

Scientific Mashups: The Issue of Trust in the Aggregation of Web 2.0 Content

Sara Javanmardi
University of California, Irvine
Irvine, CA, USA
sjavanma@uci.edu

Yasser Ganjisaffar
University of California, Irvine
Irvine, CA, USA
yganjisa@uci.edu

Cristina Lopes
University of California, Irvine
Irvine, CA, USA
lopes@uci.edu

Stanley Grant
University of California, Irvine
Irvine, CA, USA
sbgrant@uci.edu

ABSTRACT

The concept of scientific mashups is gaining popularity as the sheer amount of scientific content is scattered over different sources, such as databases or public websites. A variety of mashup development frameworks exist, but none fully address the needs of the scientific community. One limitation of scientific mashups is the issue of trust and attribute; especially when the content comes from collaborative information repositories where the quality of such content is unknown. In this paper, for our case study we focus on CalSWIM whose content is taken from both highly reliable sources and Wikipedia which may be less so. We will show how integrating CalSWIM with a reputation management system can help us assess the reputation of users and the trustworthiness of the content. Using user reputations, the system selects the most recent and trustworthy revision of the wiki article rather than merely the most recent revision, which might be vandalistic or of poor quality.

Keywords

Mashup, Wikis, Web 2.0

1. INTRODUCTION

Over the past decade, web search engines have been successful in organizing web content and making it accessible to everyone. However, due to exploding data volume and data heterogeneity, the search for scientific content has remained difficult [1]. A significant portion of scientific content is still kept in databases that are either not published on the web or are only accessible through custom-developed forms. This makes finding these data sources difficult, and even when a user succeeds in finding them, it is difficult to integrate data or functionality from different sources. In addition, a large amount of the scientific content is scattered over Web 2.0 platforms such as wikis, blogs, and forums. According to eMarketer's report in 2010, 64% of Internet users in the US consume user-generated content and about 46% of users actively participate in the creation of the content [2].

A well-known example of user-generated content is Wikipedia, which is in the sixth position of most popular websites [3]. Its English version alone has more than 12 million users and 3.2 million content pages [4]. Using wiki technology, Wikipedia has become the largest crowdsourcing project and the main online encyclopedia [5]. It has been suggested that wiki technology can harness the Internet for science; "Wikinomics" is a recent term that denotes the art and science of peer production when masses of people collaborate to create innovative knowledge resources [6]. Wikinomics has opened the flood gates and turned the stream of innovation into a raging torrent in which anyone with an idea and a computer is free to swim or be merrily transported downstream. Not only are shared efforts and goals pervading the scientific community with the continuous creation and diffusion of new models of data annotation and exchange such as Wikiproteins and Wikipathways [7, 8] but also the torrent of information and ideas pours out to the public, allowing worldwide collaboration based around ideals of openness and cooperation [9]. A very good example is the Gene Wiki portal. The Gene Wiki is an informal collection of pages on human genes and proteins deposited into Wikipedia in order to reach specific aims: (1) to provide a well written and informative Wikipedia article for every notable human gene; (2) to invite participation by interested lay editors, students, professionals, and academics from around the world; (3) to integrate Gene Wiki articles with existing Wikipedia content through the use of internal wiki links increasing the value of both. Although the open editing model of Wikipedia and high user contribution facilitate reaching these goals, trust still remains a big concern in the wiki community, especially when wikis are used for scientific purposes [10, 11].

Trust and attribution are fundamental concepts in science. When reviewing a scientific model or a dataset, trust is the degree to which the assumptions of the model or the accuracy of the data are accepted. Lack of trust is not necessarily bad; what is important is that the degree of trust can be assessed and properly accounted for. This is where attribution comes in. Models and datasets are associated with specific groups of individuals through peer reviewed scientific publications. By carefully reviewing the published literature, scientists can make informed decisions about what new research to perform [12].

To address the problem of trust, we are developing a sci-

Copyright is held by the authors.

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

entific mashup called CalSWIM [13], where the trustworthiness of the content fetched from Web 2.0 can be assessed. CalSWIM is an information and management tool designed both as a public forum for exploring watersheds and as a web location for professionals to acquire data. Leveraging the power of “crowdsourcing”, CalSWIM provides a specialized view of Wikipedia’s articles related to Water Resources. To smooth out the trust challenges for the content fetched from Wikipedia, we have integrated the mashup with a reputation management system that can automatically assign reputation to the contributors of the wiki articles and estimate trustworthiness of the content. This feature helps CalSWIM users interested in Wikipedia articles have access to the most recent reliable revision of the article (as opposed to Wikipedia’s normal practice of showing merely the most recent revision).

The remainder of this paper is as follows: Section 2 provides a brief overview of CalSWIM. Section 3 describes the reputation management system for Wikipedia. Finally, Section 4 draws some conclusions and provides some direction for future investigation.

2. CALSWIM

The California Sustainable Watershed/Wetland Information Manager (CalSWIM) allows users to geographically locate watershed-related information on the web, fetched from a variety of data sources and Web 2.0 applications such as Wikipedia. Figure 1 shows a screenshot of how data from different sources is aggregated and presented to the user, based on his geographical query.

Mining Wikipedia’s category network, we were able to extract and index more than twenty thousands water-related articles. The geographical distribution of these articles is unevenly distributed across the United States (Fig. 2). Certain regions of the country have a very low density of water-related Wikipedia articles; e.g., a swath of the mid-west stretching from Montana and North Dakota in the north, to Texas and New Mexico in the South. The highest densities of water-related articles are located near major population centers (San Francisco, Boston, New York, Washington DC, and Pittsburgh), while others are centered around regionally important (Columbia River in Oregon and Washington State) or unusual water-bodies such as Yellowstone National Park (see call out maps in Fig. 2). The CalSWIM mashup is being used to scan Wikipedia articles in these “hot zones”, and to locate sources of information or raw data that might shed light on the degree to which local perceptions of water quality are based on sound science. Our hope is that tools such as CalSWIM may promote collaborative, and thus sustainable, water management in the face of global climate change.

In order to extract water-related articles in Wikipedia, we have used Wikipedia’s API. Wikipedia articles are typically ordered into semantically related categories that can be more specifically divided into sub-categories. This category network can be used for extracting semantically related articles. In this particular case, we first extracted a list of general water-related categories such as *Water*, *Water pollution*, *Hydrology*, *Bodies of water* and *Water Supply*. Then, we recursively extracted the pages which are accessible from these general categories by traversing their sub-category networks. As an example, Figure 3 shows how the article entitled “Grand Canyon can be reached by traversing

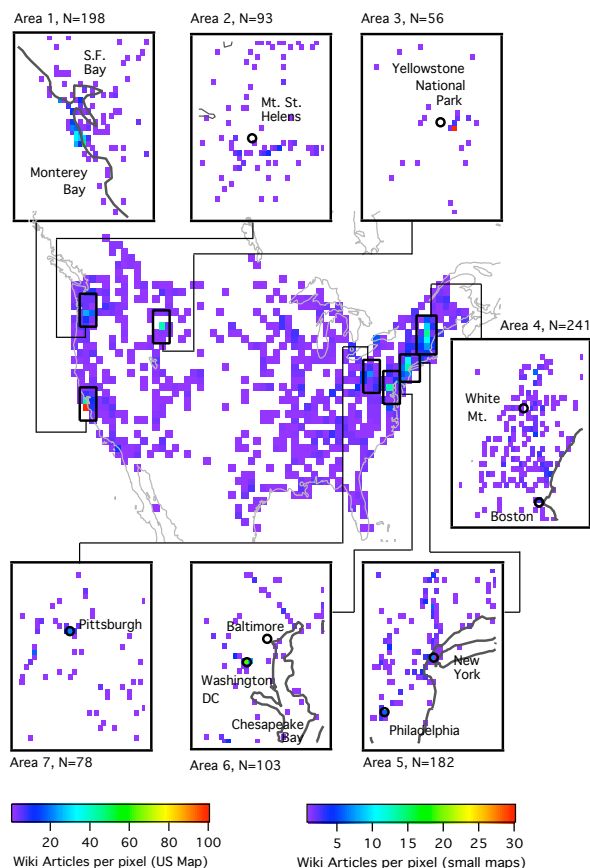


Figure 2: Density of Water-related Wikipedia articles across the United States

sub-categories starting from “bodies of Water.

The above process might result in the extraction of non water-related articles. For example, an article about a route crossing a river might be extracted. We used a set of relevant keywords such as “river”, “watershed”, “wetland”, “lake”, “creek” and “swamp” to filter out irrelevant articles.

Our final dataset contains a list of 20,824 water-related articles. Wikipedia articles can use a special markup for specifying the latitude and longitude of the concept that they are describing. We processed the text of all the extracted articles and were able to locate the latitude/longitude information for 5,270 articles.

3. TRUSTWORTHINESS OF THE CONTENT

Though user-generated content is publicly available through open Web 2.0 applications, and its volume is growing very fast, when it comes to usage by the scientific community there is always a concern for the trustworthiness of the content. Users need to make sure the content they see in the mashup is reliable, especially when it is coming from an unknown blogger or a wiki article that anyone can edit.

To address this concern, we have adopted Wikipedia as a case study. We have designed a system for the automatic assessment of user reputation and trustworthiness of the content. The reputation of a user can be viewed as the probability of him producing a high-quality contribution. This probability is computed by methods developed in

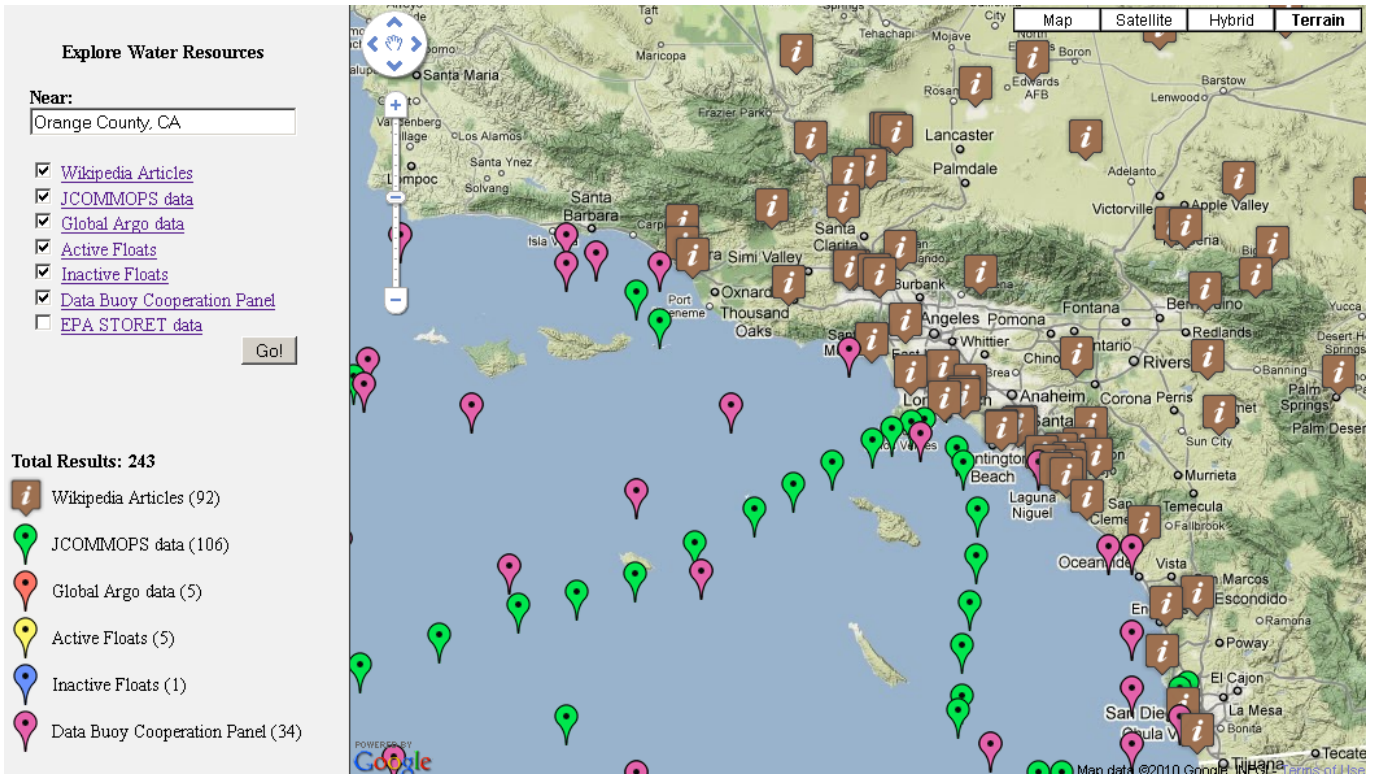


Figure 1: A Screenshot of CalSWIM project which shows how data from different sources is aggregated in a simple interface

[14]. The heuristic behind this reputation assessment is that high-quality contributions tend to survive longer in wiki articles. This heuristic is also supported by other work [15, 16]. Assume that user i has inserted $N_i(t)$ tokens in the system before time t and $n_i(t)$ of these tokens are not deleted yet. At the time t , he inserts $c_i(t)$ new tokens where $g_i(t)$ of them remain in the wiki article, and the rest are deleted by other users. Reputation of user i is updated based on the following formula:

$$R_i^+(t) = \max \left(0, \frac{n_i(t) + g_i(t) - \sum_{d=1}^{p_i(t)} R_{j(t_d)} e^{-\alpha(\Delta r)}}{N_i(t) + c_i(t)} \right) \quad (1)$$

where $R_{j(t_d)}$ is the reputation of the deleter at the time of deletion, $p_i(t)$ is the number of deleted tokens, and Δr is the number of revisions submitted between insertion and deletion of the tokens.

When used as a classifier, the model produces an area under the ROC curve of 0.98. Furthermore, we assess the reputation predictions generated by the models on other users, and show that the models can be used efficiently for predicting user behavior in Wikipedia. The effectiveness and efficiency of the model and its comparison with related work is discussed in [14].

Since Wikipedia is a dynamic system, the articles can change very frequently. An article may contain high and low quality content in different periods of its lifetime. Therefore, the quality of an article is a time-dependent function. To model quality, we consider two states for article revisions: low ($q = 0$) and high ($q = 1$) quality. In order to assess the

quality q of a revision, we consider the reputation of the author. If the reputation of the author of a revision is r , then with the probability of r the new revision will be in $q = 1$ state, and with the probability of $1 - r$, the new revision will be in $q = 0$ state. To estimate the proportion of time during which an article is in a high-quality state, we define the duration $QD(T)$ by:

$$QD(T) = \frac{\sum_{i=1}^n (t_{i+1} - t_i) q(t_i)}{T - t_1}$$

To evaluate this model, we processed the history of English Wikipedia articles. Featured articles (known “high quality” articles) on average contain high-quality content 86% of the time. Interestingly, this value increases to 99% if we only consider the last 50 revisions of the articles. The same statistics for non-featured articles show that they have high-quality content 74% of the time [17]. These results suggest that, although low-quality content appears in Wikipedia articles, it typically has a short life span. According to these results, user reputation can be a good metric for assessing quality of an article and its revisions. Therefore, quality can be considered a new feature for content search in CalSWIM.

Having evaluated user reputations, we can then rank the recent revisions of an article according to the trustworthiness of their contributors. Then, it is possible to suggest the latest reliable revision of an article to the user. To evaluate the effectiveness of this idea, we calculated the reputation of users contributing to the water-related Wikipedia articles extracted in CalSWIM. Table 1 shows properties of the dataset.

Our study on the Entire English Wikipedia in Septem-

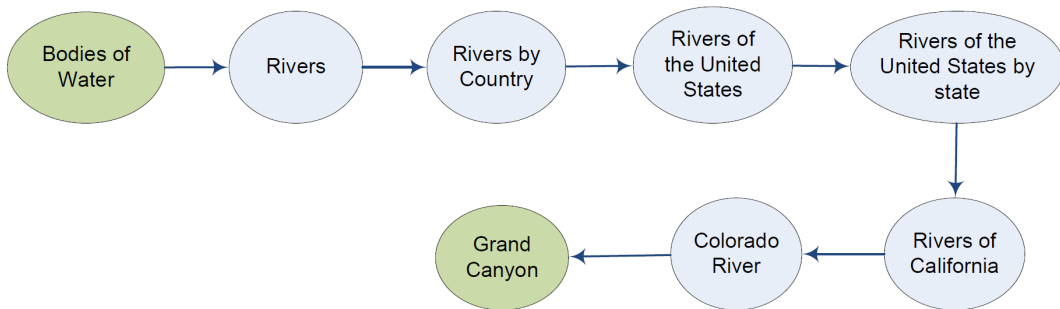


Figure 3: Traversing Wikipedia’s Category Network for Extracting Water-related Articles

Number of Articles	20,824
Number of Registered Users	101,465
Number of Anonymous Users	302,324
Number of Revisions by Registered Users	1,236,642
Number of Revisions by Anonymous Users	581,804
Average Reputation of Registered Users	0.6967
Average Reputation of Anonymous Users	0.4202

Table 1: Properties of the Dataset

	$n = 1$	$n = 2$	$n = 3$	$n = 5$
Rep > 0.82	72.61%	86.28%	92.48%	96.93%
Rep < 0.22	1.0%	0.12%	0.05%	0.01%

Table 2: Percentage of articles with high reputation and low reputation users in their last n revisions. When $n > 1$, results are based on the maximum of the reputation of users contributing the last n revisions.

ber 2009 shows that the average reputation for good users –users who contribute high-quality content– is 82% while this is 22% for vandals [14]. If we use the same settings here and assume that users with reputation more than 82% contribute high quality content and users with reputation less than 22% contribute low quality content, we can estimate the percentage of high quality content in the most recent revisions of articles. Table 2 summarizes our results. When considering only the last revision of articles ($n = 1$), about 73% of them are of high quality in their last revision and 1% are of low quality (the rest of articles have average quality revisions). When considering the last five revisions of articles ($n = 5$), we found that for almost 97% of the articles, at least one revision had been submitted by a high reputation user. Therefore, it is more beneficial to show this revision to users rather than merely the most recent one. Figure 4 shows the full distribution of reputation of the users contributing to the last five revisions of the articles.

4. DISCUSSION AND FUTURE WORK

Our initial study on the problem of trust in the CalSWIM mashup shows that the idea of showing the most trustworthy and recent revision of an article to a user can be beneficial for fetching content from wikis. However, it is important to

note that assessing trustworthiness of content based only on the reputation of the contributor has some limitations:

- **Data sparsity:** for a considerable number of users in Wikipedia we do not have enough information for accurate reputation estimation. The models that we use for estimation of user reputation are based on the observed behavior of users and how other users react to the contributions of these users. Therefore, in cases in which a user is new to the system, we do not have a stable reputation estimate for him.
- **Anonymity:** a significant number of users contribute to Wikipedia articles anonymously and they are only identified by their IP addresses. However, there is a loose correspondence between the IP addresses and the real-world users.
- **Expertise:** quality of the contribution of a user to a topic depends on the expertise of the user on that topic. Having one reputation value may not be a perfect representative for quality of the contributions of the user on different topics. In case of CalSWIM we tried to alleviate this problem by estimating the reputation of users based only on their contributions to water-related articles.

In addition to the above limitations, there is no guarantee that users will not change their behavior in the future. So, a user who has contributed high quality content in the past, might contribute low quality content in the future. In addition, when a new user comes to the article and contributes high quality content, the system sacrifices freshness for trustworthiness, only because it does not have an accurate estimate of the user’s reputation. This problem becomes worse for articles that are updated less frequently. In the case of our CalSWIM mashup, some articles get updated very infrequently. The average timespan between submission of the last two revisions of articles is 29 days. However, our study on Wikipedia featured articles shows that the update rate for an article increases significantly as it gains more visibility [17]. According to this observation, our conjecture is that mashups like CalSWIM can help these articles gain more visibility and thereby enjoy more frequent updates.

To overcome the limitations caused by inaccurate user reputation, in future work we aim at processing the changes done in newly submitted revisions of an article to see if it is vandalistic or not. Inspired by [18], we categorize Wikipedia vandalism types and build a statistical language models,

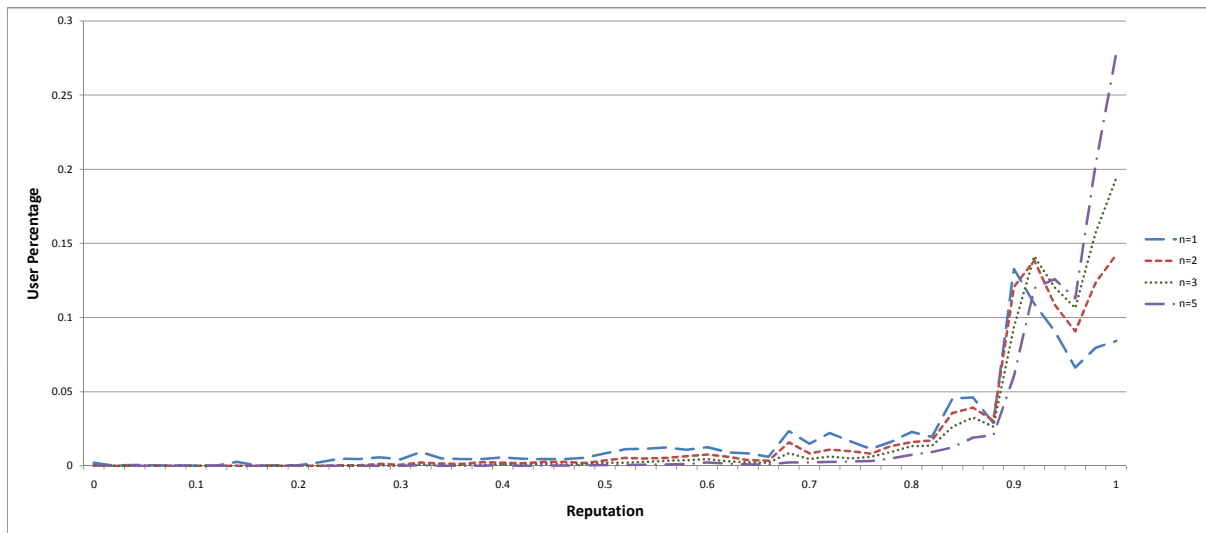


Figure 4: Distribution of user reputation for the last n revision of water-related articles in Wikipedia

constructing distributions of words from the revision history of Wikipedia articles. As vandalism often involves the use of unexpected words to draw attention, the fitness (or lack thereof) of a new edit, when compared with language models built from previous revisions, may well indicate that an edit is the product of vandalism. One of the main advantages of this technique is that it is extendable, even to other Web 2.0 domains such as blogs.

5. ACKNOWLEDGMENTS

This work has been partially supported by NSF grant OCI-074806.

6. REFERENCES

- [1] B. Howe, H. Green-Fishback, and D. Maier, “Scientific mashups: Runtime-configurable data product,” *Scientific and Statistical Database Management*, vol. 5566, pp. 19–36, 2009.
- [2] February 2009: User generated content hard to monetize. [Online]. Available: http://www.iab.net/insights_research/947883/1675/669304
- [3] Top 10 websites. [Online]. Available: <http://www.alexa.com/>
- [4] Wikipedia statistics. [Online]. Available: <http://en.wikipedia.org/wiki/Special:Statistics>
- [5] J. Zittrain, *The Future of the Internet—And How to Stop It*. Yale University Press, 2008.
- [6] D. Tapscott and A. Williams, *Wikinomics: How Mass Collaboration Changes Everything*. Penguin Group, 2006, pp. 70–77.
- [7] B. M. et al, “Calling on a million minds for community annotation in wikiproteins,” *Genome Biol*, no. 9:R89, 2008.
- [8] A. Pico and T. K. et al., “Wikipathways: Pathway editing for the people,” *PLoS Biol*, no. 6: e184, 2008.
- [9] A. Rinaldi, “Science wikinomics,” *Nature: EMBO reports*, vol. 10, pp. 439–443, 2009.
- [10] Editorial, “Standardizing data,” *Nature Cell Biology*, no. 10, pp. 1123 – 1124.
- [11] Portal:gene wiki. [Online]. Available: http://en.wikipedia.org/wiki/Portal:Gene_Wiki
- [12] Scientific wikis. [Online]. Available: <http://rosettadesigngroup.com/blog/373/scientific-wikis-part-i/>
- [13] Calswim mashup. [Online]. Available: <http://nile.ics.uci.edu/calswim/>
- [14] S. Javanmardi, C.Lopes, and P.Baldi, “Modeling user reputation in wikipedia,” *Journal of Statistical Analysis and Data Mining (accepted)*, vol. 3, no. 2, pp. 126–139, March 2010.
- [15] M. Hu, E. Lim, A. Sun, H. Lauw, and B. Vuong, “Measuring article quality in wikipedia: models and evaluation,” in *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 243–252.
- [16] B. T. Adler and L. de Alfaro, “A content-driven reputation system for the wikipedia,” in *WWW '07: Proceedings of the 16th international conference on World Wide Web*. New York, NY, USA: ACM, 2007, pp. 261–270.
- [17] S. J. Y. Ganjisaffar, , C. Lopes, and P. Baldi, “Statistical measure of the effectiveness of the open editing model of wikipedia,” in *CWSM Data Challenge*, May 2010.
- [18] R. Lopes and L. Carriço, “Using language models to detect wikipedia vandalism,” in *WSDM*, 2010.