A Content-Context-Centric Approach for Detecting Vandalism in Wikipedia

Lakshmish Ramaswamy
Dept. of Computer Science
The University of Georgia
Athens, GA 30602
laks@cs.uga.edu

Raga Sowmya Tummalapenta
Dept. of Computer Science
The University of Georgia
Athens, GA 30602
sowmya@uga.edu

Kang Li
Dept. of Computer Science
The University of Georgia
Athens, GA 30602
kangli@cs.uga.edu

Calton Pu
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332
calton@cc.gatech.edu

***Abstract*—Collaborative online social media (CSM) applications such as Wikipedia have not only revolutionized the World Wide Web, but they also have had a hugely positive effect on modern free societies. Unfortunately, Wikipedia has also become target to a wide-variety of vandalism attacks. Most existing vandalism detection techniques rely upon simple textual features such as existence of abusive language or spammy words. These techniques are ineffective against sophisticated vandal edits, which often do not contain the tell-tale markers associated with vandalism. In this paper, we argue for a context-aware approach for vandalism detection. This paper proposes a content-context-aware vandalism detection framework. The main idea is to quantify how well the words contained in the edit fit into the topic and the existing content of the Wikipedia article. We present two novel metrics, called WWW co-occurrence probability and top-ranked co-occurrence probability for this purpose. We also develop efficient mechanisms for evaluating these two metrics, and machine learning-based schemes that utilize these metrics. The paper presents a range of experiments to demonstrate the effectiveness of the proposed approach.**

Keywords*—*Collaborative online social media, vandalism detection, content-context, WWW co-occurrence probability, top-ranked co-occurrence probability

I. Introduction

Collaborative online social media applications (a.k.a social information systems) such as Wikipedia are radically transforming the World Wide Web (WWW). These applications have elevated the end-users from being passive consumers of information to ones that actively participate in generation, organization and propagation of information on the web. By facilitating *democratization of information* and *collective intelligence*, collaborative online social media (CSM) applications have had a hugely positive impact on modern free societies. End-user anonymity and low barrier for information sharing are among the prominent features that have made Wikipedia and other CSM applications widely popular.

Considering the increasingly important role that Wikipedia is playing in the modern world, it is important to ensure the trustworthiness of the information that gets shared on it. Unfortunately, the very foundational features of Wikipedia namely end-user anonymity and low information sharing barrier have made it susceptible to a variety of *vandalism attacks*. These include injection of false information into Wikipedia articles, removal of legitimate information from articles, and spamming (for commercial, ideological or other purposes). Studies show that around 5% of Wikipedia edits involve vandalism. Some of these edits were not rectified for several hours (in some, albeit infrequent, cases even days). In addition to exposing false information to Wikipedia users, vandalism has the potential to inflict wider damage. It can cause progressive degradation of quality of information [1] which can lead to frustration among honest contributors, some of whom may loose interest in contributing content and participating in Wikipedia activities. More importantly, vandalism can create social tensions and may even lead to violence in certain regions of the world. Thus, it is important to develop effective techniques for detecting vandalism in Wikipedia as well as other CSM applications.

Most existing works in this area focus on utilizing simple textual features for identifying vandalism. They work by considering whether an edit contains features that have statistically high likelihood of being associated with vandalism. Examples of such features include abusive/obscene words, spammy words/phrases (e.g., Viagra, Gucci watches), and certain URLs. These simple approaches, however, have had limited success in combating sophisticated vandal edits often referred to as *elusive vandalism* [2]. These type of vandal attacks are not likely to contain the tell-tale textual features associated with vandalism, and hence they evade common vandalism filters. Studies have shown that elusive vandalism is a growing problem .

This paper argues for a *context-aware approach* for detecting vandalism in Wikipedia. The main motivation for considering context is our important observation that the edits in Wikipedia and other CSM applications are not isolated pieces of text. Rather, they happen in a specific *context*. This is in fact a key feature of Wikipedia, and hence it can be highly effective in detecting vandalism. The context of a Wikipedia edit can have multiple distinct aspects such as the relationship of the edit to the article, whether the edit was performed by a registered or an unregistered user, the identity (or the IP address) of the user performing the edit, and the geographical location from where the edit was performed. The challenge however lies in designing vandalism detection techniques that can effectively harness these various contextual attributes.

In this paper, we focus on a specific aspect of context, namely, the relationship between an incoming edit and the Wikipedia article at a syntactic level. We refer to this as *content-context*. In a nutshell the main idea is to check how well the words contained in the edit fit into the topic and the existing content of the Wikipedia article. Intuitively, if the

COLLABORATECOM 2013, October 20-23, Austin, United States

DOI 10.4108/icst.collaboratecom.2013.254059

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.

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

words contained in the edit are unrelated to the topic and the current content of the Wikipedia article, it is a strong indicator of vandalism.

The technical contributions of this paper are three fold.

- We present two unique content-based metrics for quantifying how compatible an edit is with the context of a Wikipedia article. The first metric, called *WWW co-occurrence probability* quantifies how often the words in the edit and words in the document appear together (i.e., in the same document) in the corpus of World Wide Web (WWW) documents. The second metric, called *top-ranked co-occurrence probability* is based upon a similar theme, but the corpus is limited to top-ranked (hence, presumably high-quality) WWW documents.
- We develop efficient mechanisms for computing the above metrics, and machine learning-based vandalism detection techniques that utilize these metrics.
- We present a detailed experimental study evaluating the accuracy of the proposed content-context-centric classifiers over the Wikipedia vandalism PAN corpus and using automatic labeling strategy.

The rest of the paper is organized as follows. Section II discusses background material on Wikipedia vandalism, and motivates the need for context-based approaches for vandalism detection. In Section III, we discuss our two content-context-centric vandalism detection techniques. Section IV presents our experimental evaluation. In Section V, we review related works and conclude the paper in Section VI.

II. Background and Motivation

Wikipedia itself defines vandalism "as an act that is intentionally disruptive" [3]. It can also be defined as a deliberate act aimed at lowering the quality of information on Wikipedia. Vandalism may involve addition of false information (including unverifiable rumors), injection of abusive/obscene material, removal of legitimate information and spamming. While vandalism can appear in any Wikipedia page, articles pertaining to controversial topics and personalities are likely to be the targets of a large fraction of vandal edits.

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.

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 as *elusive vandalism*, often do not contain the textual features associated with vandalism. For example, they may not have any abusive/obscene words even when the intent is to belittle the topic of a Wikipedia article.

For example, Fig. 1 shows the Wikipedia page of "Liberalism" as it appeared on 06/05/2010 at 11:05 GMT. The version shown in the figure was the result of a vandal attack that introduced the sentence "Liberalism is the belief in the importance of big daddy government". Similarly, on 02/23/2010 at 15:49 GMT, the Wikipedia page on Geriatrics was vandalized by changing a section heading 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. Thus, traditional vandalism filters fail in these and such other instances.

A. Why Consider Context?

One of the central limitations of traditional vandalism detection techniques is that most of them treat edits as independent and isolated pieces of texts. Because of this, most of them just focus on the text that appears in the edit. However, edits in Wikipedia are not isolated pieces of text. They occur in certain *context*, and and hence the contextual attributes are an integral part of an edit's characteristics. For instance, an edit occurs on a certain version of an article. Thus, the edit cannot be completely characterized without including the content of the article at the time the edit occurred. In fact, the edit may become meaningless if it were to be performed on a different article or a different version of the same article.

In addition to article and version, an edit carries with it several other powerful contextual attributes. These include the identity (or lack thereof) of the user performing the edit, the previous history of edits performed by the user, the geographical location from where the edit originated, and the time at which the edit was performed. 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. For example, most humans will immediately identify an edit containing the word "Nazi" as vandalism if the edit is on, say, President Obama's Wikipedia page, whereas they will not classify the same edit as vandalism if it is on Goebbels' page. The human is implicitly relying on whether the edit fits into the overall context of the article to determine whether it is vandalism.

There are many challenges to utilizing context for vandalism detection. First, we need to identify contextual attributes that have strong distinguishing capabilities. Second, 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 measurement mechanisms. Third, we need to design efficient and scalable vandalism detection techniques that utilize these quantifiable contextual attributes.

In this paper, we focus on a specific type of context that we refer to as *content-context*. We discuss our strategies to address the above challenges with respect to content-context, and we present machine-learning-based vandalism detection techniques that utilize content-context.

III. Content-Context-Centric Vandalism Detection

At a very high level, our idea 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 represent the current version of a Wikipedia document and let E represent an incoming edit on D. The idea is to check how well the content being introduced by E gels with content existing in D. The central observation is that if the edit E is legitimate (non-vandal), the content of E will fit well into the content of D, 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. The challenge, however, is to devise a precise metric for measuring the extent to which the content of an incoming edit fits into the context of the existing article.

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* [4], and hence impractical. In this paper, 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 exiting 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*.

A. WWW co-occurrence Probability Metric

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 $W(D) = \{wd_1, wd_2, \ldots, wd_n\}$ be the set of keywords in the current (non-vandalized) version of the document. (i.e., $W(D)$ is the current context of the document D) and $W(E) = \{we_1, we_2, \ldots, we_n\}$ denote the set of words that the edit E is seeking to introduce in the next version of the document (i.e., $W(E)$ is the edit's context). The co-occurrence probability of the arbitrary keyword pair (we_i, wd_j) is defined as the ratio of the probability that both we_i and wd_j occur in an arbitrary WWW document to the ratio that at least one of them occurs in a WWW document. Mathematically,

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

In the above equation, DC denotes an arbitrary WWW document. The denominator in Equation 1 is a normalization

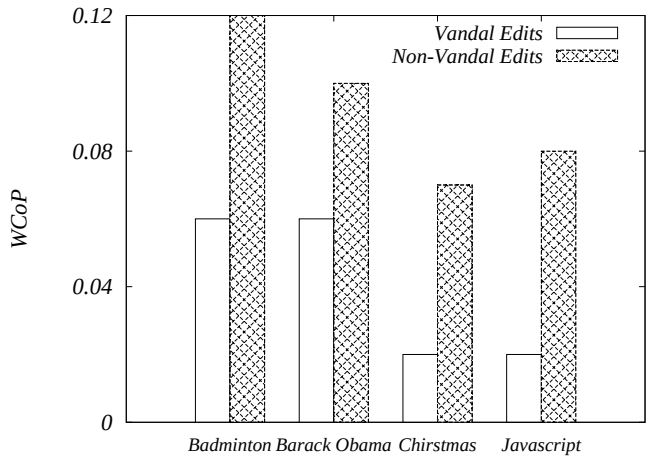


Fig. 2: WCoP Values for Vandal and Non-vandal edits

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.

$$WCoP(E,D) = \underset{we_i \in W(E), wd_j \in W(D)}{\text{argmin}} (CoP(we_i, wd_j)) \quad (2)$$

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.

In order to validate the distinguishing capabilities of the WWW co-occurrence probability 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 WCoP value between the edit and version that was existing before the edit happened. In Fig. 2, 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 WWW co-occurrence probability can be a powerful factor in distinguishing vandal edits from non-vandal ones.

Efficient Estimation Technique:

We need an efficient mechanism for computing the WWW co-occurrence probability metric. The central issue here is to estimate the CoP between various we_i-wd_j 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 $CoP(we_i, wd_j)$. We first issue a search query for documents containing both we_i and we_j (i.e, the search query will be we_i + wd_j). 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_i and wd_j 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_i - wd_j and wd_j - we_i. Let Ne_i and Nb_j be the estimates on the number of search results for these two queries respectively. Now $CoP(we_i, wd_j)$ is estimated as $\frac{Nb}{(Ne_i+Ne_j+NB)}$.

An associated problem in computing the WWW co-occurrence probability metric is that the keyword set corresponding to the current version of the document ($W(D)$) 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 $W(D)$ to the keywords in the title of the article and its introductory paragraphs. In our experiments (see Section IV), we limit $W(D)$ to the keywords in the document's title.

B. 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 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. In the interest of brevity, we do not provide the formal definition here. Instead, we focus on the technique to estimate the top ranked co-occurrence probability. Suppose we want to estimate the top ranked co-occurrence between the edit-document keyword pair we_i and wd_j. We issue separate search queries for we_i and we_j. Let $Tr^K(we_i)$ and $Tr^K(wd_j)$ denote the top K search results for we_i and we_j (K is a configurable parameter). The top K co-occurrence probability of the keywords we_i and wd_j is defined as $TrCoP^K(we_i, wd_j) = \frac{|Tr^K(we_i) \cap Tr^K(wd_j)|}{|Tr^K(we_i) \cup Tr^K(wd_j)|}$. Note that $(Tr^K(we_i) \cap Tr^K(wd_j))$ denotes the set of top K search results that contain *both* we_i and wd_j.

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

$$TCoP^K(E,D) = \underset{we_i \in W(E), wd_j \in W(D)}{\text{argmin}} (TrCoP^K(we_i, wd_j)) \quad (3)$$

TABLE I: Wikipedia Domains and Sample Pages

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

As with WWW co-occurrence probability, in order to reduce computational overheads, $W(D)$ can be limited to the keywords in the title of the article and its introductory paragraphs.

C. Vandalism Detection Algorithm

Our algorithm employs machine learning-based classifiers for detecting vandalism. For each incoming edit, we extract the keywords of the incoming edit and the keywords from the existing version to construct W(E) and W(D) respectively. We use a popular search engine to compute the WCoP and TCop values. These values are fed into machine learning-based classifiers that have been trained on known vandal and non-vandal edit instances. The machine learning-based classifiers determine whether the edit is vandalism.

In addition to WCoP/TCop, the machine language-based 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 [2]. 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.

IV. Experiments and Results

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

A. 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 [5]. 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” 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 [2].

Since our technique involves quantifying the content-contexts of edits with respect to the corresponding article versions, we need the entire edit histories of article (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* [2], 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 .*

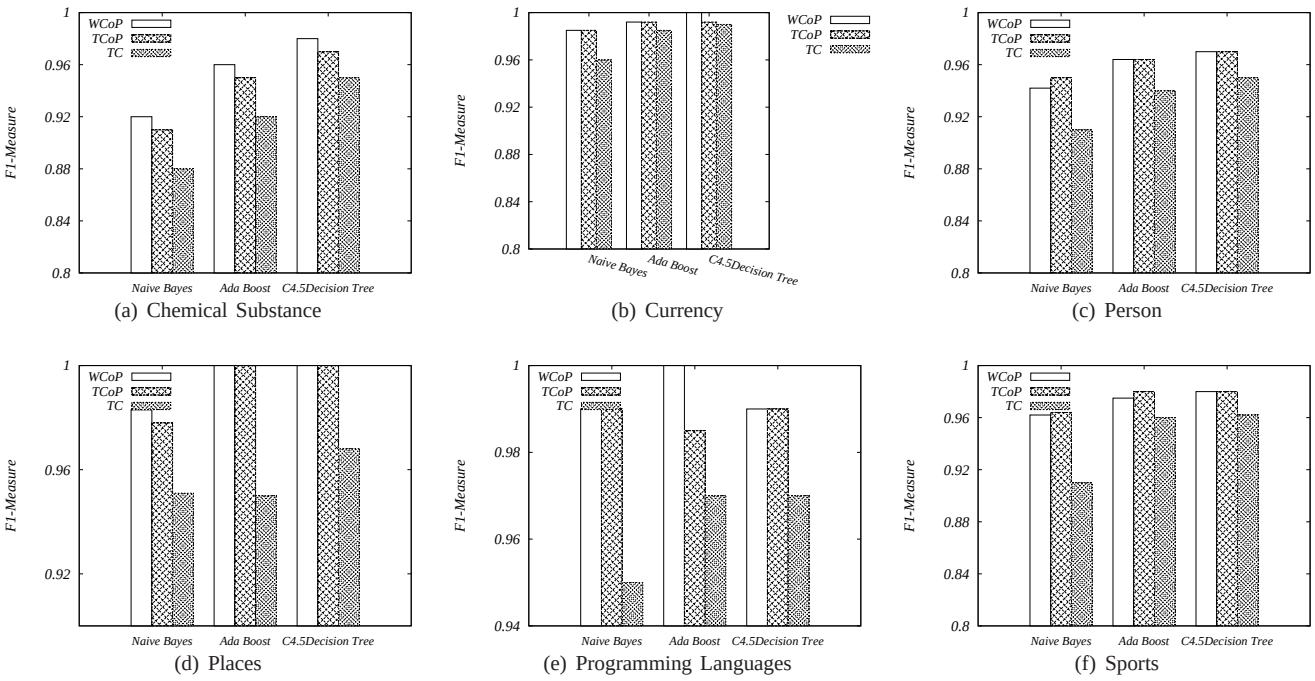
If an edit was contributed by a super user or if the version was not reverted or if the comments for the version does not contain the above patterns, then it is considered to be a regular edit.

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 6 different domains, namely, Chemical Substances, Currencies, Places, Persons, Programming Languages and Sports. Sample pages from each domain are listed in Table I. For each page, we select the 100 most recent vandal versions and 100 most recent regular versions.

B. 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 III) 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



(a) Chemical Substance (b) Currency (c) Person

(d) Places (e) Programming Languages (f) Sports

Fig. 3: Comparison of WCoP, TCoP and Text Classification Performance on Various Wikipedia Domains

co-occurrence probability, top ranked co-occurrence probability and the textual classifier).

C. Results

Fig. 3(a) through Fig. 3(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.

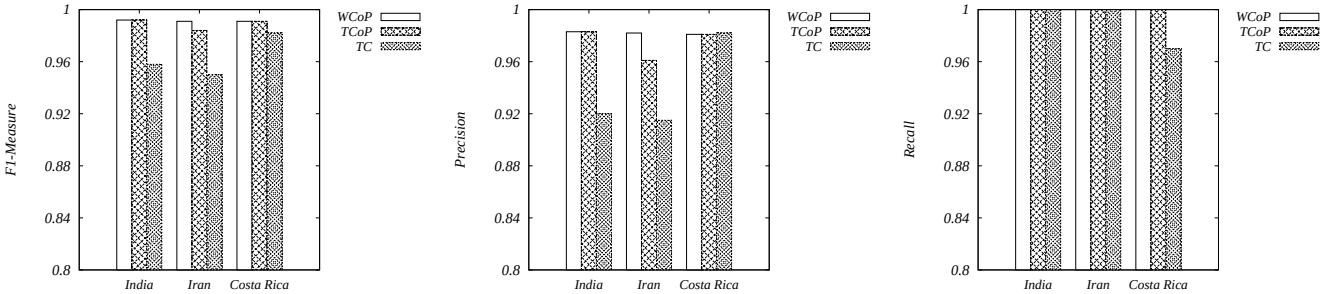
From these results it can be seen that WCoP and TCoP consistently outperform TC on all domains and on all classifiers. For example, both WCoP and TCoP yield 6.5% higher F1 scores when compared with TC on the "Sports" domain with Naive Bayes classifier. Note that a large fraction of the vandal edits in 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 two domains namely, "places" and "programming languages". These experiments were done using the C4.5 Decision tree classifier with 10-fold cross validation. Fig. 4(a), Fig. 4(b) and Fig. 4(c) respectively indicate the F1 score, precision and recall for three pages from the "places" domain. Similarly, Fig. 5(a), Fig. 5(b) and Fig. 5(c) respectively indicate the F1 score, precision and recall for two pages from the "Programming Languages" 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.

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

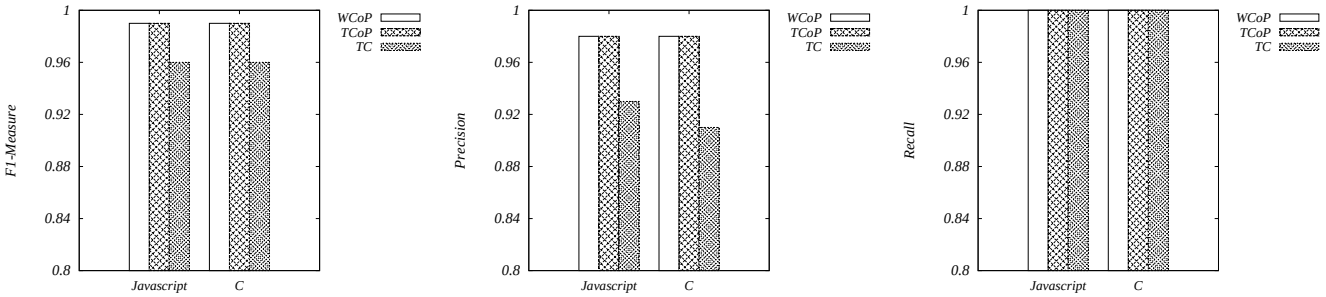
V. Related Work

Existing work on Wikipedia vandalism detection can be broadly classified into two categories, namely, content-based and behavior-based approaches. Both of these approaches use either rule-based or machine learning-based classifiers in the background. 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, and changing numbers in articles [6], [7], [8], [2]. Chin et al. have used statistical language models have for vandalism detection [9]. In a recent work, Wu et al. have proposed a text-stability-based approach for identifying vandalism [2]. 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.



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

Fig. 4: F1, Precision and Recall of WCoP, TCoP and Text Classification on sample pages of "places" domain



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

Fig. 5: F1, Precision and Recall of WCoP, TCoP and Text Classification on sample pages of "Programming Languages" domain

The behavior-based approach relies upon Wikipedia revision history to generate user behavior models which are later used to classify edits [10], [11], [12]. Reputation-based techniques form an important stream of work in this direction [13], [14], [15]. Reputation based models implicitly assume that users with good contribution histories will not indulge in vandalism. However, reputation alone is not always a reliable indicator of vandalism.

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 [16], [17], [18], [19], [20], [21]. 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.

In summary, while there have been several works on identifying vandalism in Wikipedia, very few of them consider context. Our work is unique in the sense that we demonstrate that content-context is a powerful feature for identifying vandalism.

VI. Conclusion

Vandalism is a growing problem for Wikipedia and other collaborative social media applications. Vandalism detection techniques that are based upon simple textual features have not been very effective in combating sophisticated vandal attacks. In this paper, we have proposed a content-context-centric approach for vandalism detection in Wikipedia. The main idea is to measure the compatibility of the incoming edit's content with the context of the existing article. We have presented two metrics, namely, WWW co-occurrence probability and top ranked co-occurrence probability, to measure the compatibility of the edit's keywords with the keywords of the existing article. The paper also provides efficient mechanisms for estimating these metrics. These features are used in machine learning-based classifiers. Our experiments on Wikipedia vandalism PAN corpus have demonstrated that the content-context features significantly improve vandalism detection accuracy when compared with simple textual features.

Acknowledgment

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] R. Priedhorsky, J. Chen, S. T. K. Lam, K. Panciera, L. Terveen, and J. Riedl, "Creating, destroying, and restoring value in wikipedia," in *GROUP '07: Proceedings of the 2007 international ACM conference on Supporting group work*. New York, NY, USA: ACM, 2007, pp. 259–268.

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

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

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

[5] M. Potthast, "Crowdsourcing a wikipedia vandalism corpus," in *Proceedings of SIGIR*, 2010.

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

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

[8] 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.

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

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

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

[12] A. Halfaker, A. Kittur, R. Kraut, and J. Riedl, "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, 2009, pp. 1–10.

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

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

[15] H. Zeng, M. Alhossaini, R. Fikes, and D. L. McGuinness, "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)*, May 2006.

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

[17] S. Webb, S. Chitti, and C. Pu, "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)*, December 2005.

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

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

[20] E. Damiani, S. D. C. di Vimercati, S. Paraboschi, and P. Samarati, "P2p-based collaborative spam detection and filtering," in *The Fourth International Conference on Peer-to-Peer Computing*, August 2004. [Online]. Available: citeseer.ist.psu.edu/721025.html

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