *quality of users* macro-area seems not to be important in the trust computation. Among the featured articles, only *Nepal* (7), *Australia* (10) and *Bhutan* (41) seem to have a good rank. Since the quality of the users writing the article should be a strong factor for its trustworthiness, the evaluation phase suggests that our formulas extracted from the application model failed and we need to go deeper in the analysis.

Table 7. Quality of Editing ranking.

R	Article	T. V.	R	Article	T. V.
1	Australia	1	9	U.K.	0.902
2	U.S.A.	0.987	10	Israel	0.891
3	Portugal	0.97			
4	S. Africa	0.968	34	India	0.842
5	Germany	0.967	37	Nepal	0.751
6	Singapore	0.951	47	Bhutan	0.714
7	Turkey	0.936	52	Cambodia	0.685
8	Belgium	0.922			

Table 7 shows the *quality of editing*. This factor is very effective: featured articles are among the more referenced, they have the right length (which is about 5000-6000 words), balanced sections and images. It is interesting that many articles, good in the others factors, cannot survive the quality of editing analysis.
We observed that *Portugal* is the longest article, with double the text of *France* and 30% more than *United Stated*. The only comparable one is *Cuba*. If we look at the *dictatorship effect* (table 6) we can think that this is the result of lack of control on single users' edits.

Table 8. Article's stability ranking.

R	Article	T. V.	R	Article	T. V.
1	Belgium	1	9	Paraguay	0.877
2	Saudi Arabia	0.99	10	Austria	0.86
3	Liberia	0.954			
4	Fiji	0.95	23	Cambodia	0.782
5	Honk Kong	0.914	73	Bhutan	0.488
6	Australia	0.904	97	S. Africa	0.357
7	China	0.895	106	India	0.281
8	Madagascar	0.886	146	Nepal	0.104

Table 8 shows the *stability* ranking. The two most stable articles are *Belgium* and *Saudi Arabia*. Featured articles like *S. Africa*, *India* and *Nepal* show a bad degree of stability. Regarding *Article's activity ranking*, as expected the most important and influent countries appear at the top of the table. This factor should be considered as a condition to test stability: stable articles with less than 0.5 degree of activity are considered not edited rather than stable; the instability of an article is more dangerous if it has a high degree of activity and controllability.

Table 9. Overall Ranking.

R	Article	T. V.	R	Article	T. V.
1	Australia	1	9	Portugal	0.87
2	Belgium	0.93	10	U.S.A.	0.86
3	Singapore	0.91			
4	China	0.90			
5	Germany	0.89	16	S. Africa	0.84
6	H. Kong	0.89	20	Cambodia	0.83
7	India	0.88	37	Bhutan	0.76
8	Japan	0.88	52	Nepal	0.71

Table 10. Warning among Table 12 articles

R	Article	Warning reason
1	Portugal	Dictatorship effect. Article too long
2	India	Instability
3	Nepal	Instability

Joining all the previous factors, we can estimate our trust value. *Australia* is the more trustworthy article, 4 out of 8 featured article are in the top 10 position, 6 out of 8 with a trust value higher than 83%. *Nepal*, the worst among them, scored 71.3%. *Nepal* was a featured article in a previous version that was almost 20% different from the current one, situation that is underlined by the warning on stability. *Belgium* had a warning on its activity rate but, due to the quality of the editing and its higher stability, the warning could be interpreted as an evidence that the article has reached a reasonably complete and satisfying state. Regarding the non-featured articles in the top-ten list, *U.S.A.* and *Japan* have the *good article* status, while *Singapore* is a *former featured article.*

6.2 Study-Case 2: 8 000 articles

The experiment was conducted on the 17th of March 2006 on 7 718 Wikipedia articles.

These articles include all 846 featured articles plus the most visited pages with at least 25 edits. These articles represent 65% of the editing activity of Wikipedia and the vast majority of its access, making it a significant set. The results are summarized in figure 3. The graph represents the distribution of the articles on the base of their trust values. We have isolated the featured articles (grey line) from standard articles (black line): if our calculation is valid, featured articles should show higher trust values than standard articles. Results obtained are positive and encouraging: the graph clearly shows the difference between standard articles distribution, mainly around a trust value of 45-50%, and featured articles distribution, around 75%.

Among the featured articles, 77.8% are distributed in the region with trust values > 70%, meaning that they are all considered good articles, while only 13% of standard articles are considered good. Furthermore, 42.3% of standard articles are distributed in the region with trust values < 50%, where there are no featured articles, demonstrating the selection operated by the computation. Only 23 standard articles are in the region >85%, where there are 93 featured ones. The experiment, covering articles from different categories, was conducted on an absolute scale, and it shows a minimal imprecision if compared with a previous experiment conducted on a set of 200 articles taken all from the category "nations" [22], where we could rely on relative comparisons of similar articles. This shows that the method has a promising general validity.

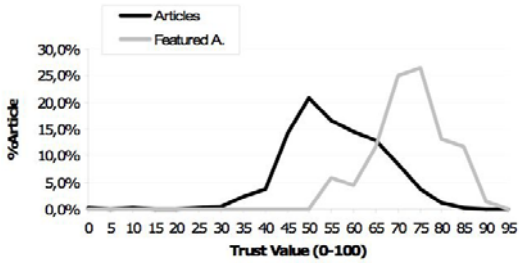

Figure 3 Expertise-based computation

Table 11: expertise-based computation. TV = trust value; SA = standard articles, FA = featured articles

Correlation	18.8 %		
	% of FA	% of SA	GAP
Bad: TV < 50	0	42.3%	42.3%
Average: 50 < TV < 70	22.2 %	54.7 %	32.5%
Good: TV > 70	77.8 %	13 %	64.8 %
Very Good: TV > 85	13.2 %	23 articles	13.2 %

6.3 Study-Case 3: Cluster Analysis

In this experiment we performed a pre-processed cluster analysis over Wikipedia articles after identifying a subset of principal articles characteristics. The scope of this experiment is to verify the value added by the expertise by comparing the results obtained in the two cases.

In previous experiments we exploited some aggregated and non-intuitive trust factors, justified and derived by expertise in areas relevant to Wikipedia. We defined some formulae that in general were not intuitive but achieved relying on domain specific expertise.

In this experiment we perform a cluster analysis to automatically divide featured and standard articles based on common features among articles. The comparison of the two results will show if the application of expertise has added value to the quantitative value of the predictions or has only a negligible effect.

Cluster analysis is an unsupervised learning technique, but in our experiment before applying data clustering we trained the system in order to identify, among a set of basic article characteristics, the most important one for a classification of articles.

The key difference with previous experiments is that we now need limited expertise, because we have replaced it by training the system (using a subset of articles of known quality) and by relying on the common featured identification of the clustering algorithm.

The hypothesis is that featured articles are recognizable by simple characteristics that do not require the application of complex expertise. Of course, some kind of knowledge is needed in order to identify a set of article components to be used by the training and by the cluster algorithm, but we avoided complex or derived trust evidence. Note that all of the elements used in this experiment have been considered also by the expertise-based computation (in their simple form or as part of a formula); this is perfectly in line with the aim of the experiment, that is to test the added value of those characteristics that are clearly expertise-derived. In other words we test if, by using only loosely expertise-dependent elements, we will still have valid results. We remind that, in any case, implicit approaches cannot give justifications like our expertise-based method does.

We performed the experiment in two phases. We started by selecting 13 base characteristics of a Wikipedia article. Elements cover both text and editing characteristics. We performed a pre-processing of the data in order to increase the validity of the experiment: we discarded trivial and out-of-standard articles, and we normalized some of the characteristics (like article length, number of images...) using the number of ingoing links of an article, representing a good estimation of its importance (note that we used the same mechanism for the expertise-based computation). These are common procedures and not domain-specific expertise.

Article characteristics are listed in table 15. All the 13 characteristics have low correlation to each other. In the first phase we simplified the list of elements by identifying the principal components. We used 30% of the featured and standard articles to train the system and the rest to perform the experimentation. Using this sample of articles we computed how each characteristic is correlated to featured article and standard article status, i.e. how effective it is in separating the two types of articles. The results are listed in the second column of table 12, and 5 principal components were identified.

In the second phase we performed a cluster analysis of the articles using the 5 principal components identified. We used the well-know k-mean clustering algorithm [13] to identify 2 clusters (featured and standard articles).

Table 12: Components of an Article

N.	Components	Importance
1	Average number of edits per author	Low
2	**Variance of edits length**	**High**
3	**Percentage of Registered Authors**	**High**
4	Percentage of Contributions by Registered Authors	Medium
5	Percentage of Reverted edits	Low
6	Average Length of editing	Medium
7	**Variance of sections**	**High**
8	Average length of a section	Medium

9	Number of discussions	Low
10	**Number of Images**	**High**
11	**Length of the article**	**High**
12	Number of Section	Medium
13	Number of references, external link	Medium

A graphical representation of the results is displayed in Figure 4. The graph represents the distribution of the *featured articles* (grey line) and *standard articles* (black line) according to the normalized difference of the distances between the article and the two centroids (centres of each cluster). A value of 0 on the horizontal axis means that the article has the same distance from the two centroids; a negative value means that the article is closer to centroid 1 – standard article cluster – while a positive value means that the article is closer to centroid 2 – featured article cluster. Articles whose values are less than -1 or greater than 1 are accumulated at the border of the graph.

Referring to table 16a, the two clusters have a recognizable separation: 76.4% of standard articles are in one cluster (region of negative values) while 78% of featured articles are in the other cluster. 23.6% of standard articles fall into featured articles cluster, while in the expertise-based computation they were only 13.2%. A portion of 2.5% standard articles is very close to the featured articles centroid, while only 23 standard articles (less than 0.03%) had a trust value >85 in the expert computation. In general, the results still show an interesting validity partially comparable with previous results. The value added by the expertise results more evident if we consider the uncertainty of the predictions. We divided the algorithm results in 3 zones: featured articles (-1,-0.33), standard articles (0.33,1) and an intermediate zone (-0.33,0.33) where a decision cannot be taken. Referring to table 13b, 45.6% of featured articles and 39.6% of standard articles fall into the intermediate region. This means that in almost half of the cases the algorithm predictions are highly uncertain. Only 10.8% of standard articles are in the featured cluster (slightly better than expertise computation), but 6.1% of articles are in the standard articles cluster compared to none in the expertise case.

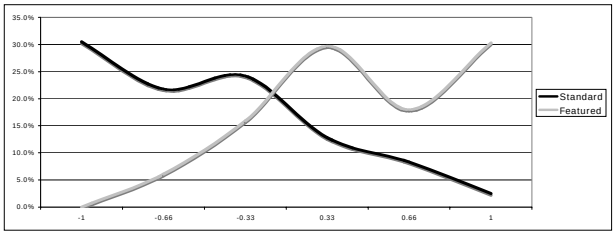

Figure 4 Graphical representation of Cluster Analysis.

Table 13: Cluster Divisions

CASE A			
	Cluster 1	**Cluster 2**	
S. Articles	76.40%	23.60%	
F. Articles	22%	78%	
CASE B			
	Cluster 1	**Intermediate**	**Cluster 2**
S. Articles	52.30%	37.00%	10.80%
F. Articles	6.10%	45.60%	48.30%

If we compare these results with the expert-based computation, we see that by using expertise-derived trust evidence the main added value is the reduction of uncertainty: 77.2% of featured articles have a clear high trust value against only 48.3% in the cluster computation. Moreover, in the expertise case not a single featured article was placed in the region with trust value <50%. This means that, if an article is in that region, it is almost certainly a standard one. On the contrary, 6.1% of featured articles in the cluster computation fall into the region of standard articles (cluster 1). The value added is due especially to the analysis of user distribution and stability. Thanks to these aggregated expertise-justified functions, more certain predictions can be done in borderline cases, and it is possible to capture characteristics that an implicit approach fails to identify.

7 Conclusions

In this paper we have proposed a transparent, non-invasive and automatic method to evaluate the trustworthiness of Wikipedia articles. The method was able to estimate the trustworthiness of articles relying only on their present state, a characteristic needed in order to cope with the changing nature of Wikipedia. After having analyzed what brings credibility and expertise in the domains composing Wikipedia, i.e. content quality and collaborative working, we identified a set of new trust sources, trust evidence, to support our trust computation. The experimental evidence that we collected from almost 8 000 pages covering the majority of the encyclopaedia activity leads to promising results. This suggests a role for such a method in the identification of trustworthy material on the Web. The detailed study case, conducted by comparing a set of articles belonging to the category of *"national country"* shows how the accuracy of the computation can benefit from a deeper analysis of the article content. In our final experiment we compared our results with the results obtained using a pre-processed cluster analysis to isolate featured and standard articles. The comparison has shown the value added by explicitly using domain-specific expertise in a trust computation: a better isolation of articles of great or low quality and the possibility to offer understandable justifications for the outcomes obtained.

References

[1] Ciolek, T., Today's WWW, Tomorrow's MMM: The specter of multi-media mediocrity, IEEE COMPUTER, Vol 29(1) pp. 106-108, January 1996

[2] *How much do you trust wikipedia?* (March 2006) http://news.com.com/20091025_3-5984535.html

[3] Gales, J. *Encyclopaedias goes head a head*, Nature Magazine, issue N. 438, 15, 2005

[4] Romano, D. M., *The Nature of Trust: Conceptual and Operational Clarification*, Louisiana State University. PhD Thesis, 2003
[5] Blaze, M., Feigenbaum, J. and Lacy, J., Decentralized Trust Management. Proceedings of IEEE Conference on Security and Privacy, 1996
[6] Chu, Y., Trust Management for the World Wide Web, Master Thesis, MIT, 1996
[7] Terzis, S., Wagealla W. *The SECURE Collaboration Model*, SECURE Deliverables 2.1, 2.2, 2.3, 2004
[8] Marsh, S. Formalizing Trust as a Computational Concept. PhD thesis, University of Stirling, Scotland, 1994
[9] V. Cahill, et al. Using Trust for Secure Collaboration in Uncertain Environments. IEEE Pervasive Computing Magazine, 2003
[10] Golbeck, J., Hendler, J., Parsia, B., *Trust Networks on the Semantic Web*, University of Maryland, USA, 2002
[11] www.foaf-project.com, FOAF project website
[12] Ziegler, C., Golbeck, J., *Investigating Correlations of Trust and Interest Similarity*, Decision Support Services, to appear.
[13] MacQueen J., *Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics, Berkeley, University of California Press, USA, 1967
[14] Bucher, H., Crisis Communication and the Internet: Risk and Trust in a Global Media, First Monday, Volume 7, Number 4 2002)
[15] Huang, J., Fox, S., *Uncertainty in knowledge provenance*, Proceedings of the first European Semantic Web Symposium, Heraklio, Greece, 2004.
[16] Burke, R., *Knowledge-based Recommender Systems*. Encyclopaedia of Library and Information Systems.Vol. 69, Supplement 32, New York 2000
[17] Fogg, B. J., *How Do Users Evaluate The Credibility of Web Sites?* Proceedings of the conference on Designing for user experiences, ACM Press, USA, 2003
[18] Alexander, J., Tate, M., *Web Wisdom: How to Evaluate and Create Information Quality on the Web*, Lawrence Eribaum Associates Inc, New Jersey, USA, 1995
[19] Cassel. R., *Selection Criteria for Internet Resources*, College and Research Library News, N. 56, pagg. 92-93, 1995
[20] Standford Web Credibility Guidelines web site, http://credibility.stanford.edu/guidelines
[21] Roberts, T., Online Collaborative Learning, Theory and Practice, Idea Group Pub, USA, 2004
[22] http://download.wikimedia.org, download site of the Wikipedia project