

Quantifying the trustworthiness of social media content

Sai T. Moturu · Huan Liu

© Springer Science+Business Media, LLC 2010

Abstract The growing popularity of social media in recent years has resulted in the creation of an enormous amount of user-generated content. A significant portion of this information is useful and has proven to be a great source of knowledge. However, since much of this information has been contributed by strangers with little or no apparent reputation to speak of, there is no easy way to detect whether the content is trustworthy. Search engines are the gateways to knowledge but search relevance cannot guarantee that the content in the search results is trustworthy. A casual observer might not be able to differentiate between trustworthy and untrustworthy content. This work is focused on the problem of quantifying the value of such shared content with respect to its trustworthiness. In particular, the focus is on shared health content as the negative impact of acting on untrustworthy content is high in this domain. Health content from two social media applications, Wikipedia and Daily Strength, is used for this study. Sociological notions of trust are used to motivate the search for a solution. A two-step unsupervised, feature-driven approach is proposed for this purpose: a feature identification step in which relevant information categories are specified and suitable features are identified, and a quantification step for which various unsupervised scoring models are proposed. Results indicate that this approach is effective and can be adapted to disparate social media applications with ease.

Communicated by Anupam Joshi.

All the work of S.T. Moturu was performed while he was at ASU.

S.T. Moturu · H. Liu
School of Computing, Informatics and Decision Systems Engineering, Arizona State University,
Tempe, AZ 85287, USA

H. Liu
e-mail: huan.liu@asu.edu

Present address:
S.T. Moturu (✉)
Media Lab, Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: smoturu@mit.edu

Keywords Trust evaluation · Trustworthiness · Social media · Content · Quality

1 Introduction

The advent of Web 2.0 has changed the average web user from a consumer to a content creator. Social media applications like blogs, wikis and social networks are generating a large expanse of content, driven by user contributions including the sharing of opinions, experiences and acquired knowledge. Articles about politics, history, people and health are available on wikis and blogs. Advice is being shared on question-answer sites and social networks. Opinions are being conveyed on blogs, microblogs and social networks. In such a scenario, users not only need a way to sift through this data but also a method to quantify the value of content. Is the information trustworthy? Is the data accurate? While search engines can provide a way to find data, they cannot provide an answer to these questions. Hence, one has to rely on their own intellect, knowledge, analytical capabilities, and limited visible information to answer them.

As a motivating example, consider the search for “How to prevent Restless Leg Syndrome”, performed on Google on June 18, 2008¹. The top result was from a popular social media site with collaboratively contributed content, wikiHow. The article claims that the condition can be caused by drinking large quantities of orange juice due to their possible insecticide content. Could this claim be true?

Since the article is the first result on Google and from a relatively popular website, one is tempted to trust the article without further inquiry and assume that its content is trustworthy. However, a quick check using other search results negates this claim. There are two problematic assumptions here. The first one is the excessive trust placed on search results because search relevance does not imply the trustworthiness of content. The second one is the trust placed on a website. While this might have worked earlier, it is not appropriate for user-driven social media sites with content contributed by numerous authors.

The examples discussed above establish the need for trust assessment and its necessity as an addition to search relevance for social media portals. The presence of such assessments can change the way people perceive and utilize information from social media. This is of greater importance for domains such as health where the negative impact of acting on untrustworthy information is high. The web has its share of useful health information in combination with marketing, vandalistic and exploitative [9] activities and concerns about such information [5] are understandable. Surveys indicate that 81% of all adults who are online have used the Internet for health information² and it was the most widely used resource for health information³ beating doctors, relatives, and friends. Further, the latter survey finds that Wikipedia is the most popular social media health resource followed by forums and social networks.

This paper addresses the trust assessment problem and proposes a relevant solution that is not restricted to a specific social media application. The approach is tested

¹<http://www.webcitation.org/5Yg4uiE2K>.

²http://www.harrisinteractive.com/harris_poll/index.asp?PID=937.

³<http://www.icrossing.com/research/how-america-searches-health-and-wellness.php>.

using two data sets consisting of health articles from Wikipedia and advice shared on an autism forum on the health social network, Daily Strength.

2 Related work

A considerable number of works in recent years have been devoted to studying different aspects of social media content including quality and trust. Here, we discuss a subset of these focusing on the quantitative assessment of content. Agichtein et al. [2] embark upon the task of quality assessment using data from a community question-answer website. Semantic features based on intrinsic content quality as well as features derived from user relationships and usage statistics are used to classify answers. While this study uses the Answers domain, most others focus on Wikipedia.

The simplest of these uses just a single feature, article size, to evaluate Wikipedia articles [4]. McGuinness et al. [11] base their assessment of trust on the occurrences of the encyclopedia term in a Wikipedia article. The same group studied the possibility of using revision history to assess trust [16]. Revision history has also been used to assess and depict the varying trustworthiness of different parts of the text of a Wikipedia article [1]. Hu et al. [7] base their study of article quality on the assumption that revisions involve peer review of at least part of the content. Dondio and Barrett [6] use objectivity, completeness and pluralism as the hallmarks of good information to select relevant features and predict a trust score for Wikipedia articles.

Our work differs from existing studies in two major aspects. First, the term “trust” is used loosely even though it is a thoroughly studied concept in sociology and other domains. Hence, we use existing sociological theory to motivate our approach to trust assessment. Second, we propose a more generalized approach to quantify the trustworthiness of content across social media unlike most existing studies that focus on a single application. This approach includes two levels of analysis: feature identification and trust quantification. We derive relevant features indicative of content trustworthiness for the specific application using categories identified from sociological theory. We then use unsupervised trust scoring models that are not application-dependent to appraise these features and output a trust score. This process is intended for use across disparate social media with the feature extraction repeated for each application while the trust scoring approach remains constant, allowing for easier application to different sites.

The rest of the paper is arranged in the following manner. We present relevant sociological notions on trust in the following section and adapt these notions for the analysis of social media content. We then describe the data and features, followed by our trust scoring models and evaluation approaches. Finally, we discuss the experimental results and draw conclusions.

3 Trust

3.1 Theory of trust

Trust is an important sociological concept that has been studied in depth by many researchers for a number of years [14]. It is defined with respect to a transaction

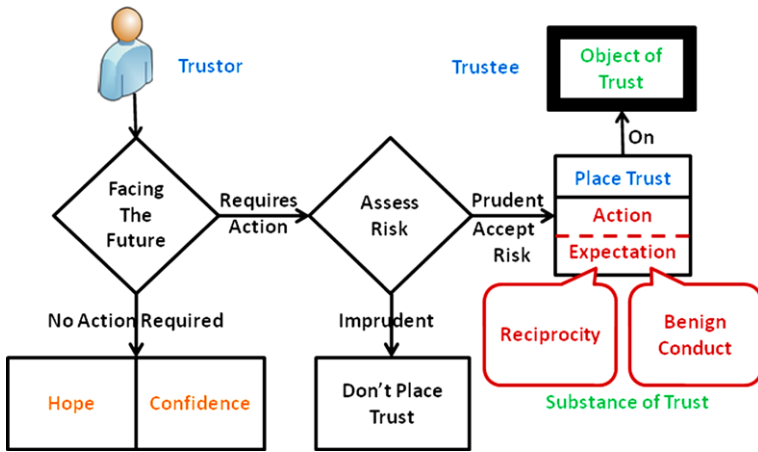


Fig. 1 Trust: A Path of Action

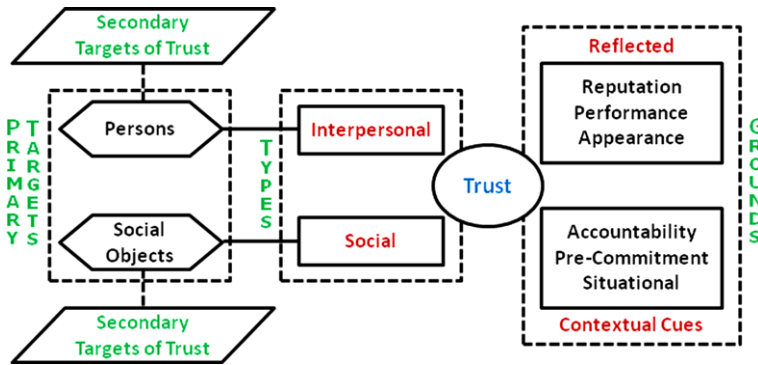


Fig. 2 Trust: Targets and Grounds

between two entities, the trustor and the trustee. Trust can be defined as the perception of the trustor about the degree to which the trustee would satisfy an expectation about a transaction constituting risk. Trustworthiness can be defined from the perspective of both these entities. In this work, we will only consider the perspective of the trustor, which defines this property to be the amount of trust associated with the trustee [3]. For the purpose of this work, we use the term *trustworthiness* as a property that quantifies one aspect of social media content.

Trust is a tool that is used to reduce social complexity [13]. It appears only in the context of human actions, necessitated by the presence of risk and uncertainty. Apart from a commitment through action, it also involves specific expectations about others' actions and/or the future. Trust is a solution for situations involving risk. Placing trust involves the suspension or discounting of risk [14]. In the following discussion, we endeavor to delineate the aspects to be considered for trust assessment. The notions of trust described here are encapsulated in Figs. 1 and 2.

There are two major types of trust. The most commonly studied type is interpersonal trust. This involves face-to-face commitment between persons. The second type is social trust, which involves faceless commitments towards social objects that may involve individuals, likely unknown to us, in the background.

There are two types of targets for Trust. Primary targets are the objects that are the main source of risk and uncertainty. With interpersonal trust, these are the persons with whom we come in direct contact. With social trust, these may be one of a number of varied social objects including social categories (race), social roles (priest), institutions (banks) and systems (technological, social). Secondary targets are derivative targets that are used in the process of justifying or placing trust towards the primary objects.

The targets of trust specify the entities under consideration while the expectations generated from the act signify its substance. One such expectation is that of reciprocity wherein we expect our act to be reciprocated with a similar one. The act of placing trust might also bring about expectations of benign conduct that could include regularity (orderliness, coherence, consistency), reasonableness (justification, acceptance), efficiency (competence, effectiveness), moral responsibility, kindness, authenticity, fairness and fiduciary conduct.

Once the target and substance of trust are understood, we need to know the reasons and rules that can be used to grant or withdraw trust. Grounds for trust are dependent on the individual or object and the situation. Some people are more trusting than others by nature or due to societal and cultural conventions. Further, the situation dictates whether or not trust can be placed. While such grounds may never be conclusive or foolproof, they are necessary to assess trustworthiness.

The first category for such grounds is called Reflected Trustworthiness. This category includes characteristics derived from the immanent qualities of the object reflecting trustworthiness. Such traits fall under three primary classes: reputation, performance and appearance. Reputation reflects the record of past deeds while performance refers to the impact of actual deeds and present conduct. Appearance includes external features such as personality or status. The second category involves Contextual Cues and includes properties dependent on aspects of the external context in which the actions take place [14].

3.2 Social media content and trust assessment

A significant amount of social media content is being assimilated by a large audience that trusts it and acts based on it from time to time. This is particularly relevant for a domain such as health, where users are turning to social media for advice and recommendations for everything from drugs to treatments and doctors. Users assimilating such information face risk and uncertainty when they act based on this information. Certain knowledge is required to evaluate the risk before placing trust on the content and acting on it. However, much of the knowledge used to evaluate such information in the real world may not be available in the virtual one, making it difficult to assess the prudence of taking a risk. This brings about a need for the trust assessment of social media content.

Using the sociological concepts described in the previous section, we will define this problem. The goal of this work is to devise models that would assess the risk

involved in trusting a piece of information taken from social media and quantify it through the generation of trust scores. Here, the trustee is the content while the trustor is an end user who would use the proposed models. The trust score represents a quantification of the prudence of taking a risk and would ultimately allow the end user to decide on the amount of trust that can be placed on the content, i.e. its trustworthiness.

This is a case of social trust, wherein the social object is the content. The substance of this trust problem is an expectation of benign response including regularity, reasonableness, efficiency, moral responsibility and authenticity. The primary target of our trust is the content. However, assessing the trustworthiness of the content directly would entail complicated semantic analyses driven by a large knowledge base. Finding a suitable solution along these lines that can work for all social media can be a nearly insurmountable task. Instead, we can make use of secondary targets to generate features that can help assess the trustworthiness of the primary target. We intend to use accepted grounds for trust such as reputation, performance and appearance as trust feature categories from which such features are derived. These categories and the proposed trust models are discussed in the following sections.

4 Trust assessment

4.1 Intuition

One of the primary challenges associated with research on social media is the difference among various applications and even different portals for the same application. While there is no scope to develop a single model that can work across all social media portals, our intent is to develop an approach that can be easily adapted across applications. In such a scenario, the question that remains is about the commonality between these applications that can be exploited to develop such an approach. The availability of metadata across social media sites with respect to the creation and dissemination of the content as well as the responses it generates presents us with this commonality. Factors derived from such metadata, in addition to the content itself, can be useful in summarizing the underlying qualities of the content. While they may differ across social media portals, the availability of these factors can be utilized to develop an easily adaptable approach to trust assessment in social media.

In the real world, when we are faced with a situation where a decision is to be made regarding the trustworthiness of a person or a social object, we assess it based on numerous factors including the past, present and perceived qualities. The same intuition is applied here. We can derive numerous suitable features about the content (primary target of trust) depicting reputation (past), performance (present) and appearance (perceived) from the external characteristics of the content as well as the associated metadata (secondary targets of trust). These features can together help predict the trustworthiness of content.

While these features can be useful predictors, there is still a need for scoring models that can quantify the trustworthiness of content. As mentioned above, the features are different across social media applications even if the process to derive them is similar. In order to keep the process from being more complicated and less adaptable,

we need to keep the scoring process unchanged across social media. In order to do this, the scoring models need to have two qualities. The first one is that the models are unsupervised. This requirement keeps them independent of the dataset. In addition, there may be no class information available for most social media applications. This requirement alleviates the need for such information making it adaptable for all social media. The second requirement is that the model is feature-driven. As described above, the intuition is to use multiple suitable predictors to assess trustworthiness as we would in the real world. This requirement means that the models can be used with any features but ultimately, performance is driven by the features that are utilized.

Among the other challenges for analysis are the lack of a ground truth and suitable baselines for performance comparison. To solve the first issue, manual evaluation of content can be performed to provide a ground truth for performance evaluation. Further, suitable baselines need to be created to help assess whether a model is useful or not. This discussion provides an overview and lays the ground for a more detailed description of our approach to content assessment.

4.2 Feature categories

As discussed earlier, features derived from secondary targets can help assess the trustworthiness of our primary target, the content. Such features can be extracted based on the various grounds for trust. One group of such targets can be derived from the content. For example, features that ascertain qualities such as the structure or size can be derived from the content. These features are only indirect evaluators and are therefore considered secondary targets. In addition, social media have a layer of commonality—the availability of metadata or meta-information about the creation and/or evolution of content as well as the response to it. Metadata about users is available from their contribution history and activity patterns. Such data can be of immense value for trust evaluation. Hence, features derived from such metadata from another group of targets. The various grounds for truth described earlier provide a basis for feature extraction from these targets. We describe these feature categories in the following subsections.

4.2.1 Reputation

An individual's reputation, revealed through his past actions, is indicative of his trustworthiness [14]. A suitable parallel for content assessment is the reputation of the content creator through his past actions on the social media portal such as content creation or consumption, responses to content, generic interactions with others, functions such as social networking and other activities on the site. Information about user activity is available across social media applications. Much like the real world, where one's past is used to evaluate them, this data can help assess one aspect of trust. Examples of useful variables can include information about past contributions of authors including number and frequency. Feedback from other users about the author and/or his contributions could be a useful indicator of trust that is directly derived from the social capabilities of the medium. Further, the author's social connections, related interactions and his social status on the site could help lend credence to his contributions.

4.2.2 Performance

The performance of an individual, as evinced by his present conduct and current actions can be used to estimate his trustworthiness [14]. With respect to content, performance can be determined from user actions towards the development of content and the responses to it. Properties signifying performance would vary significantly based on the social media application. Examples of useful features can include the number and tenor of responses to a blog post, the direct evaluation of the content from audience members on a social network, and the number of edits in the development of a wiki article. Features based on references to the sources of information and citations to them in text are pertinent as well.

4.2.3 Appearance

External characteristics that symbolize an individual's appearance and demeanor through indicators of personality, status and identity can be used to assess his trustworthiness [14]. In a similar vein, outwardly visible characteristics of content, such as style, size and structure, which are useful in judging its quality, can be useful cues in quantifying the amount of trust that can be placed on it. Such features can be collected from the content characteristics. For example, the length of a blog post or wiki article could indicate the seriousness of the author or the comprehensiveness of the article and the structure and language of the content could indicate its quality.

4.3 Trust models

Our approach to trust evaluation is divided into two tasks. The first task, described above, is the identification of features to assess trustworthiness. The second task, described below, involves the creation of feature-driven trust evaluation models.

4.3.1 Problem formulation

The creation of appropriate trust evaluation models is a considerable challenge. The aim is to develop models that can furnish a score to help users evaluate the relative trustworthiness of content with ease. The goal is to create models that are unsupervised, feature-driven and independent of the applications.

Given a dataset, D , with social media content, I , and associated metadata, M , we intend to quantify the amount of trust that can be placed on I through a trust score, $T_X(i)$, derived from feature values. To achieve this, we use each feature as a predictor. Following the manual selection of n features (f_j) suitable for trust evaluation, the value of each feature f_{ij} corresponding an article I is expected to contribute positively or negatively to its overall trust score $T_X(i)$.

The feature score that each feature f_{ij} contributes, S_{ij} , is dependent on the trust evaluation model X . This feature score is either continuous or discrete. In the case of the latter, the set of assigned values will belong to any one of p score classes. In the case of the former, the feature score may be derived from the differences in feature values and their corresponding importance.

Following feature score assignment, a system is required to combine these values to create an overall trust score, $T_X(i)$, for article I . The simplest way to achieve this is by simply adding together feature scores f_j assuming variables are equally important. To allow for variables with different levels of importance, a weighted sum can be used, with the weights, w_j , implying varied significance.

4.3.2 Intuition

The models formulated here make use of the fact that all features are one-dimensional, continuous and ordered. Let $v_1 \leq v_2 \leq \dots \leq v_n$ be the ordered values for a feature. The goal is then to divide these values into k contiguous clusters $(v_1, \dots, v_{c1}), (v_{c1+1}, \dots, v_{c2}), \dots, (v_{ck-1}, \dots, v_{ck})$. This first step is akin to unsupervised univariate discretization [10]. Following this step, a score is assigned to a feature based on the cluster in which it falls. The ordering and/or distribution of clusters are to be considered while assigning the scores in order to encapsulate the degree of difference between clusters. A feature score, S_{ij} , is output by a trust model for each feature, based on the cluster in which the value falls. The overall trust score, $T_X(i)$, output by model X allows us to rank the documents. In the following subsections, we propose trust scoring models using this intuition.

4.3.3 Simple aggregated score

The Simple Aggregated Score (SAS), as the title suggests, is a simple scoring model that we intend to use for performance comparison with the more sophisticated models to follow. As discussed above, we want to use each feature as an individual predictor that contributes to the overall score. The simplest way to do this is to use the feature values as relative indicators of trust. A high feature value for an article makes it more trustworthy (based on that feature) than another with a lower value (when a higher value indicates greater trustworthiness). The simplest system to create a trust score for an article is to sum these values to provide an overall trust score. To avoid giving undue importance to features with values of greater magnitude, we use feature values normalized between 0 and 1 while calculating the trust scores (the normalized value is subtracted from 1 when higher values indicate lower trust). The following equations describe the model.

$$S_{ij} = \frac{f_{ij} - \min(f_j)}{\max(f_j) - \min(f_j)} \tag{1}$$

$$T_{SAS}(i) = \sum_{j=1}^n S_{ij} \tag{2}$$

The normalized value of feature f_{ij} corresponding to article I , is its feature score, S_{ij} . The sum of these scores provides the final trust score, T_{SAS} . The final trust scores provide a relative viewpoint of trustworthiness, with higher values indicating greater trust. The SAS model is intuitive and makes use of the feature values to indicate the differences in trustworthiness. However, its use of absolute values means that it does not take the distribution of feature values into account. This means that a difference

Algorithm 1 Cluster Rank Score

```

Input: Set of data points  $d_i$  with associated features  $f_{ij}$ 
Output: Trust score for all data points
foreach Feature  $f_{ij}$  do
  Order feature values
  Initialize cluster list with a single cluster containing ordered
  feature values
  repeat
    Select cluster with maximum SSE for division
    for  $i=1$  to (number of values-1) do
      Bisect the cluster at  $i$ 
      Find total SSE for the resulting clusters
    Select the clusters with lowest total SSE
    Replace the original cluster in the cluster list with the two
    new clusters
  until  $K$  clusters are created
  Assign a feature score,  $s_{ij}$  to each cluster, where the scores reflect
  the order
foreach data point  $d_i$  do
  Calculate trust score  $T_{CRS}$  by combining feature scores  $s_{ij}$  for
  each feature  $f_{ij}$ 
  
```

of X is the same anywhere in the range. We seek to build on the basic idea of the SAS model and overcome this shortcoming.

4.3.4 Cluster rank score

The Cluster Rank Score (CRS) model goes beyond the SAS model and attempts to cluster the ordered feature values from v_1 to v_n using a divisive clustering approach such that the sum of squared error (SSE), defined in (3) is minimized.

$$\begin{aligned}
 \text{SSE} &= \sum_{i=1}^k \sum_{j=m_{i-1}+1}^{m_i} (v_j - \bar{v}_i)^2 \quad \text{where} \\
 \bar{v}_i &= \frac{1}{m_i - m_{i-1}} \sum_{j=m_{i-1}+1}^{m_i} v_j
 \end{aligned}
 \tag{3}$$

Algorithm 1 depicts the procedure. The cluster with maximum SSE is selected for division. From all the possible divisions, the cluster pair with lowest total SSE is kept. After the creation of k clusters, a score from 1 to k is assigned to each of the clusters based on ordering of the values that they contain. The trust score for a data point, T_{CRS} , is created by combining the features scores assigned to the clusters in which the feature values fall and a larger value indicates greater trustworthiness. As in the case of many clustering approaches, the selection of the best k presents a significant challenge [15]. For the purpose of this work, we do not delve into this problem as it has been thoroughly studied.

4.3.5 Dispersion degree score

In contrast to the CRS model, the Dispersion Degree Score (DDS) model utilizes the dispersion of a feature value from its mean to derive its relative importance. The assumption is that the farther a feature value is from its mean, the greater its effect on

trustworthiness. Feature values are clustered using this dispersion, with cluster scores dependent on order. The following equations describe the model.

$$s_{ij} = x + 1 \quad \text{if } m_i - d_i + cxd_i < f_{ij} < m_i - sd_i + c(x + 1)sd_i \quad (4)$$

$$0 \quad \text{if } f_{ij} < m_i - sd_i$$

$$11 \quad \text{if } f_{ij} < m_i + sd_i$$

$$S_{\text{DDS}}(i) = \sum_{j=1}^n S_{ij} \quad (5)$$

A score, s_{ij} , is assigned to each feature f_{ij} based on the dispersion of its value from the mean, m_i , as measured by the standard deviation sd_i . Each feature can fall in one of twelve trust classes with scores from 0 to 11. The constant, c , is used to define the class interval and a value of 0.2 is used here. The sum of feature provides the trust score, T_{DDS} , with a larger value being preferred as earlier.

4.3.6 Weighted cluster rank score

The CRS and DDS models use different approaches to clustering feature values. The CRS model considers the local distribution of feature values while clustering. The notion of maximizing intercluster distance and minimizing intracluster distance is utilized in this model but the degree of difference between clusters on a global level is not considered. On the other hand, the DDS model considers the global distribution of the feature values to group them based on their dispersion from the mean but the local differences are not accounted for. While both models are intuitive and useful, they leave scope for improvement. We seek to bridge this gap and combine both these models with the Weighted Cluster Rank Score (WCRS) model. First, the CRS model is used to cluster data points based on the local distribution and keeping the order of data points in mind. Then, the DDS approach is applied to take the dispersion between the clusters into account.

WCRS follows Algorithm 1 but differs from the CRS model at the penultimate step, where feature scores are assigned. Here, we use the DDS model. The cluster means are used in place of f_{ij} in (2) and the resulting score, s_{ij} , is used as a weight and is multiplied by the scores generated from the CRS model to provide the feature score for each cluster. This is aggregated as described in the final step to provide the trust score, T_{WCRS} .

4.4 Evaluation

4.4.1 Normalized discounted cumulative gain

The popular Normalized Discounted Cumulative Gain (NDCG) evaluation metric [8] was originally designed to test the ability of a document retrieval query to rank documents that are more relevant highly. This metric has since been used to evaluate quality and trustworthiness predictions of Wikipedia articles [7]. The trust scores output from each of our models can be used to rank responses. Though we are not concerned

with retrieving relevant responses, we require responses that are more useful or more trustworthy to be ranked highly. Therefore, NDCG is a suitable evaluation measure for our models.

Equation (6) is used to calculate the discounted cumulative gain (DCG) for the top k articles. The numerator in (6) defines the gain where $s(r)$ denotes the score for an article ranked r . Consider the case where three classes have scores of 10, 5, and 1 representing the proximity of the classes. The gain for an article from the top class is $2^{10} - 1$ and only $2^1 - 1$ for an article from the bottom class. The sum of this gain term for k articles defines their cumulative gain. The denominator in (6) is used to discount gain as the rank increases. Discounted gain for a featured article with ranks 1 and 2 will differ based on their position. While the former has a discounted gain of 1023, the latter's gain is discounted from 1023 to 645.44.

$$DCG_k(S_m) = \sum_{r=1}^k \frac{2^{s(r)} - 1}{\log_2(1 + r)} \tag{6}$$

$$NDCG_k(S_m) = \frac{DCG_k(T_m)}{DCG_k(T_p)} \tag{7}$$

$$DCG_k(T_m) = \sum_{r=1}^k \left(\left(\frac{1}{n_i} \sum_{j=i+1}^{t_{i+1}} (2^{s(r)} - 1) \right)^{\min(t_{i+1}, k)} \sum_{j=i+1}^{\min(t_{i+1}, k)} \frac{1}{\log_2(1 + r)} \right) \tag{8}$$

The NDCG function in (7) normalizes the DCG value calculated from (1) by dividing it with the DCG obtained for a perfect ranking using the same formula. This helps us obtain an NDCG value between 0 and 1. As the preference would be to obtain a ranking as close to the perfect ranking as possible, an NDCG value closer to 1 indicates a high accuracy in prediction.

While, it is a popular measure, NDCG does not take into account the effect of tied scores when multiple possibilities exist for result ordering. We use a tie-oblivious version of NDCG [12] in our evaluations. Equation (8) defines the new discounted cumulative gain function that averages the gain across each position in a tied group. The NDCG formula in (7) remains the same.

4.4.2 Baselines

We propose the following two simple baselines for performance comparison, keeping the NDCG evaluation measure in mind. These baselines provide a simple set of feature independent rankings reflecting the worst and average case scenarios. These baselines along with the SAS score provide a set of performance measures that will serve as a reference point based on which the performance of our proposed trust scoring models will be assessed. In particular, NDCG performance can vary based on the scores assigned to the information classes. The baselines, therefore, provide a suitable base performance to compare against.

The Reverse Baseline Score (RBS) is a baseline approach that represents the worst case. In this approach, the responses are ranked in reverse order of trust and these ranks are used as the trust score. This would mean that the best responses are at the

bottom and the worst at the top, resulting in the worst possible performance. The Equal Baseline Score (EBS) is a baseline approach that represents the average case. Here, each article is assigned the same score irrespective of its feature values. As all articles are equally important, the performance of this model would always be much better than the RBS model. Our intent is to create models that improve considerably on these simple baselines.

5 Case studies

We study two varied social media applications: the collaboratively edited online encyclopedia, Wikipedia, and the health social network, Daily Strength. We perform thorough experimental analysis on data from these sources to confirm the usefulness of our approach and our scoring models as described below.

5.1 Data and features

5.1.1 Wikipedia

One of the primary reasons for the selection of Wikipedia as a data source is the presence of a small proportion of articles that are manually classified by users. Such a classification provides us with an opportunity to evaluate our experimental results. Wikipedia articles that are of high quality can be nominated by a user to be recognized as a featured article or a good article. Such articles are then evaluated by a group of impartial users against stringent requirements judging accuracy, neutrality, completeness, and style to assess whether they are of suitable quality to be classified as featured or good. Featured articles are of the highest quality, closely followed by good articles. Consequently, only a small proportion of articles are classified as featured (about 1 in 1140) or good (about 1 in 420).

Further, users also have the capability to tag articles as cleanup or stub. An article is tagged as a stub when it has very limited content and does not provide reasonable coverage of the topic. Articles are tagged for cleanup when users specifically identify areas for improvement. Some common reasons for tagging an article as cleanup include the lack of sources, a need for expansion or a lack of clarity. A small proportion of Wikipedia articles are tagged for cleanup (1.42%). Many articles do not fall under any of these categories and are considered standard articles. These articles have more content than stubs, have not (yet) been marked for improvement and/or were not worthy enough to be considered good or featured. The dataset consists of 230 health-related articles from Wikipedia that fall under the featured (25), good (25), standard (105), cleanup (25) and stub (50) classes. Next, we describe the features from each feature category and the intuition behind their selection.

We begin with the Reputation category where we can consider features based on the editors' registration status in combination with the quantity and frequency of contributions to this and/or other articles. However, there is a caveat to be considered. Stubs have a higher percentage of trustworthy registered editors with because they

create and setup such articles for others to develop. Similarly, cleanup articles also receive more attention from such editors who contribute to their betterment. As a result, our intuitive association of reputed and credible editors with trustworthy articles, while reasonable, is not in tune with content development in this domain. Hence, no features are selected in this category.

The first feature in the Performance category is the number of edits or the revision count of the article that results in its current state. Collaborative editing is dependent on the constant improvement of articles. Therefore, it is reasonable to assume that an article with a large number of revisions has improved with age and is therefore more trustworthy. Another suitable feature is the median time between edits for the article. A shorter time between edits indicates that the actions related to the development of the article are more frequent and hence, unacceptable changes may be found sooner. The median edit length can also be helpful assuming that trustworthy articles would contain more purposive and comprehensive edits resulting in a larger value for this feature.

Wikipedia allows users to revert bad edits to restore articles to an earlier state. These actions are very relevant to the development of a trustworthy article. Users can also tag such edits as vandalism if they consider them so. Reverted edits indicate that the editors are vigilant and able to remove material that is detrimental to article quality. Therefore, the proportion of reverted edits is selected as a feature. It is assumed that a smaller number indicates that the article is not well maintained. A related feature, the proportion of reverted edits that were not considered vandalism, is also selected as it too illustrates article maintenance.

References are available for most Wikipedia articles. These additions to the content point to the sources of information rendering it more trustworthy. It is possible to differentiate between sources and use their relative importance for assessment. Journal articles and other peer-reviewed publications are assumed more important than web articles. A weighted reference score is calculated with publications receiving double importance. Larger articles would have more references. To remove this effect, the reference score is divided by the number of paragraphs in the article. A larger value for the resulting variable indicates a more well-referenced article. For such an article, we also expect that a large part of the content cite relevant sources. Citing sources helps ascertain the origin of each piece of information and therefore, these actions make the article more trustworthy. Hence, we use the proportion of paragraphs with citations as a relative indicator of the citation quality of the article. The higher this number, the better sourced the article is.

Finally, we look at features from the Appearance category. Content size is an external characteristic that has been found to be useful in judging Wikipedia article quality [7] as larger size could symbolize the content creator's efforts as well as the content's comprehensiveness, thereby indicating trustworthiness. In addition, an article with content divided into logical sections can be considered well structured. Hence, features based on the structure (sections and paragraphs) could be useful. However, they are excluded as they are similar in nature to content size and our tests indicate that the latter is a marginally better predictor.

5.1.2 Daily Strength

Communications in online social networks are another major area where trust assessment is particularly relevant. Daily Strength is an online health social network that facilitates user discussions, allowing them to connect with others who share similar health conditions and seek advice and suggestions regarding medical conditions, drugs, treatments and doctors and gain some much needed emotional support from others. Unlike the Wikipedia dataset, this data does not contain any manual evaluations or classification of responses to act as the ground truth for comparison. To provide such a ground truth, we conducted a survey where users were asked to evaluate the trustworthiness of the responses.

For this study, we select data from an Autism-Autism-Spectrum support group. From the numerous discussions on the site, we selected several threads with five to eight responses where advice was solicited. Thirty nine participants were recruited to take part in the survey. Each participant evaluated every response in the discussions allotted to them. Over two hundred discussions were used in the survey with each discussion assigned to 3 different participants. Out of these, only those discussions that had all three evaluations were used for the final study which resulted in 156 discussions with 853 responses. For each response, the participants were asked to classify the responses as trustworthy (702), untrustworthy (32) or unclear (119). The consensus was used to categorize the data. Responses that did not receive a consensus were also placed in the unclear category. Next, we describe the features and the intuition behind their selection.

A suitable feature indicating Reputation for this dataset is related to the user's social network activities. We measure the social connectedness of the author in the social network through a simple measure that averages number of friends for each friend that the author has in the network. A larger number indicates that the author is more well-connected and therefore more credible.

The second feature for this dataset is the number of internal links and it falls in the Performance category. When responses contain names of healthcare related terms such as drugs and treatments, links to pages related to these terms on the website are automatically added. Such internal links indicate the usage of healthcare terms mean that the content is related to health issues and not just responses that provide emotional support or make conversation. In addition, it can also be used to detect trustworthiness as the usage of such terms depicts the intent of the user in developing useful content. Links to external sources can also help ascertain the source of information. As in the case of Wikipedia, where references and citations were useful in selecting suitable trust indicators, external links are useful features for the assessment of trustworthiness in these responses.

For the Appearance category, content size is once again a suitable feature. Considering the informal nature and brevity of advice shared on such sites, we do not select any features based on content structure or language style.

Table 1 Wikipedia Trust Scores: Model Performance with All Features

NDCG Performance				
Model	Top 50	Top 100	Top 150	All
RBS	0.0014	0.0195	0.0412	0.4500
EBS	0.2096	0.3315	0.4353	0.5943
SAS	0.7789	0.8879	0.8927	0.9000
DDS	0.8344	0.9286	0.9340	0.9398
CRS	0.8347	0.9300	0.9345	0.9404
WCRS	0.8851	0.9487	0.9450	0.9507

5.2 Experimental results

5.2.1 Wikipedia

Table 1 presents the NDCG performance results for this dataset. As expected, the average case EBS baseline model performs much better than the worst case RBS model (0.2096 vs. 0.0014 for the top 50 articles and 0.5943 vs. 0.45 for all articles). It is easily apparent that our models work much better than the EBS model which assumes that all articles are equally reliable. When all features are used, the WCRS model produces the best performance for $k = 50$ with an NDCG of 0.8851. The CRS model is a distant second with an NDCG of 0.8347 followed by the DDS model with 0.8344. The same trend is observed for other values of k though the magnitude of the difference is reduced as k increases. These three models perform much better than the simpler SAS model, which serves to indicate the level of performance improvement achieved. While not comparable due to differences in the data used, it is interesting to note that the quality models proposed by Hu et al. [7] result in lower NDCG (<0.9 for all articles, $k = 242$ and <0.8 for $k = 45$) with a similar number of Wikipedia articles.

To illustrate the success of these models and to compare them, we provide another view of the results from this experiment. The trust scores are normalized between 0 and 100 and discretized into bins of equal frequency. The proportion of data from a certain class that falls into a bin is represented by a suitably sized bubble and plotted in Fig. 3 for the WCRS and DDS models. In both cases, it is easily apparent that the trust scores clearly differentiate between the article classes. More importantly, the figures show the difference in performance between the two models. The WCRS model performs much better with trust scores showing limited overlap between classes while the trust scores from the DDS model show a greater spread due to the inaccurate scoring of few data points, resulting in a larger overlap between the classes. This useful illustration provides a much clearer picture of the difference in the performance of these two models that is not as tangibly perceptible from the NDCG performance values.

In the past, some studies have found that particular features such as article size can improve the performance of some quality assessment models [7]. One study has even used this feature exclusively for quality assessments [4]. However, an approach that is overly dependent on one feature is more a reflection on the usefulness of the feature

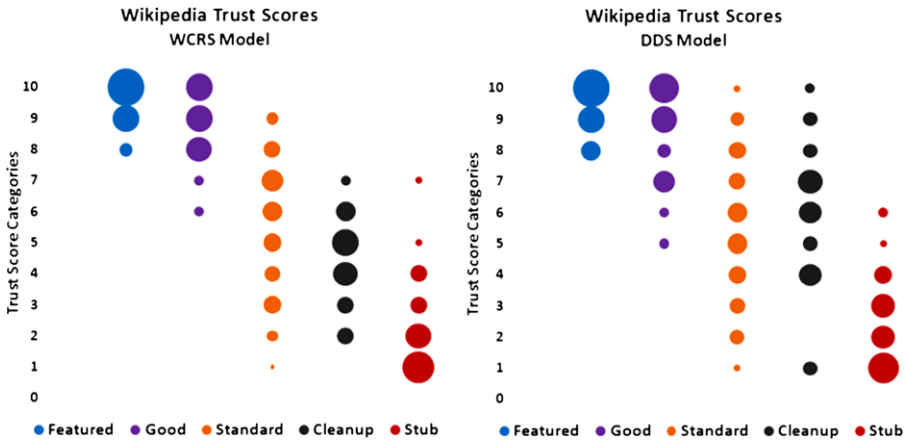


Fig. 3 Distribution of Wikipedia Trust Scores

than on the success of the approach. Further, such an approach might not work well when extended to other social media as we intend to do. Another problem is that a model dependent on one or two features would provide vandals with an opportunity to manipulate content in an attempt to game the system. Therefore, a good model should not be overly dependent on any feature. To test whether the performance of our models is dependent on any one individual feature, we exclude one feature at a time and test the performance. A part of these results is tabulated in Table 2.

When article size is excluded, the performance drops very slightly with the DDS model. With the CRS and WCRS models, the performance actually shows a mild improvement. This clearly depicts that article size, while useful, is not a blockbuster feature that a trust prediction model for Wikipedia cannot do without as results from some previous studies seem to indicate. The exclusion of other features shows no major changes in performance. The largest performance drop is observed when the reference score per paragraph feature is excluded and even then, the performance is much better than that of the SAS model using all features, which we use to indicate the minimum reasonable performance level. This insignificant change in performance clearly demonstrates that the success of our approach is not dependent on any single feature. Instead, it is based on the use of multiple useful predictors that are powerful trust indicators when used together.

As many existing studies utilize limited classes of information [4, 6], it may not be suitable to compare our results with them. The reason for this is that some of the classes of information used are very close to each other. A major difference does not exist between featured and good articles or between standard and cleanup articles. Hence, there is a greater chance of the predicted trustworthiness being out of place. This is akin to misclassification between similar data classes. To provide a better comparison with existing studies that use limited classes, we perform a set of experiments by eliminating some of the classes, as depicted in Table 3.

An improved performance is observed, as expected, with the removal of good and cleanup articles. Due to the elimination of these classes, which are very close in terms of reliability to featured and standard articles respectively, there are fewer

Table 2 Wikipedia Trust Scores: Model Performance with Some Features Excluded

NDCG Performance					
Excluded Features	Model	Top 50	Top 100	Top 150	All
None	DDS	0.8344	0.9286	0.9340	0.9398
	CRS	0.8347	0.9300	0.9345	0.9404
	WCRS	0.8851	0.9487	0.945	0.9507
Article Size	DDS	0.8347	0.9273	0.9328	0.9388
	CRS	0.8380	0.9326	0.9368	0.9428
	WCRS	0.8902	0.9531	0.9491	0.9550
Reference	DDS	0.7941	0.9106	0.9160	0.9219
Score Per	CRS	0.8293	0.9271	0.9283	0.9375
Paragraph	WCRS	0.8635	0.9339	0.9299	0.9366
Proportion of Paragraphs with Citations	DDS	0.8093	0.9253	0.9262	0.9362
	CRS	0.8279	0.931	0.9319	0.9412
	WCRS	0.8773	0.9435	0.9440	0.9496
Revision	DDS	0.8142	0.9269	0.9319	0.9382
Count	CRS	0.8364	0.9287	0.9335	0.9393
	WCRS	0.8900	0.9492	0.9453	0.9512

Table 3 Wikipedia Trust Scores: Model Performance with Fewer Classes

NDCG Performance			
Comparison	Model	Top 50	All
Featured, Standard, Stubs	DDS	0.9666	0.9676
	CRS	0.9725	0.9733
	WCRS	0.9768	0.9774
Featured, Standard	DDS	0.9666	0.9676
	CRS	0.974	0.9751
	WCRS	0.9799	0.9804
Featured, Stubs	DDS	1	1
	CRS	1	1
	WCRS	0.9997	0.9997

misclassifications. Further, removing stubs shows little improvement in performance as they are easily differentiable from the remaining two classes. Finally, a maximum possible NDCG of 1 is observed when only featured articles and stubs are used.

5.2.2 Daily Strength

Table 4 depicts the results from our experiments with the Daily Strength dataset. As earlier, the RBS model has the worst performance while the EBS model performs much better. The NDCG is reasonably high due to the fact that only 3.8% of the re-

Table 4 Daily Strength Trust Scores: Model Performance with All Features

NDCG Performance				
Model	Top 100	Top 200	Top 300	All
RBS	0.1391	0.4617	0.6044	0.8989
EBS	0.8908	0.8908	0.8908	0.978
SAS	0.9504	0.9339	0.9267	0.987
DDS	0.9862	0.9764	0.9704	0.9927
CRS	0.9816	0.9795	0.9762	0.9937
WCRS	0.9938	0.9911	0.982	0.9958

sponses are unreliable and only 14.49% of the responses are classified as unclear. Due to the highly skewed nature of the dataset, with the presence of a high proportion of trustworthy responses that result in high gains, high NDCG values are observed even for baseline models when many articles are considered. As a result, a near perfect performance is expected from our models.

Such performance is observed from the three models. There is little to choose between the models but the WCRS model once again proves to be the best. In comparison to the baselines, there is considerable improvement with all models, especially for the top few hundred articles. Despite the high values for the NDCG performance when all articles are considered, there is still a considerable improvement with our models over the baseline scores as well as a notable improvement over the SAS model. While, the skewed nature of the data makes it a relatively difficult dataset to analyze, our models once again prove useful in predicting the trustworthiness of these user responses. A visual plot of this performance is generated as before for the WCRS model and depicted in Fig. 4. While there is a clear difference between the classes, the spread of trust scores is wider and results in greater overlap across classes. This is primarily an effect of the highly skewed data and partly an artifact of the equal frequency discretization.

We once again test the dependence of our models on a particular feature by excluding one feature at a time and performing our experiments. These results are presented in Table 5. Once again there are no major changes observed when any single feature is excluded. There is either a small drop or increase in performance. The largest performance drop is observed when the response size is excluded, which is still better than the performance of the SAS model using all features. As has been indicated in previous studies on social media, content size is once again a useful predictor. However, the models are still very resilient despite the exclusion of a useful predictor. This is especially notable because the total number of predictors is quite small in this case. These experiments once again indicate the power of using multiple useful predictors and the robustness of our approach to the exclusion of useful features. While the data limits us to an extent, these results are nonetheless promising and depict the ease with which this approach was applied to two dissimilar social media.

Fig. 4 Distribution of Daily Strength Trust Scores

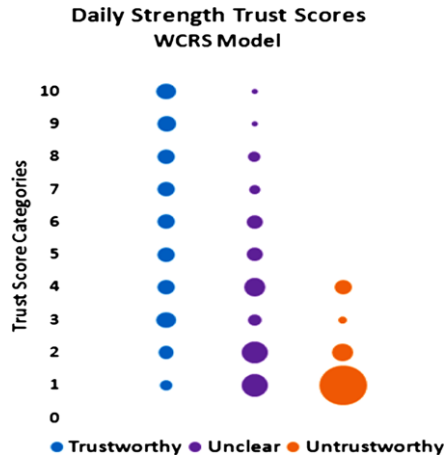


Table 5 Daily Strength Trust Scores: Model Performance with Some Features Excluded

NDCG Performance					
Excluded	Model	Top 50	Top 100	Top 150	All
None	DDS	0.9862	0.9764	0.9704	0.9927
	CRS	0.9816	0.9795	0.9762	0.9937
	WCRS	0.9938	0.9911	0.9820	0.9958
Response Size	DDS	0.9691	0.9644	0.9508	0.9895
	CRS	0.9611	0.9479	0.9365	0.9881
	WCRS	0.9940	0.9776	0.9688	0.9936
External Links	DDS	0.9842	0.9747	0.9728	0.9924
	CRS	0.9871	0.9754	0.9698	0.9928
	WCRS	0.9877	0.9866	0.9795	0.9946
Internal Links	DDS	0.9796	0.9703	0.9581	0.9922
	CRS	0.9747	0.9652	0.9581	0.9915
	WCRS	1.0000	0.9967	0.9696	0.9953
Author Social	DDS	0.9783	0.9813	0.9730	0.9931
Connectedness	CRS	0.9836	0.9786	0.9748	0.9935
	WCRS	0.9935	0.9866	0.9838	0.9960

6 Conclusions

Social media has quickly garnered tremendous popularity in recent years. User-developed content shared on social media can be quite useful for individuals to learn and improve their knowledge on various topics and such content is being utilized prominently by consumers. Health information on the web, in particular, can be beneficial for patient education, health promotion and preventive care. With the rising popularity of social media and the widespread use of user-generated content, there

is a pressing need for the assessment of trust to guide users and prevent harm from inaccurate information.

In this work, we aim to solve an important content assessment problem. We provide illustrative examples that distinctly depict the need for trust assessment of social media content. Unlike existing work that focus on a single application, we aim to solve this problem with a generalized solution that can be adapted with ease for trust assessment across diverse social media applications. To achieve this, we divide the problem into two tasks: feature identification and quantification. We build our approach such that the second task remains constant and the first task is repeated across various applications. We develop an unsupervised approach as there is limited or no content class information available with respect to trustworthiness. We design it to be feature-driven such that it draws power from the use of multiple pertinent predictors in order to be a better indicator of trust and robust enough for use in the presence of limited features or in the absence of some useful features.

One of the important contributions of this work is in adapting existing sociological theories on trust to this domain of informational trust for user-generated content. We define the problem, describe sociological notions of trust and use them to provide a broad framework to identify relevant features. We propose multiple trust scoring models and test them using data from two diverse social media applications. Promising experimental results render our approach and models sound. The experiments also depict that the models draw power from the use of multiple useful predictors and are not dependent on a single blockbuster feature. Further, this makes it more difficult to manipulate articles in order to achieve higher scores. In addition, the models can work well with limited features. The generality of our approach in that it can be easily adapted and applied across social media is its most distinguishing characteristic.

Acknowledgements This work is sponsored, in part, by grants from ONR (N000140810477) and AFOSR (FA95500810132).

References

1. Adler, B., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I., Raman, V.: Assigning trust to Wikipedia content. In: 4th Intl Symposium on Wikis, Wikisym 2008. (2008)
2. Agichtein, E., Castillo, C., Donato, D., Gionis, A., Mishne, G.: Finding high-quality content in social media. In: Proc. of the Intl. Conf. on Web Search and Web Mining, pp. 183–194. (2008)
3. Bailey, B.P., Gurak, L.J., Konstan, J.A.: Trust in cyberspace. In: Ratner, J. (ed.) Human Factors and Web Development, 2nd ed., pp. 311–321. Lawrence Erlbaum, New Jersey (2002)
4. Blumenstock, J.: Size matters: word count as a measure of quality on Wikipedia. In: Proc. of the 17th Intl. Conf. on World Wide Web (WWW) 2008, pp. 1095–1096. ACM, New York (2008)
5. Childs, S.: Judging the quality of Internet-based health information. *Perform. Meas. Metr.* **6**(2), 80–96 (2005)
6. Dondio, P., Barrett, S.: Computational trust in web content quality: a comparative evaluation on the Wikipedia project. *Informatica* **31**(2), 151–160 (2007)
7. Hu, M., Lim, E.P., Sun, A., Lauw, H.W., Vuong, B.Q. Measuring article quality in Wikipedia: models and evaluation. In: Proc. of the 16th ACM Conf. on Information and Knowledge Management, CIKM 2007, pp. 243–252. (2007)
8. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Sys.* **20**(4), 422–446 (2002)
9. Korp, P.: Health on the Internet: implications for health promotion. *Health Educ. Res.* **21**(1), 78–86 (2005)

10. Liu, H., Hussain, F., Tan, C., Dash, M.: Discretization: an enabling technique. *Data Min. Knowl. Discov.* **6**(4), 393–423 (2002)
11. McGuinness, D., Zeng, H., Da Silva, P., Ding, L., Narayanan, D., Bhaowal, M.: Investigations into trust for collaborative information repositories: a Wikipedia case study. In: *Proc. of the Workshop on Models of Trust for the Web 2006*, pp. 3–131. (2006)
12. McSherry, F., Najork, M.: Computing information retrieval performance measures efficiently in the presence of tied scores. *Lect. Notes Comput. Sci.* **4956**, 414–421 (2008)
13. Siegrist, M., Cvetkovich, G.: Perception of hazards: the role of social trust and knowledge. *Risk Anal.* **20**(5), 713–720 (2006)
14. Sztompka, P.: *Trust: A sociological theory*. Cambridge Univ Press, Cambridge (1999)
15. Tan, P., Steinbach, M., Kumar, V.: *Introduction to data mining*. Addison-Wesley/Longman, Boston (2005)
16. Zeng, H., Alhossaini, M., Ding, L., Fikes, D., McGuinness, D.L.: Computing trust from revision history. In: *Proc. of the 2006 Intl. Conf. on Privacy, Security and Trust* (2006)