ICST Transactions on Collaborative Computing Research Article ICST.ORG

Harnessing Context for Vandalism Detection in Wikipedia

Lakshmish Ramaswamy*[1], Raga Sowmya Tummalapenta[1], Deepika Sethi[1], Kang Li[1], Calton Pu[2]

[1] Computer Science Department, The University of Georgia, Athens, GA 30602, USA
[2] College of Computing, Georgia Institute of Technology, Atlanta, GA 30332, USA

Abstract

The importance of collaborative social media (CSM) applications such as Wikipedia to modern free societies can hardly be overemphasized. By allowing end users to freely create and edit content, Wikipedia has greatly facilitated democratization of information. However, over the past several years, Wikipedia has also become susceptible to vandalism, which has adversely affected its information quality. Traditional vandalism detection techniques that rely upon simple textual features such as spammy or abusive words have not been very effective in combating sophisticated vandal attacks that do not contain common vandalism markers. In this paper, we propose a context-based vandalism detection framework for Wikipedia. We first propose a context-enhanced finite state model for representing the context evolution of Wikipedia articles. This paper identifies two distinct types of context that are potentially valuable for vandalism detection, namely content-context and contributor-context. The distinguishing powers of these contexts are discussed by providing empirical results. We design two novel metrics for measuring how well the content-context of an incoming edit fits into the topic and the existing content of a Wikipedia article. We outline machine learning-based vandalism identification schemes that utilize these metrics. Our experiments indicate that utilizing context can substantially improve vandalism detection accuracy.

Received on 04 March 2013; accepted on 01 May 2014; published on 27 May 2014

Keywords: Collaborative Social Media, Vandalism, Content-context, Contributor-context

doi:10.4108/cc.1.1.e6

1. Introduction

Collaborative online social media (CSM) applications form an important category of Web 2.0 applications. In recent years, CSM applications such as Wikipedia have radically transformed the World Wide Web (WWW) landscape by enabling end-users to actively engage in the creation, organization and propagation of web content. *Democratization of information* and *collective intelligence* are the two core principles of Wikipedia, and it tries to achieve them through a model that permits contributors to freely create and edit content.

The importance of Wikipedia to modern societies is reflected in the exponential growth of people who rely upon it as a source of information. A study by the Pew research center indicates that 53% of American Internet users regularly look up information on Wikipedia [1]. Thus, it is important to ensure the trustworthiness and quality of information that is available on Wikipedia. Over the past several years, vandalism has emerged as a significant threat to the quality as well as trustworthiness of Wikipedia information. Vandalism attacks on Wikipedia include, but are not limited to, creation of false information, presentation/interpretation of facts in a deliberately biased manner, using Wikipedia articles as propaganda tools (e.g., spamming), and blocking certain information/opinions (e.g., removing content from Wikipedia pages). Vandalism not only undermines the core philosophies of Wikipedia, namely, information democratization and collective intelligence, but can also cause wider damage. First, progressive degradation of information resulting from vandalism can lead to frustration among honest contributors, some of whom may lose interest in contributing content and participating in Wikipedia activities. Second, vandalism not only undermines the credibility

*Corresponding author. laks@cs.uga.edu

This article discusses the ideology of liberalism. Local differences in its meaning are listed in Liberalism worldwide. For other uses, see Liberal.

Liberalism (from the Latin *liberalis*, "of freedom"[1]) is the belief in the importance of dependency on big daddy gov't and equality.[2][3] Liberals espouse a wide array of views depending on their understanding of these principles, but most liberals support such fundamental ideas as constitutions, liberal democracy, free and fair elections, human rights, free trade, secularism, and the market economy. These ideas are often accepted even among political groups that do not openly profess a liberal ideological orientation. Liberalism encompasses several intellectual trends and traditions, but the dominant variants are classical liberalism, which became popular in the 18th century, and social liberalism, which became popular in the 20th century.

Figure 1. Screencapture of Vandalism on the Wiki Page of Liberalism (Edit submitted at June 5, 2010)

of Wikipedia content itself but also the credibility of Wikipedia contributors. Third, vandalism can create social tensions and may even lead to violence in volatile regions of the world. Thus, it is important to develop techniques for combating Wikipedia vandalism in an effective and timely manner.

Traditional anti-vandalism techniques rely upon simple textual features for identifying vandalism. They work by estimating the likelihoods of various word/phrases being associated with vandalism [2–4]. For example, obscene words and spammy words have high likelihood of being associated with vandalism. This information is used for identifying vandalism in incoming edits. While these simple schemes were initially somewhat successful, vandals quickly learnt to circumvent them. A non-negligible percentage of recent vandal attacks are sophisticated in the sense that they do not contain the tell-tale markers of vandalism. This type of vandal edits is also referred to as *elusive vandalism* [5]. Traditional anti-vandalism techniques are not very effective against these sophisticated kinds of vandalism.

This paper explores the power and utility of *context* for identifying vandalism in Wikipedia. Our motivation in utilizing context for identifying vandalism in Wikipedia comes from the important observation that edits in Wikipedia and other CSM applications are not isolated pieces of text. Rather, they happen in a specific *context*. Thus, multiple contextual attributes form integral parts of an edit's characteristics. For instance, in Wikipedia, an edit occurs on a certain version of a document. Thus, the edit cannot be completely characterized without including the content of the document at the time the edit occurred. Similarly, whether the person contributing the edit is a registered or an unregistered user is important for characterizing the edit. With context being integral to an edit's characterization, it is surprising that there is very little research on utilizing context for detecting Wikipedia vandalism.

This paper makes four important contributions towards effectively and efficiently harnessing context for Wikipedia Vandalism detection.

- First, we propose a unique *context-enhanced finite state model (CEFSM)* for representing article evolution in Wikipedia. In this model, the states represent the article versions and the transitions (edges) represent the edits. Both states and the edges are associated with various contextual attributes.

- Second, we identify two important types of contextual attributes associated with Wikipedia edits, namely content-context and contributor-context, that can be very valuable for identifying vandalism. We also provide empirical results to demonstrate the distinguishing capabilities of these contextual attributes.

- Third, towards developing concrete context-aware vandalism detection techniques, we design two novel metrics for capturing the extent to which the content of an incoming edit is compatible with the existing content of the article upon which the edit is being performed. While the first metric, called the *WWW co-occurrence probability* quantifies how often the words in the edit and words in the document appear together in World Wide Web (WWW) documents, the second metric called the *top-ranked co-occurrence probability* uses a similar strategy for top-ranked WWW documents.

- Fourth, in addition to developing cost-effective mechanisms for computing the *WWW co-occurrence probability* and the *top-ranked co-occurrence probability*, we discuss how these mechanisms can be utilized in conjunction with a machine-learning framework for identifying vandalism.

This paper also reports several sets of experiments over the Wikipedia vandalism PAN corpus to evaluate the efficacies of the proposed techniques. The remainder of the paper is organized as follows. Section 2 provides background on Wikipedia vandalism. In Section 3, we motivate our research by discussing the role of context in Wikipedia, and we also present our context-enhanced finite state model for Wikipedia. Section 4 discusses the content and contributor contexts and provides empirical results to highlight their distinguishing capabilities. Section 5 outlines our vandalism detection algorithm. In Section 6 we discuss the experimental evaluation. Section 7 discusses the related work and we conclude the paper in Section 8.

2. Wikipedia and Vandalism

Wikipedia is one of the most popular Web 2.0 applications. It is a free online encyclopedia whose contents are generated and managed in a collaborative manner. Wikipedia has an *open-edit* policy in which most Wikipedia articles can be edited by anyone. While the open-edit policy is inherent to Wikipedia's philosophy of information democratization, it has also made Wikipedia susceptible to vandalism. Wikipedia itself defines vandalism "as an act that is intentionally disruptive" [6]. It can also be defined as a deliberate act aimed at lowering the quality of information on Wikipedia. In this sense, vandalism can also be regarded as a type of denial of information (DoI) attack [7].

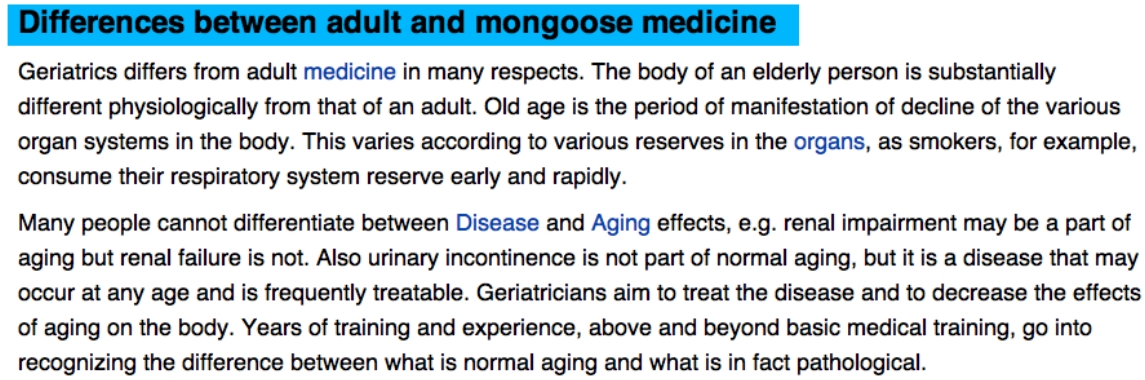
While vandalism can appear in any Wikipedia page, articles pertaining to controversial topics and personalities are more likely to be vandalized. Persistent vandalism has forced Wikipedia to modify its open edit policy - several levels of *protections* have been introduced to prevent vandalism. For example, *semi-protection* prevents the page from being edited by unregistered users (and users whose accounts are yet to be confirmed), while *full-protected* pages can only be edited by Wikipedia administrators. Introducing protection levels, in some sense, runs contrary to the open-edit policy of Wikipedia. Thus, it is evident that vandalism has affected the fundamental philosophy of information democratization.

Vandalism in Wikipedia can be of various types. Some of the prominent types of vandalism include tags abuse, illegitimate blanking, image vandalism, illegitimate page creation, and talk page vandalism [6]. These different types of vandalism vary in terms their target (Wikipedia articles, talk pages, etc.) and their mechanisms (adding content, removing content, relocating content, etc.). In this paper, our focus is primarily on vandalism that targets Wikipedia articles. Injection of abusive and obscene materials and spamming were among the earliest forms of vandalism. Even now, they constitute a substantial percentage of vandal edits. Thus, it is not surprising that the earliest works on vandalism detection were based upon identifying and utilizing textual features that have high likelihood of being associated with vandalism. However, vandal attacks are increasingly becoming subtle. These sophisticated attacks, called *elusive vandalism*, often do not contain the textual features associated with vandalism [5]. For example, they may not have any abusive/obscene words even when the intent is to belittle the topic of a Wikipedia article.

Figures 1 and 2 show examples of subtle vandalism. In Figure 1, the Wikipedia article on "Liberalism" has been vandalized by introducing the sentence "Liberalism is the belief in the importance of big daddy government". This vandal edit occurred on 06/05/2010 at 11:05 GMT. Figure 2 shows the Wikipedia articles on "Geriatrics" as it appeared on 02/23/2010 at 15:49 GMT. Here the a section heading has been changed from "Differences between adult and geriatric medicine" to "Differences between adult and mongoose medicine". Notice that although both of them are obvious cases of vandalism neither of them contain explicit features associated with vandalism. The words "importance", "big daddy", "government" or "mongoose" are neither abusive nor spammy. Clearly, anti-vandalism approaches that exclusively rely upon such textual features will not be identifying these subtler forms of vandalism.

3. Context in Wikipedia

One of the fundamental drawbacks of traditional anti-vandalism techniques is that most of them treat edits as independent pieces of text. Because of this, the traditional techniques limit themselves exclusively to the textual features of the edit. In reality, however, edits in Wikipedia are not isolated pieces of text. The text that is being added/removed in an edit does not completely characterize the edit. The edits in Wikipedia occur in certain *context*. For instance, an edit occurs on a certain version of an article. Similarly, the edit is performed by a certain person who might be a registered user or an un-registered user, and the edit is performed at a certain time. The edit cannot be completely characterized without considering these and other such contextual attributes.



Differences between adult and mongoose medicine

Geriatrics differs from adult medicine in many respects. The body of an elderly person is substantially different physiologically from that of an adult. Old age is the period of manifestation of decline of the various organ systems in the body. This varies according to various reserves in the organs, as smokers, for example, consume their respiratory system reserve early and rapidly.

Many people cannot differentiate between Disease and Aging effects, e.g. renal impairment may be a part of aging but renal failure is not. Also urinary incontinence is not part of normal aging, but it is a disease that may occur at any age and is frequently treatable. Geriatricians aim to treat the disease and to decrease the effects of aging on the body. Years of training and experience, above and beyond basic medical training, go into recognizing the difference between what is normal aging and what is in fact pathological.

Figure 2. Screencapture of Vandalsim on the Wiki Page of Geriatrics (Edit submitted at February 23, 2010)

Many of these contextual attributes can be very powerful features in identifying vandalism. The importance of context is evident by the fact that even humans (implicitly) rely upon context when identifying vandalism. In the example depicted in Figure 2, most humans will immediately identify the edit as vandalism. This is because the word "mongoose", although not abusive, is irrelevant to the topic (Geriatrics). However, if this same word "mongoose" is introduced by an edit into the article on Snakes, it will not be considered as vandalism because the word being introduced is relevant to the topic (mongooses are predators of snakes). Similarly, if an edit on President Obama's Wikipedia page contains the word "Nazi", it will be recognized as vandalism, whereas the same works may not constitute vandalism if it is on Goebbels' Wikipedia page.

Harnessing context for Wikipedia vandalism detection poses several important challenges. First and foremost, we need a conceptual model for Wikipedia article evolution that captures various aspects of context. Second, we need to identify contextual attributes that have strong distinguishing capabilities. Third, context is often an abstract concept, and for machines to understand and process it, context has to be made *quantifiable*. This means that we have to not only invent meaningful metrics for various contextual attributes, but also devise efficient measurement mechanisms. Fourth, we need to design efficient and scalable vandalism detection techniques that utilize these quantifiable contextual attributes.

3.1. Context-Enhanced Finite State Model for Wikipedia Evolution

In this paper, we introduce a conceptual, context-enhanced finite state model (CEFSM) to represent and analyze the evolution of each Wikipedia article. Our CEFSM helps us to continually capture and analyze the context of the edits and Wikipedia articles as they

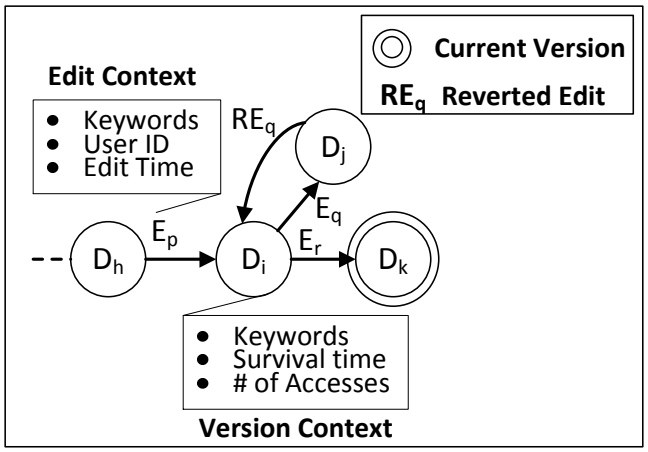


Figure 3. Context-Enhanced FSM for Wikipedia

evolve over time. Each version of the article that was installed forms a state (or node) of the article's CEFSM, with the last state representing the current version. The edits (which may involve content addition, modification or deletion) form the labels of the transition edges of the CEFSM. In essence, the article transitions from one version to the next through the corresponding edit. In Wikipedia, an edit can be *reverted* in which case the previous version will be restored back and made the current version. Our model provides for a rollback operation to represent this feature. When an edit is rolled back, the article transitions to its previous state in the corresponding FSM.

In our model, both nodes (i.e., article versions) and edges (i.e., edits) are associated with various contextual attributes. For example, the contextual attributes of a version can include its topic/category, content (e.g., keywords), links with other documents, and the time duration for which the version remained current. The contextual attributes of an edit can include the modification carried out by the edit (i.e., added and deleted key words), the time instance at which the edit

occurred, the ID or any other identifying information (such as IP address) on the contributor performing the edit. For conceptual clarity, we classify the contextual attributes into two broad categories – *Content-based context* in which contents of documents/edits (at granularity of keywords, sentences or semantic units) form the context sources; and *meta-context* which comprises of certain important meta-data pertaining to documents/edits (e.g., time of an edit, user contributing the edit, interlinks among documents, etc., as discussed above). Figure 3 illustrates part of an article's CEFSM. In this figure, the version represented by the state E_q has been reverted.

With our CEFSM, the problem of vandalism detection can be conceptualized as whether a particular transition (usually the last transition) leaves the article in an inconsistent state. In theory, context-aware vandalism detection techniques may utilize the entire contextual history (i.e., the contextual information associated with all previous states and transitions) in determining whether the current transition is a vandal edit. However, it is often impractical to take into account such large amounts of contextual information. Thus, our approach takes into account the contextual information associated with the edit (transition) that is being tested for vandalism and the contextual information of the article version (state) upon which the edit was performed.

4. Context for Vandalism Detection

In this section, we identify contextual attributes that can be harnessed for vandalism detection in Wikipedia. A contextual attribute is ideally suited to be utilized for vandalism detection if it provides two properties. First, it should exhibit strong distinguishing capabilities with regard to vandal and non-vandal edits. Second, it should be readily available or easily computable. We identify two contextual attributes, namely, contributor context and content context. For each attribute, we discuss its distinguishing capabilities and how it can be obtained.

4.1. Content-Context

Towards utilizing content-context for vandalism detection, our main idea is is to analyze *how well the content of an incoming edit fits into the context of the existing version (i.e., existing content) of the document.* Let D_j represent the current version of a Wikipedia document and let E_r represent an incoming edit on D_j. The idea is to check how well the content being introduced by E_r gels with content existing in D_j. The central observation is that if the edit E_r is legitimate (non-vandal), the content of E_r will fit well into the content of D_j, and vice-versa. For example, consider the edit that contains the following sentence: "He was a close associate of Adolf Hitler". Note that this edit fits well into the context of Goebbels' Wikipedia page because the page is likely to contain quite a bit of material about Nazism and the Third Reich. Also note that this edit will be legitimate (non-vandal). On the other hand, if the same edit were to happen on President Obama's Wikipedia page, it will certainly be out of context (because the page will not contain any material that is even remotely connected with Nazism), and it will be readily recognized as vandalism by humans. Note that our content-context-based approach utilizes the context associated with the incoming edit as well as the context of the current version of the document.

Unlike contributor context (to be discussed later in this section), content-context is not readily available. In fact, an important challenge is to *quantify* the compatibility of the content-context of the incoming edit with that of the existing version of the document. Contextual analysis can be performed at various levels of textual understanding. For instance, one can adopt *language-based analysis* which is based upon *natural language understanding (NLU)*. However, NLU is one of the *AI-complete problems* [8], and hence impractical. We adopt a *bag-of-words* approach in which the contexts of the edit as well as the version on which the edit is performed are captured as sets of respective keywords and phrases. In other words, we analyze how well the keywords of the edit fit with the keywords of the existing Wikipedia page. For performing the analysis, our strategy does not understand or rely upon the word meanings. Instead, it uses statistics regarding co-occurrence of words in documents to determine whether a particular edit is vandalism. We propose two metrics in this regard namely, *WWW co-occurrence probability* and *top-ranked co-occurrence probability*.

WWW Co-Occurrence Probability for Quantifying Content–Context. The overall idea here is to measure the likelihood of the keywords of an incoming edit and the keywords of the existing version of the document occurring together (in the same document) in the World Wide Web (WWW) corpus of documents. The rationale is that if an incoming edit (represented as E) fits well into the context of the existing version of the Wikipedia page (represented as D), then the keywords of E and D should occur together in a non-negligible fraction of WWW documents.

Let $W(D_j) = \{wd_1, wd_2, \ldots, wd_p\}$ be the set of keywords in the current (non-vandalized) version of the document. (i.e., $W(D_j)$ is the current context of the document D) and $W(E_r) = \{we_1, we_2, \ldots, we_q\}$ denote the set of words that the edit E_r is seeking to introduce in the next version of the document (i.e., $W(E_r)$ is the edit's context). The co-occurrence probability of the arbitrary keyword pair (we_l, wd_m) is defined as the ratio of the probability that both we_l and wd_m occur in an arbitrary

WWW document to the ratio that at least one of them occurs in a WWW document. Mathematically,

$$CoP(we_l, wd_m) = \frac{P(we_l \in DC \wedge wd_m \in DC)}{P(we_l \in DC \vee wd_m \in DC)} \quad (1)$$

In the above equation, DC denotes an arbitrary WWW document. The denominator in Equation 1 is a normalization term that has been introduced to account for the popularity variations among keywords.

The WWW co-occurrence probability is defined as the minimum of the CoPs over all the edit-document keyword pairs.

$$WCoP(E_r, D_j) = \underset{we_l \in W(e_r), wd_m \in W(D_j)}{\text{argmin}} (CoP(we_l, wd_m)) \quad (2)$$

The reason we use argmin in Equation 2 is that an edit can have only a single vandal word/phrase (i.e., all other words of the edit may be completely legitimate). Thus, we are interested in the contextual fitness (measured by CoP) of the least contextually appropriate word among all the keywords of the edit.

Top Ranked Co-occurrence Probability Metric. Our second content-based contextual analysis metric, called the top ranked co-occurrence probability metric is thematically similar to the WWW co-occurrence probability metric. The key difference however, is that instead of using the entire WWW document corpus, this metric uses only the top-ranked WWW documents (as determined by a popular search engine). The rationale for using the top-ranked documents is that these documents are typically perceived to be reliable and trustworthy information sources.

The formal definition of top ranked co-occurrence probability metric is analogous to that of the WWW co-occurrence probability except that the corpus is limited to top-ranked web documents. Formally, Let $W(D_j) = \{wd_1, wd_2, \ldots, wd_p\}$ be the set of keywords in the current (non-vandalized) version of the document and $W(E_r) = \{we_1, we_2, \ldots, we_q\}$ denote the set of words that the edit E_r is seeking to introduce in the next version of the document. Let TCP^K denote the corpus of K top-ranked documents containing at least one word from $W(D_j) \cup W(E_r)$ and let TC denote an arbitrary document in TCP^K. The top K co-occurrence probability of the keywords we_l and wd_m is defined as follows.

$$TrCoP(we_l, wd_m) = \frac{P(we_l \in TC \wedge wd_m \in TC)}{P(we_l \in TC \vee wd_m \in TC)} \quad (3)$$

The top ranked co-occurrence of the edit E_r with respect to the document version D_j is the minimum TrCoP over all the edit-document keyword pairs.

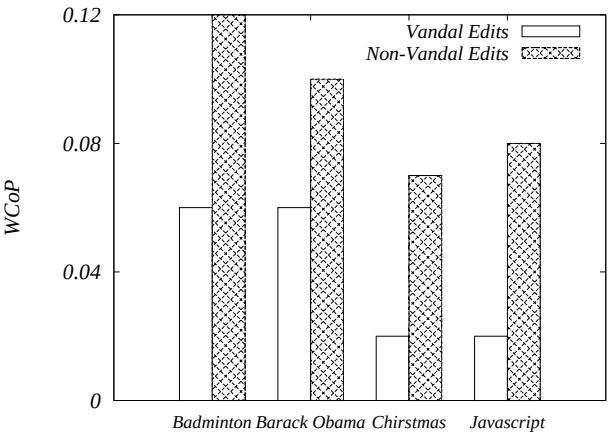


Figure 4. WCoP Values for Vandal and Non-vandal edits

$$TCoP^K(E_r, D_j) = \underset{we_l \in W(E_r), wd_m \in W(D_j)}{\text{argmin}} (TrCoP^K(we_l, wd_m)) \quad (4)$$

Computing the WWW co-occurrence probability and top-ranked co-occurrence probability metrics is challenging. We address this issue in Section 5.

Distinguishing Capabilities of Content-Context. In order to validate the distinguishing capabilities of content-context in detecting vandalism, we report the results from a small experiment. We have chosen 4 Wikipedia pages, namely “Badminton", “ Barack Obama", “Christmas" and “ Javascript". For each page we have randomly chosen 1000 edits that are known (human-validated) cases of vandalism and 1000 edits that are known to be legitimate. For each edit, we have computed the WWW co-occurrence probability (WCoP) value between the edit and version that was existing before the edit happened. In Figure 4, we plot the average WCoP values for the 1000 vandal and the 1000 legitimate edits for each page. The results indicate that the average WCoP values of non-vandal edits are 1.7 to 4 times higher than the corresponding values for vandal edits. This shows that content-context can be a powerful factor in distinguishing vandal edits from non-vandal ones.

4.2. Contributor-Context

The second type of context that we explore for vandalism detection is with respect to the person contributing an edit. Several features concerning an edit contributor can be very useful in identifying vandalism. The feature that is simplest to obtain is whether the contributor of an edit is a registered Wikipedia editor or he is an unregistered user. Wikipedia logs the information with respect to the person performing each edit. If the edit is from a registered user, editor id (user

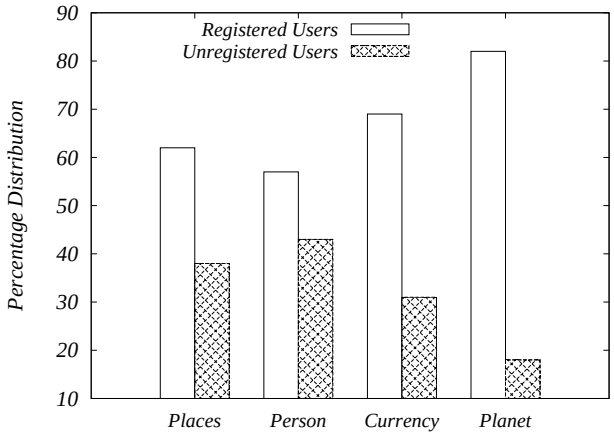


Figure 5. Registered and Unregistered Contributions for Legitimate Edits

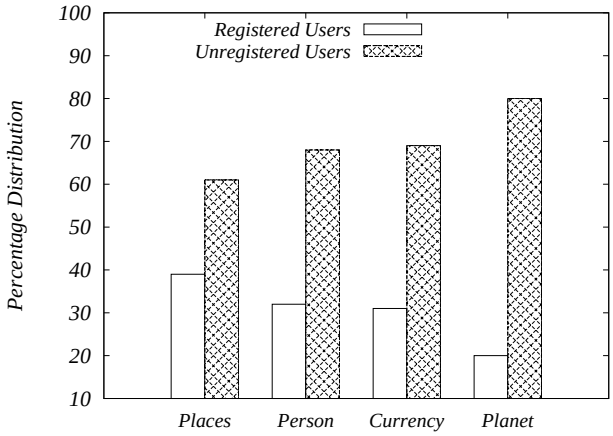


Figure 6. Registered and Unregistered Contributions for Vandal Edits

name) is maintained. If the edit is from an unregistered user, the ip address of the machine from where the edit was performed is maintained. Our study validates that the registration status of the edit contributor (registered vs. unregistered) has very strong vandal edit vs. non-vandal edit distinguishing capabilities.

To demonstrate the distinguishing capabilities of registration status, we perform the following experiment. We select 20 wikipedia articles each from five top-level Wikipedia domains, namely, "Places", "Person", "Currency", and "Planet" [1] . For each article, we randomly select 500 edits that are manually annotated as legitimate edits and 500 edits that are annotated as vandal edits and create a corpus. For each article, we compute the percentage of legitimate edits contributed by registered and unregistered users. Similarly, we also compute the percentage of contributions from registered and unregistered users for vandal edits. Figure 5 shows the mean percentage of legitimate edits contributed by registered and unregistered users for the articles in each of the five domains, and Figure 6 shows the mean percentage of vandal edits contributed by registered and unregistered users. These results clearly indicate that large fractions of legitimate edits are done by registered users whereas it is quite the opposite for vandal edits. Thus, registration status of edit contributors can be a very powerful factor in identifying vandalism.

Another contributor context attribute that can be useful for vandalism detection is the contributor reputation. For example, it is unlikely that a user who has consistently contributed high-quality edits for a significant duration of time will suddenly indulge in vandalism. On the other hand, Wikipedia notes several instances of *repeat vandalism* wherein the same user id (or IP address) is associated with multiple instances of vandalism. This suggests that edits coming from a user who has indulged in vandalism in the recent past needs to be carefully scrutinized to ensure that they are not vandal edits [9–11]

[1] Please see Section 6 for a description about the domains in Wikipedia

5. Vandalism Detection Algorithm

In this section, we explain our machine learning-based, context-centric algorithm for vandalism detection. We first discuss computationally efficient strategies for estimating WCoP and TCoP. Low overhead techniques for computing WCoP and TCoP are critical for ensuring the scalability of context-centric vandalism detection paradigm. The central issue in estimating WCoP is to compute the CoP between various we_l-wd_m keyword pairs. Our technique for estimating the CoP values works as follows. Our technique relies upon a popular search engine for estimating the CoP values (we use "Bing" in our experiments). Suppose we want to estimate $CoP(we_l, wd_m)$. We first issue a search query for documents containing both we_l and we_m (i.e, the search query will be we_l + wd_m). Most search engines indicate an estimate on the number of search results (the number of web documents containing both terms). Let the number of search results containing both we_l and wd_m be represented as Nb. We also issue queries for documents that exclusively contain each one of the search terms. In other words, we search for we_l - wd_m and wd_m - we_l. Let Ne_l and Nb_m be the estimates on the number of search results for these two queries respectively. Now $CoP(we_l, wd_m)$ is estimated as $\frac{Nb}{(Ne_l+Nb_m+NB)}$.

Our technique for computing TCoP works as follows. Suppose we want to estimate the top ranked co-occurrence between the edit-document keyword pair

we_l and wd_m. We issue separate search queries for wd_l and we_m. Let $Tr^K(we_l)$ and $Tr^K(wd_m)$ denote the top K search results for we_l and wd_m (K is a configurable parameter). The top K co-occurrence probability of the keywords we_l and wd_m is defined as $TrCoP^K(we_l, wd_m) = \frac{|Tr^K(we_l) \cap Tr^K(wd_m)|}{|Tr^K(we_l) \cup Tr^K(wd_m)|}$. Note that $(Tr^K(we_l) \cap Tr^K(wd_m))$ denotes the set of top K search results that contain *both* we_l and wd_m. The top ranked co-occurrence of the edit E_r with respect to the document version D_j is the minimum TrCoP over all the edit-document keyword pairs.

An associated problem in computing the WCoP and TCoP metrics is that the keyword set corresponding to the current version of the document ($W(D_j)$) is typically quite large. While edits usually contain a few keywords and phrases, document versions can be quite large. Thus computing CoP values for each edit-document keyword pair becomes prohibitively expensive. This overhead can be alleviated by limiting $W(D_j)$ to the keywords in the title of the article and its introductory paragraphs. In our experiments (see Section 6), we limit $W(D_j)$ to the keywords in the document's title.

Our vandalism detection algorithm works as follows. We employ machine learning-based (ML) classifiers for detecting vandalism. The ML classifiers are trained using known (human annotated) vandal and non-vandal edits as well as the respective article versions. Once the ML classifiers are trained the algorithm will be ready for vandalism detection. For each incoming edit, we extract/compute the selected contextual parameters (the algorithm can be configured to use a selected subset of contextual attributes). For example, if WCoP/TCoP parameters are to be employed by the ML algorithm, we extract the keywords from the incoming edit as well as the existing version and then use a popular search engine to compute the WCoP and TCoP values as described above. The selected contextual parameters are fed into the ML classifiers which determine whether the edit is vandal or legitimate edit.

In addition to the contextual attributes, the ML classifiers utilize one additional feature, namely, whether the edit involves inversion of statement meanings. This feature has been considered by prior works on Wikipedia Vandalism detection [5]. The reason for using the *statement inverse* feature is that previous studies have shown that a significant fraction of vandal edits just invert the meaning of one or more sentences by inserting or removing words and prefixes such as "not", "none", "un-", and "dis-". However, these are very common words and prefixes. Hence, they would not be part of keyword sets. Thus, in order to identify these vandal edits, it is necessary to consider statement inverse as a separate feature for the machine learning-based classifiers.

5.1. Discussion

We now discuss two issues that can further enhance the efficacy of context-driven vandalism detection. First, notice that currently our technique captures compatibility of an incoming edit's content with that of the existing version in terms of the co-occurrence probabilities of words. This can be viewed as a *syntactic approach* for capturing content-context. Currently our system does not analyze the meanings of words or relationships among them. A syntactic approach, by its very nature, cannot account for factors such as synonyms and homonyms. This can affect vandalism detection accuracy. We believe that performing the compatibility analysis at the semantic level can help alleviate these limitations. Such an approach should ideally take into account not only the meanings of words but also the inter-relationships between the words in the edit and the words in the existing version of the document. One way to accomplish this will be to use an ontology and capture inter-relationships through the semantic distances between the words. Wikipedia-based ontologies such as DBpedia and Yago are potential candidates in this regard [12, 13].

The second issue with regard to enhancing the efficacy of context-driven vandalism detection is that of *context evolution* or *context drifting*. Any ML-based context-driven vandalism detection scheme makes the inherent assumption with respect to *stability of context*. In other words, these schemes assume that the context attributes of incoming edits that need to be classified are not very different than those used for training the ML classifiers. However, contextual attributes in a collaborative system like Wikipedia is dynamic and it evolves over time. This evolution or drifting of context can adversely impact vandalism detection accuracy. This can be partially addressed by continuously updating the context training sets and re-training the ML classifiers. In effect, the context attributes derived from more recent edits and article versions receive more weight rather than the context attributes derived from older edits and documents. We believe this can address context drifts that are not drastic. In some, albeit rare, instances, context does undergo drastic changes. These are usually driven by real-world events. Dealing with these sort of drastic events is a challenge even to the human editors of Wikipedia. For example, when the singer Michael Jackson died on June 25, 2009, the user "Qc" added June 25, 2009 as the date of death to Michael Jackson's Wikipedia page. However, this edit was mistaken to be vandalism by a human editor who promptly reverted it. This highlights the challenge in dealing with drastic context changes. One possible way to address this challenge is to utilize information from realtime event sources such as Twitter and news feeds.

Developing concrete techniques for the above two issues requires comprehensive study and significant research, and it is beyond the scope of the current paper.

6. Experiments and Results

In this section, we discuss the experiments we performed to study the efficacy of content-context-centric vandalism detection technique.

6.1. Data Set

For our experiments, we use the PAN Wikipedia vandalism corpus 2010 (PAN-WVC-10). This corpus was compiled by Potthast at Bauhas-Universitat Weimar [14]. The corpus contains 32452 human-annotated edits on 28468 Wikipedia articles. The corpus has been annotated using Amazon's Mechanical Turk. Each edit has been annotated by at least three humans. Based on these annotations, each edit is labeled either as a "regular edit" (legitimate edit) or a "vandal edit". PAN-WVC-10 and its previous versions have been used as "gold standards" in several previous Wikipedia vandalism detection research projects [5].

Since our technique involves quantifying the content-contexts of edits with respect to the corresponding article versions, we need the entire edit histories of articles (including the labels for each version). For this purpose, we fetched the entire history of each article in the PAN-WVC-10. These additional edits are unlabeled. These additional edits are labeled using the *automatic data instance labeler* [5], which we briefly explain below.

The automatic data instance labeler uses the revision history (specifically, the revert and rollback history) to label edits as vandalism or regular edit. The automatic labeler marks a version as vandalism if the following conditions are satisfied. *(1)* It was contributed by an unregistered user; *(2)* the version was reverted by a super user or a bot and *(3)* the revert commentary on the article contains either of the following two patterns:

- Sensitive keywords: (?i).*vandal.*|(?i)rvv|(?i)rvv .*|(?i).* rvv .*|(?i).* rvv
- Signatures of anti-vandalism programs: (?i)Reverted edits by .* to last version by .*

If an edit was contributed by a super user or if the version was not reverted or if the comments for the version does not contain the above patterns, then it is considered to be a regular edit.

Wikipedia organizes articles into top-level *domains*. The prevalence and nature of vandalism varies significantly across domains. In our experimental evaluation, we study the efficacy of the proposed techniques for 7 different domains, namely, Chemical Substances, Currencies, Places, Persons, Programming Languages and Sports. Sample pages from each domain are listed in Table 1. For each page, we select the 100 most recent vandal versions and 100 most recent regular versions.

6.2. Experimental Setup

In our experimental study, we use the Bing search engine (`www.bing.com`) for calculating the WWW co-occurrence-probability and the top-ranked co-occurrence probability. We calculate the top-ranked co-occurrence probability based upon the top 250 search results returned by the search engine. In other words, in our experiments the configurable parameter K (see Section 5) is set to 250. We compare the WWW co-occurrence-probability-based and the top-ranked co-occurrence probability-based vandalism detection methods to a textual classifier. This text-based classifier assigns vandalism likelihoods for various keywords (using training data), which is then used for edit classification.

We use the Weka machine learning toolkit for classification. We have experimented with various classifiers including Naive Bayes, AdaBoost, and C4.5 Decision Tree. We measure precision, recall and F-1 measure of all three schemes (WWW co-occurrence probability, top ranked co-occurrence probability and the textual classifier).

6.3. Results

Figures 7(a) through 7(f) indicate the average F1 scores of the three vandalism detection techniques (WWW co-occurrence probability, top-ranked co-occurrence probability and text-based classification) for the six Wikipedia domains with 3 different classifiers, namely, Naive Bayes, AdaBoost and C4.5 Decision tree. WWW Co-occurrence probability technique, top-ranked co-occurrence probability technique and text-based technique are represented as "WCoP", "TCoP" and "TC" respectively. Each bar indicates the mean F1 score over the pages considered for that domain.

From these results it can be seen that WCoP and TCoP consistently outperform TC on all domains and on all classifiers. For example, both WCoP and TCoP yield 6.5% higher F1 scores when compared with TC on the "Sports" domain with Naive Bayes classifier. Note that a large fraction of the vandal edits in this data set are instances of regular vandalism (involving additions of swear words, massive spamming, etc.). For these cases, TC performs reasonably well. Thus the F1 measure of TC is also reasonably high. However, WCoP and TCoP are successful in detecting sophisticated instances of vandalism for which TC fails. In most cases, the F1 scores of WCoP and TCoP are above 0.95.

In order to give better insight into the performance of WCoP and TCoP, we plot the F1 score, precision and recall for sample pages from three domains namely,

No.	Domain Name	Sample Pages
1	Chemical Substance	Acetic Acid, Folic Acid, Phosphorous pentachloride
2	Currency	US Dollar, Canadian Dollar, Philippine Dollar, North Korean Won
3	Persons	Barack Obama, Jimmy Carter, Golda Mier, George W. Bush, Albert Einstein
4	Places	Canada, Costa Rica, India, Iran, United Kingdom
5	Programming Language	Javascript, C, Logo, Ada, True basic
6	Sports	Badminton, Tennis, National Rugby League, Golf

Table 1. Wikipedia Domains and Sample Pages

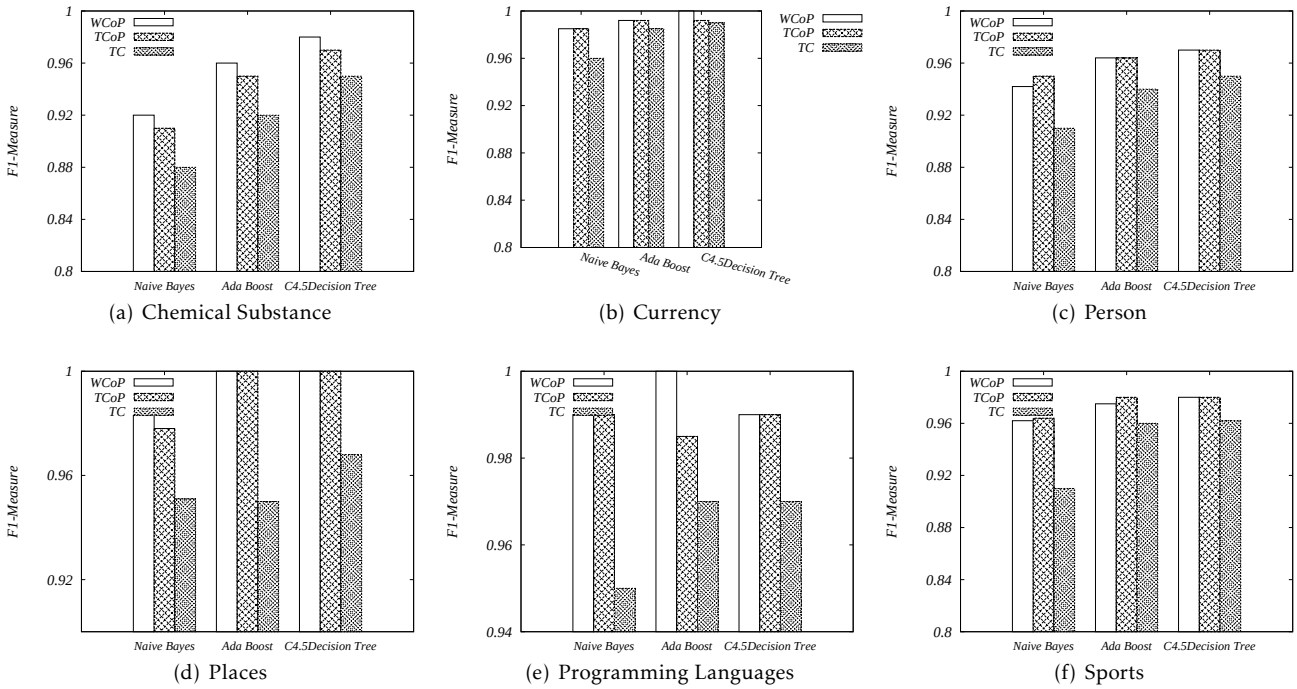
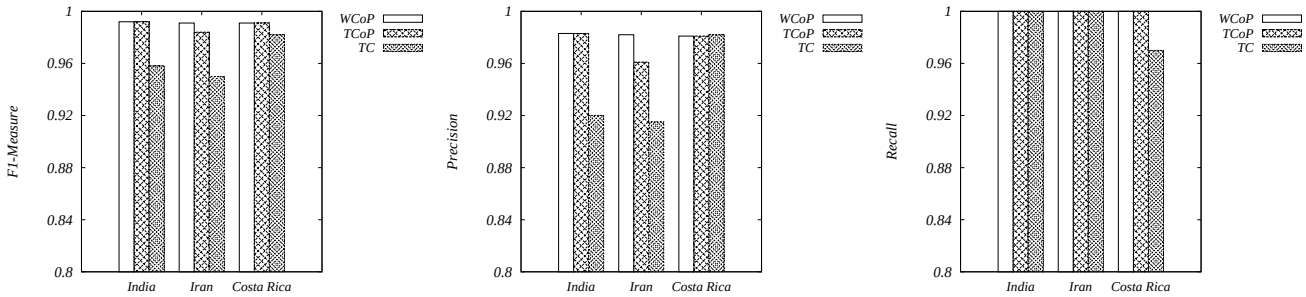


Figure 7. Comparison of WCoP, TCoP and Text Classification Performance on Various Wikipedia Domains

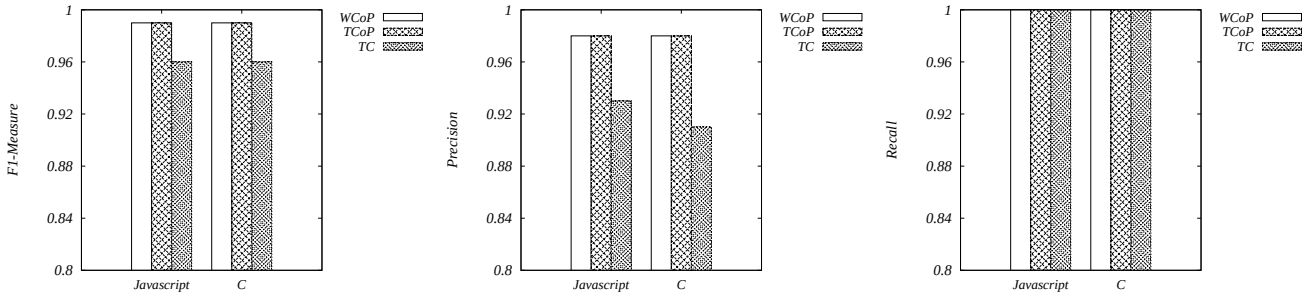
"places", "programming languages" and "currencies". These experiments were done using the C4.5 Decision tree classifier with 10-fold cross validation. Figures 8(a), 8(b) and 8(c) respectively indicate the F1 score, precision and recall for three pages from the "places" domain. Similarly, Figures 9(a), 9(b) and 9(c) respectively indicate the F1 score, precision and recall for two pages from the "Programming Languages" domain, and Figures 10(a), 10(b) and 10(c) show the F1score, precision and recall for two pages from the "currencies" domain. In most cases, WCoP and TCoP yield higher precision values than TC, while the recall values for the three schemes are quite comparable. Thus, higher F1 scores are a direct result of better precision.

Below, we provide a brief analysis of the characteristics of the edits that cause false positives and false negatives with our context-based vandalism detection system. False positives are legitimate edits that our system incorrectly marks as vandal edits. False negatives, on the other hand, are vandal edits that are not detected by our approach. In our system, false positives typically occur in three scenarios. The first is when an edit introduces factually correct statement that are not widely known. These sorts of statements can contain words/phrases that may seem out of context, and thus may be marked as vandalism. These kinds of edits are not very common in Wikipedia. The second scenario is when an edit contains words that colloquial, regional or even from other languages but written in the English script. These kinds of words commonly occur in pages pertaining to cultures, cuisines and personalities from remote, non-English speaking regions of the world. Since our system uses co-occurrence probability for



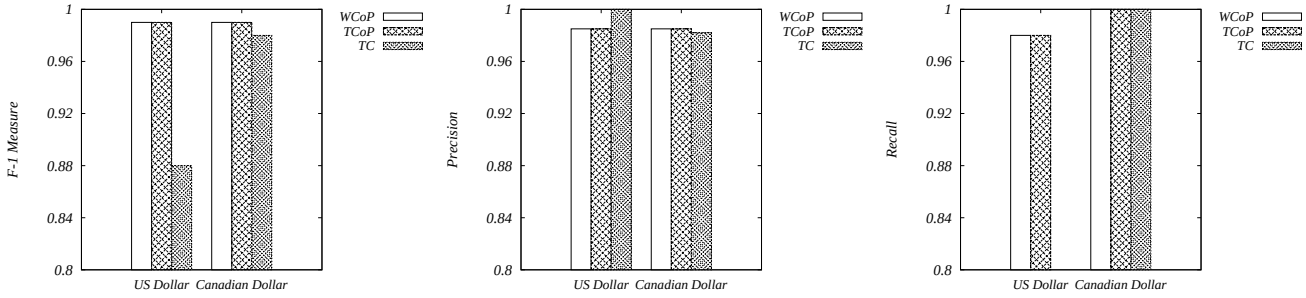
(a) F1 Measure (b) Precision (c) Recall

Figure 8. F1, Precision and Recall of WCoP, TCoP and Text Classification on sample pages of "places" domain



(a) F1 Measure (b) Precision (c) Recall

Figure 9. F1, Precision and Recall of WCoP, TCoP and Text Classification on sample pages of "Programming Languages" domain



(a) F1 Measure (b) Precision (c) Recall

Figure 10. F1, Precision and Recall of WCoP, TCoP and Text Classification on sample pages of "Currencies" domain

measuring context, these words can cause false positives. Finally, when a page undergoes sudden and drastic context change (e.g., death of a person, revolutions in countries, etc.), our system might wrongly mark the edit reflecting the context change as vandalism. As remarked earlier, one way to address the last scenario is to utilize from realtime event sources such as Twitter and news feeds.

Our system might fail to detect vandal edits that do not contain contextually-mismatched attributes. For example, a vandal edit that removes certain key sentences in the document (which may be 'inconvenient truths' from the perspective of the user performing the edit) will not be identified by content-context-centric technique. The content-context-centric may also fail to detect edits that vandalize Wikipedia articles

using common words (e.g., 'death', 'divorce', etc.) or words that have multiple connotations (e.g., disaster, etc.). Using these kinds of words/phrases, a smart and determined adversary can construct sentences that bring disrepute to the article's topic without being detected by our system. These sort of vandal edits are not common, but nevertheless do occur. Furthermore, we believe using other types of contexts (spatial and contributor context) can help mitigate some of these false negatives.

In summary, our experiments demonstrate that utilizing context provides significant improvement in vandalism detection accuracy.

7. Related Work

In recent years, various aspects of Wikipedia have been extensively studied, including its efficacy as a collaborative knowledge sharing platform, the demographies and behaviors of its user population, quality and trustworthiness of its information, semantic-analysis and ontology development for Wikipedia, and the effect of Wikipedia on various societies around the world [1, 13, 15–21]. The study by Kittur et al. [15] indicates that n the early days of Wikipedia, a core group of editors performed bulk of the editing. However, as Wikipedia became more popular, the contributions from common users has drastically increased. In another study, Kittur and Kraut [19] distinguish between implicit and explicit collaborations in Wikipedia and conclude that explicit collaboration (through discussion and talk pages) yields better quality content than implicit collaboration.

A number of researchers have studied vandalism in Wikipedia. Preidhorsky et al. [3] attempt to estimate the value of Wikipedia content. In this context, they analyze the damage done by vandal edits in terms of the length of time the article was in vandalized status and the number of views on the article when it was in the vandalized state. Existing Wikipedia vandalism detection techniques can be broadly classified into two categories, namely, content-based and behavior-based approaches [22]. Both of these approaches use either rule-based or machine learning-based classifiers in the background [23]. Features that are typically used in content-based approaches include edit types (such as complete or partial *blanking*, inclusion of repetitive text) insertion of obscene words, spammy words, or spammy URLs, inversion of statement meanings, replacement of article titles and sub-titles, changing numbers in articles, length of comments, size of edit, and character diversity of edit [2–5, 24]. Chin et al. have used statistical language models have for vandalism detection [25]. The work by Wang and McKeown [26] utilizes lexical features such as misplaced punctuations and slangs for detecting vandalism. In a recent work, Wu et al. have proposed a text-stability-based approach for identifying vandalism [5]. The main idea here is to quantify the stabilities of various parts of a Wikipedia article (in terms of number of versions, number of views and amount of time since last modification), and use them to predict the likelihood of these parts being modified through legitimate edits.

The behavior-based approach relies upon Wikipedia revision history to generate user behavior models which are later used to classify edits [27–29]. Reputation-based techniques form an important stream of work in this direction [9–11]. Adler et al. [30] propose a vandalism detection technique that combines computation linguistics with contributor reputation. Reputation-based techniques are similar to our approach of utilizing contributor-context for vandalism detection. A closely related stream of work is that of user community-based trust enhancement techniques for collaborative social media [31, 32].

Spamming, while not being the sole motivation for vandalism, certainly contributes to a considerable portion of it. Researchers have proposed many spam resistance approaches, including white and black lists, statistical filtering, network analysis, and sender authentication, and coordinated real-time spam filtering [33–39]. However, the anti-spam work does not completely address the vandalism problem because while spam is mostly driven by financial interests, vandalism can be generated by a variety of causes.

Context-awareness has been widely studied in the pervasive computing and human-computing interaction domains [40–43]. Several issues including developing infrastructures for capturing and maintaining context, analysis of context and security and privacy aspects of context have been explored. Our work is unique in that it uses context for vandalism detection in CSM applications.

8. Conclusions

In recent years, vandalism has emerged as a significant threat to information quality and trustworthiness of collaborative social media application such as Wikipedia. Many of the existing vandalism detection techniques rely upon simple textual features, and hence are not very effective in dealing with sophisticated vandal attacks. In this paper, we proposed harnessing context for vandalism identification. We presented a unique context-enhanced finite state model for Wikipedia article evolution which helps us capture and analyze various contextual attributes. This paper studies the distinguishing capabilities of two important types of context namely content-context and contributor-context. We have designed two metrics, namely, WWW co-occurrence probability and top

ranked co-occurrence probability, to measure the compatibility of an edit's content-context with the content-context of the existing article. In addition to providing efficient mechanisms for estimating these metrics, we have discussed how these metrics can be utilized in machine learning-based classifiers. This paper also reports several experiments on the Wikipedia PAN corpus that demonstrate that utilizing context significantly improves vandalism detection accuracy when compared with simple text-based techniques.

Acknowledgement

This research is partially supported by the National Science Foundation under grants CNS-1338276, DUE-1318881, OCI-1127195, CNS/SAVI-1250260, IUCRC/FRP-1127904, CISE/CNS-1138666, RAPID-1138666, CISE/CRI-0855180, NetSE-0905493 and gifts, grants, or contracts from Intel Corp, DARPA/I2O, Singapore Government, Fujitsu Labs, and Georgia Tech Foundation through the John P. Imlay, Jr. Chair endowment. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation or other funding agencies and companies mentioned above.

References

[1] Pew Research Center's Project for Excellence in Journalism (2011), The State of the News Media-2011 (An Annual Report on American Journalism), http://stateofthemedia.org/.

[2] Wikipedia (Revision as of 20:29, 22 May 2010), Cluebot, http://en.wikipedia.org/wiki/User:ClueBot.

[3] Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L. and Riedl, J. (2007) Creating, destroying, and restoring value in wikipedia. In *Proceedings of the International ACM Conference on Supporting Group Work*: 259–268.

[4] Adler, B.T. and de Alfaro, L. (2007) 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): 261–270. doi:http://doi.acm.org/10.1145/1242572.1242608.

[5] Wu, Q., Irani, D., Pu, C. and Ramaswamy, L. (2010) Elusive vandalism detection in wikipedia: a text stability-based approach. In *CIKM*.

[6] Wikipedia, Vandalism on Wikipedia (retrieved on Aug 01, 2013), http://en.wikipedia.org/wiki/Vandalism_on_Wikipedia.

[7] Ahamad, M., Mark, L., Lee, W., Omicienski, E., dos Santos, A., Liu, L. and Pu, C. (2002) Guarding the next Internet frontier: countering denial of information attacks. In *NSPW*.

[8] Wikipedia, Wikipedia Article on AI Complete Problem, http://en.wikipedia.org/wiki/AI-complete.

[9] Adler, B.T., Benterou, J., Chatterjee, K., de Alfaro, L., Pye, I. and Raman., V. (2007) Assigning trust to wikipedia content. In *Technical Report, School of Engineering, University of California, Santa Cruz*.

[10] Javanmardi, S. and Lopes, C. (2007) Modeling trust in collaborative information systems. In *Proceedings of the 3rd International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2007)*: 299–302.

[11] Zeng, H., Alhossaini, M., Fikes, R. and McGuinness, D.L. (2006) Mining revision history to assess trustworthiness of article fragments. In *Proceedings of the 4th International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2006)*.

[12] Bizer, C., Lehmann, J., Kobilarov, G., Auer, S., Becker, C., Cyganiak, R. and Hellmann, S. (2009) Dbpedia: A crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web* **vol.7**(3): 154–165.

[13] Suchanek, F.M., Kasneci, G. and Weikum, G. (2008) Yago: A large ontology from wikipedia and wordnet. *Web Semantics: Science, Services and Agents on the World Wide Web* **6**(3): 203–217.

[14] Potthast, M. (2010) Crowdsourcing a wikipedia vandalism corpus. In *Proceedings of SIGIR*.

[15] Kittur, A., Chi, E., Pendleton, B.A., Suh, B. and Mytkowicz, T. (2007) Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. *World Wide Web* **1**(2).

[16] Suh, B., Ch, E.H., Kittur, A. and Pendleton, B.A. (2008) Lifting the veil: improving accountability and social transparency in Wikipedia with wikidashboard. In *CHI*.

[17] Suh, B., Ch, E.H., Pendleton, B.A. and Kittur, A. (2007) Us vs. Them: Understanding Social Dynamics in Wikipedia with Revert Graph Visualizations. In *IEEE VAST*.

[18] Kittur, A., Suh, B. and Chi, E.H. (2008) Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. In *CSCW*.

[19] Kittur, A. and Kraut, R. (2008) Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *CSCW*.

[20] Gabrilovich, E. and Markovitch, E. (2007) Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *IJCAI*.

[21] Ponzetto, S.P. and Navigli, R. (2009) Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia. In *IJCAI*.

[22] Potthast, M., Stein, B. and Gerling, R. (2008) Automatic vandalism detection in wikipedia. In *Proceedings of Advances in Information Retrieval*: 663–668.

[23] Smets, K., Goethals, B. and Verdonk, B. (2008) Automatic vandalism detection in wikipedia: Towards a machine learning approach. In *Proc. of AAAI workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy* (AAAI): 43–48.

[24] Mola-Velasco, S.M. (2011) *Wikipedia Vandalism Detection*. Master's thesis, Polytechnic University of Valencia.

[25] chi Chin, S., Srinivasan, P., Street, W.N. and Eichmann, D. (2010) Detecting wikipedia vandalism with active learning and statistical language models. In *Proceedings of 4th Workshop on Information Credibility on the Web*.

[26] Wang, W.Y. and McKeown, K.R. (2010) "Got You!": Automatic Vandalism Detection in Wikipedia with Web-based Shallow Syntactic-Semantic Modeling. In *COLING*.

[27] Hu, M., Lim, E., Sun, A., Lauw, H.W. and Vuong., B. (2007) Measuring article quality in wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM)*: 243–252.

[28] Lim, E., Vuong, B., Lauw, H.W. and Sun, A. (2006) Measuring qualities of articles contributed by online communities. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*: 81–87.

[29] Halfaker, A., Kittur, A., Kraut, R. and Riedl, J. (2009) A jury of your peers: quality, experience and ownership in wikipedia. In *WikiSym '09: Proceedings of the 5th International Symposium on Wikis and Open Collaboration* (New York, NY, USA: ACM): 1–10. doi:http://doi.acm.org/10.1145/1641309.1641332.

[30] Adler, B.T., de Alfaro, L., Mola-Velasco, S.M., Rosso, P. and West, A.G. (2011) Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In *CICLing*.

[31] Caverlee, J., Cheng, Z., Eoff, B., Hsu, C.F., Kamath, K., Kashoob, S., Kelley, J. *et al.* (2010) Socialtrust++: Building community-based trust in social information systems. In *CollaborateCom*.

[32] Caverlee, J., Cheng, Z., Eoff, B., Hsu, C.F., Kamath, K.Y. and McGee, J. (2011) Crowdtracker: enabling community-based real-time web monitoring. In *SIGIR*.

[33] Androutsopoulos, I., Koutsias, J., Chandrinos, K., Paliouras, G. and Spyropoulos, C. (2000) An evaluation of naive bayesian anti-spam filtering. In *Proceedings of the workshop on Machine Learning in the New Information Age, 2000.*: 9–17. URL citeseer.ist.psu.edu/androutsopoulos00evaluation.html.

[34] Webb, S., Chitti, S. and Pu, C. (2005) An experimental evaluation of spam filter performance and robustness against attack. In *Proceedings of the 1st International Conference on Collaborative Computing: Networking, Applications and Worksharing (CollaborateCom 2005)*.

[35] Schryver, V., Distributed checksum clearinghouse. http://www.rhyolite.com/anti-spam/dcc/ Last accessed Nov 2, 2005.

[36] Gray, A. and Haahr, M. (2005) Personalised, Collaborative Spam Filtering. In *Proceedings of the Second Email and SPAM conference (CEAS)*.

[37] Damiani, E., di Vimercati, S.D.C., Paraboschi, S. and Samarati, P. (2004) P2p-based collaborative spam detection and filtering. In *The Fourth International Conference on Peer-to-Peer Computing*. URL citeseer.ist.psu.edu/721025.html.

[38] Ramachandran, A., Feamster, N. and Vempala, S. (2007) Filtering spam with behavioral blacklisting. In *ACM Conference on Computer and Communications Security (CCS)*.

[39] Ramachandran, A. and Feamster, N. (2006) Understanding the Network-Level Behavior of Spammers. In *Proceedings of ACM SIGCOMM 2006*.

[40] Chen, G. and Kotz, D. (2000) *A survey of context-aware mobile computing research*. Tech. rep., Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College.

[41] Dey, A.K. (2001) Understanding and using context. *Personal and ubiquitous computing* **5**(1).

[42] Hong, J.I. and Landay, J.A. (2001) An infrastructure approach to context-aware computing. *Human-Computer Interaction* **16**(2).

[43] Smailagic, A. and Kogan, D. (2002) Location sensing and privacy in a context-aware computing environment. *Wireless Communications, IEEE* **9**(5): 10–17.