Quantifying Utility and Trustworthiness for Advice Shared on Online Social Media

Sai T. Moturu
Computer Science and Engineering
Fulton School of Engineering
Arizona State University
Tempe, AZ 85282
Email: smoturu@asu.edu

Jian Yang
Computer Science and Engineering
Fulton School of Engineering
Arizona State University
Tempe, AZ 85282
Email: jyang20@asu.edu

Huan Liu
Computer Science and Engineering
Fulton School of Engineering
Arizona State University
Tempe, AZ 85282
Email: huan.liu@asu.edu

***Abstract*—The growing popularity of social media in recent years has resulted in the creation of an enormous amount of user-developed content. While information is readily available, there is no easy way to find the most useful content or to detect whether it is trustworthy. A casual observer might not be able to differentiate between the useful and the useless or the trustworthy and the untrustworthy. In this work, we wish to study the problem of quantifying the value of such user-shared content. In particular, we are focussed on health content as the negative impacts are higher for this domain. We use advice shared on a health social network, Daily Strength, for this study. We describe and define the notions of trustworthiness and utility for social media content. We identify the necessity and challenges for their assessment, and propose a framework that helps address these challenges by identifying relevant features and providing empirical means to meet the requirements for such an evaluation. We select relevant variables and perform numerous experiments to evaluate our models. The results demonstrate promising performance that could possibly be replicated with other social media applications.**

I. Introduction

The proliferation of Social Media portals in recent years has resulted in a torrent of user-shared content on the web. Articles about politics, history, and health are available on wikis and blogs. Advice is being solicited on question-answer sites and social networks. Opinions are being shared 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? How useful is it? While search engines can provide relevant results, they cannot provide an answer to these questions. Nevertheless, search engines still serve as the starting point for knowledge seekers. To quantify the data, one has to depend on their own intellect, knowledge and analytical capabilities but this task is not simple.

Relevant information found on search engines may not always be useful or trustworthy. As a motivating example, consider the search for "How to prevent Restless Leg Syndrome", performed on June 18, 2008 on Google. 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. 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 useful. 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. It has to be understood that search relevance does not imply content reliability (trustworthiness) or usefulness (utility). 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 content that is contributed by unknown authors.

The identification of useful content from the voluminous amount of user-generated information is another major issue. Finding the best content is a time consuming task that is not always successful. That brings about the need for ways to automate the quantification of such information in terms of attributes such as utility and trustworthiness. The presence of such assessments can change the way people perceive and utilize information from social media.

In particular, we focus on shared health content as the negative impact of acting on untrustworthy content or not finding useful content is high in this domain. In this paper, we use data from Daily Strength, a web-based health social network where one of the benefits is that users can pose questions in related forums seeking advice and suggestions. While advice is provided by different users, the value of such advice or its reliability is not easily apparent. We identify relevant features and provide an intuitive scoring measure to quantify the value and trustworthiness of content. Such quantification is useful for participants in the discussion as well as for visitors who are looking to uncover useful information.

II. Related Work

A considerable number of works in recent years have been devoted to studying various aspects of Social Media. Here, we list a subset of these focusing on the quantitative assessment of Social Media content. Hu et al. [1] base their study of article quality on the assumption that revisions involve peer review of at least part of the content. Dondio and Barrett [2] use objectivity, completeness and pluralism as the hallmarks of good information. McGuinness et al. [3] base their assessment

of trust on the occurrences of the encyclopedia term in an article. The same group also studied the possibility of using revision history to assess trust using a dynamic Bayesian network[4]. Revision history has also been used to assess and depict the varying trustworthiness of different parts of the text of a Wikipedia article [5]. Agichtein et al. [6] embark upon the task of quality assessment in Social Media using data from a community question/answer domain.

While many previous studies have discussed the issue of quality, the perspective of quality might differ. In certain situations, trust and quality are used interchangeably but this is inaccurate [7]. In this paper, we focus on trustworthiness and utility of the shared content, with its quality being an important aspect in both cases. We separate our work into two tasks - feature identification and scoring trustworthiness and utility. We propose a simple hierarchy of feature categories from which relevant features can be extracted, not only for this domain, but also for other social media. We design unsupervised trust evaluation models, that are independent of the application, to generate trust scores. This means that the selected features ultimately drive trust and utility scores. Since these models can work with any set of features, they ensure the possibility of extending this work to other social media.

III. Trust and Utility

A. Trust

Trust is an important sociological concept that has been studied in depth by many researchers for a number of years. Trust can be of different types, focussed on numerous targets [8]. For the purpose of this paper, we rely on the terms *trust* and *trustworthiness* to focus on the reliability of information shared in social media.

Trust is a concept involving in a transaction 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 paper, we will only consider the perspective of the trustor, which defines this property to be the amount of trust associated with the trustee [9].

With limited personal knowledge and relationships built on virtual interactions, trust is hard to assess in cyberspace. This necessitates the creation of a trust indicator that can aid an informed decision. In the following subsections, we delineate the aspects to be considered for trust assessment.

1) Quality: Quality, in the sense used here, represents an inherent feature or essential character [10]. For any information, predictors derived from intrinsic aspects of the content can be used to define its quality. Positive predictors improve quality while negative ones reduce it. Quality is sometimes used interchangeably with trust in the context of content evaluation. Though associated, these are not identical issues [7]. In the context used for this article, quality is only one aspect of trust.

2) Credibility: Credibility is the quality of inspiring belief [11]. Factual accuracy is a suitable property of reliable content. However, content may not be enough to assess article reliability because it might not contain all the information necessary to draw conclusions. When combined with external data, however, such conclusions become possible. External cues can include information on editing patterns, development history and user behavior. Predictors derived from such metadata associated directly or indirectly with the content but not from the content itself measure the credibility of an article from the perspective of its development, deployment and response.

B. Utility

Like trust, utility is also a concept that has been studied for a long time by sociologists and economists. For the purpose of this paper, we rely on the terms *utility* and *usefulness* to focus on the value of a response contributed by a user with respect to the question asked by another.

Utility is a concept that refers to the benefit or satisfaction derived from the utilization of a commodity [12]. Here, the commodity is the user response and the benefit is the perceived value of the response with respect to answering the question suitably. Utility theory deals with an individual's preferences or values and the assumptions that enable their numeric representation [13]. Measurement is essentially the assignment of numbers to entities and utility measures are choice indicators that denote the value of an entity numerically[14].

In today's virtual world, an individual is presented with an exhaustive number of choices in the quest for knowledge but it is a difficult task to pick the most suitable. In some cases, such choices may never come to the fore due to the sheer magnitude of information. This necessitates the creation of a utility indicator. In the following subsections, we delineate the aspects to be considered for the assessment of utility.

1) Quality: The definition of quality remains unchanged. As in the case of trust, quality is only one aspect of utility. Though quality is a common aspect in both cases, features in this category need not be relevant for the evaluation of both utility and trust. While some features are useful for detecting trustworthiness, others maybe only indicative of utility.

2) Pertinence: Pertinence is the quality or state of having a clear decisive relevance to the matter at hand [15]. In the case of user advice, looking at just the quality of the response is not sufficient. A user response may be of high quality without necessarily answering the question or even being relevant to it. Hence, it is necessary that advice is not only of high quality but also pertinent. A user response is pertinent if it is relevant to the matter at hand, which is the question asked. Features indicative of pertinence are derived not only based on the content in the responses but also on the content in the question.

IV. Data Collection

A. Daily Strength Data

Daily Strength is an online health social network where users can maintain friendship networks, discuss their conditions, ask for advice, share opinions and experiences regarding drugs, treatments or doctors and gain some much needed emotional support. For this study, we select data from an

TABLE I
DATA DISTRIBUTION IN UTILITY CATEGORIES

Category	Highest	High	Medium	Low	Lowest
Responses	135	324	241	117	36

TABLE II
DATA DISTRIBUTION IN TRUST CATEGORIES

Category	Trustworthy	Unclear	Untrustworthy
Responses	702	119	32

Autism-Autism-Spectrum support group. This group consists of over 2500 members, who are either patients themselves or parents and relatives of patients. From the forums where advice is solicited, we select numerous threads with five to eight responses. Quantitative assessment of this content presents a challenge due to the lack of a suitable ground truth to compare against. A suitable solution is to perform a manual assessment of the data to reveal the ground truth which can later be used for evaluation.

B. User Evaluation

Thirty nine participants were recruited to take part in the manual assessment. 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 (excluding responses from the individual asking the question). For each response, the participants were asked to rate the response in terms of its usefulness from 1 to 5 (1 being the lowest). The scores from the three participants were then averaged and distributed into the five categories. The data distribution is presented in Table I.The participants were also asked to classify the responses as trustworthy, untrustworthy or unclear for every response. The consensus was used to categorize the data. No consensus was achieved for 45 of the responses. These were also placed in the unclear category. The data distribution is presented in Table II.

V. QUANTIFYING TRUSTWORTHINESS AND UTILITY

Our approach to quantifying trust and utility is divided into three major tasks. The first task is the identification of relevant features capable of assessing the quality, credibility and relevance of content and contributors. Next is the creation of a feature-driven scoring model that is independent of the application. The final task is the performance evaluation of these models. We detail these tasks in the upcoming subsections.

A. Features

As discussed earlier, features can be extracted from content and metadata. The identification of such features is a critical part of the evaluation of trust and utility. In the following subsections, we identify useful features, provide the intuition for their selection and discuss their trends with respect to trust and utility classes. For each feature, we judge its statistical significance in the differentiation between trust or utility categories by performing the Kruskal Wallis test for a non-parametric one-way analysis of variance. Only features showing significant differences are used for further analysis.

1) Quality: Features that ascertain the quality of information via the appraisal of information provenance and content characteristics are included in this category. A feature that can help assess information provenance is the presence of external links. In general, we do not expect content in the question-answer domain to be well-sourced as the responses are more likely to be based on personal experiences and opinions. However, when the responses include factual information, external links can provide relevant references and information sources for the shared content. Suitably referenced content is of higher quality than content where there is no way of ascertaining the source of information as the former is more trustworthy and possibly more useful.

In addition, external links could point to useful content and resources that might be useful in answering a question. As per our expectations, a significant difference ($p<0.001$) is observed in the number of external links between the various utility categories. However, there is no significant difference ($p=0.358$) in the number of external links between the trust categories. This observation could be due to the unclear reliability of some of the external links. Based on these results, this feature is only used for quantifying utility.

Another useful feature, derived from the content characteristics, is the size of the response. A larger response size could symbolize the effort made by the authors towards their contribution and would therefore indicate quality of content that is useful in assessing both trustworthiness and utility. Content size has previously been found to be useful for the prediction of Wikipedia article quality [1]. Here too, a significant difference ($p<0.001$) is observed between the reply sizes for both trust and utility categories.

The third feature in this category is the number of internal links. 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 and are indicative that the content being discussed 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. As expected, a significant difference ($p<0.001$) is observed in the number of internal links for both categories.

2) Credibility: Features in this category include those that determine author credibility. The first feature in this category is the number of friends for the author in the social network. An author with a larger number of connections is expected to be more credible as he has some reputation to maintain. However, a significant difference ($p=0.603$) in this number is not seen between the trust categories. While our intuition is acceptable, most of the users in the selected forum are expected to be credible due to their health conditions and therefore, it might

be difficult to perceive a difference. While it may be useful in another application, it is not the case here.

The second feature is related to the connectedness of the author in the social network. A simple measure for this is the average 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. A significant difference in this value (p=0.029) is observed between the trust categories. One possible reason for this result could be that less connected users are relatively less knowledgeable on the health issues involved and might therefore contribute untrustworthy content, making them less credible.

Two features that derive credibility from author contributions and responses to them are the number of journal entries by the author and the number of replies to them. While these features may delineate regular contributors with useful contributions from those who do not make contributions, no significant difference in their values is seen (p=0.314, p =0.604) for the trust categories.

3) Pertinence: The features in this category indicate the relevance of the response with respect to the question. The first feature is text similarity. This feature measures the similarity of the response to the query using term frequency-inverse document frequency (TF-IDF). The intuition here is that if the terms used in the response match some of those in the question, it indicates that the response is discussing the same topics and is therefore relevant. The greater the similarity, the higher is the relevance. A significant difference ($p<0.001$) is observed for text similarity between utility categories.

The next feature in this category is keyword similarity. As discussed earlier, internal links are created for health-related keywords. The number of such keywords featured in both the question and response is the value this feature. The intuition is the same as text similarity. The higher the similarity, the higher is the relevance. A significant difference ($p=0.035$) is observed for keyword similarity between utility categories.

B. Scoring Models

1) Reverse Baseline Score: The Reverse Baseline Score (RBS) is a simple 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.

2) Equal Baseline Score: The Equal Baseline Score (EBS) is another baseline approach that represents the average case. In this case, each article is assigned the same arbitrary score. As all articles are of equal importance, the performance of this model would always be much better than the RBS model. The motivation to use the EBS model is enhanced due to the fact that an unscored set of articles seem of equal value to a user. Our intent is to come up with a scoring system that allows the user to select the most trustworthy content and a model that performs better than the EBS model will serve that need.

3) Dispersion Degree Score: In the Dispersion Degree Score (DDS) model, each feature contributes a score s_{ij} towards the aggregated feature score S_{ij}. The dispersion of a feature value from its mean is utilized to derive its relative importance. The underlying assumption is that the farther a feature value is from its mean, the greater its effect on the quantity being scored. Eqs. 1 and 2 describe the model. 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 d_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 scores from each feature provides the final score, $S_D(i)$, with a larger value indicating a better response (in terms of utility or trustworthiness, depending on the situation).

$$s_{ij} = \begin{matrix} 0 \; if \; f_{ij} < m_i - d_i \\ x+1 \; if \; m_i - d_i + cxd_i < f_{ij} < m_i - d_i + c(x+1)d_i \\ 11 \; if \; f_{ij} < m_i + d_i \end{matrix} \tag{1}$$

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

C. Evaluation: Normalized Discounted Cumulative Gain

The popular Normalized Discounted Cumulative Gain (NDCG) evaluation metric [16] is used to evaluate the performance of our models. The measure 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 predictions of Wikipedia articles [1]. The trust and utility scores output from each of our models can 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.

$$DCG_k(S_m) = \sum_{r=1}^{k} \frac{2^{s(r)} - 1}{\log_2(1+r)} \tag{3}$$

$$NDCG_k(S_m) = \frac{DCG_k(T_m)}{DCG_k(T_p)} \tag{4}$$

$$DCG_k(S_m) = \sum_{r=1}^{k} \left(\left(\frac{1}{n_i} \sum_{j=t_i+1}^{t_{i+1}} (2^{s(r)} - 1) \right) \sum_{j=t_i+1}^{min(t_{i+1},k)} \frac{1}{\log_2(1+r)} \right) \tag{5}$$

Eq. 3 is used to calculate the discounted cumulative gain (DCG) for the top k articles. The numerator in Eq. 3 defines the gain where $s(r)$ denotes the score for an article ranked r. Consider a case where the scores used for two classes are 10 and 1 with the score differences representing the proximity of the classes. Hence, the gain for an article from the top class is $2^{10}-1$ but 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 Eq. 3 is used to discount gain as the rank increases. Discounted gain for an article from the top class 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. The NDCG function

TABLE III
UTILITY EVALUATION

	NDCG			
	Top 100	Top 200	Top 400	All
RBS	0.002	0.009	0.033	0.646
EBS	0.262	0.324	0.453	0.770
DDS	**0.519**	**0.557**	**0.680**	**0.858**

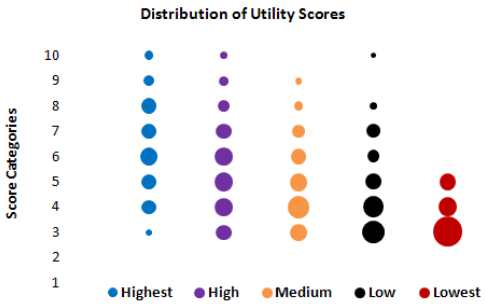

Fig. 1. Distribution of utility scores

in Eq. 4 normalizes the DCG value calculated from Eq. 3 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. Tied scores mean that multiple possibilities exist for result ordering. McSherry and Najork [17] proposed an efficient way to average the performance across all possible orderings in such cases. Eq. 5 defines the new discounted cumulative gain function that averages the gain across each position in a tied group. The NDCG formula in Eq. 4 remains the same and the normalization factor in the denominator does not change as a result of the new DCG function. We use this tie-oblivious NDCG in our evaluations.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

A. Utility

To test the value of our approach to quantifying utility, the three scoring models are evaluated using the tie-oblivious NDCG measure. All 853 data points are used in this experiment. The scores used in the NDCG measure for each class are 10 (highest), 8 (high), 5 (medium), 2 (low) and 1 (lowest). By definition, we expect a high NDCG value when the best responses are ranked highly and a low value otherwise.

Table III depicts the results from this experiment. The first model, RBS presents the worst case scenario where all the documents are ranked in reverse. This would result in the lowest possible NDCG as the best responses are ranked at the bottom. An NDCG of less than 0.05 is observed when as many as the top 400 articles are considered. This value increases to 0.646 when all articles are considered as the bottom half includes the most useful articles which generate high gains. Another contributing factor is that 53.9% of the data points belong to the top two classes that generate high gains. The next model, EBS represents an average case scenario where all the responses are ranked equally. We observe that the performance is considerably better when compared to the RBS model while considering the top ranked articles. The difference is less when considering all articles due to the aforementioned reasons.

As all responses are ranked equally, The EBS model represents the default situation for the user, where there is no way to distinguish one response from another. The final model, DDS is expected to perform better than the EBS model to be useful. This is precisely the observation from the results with the DDS model showing much better NDCG performance when considering the top articles, with the difference decreasing as the number of articles under consideration increases.

To illustrate the success of the DDS model, we provide another view of the results from this experiment. The utility scores are separated into ten bins of equal width. The proportion of articles from each class falling into these bins (indicated by the bubble size) is calculated and illustrated in Figure 1. Sixty percent of the responses in the Lowest category and 33.33% from the Low category fall into the bin at the bottom. In contrast, only 2.2% of the responses from the Highest and 13.89% from the High category fall into this bin. On the other hand, 46.67% of responses from the Highest category and 27.78% from the High category fall into the top four bins while only 12.82% from the Low category and 0% from the Lowest category fall here (note that the distribution of the most useful articles into multiple bins instead of one is an artifact of equal width binning). This illustration presents a clearer picture of the distribution of predicted scores and the utility of the DDS model. These results are impressive. The simple and intuitive DDS model shows promise and depicts the usefulness our approach to feature identification and utility measurement.

B. Trust

As with the quantification of utility, the three scoring models are evaluated using the tie-oblivious NDCG measure for trust. All 853 data points are used in this experiment. The scores used in the NDCG measure for each class are 4 (trustworthy), 2 (unclear), 1 (untrustworthy). Table IV depicts the results from this experiment. As earlier, RBS depicts the worst performance. Unlike the earlier case, the NDCG observed for even the top 400 articles is reasonably high (0.683). This is due to the fact that only 3.8% of the responses are unreliable and only 14.49% of the responses are classified as unclear. Due to the presence of high proportion of trustworthy responses that result in high gains, high NDCG values are observed even for RBS and EBS models when many articles are considered. The NDCG for the EBS model is the same for the top 100, 200 and 400 articles due to the presence of over 700 trustworthy articles. Despite the high values, the DDS model performs much better than the EBS model in relative terms, especially when considering the top ranked articles.

TABLE IV
TRUST EVALUATION

	NDCG			
	Top 100	Top 200	Top 400	All
RBS	0.139	0.462	0.683	0.899
EBS	0.891	0.891	0.891	0.978
DDS	**0.984**	**0.975**	**0.955**	**0.992**

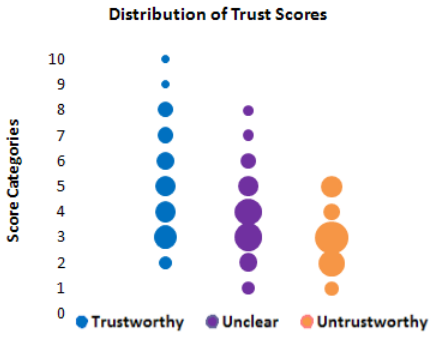

Fig. 2. Distribution of trust scores

To illustrate the success of the DDS model, we use the bubble representation in figure 2. The trust scores are separated into ten bins of equal width and the proportion of articles from each class falling into these bins (indicated by the bubble size) is calculated and illustrated. Nearly thirty-three percent of the responses in the Untrustworthy category and 17.57% from the Unclear category fall into the two bins at the bottom. In contrast, only 6.86% of the responses from the Trustworthy category fall into this bin. On the other hand, 36.14% of responses from the Trustworthy category fall into the top five bins while only 13.51% from the Unclear category and 0% from the Untrustworthy category fall here (note that the distribution of the most trustworthy articles into multiple bins instead of one is an artifact of equal width binning). This illustration reiterates the usefulness of the DDS model using a depiction of the distribution of predicted scores. While the highly skewed nature of the data limits us to an extent, these results are promising nonetheless.

VII. CONCLUSION

With the advent of social media and user-generated content, there is a pressing need for content assessment to guide users toward useful content and prevent harm from inaccurate information. In this paper, we identify and study the critical problem of quantifying trustworthiness and utility for advice shared on a health social network. We describe the problem, define the notions of trust and utility in terms of quality, credibility and pertinence and provide a framework to identify relevant features. We propose an intuitive model to quantify the utility and trustworthiness of content. We test this model using appropriate evaluation methodologies and compare the results against two suitable baselines. Promising performance renders our approach and models sound.

As our approach is feature-driven and application independent, extensions to other social media applications would only require appropriate feature identification using relevant assumptions. While a one-size-fits-all solution for the entire social web is difficult to accomplish, our framework could possibly be used across social media applications to quantify trustworthiness and utility. Currently, this approach has been successfully used to quantify the trustworthiness of Wikipedia articles. A major roadblock to the extension of our work is the lack of a suitable ground truth for many social media applications. That challenge was addressed in this work through manual evaluation of data and a similar approach can be used in the future as well. While we present a simple model to quantify trust and utility here, we intend to refine and update our existing models in the future with an eye on performance improvement and also hope to extend this work to different social media applications.

ACKNOWLEDGMENT

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

REFERENCES

[1] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong, "Measuring article quality in wikipedia: models and evaluation," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM New York, NY, USA, 2007, pp. 243–252.
[2] P. Dondio and S. Barrett, "Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project," *Informatica*, vol. 31, no. 2, pp. 151–160, 2007.
[3] D. McGuinness, H. Zeng, P. Da Silva, L. Ding, D. Narayanan, and M. Bhaowal, "Investigations into trust for collaborative information repositories: A wikipedia case study," in *Proceedings of the Workshop on Models of Trust for the Web*, 2006, pp. 3–131.
[4] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. McGuinness, "Computing trust from revision history," in *Proc. of the 2006 Intl. Conf. on Privacy, Security and Trust*. ACM, NY, USA, 2006.
[5] B. Adler, J. Benterou, K. Chatterjee, L. de Alfaro, I. Pye, and V. Raman, "Assigning trust to wikipedia content," in *WikiSym 4th Intl Symposium on Wikis*, 2008.
[6] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne, "Finding high-quality content in social media," in *Proc. of the Intl. Conf. on Web search and web data mining*. ACM, NY, USA, 2008, pp. 183–194.
[7] K. Lampe, P. Doupi, and J. van den Hofen, "Internet health resources: from quality to trust," *Methods of information in medicine*, vol. 42, no. 2, pp. 134–142, 2003.
[8] P. Sztompka, *Trust: A sociological theory*. Cambridge Univ Pr, 1999.
[9] B. Bailey, L. Gurak, and J. Konstan, "Trust in cyberspace," *Human factors and Web development*, pp. 311–21, 2003.
[10] "Quality," *Merriam-Webster Online Dictionary*, Nov 2008. [Online]. Available: http://www.merriam-webster.com/dictionary/quality
[11] "Credibility," *Merriam-Webster Online Dictionary*, Nov 2008. [Online]. Available: http://www.merriam-webster.com/dictionary/credibility
[12] G. Marshall, "Utility," *A Dictionary of Sociology*, 1998. [Online]. Available: http://www.encyclopedia.com/doc/1O88-utility.html
[13] P. Fishburn, "Utility theory," *Management Science*, pp. 335–378, 1968.
[14] A. Alchian, "The meaning of utility measurement," *The American Economic Review*, pp. 26–50, 1953.
[15] "Pertinence," *Merriam-Webster Online Dictionary*, Nov 2008. [Online]. Available: http://www.merriam-webster.com/dictionary/pertinence
[16] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 422–446, 2002.
[17] F. McSherry and M. Najork, "Computing information retrieval performance measures efficiently in the presence of tied scores," *Lecture Notes in Computer Science*, vol. 4956, p. 414, 2008.