A LARGE SCALE STUDY OF EDIT PATTERNS IN WIKIPEDIA AND ITS APPLICATIONS TO VANDALISM DETECTION

by

DEEPIKA SETHI

(Under the Direction of Prof. Lakshmish Ramaswamy)

ABSTRACT

In recent years, Web 2.0 applications such as Wikipedia have transformed the landscape of the World Wide Web by elevating the end-users from being passive consumers of information to ones that actively participate in content creation, organization and propagation. Wikipedia is a free online encyclopedia where any user can edit information with minimal restriction. Recent studies indicate that a large fraction of Internet users rely on Wikipedia for their information needs. Thus, it is immensely important to ensure the quality and accuracy of information that is shared on Wikipedia. Ironically, the open-edit nature of Wikipedia has also made it susceptible to various kinds of vandalism attacks.

In this thesis, we perform a large-scale study of the edit patterns of Wikipedia articles. The goal of this study is to identify meta-data characteristics that can help us distinguish between high-quality edits and potential vandalism attacks. Our study is unique in several different aspects. Firstly, we trace the history of edits of Wikipedia articles and study the stability of articles, their growth over time, and the nature of users who perform the edits. Secondly, we study the spatial distributions of the origin of the edits. Thirdly, we also study the commonality of content and commonality of users among various Wikipedia articles. Through this study, we show that

various types of contextual attributes of edits such as co-occurrence probabilities of words, registration status of edit contributors, and geographical region of origin of edits have strong distinguishing capabilities with regards to vandalism.

INDEX WORDS: WIKIPEDIA, VANDALISM, STUDY OF WIKIPEDIA, VANDAL CHARACTERISTICS

A LARGE SCALE STUDY OF EDIT PATTERNS IN WIKIPEDIA AND ITS APPLICATIONS TO VANDALISM DETECTION

by

DEEPIKA SETHI

B.Tech, B.C.E.T Gurdaspur, India, 2007

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA
2012

A LARGE SCALE STUDY OF EDIT PATTERNS IN WIKIPEDIA AND ITS APPLICATIONS TO VANDALISM DETECTION

by

DEEPIKA SETHI

Major Professor: Lakshmish Ramaswamy

Committee: Kang Li
Roberto Perdisci

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
December 2012

ACKNOWLEDGEMENTS

Studying in University of Georgia was a great experience and past 2.5 years are the most memorable years of my life. I learned a lot from the astute faculty. Dr. Lakshmish Ramaswamy has been a huge support throughout my degree. He always guided me in the right direction and motivated me. I would like to thank Dr. Kang Li and Dr. Roberto Perdisci for being a part of my committee. I would also like to thank my friend Raga Sowmya for her support; we started our research together on vandalism detection and it was fun working with her.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Motivation

Wikipedia is a free encyclopedia where anyone can edit information. Wikipedia includes over 22.9 million freely usable articles in 285 languages, written by over 36 million registered users and numerous anonymous contributors worldwide [14]. There are around 4,046,834 articles in English Wikipedia [13].

Wikipedia is a volunteer open source project characterized by low ties between contributors, no formal obligations and very few means for the exercise of formal sanction. The wiki technology is open, inviting many to the task and imposing low costs on participation while reducing transaction costs [15].

Wikipedians are people who write and edit the pages for Wikipedia, unlike readers who simply read the articles. To become a Wikipedian, we have to click the edit link at the top of any page, or at the beginning of each section. The number of named accounts is currently 17,701,915. Only minorities of account holders are regular contributors, and only a minority of those users interacts in discussions about the community. An unknown but relatively large number of unregistered Wikipedians also contribute to the site [16]. Users are divided into various user groups like Administrators who have access to page deletion, page protection, blocking and unblocking buttons, ability to edit protected pages, mediawiki interface, ability to grant and remove rollback, flood and ip-block exempt rights to users , Anonymous users who

have not created an account and they may create new pages and edit existing pages that are not protected, Registered users who have an account and can create new pages and edit existing pages and get automatic access into Autoconfirmed/Established users group when their account is four days old. There are other flags like Rollback, CheckUser, Oversighter, Bots etc. [17]. Wikipedia is criticized on the principle of being open for editing by everyone making Wikipedia un-authoritative and unreliable and further for allowing editors to contribute anonymously. Its critics claim that the consequences of this include a lack of authority and accountability and poor quality of discourse [20].

The open collaborative nature of Wikipedia allows it to be modified by multiple users and this helps in collaboration of knowledge of millions of people but this collaborative nature of Wikipedia leads to a big problem i.e. Vandalism.

Vandalism is a deliberate activity attempted to compromise the integrity of Wikipedia. Changes such as add, remove or edit correct information into incorrect information are done. There are various types of vandalism that occurs and hampers the integrity of Wikipedia. Several types of vandalism are blanking a page, adding malicious information into pages and creating malicious accounts, addition of obscenities or crude humor, insertion of nonsense into the pages, elusive vandalism[7] in which words added are very hard to detect, lengthening of a page which makes it difficult to load pages and several other types.

A large number of edits on Wikipedia pages are done daily and vandalism in the pages misleads the readers. If proper measures are not taken to remove vandalism, Wikipedia may lose its popularity and people may no longer rely on Wikipedia for their information needs. Hence, there is a need to study vandalism and devise various factors that are responsible for vandalism.

Wikipedia has taken many measures to address the challenges of vandalism, such as restricting the privileges of anonymous users, adopting "article validation" and using an "abuse filter" to control user activities by reacting automatically to suspicious user behaviors. Currently active tools to fight vandalism include ClueBot and VoABot II. The two anti-vandal bots provided an automatic solution to detect and revert vandal edits. There, however, exists opportunity for improvement. Research [14, 11] has shown that the current bots were limited in their extensibility as well as in their effectiveness at detecting instances of committed vandalism. Therefore, exploring additional automated measures to improve the accuracy of the vandalism detection carries numerous benefits. First, it helps alleviate manual effort required for cleaning vandalism edits. Second, it helps identify automated solutions to address the weakness of the current tools. Finally, an effective anti-vandalism tool could prevent or correct future malicious editing – thus protecting the integrity of Wikipedia articles.

Detailed analysis of vandalism needs to be done. Certain basic questions need to be answered first such as which users are responsible for vandalizing pages and their spatial characteristics such as what IP addresses do they belong to and what regions do these IP addresses belong to. Temporal characteristics need to be studied such as what are the time durations when vandalism occur the most. There are various other factors that need to be studied so that we can find out the main reasons for occurrence of vandalism.

1.2 Dissertation contributions

This dissertation concentrates on studying about Wikipedia on comprehensive scale and finding out the main reasons that are responsible for vandalism. We study the temporal characteristics for various articles of Wikipedia belonging to different countries and finding out the time durations when vandalized and non – vandalized edits occurs the most. Further we study the

spatial characteristics of various articles of Wikipedia belonging to different countries and mapping IP addresses of vandal contributors to their regions. We also collect words from various domains that are added and removed for non- vandalized and vandalized versions respectively and list the top words that are used in vandalizing the pages. We find the correlation between users viewing the page with the edits made on that page. Further we study the historic evolution of size of pages i.e. how the size of a page changes over time.

Then we study the distribution of contributions of edits among registered and unregistered users for a page and finding out the common users among various pages in various domains. Our study determines how the occurrence of events affect edits in Wikipedia. We also introduce some parameters: minimum ratio, mean and maximum ratio and use context aware approach that can be useful in detecting vandalism.

1.3 Organization

In Chapter 2, the related work is reviewed. It provides background knowledge of various studies in Wikipedia and vandalism detection methods. In Chapter 3, we introduce the dataset and brief explanation of the contents of the dataset. In Chapter 4, we do empirical study of Wikipedia. We study about the mean duration of edits, mean number of edits, evolution of size of pages and we list the most common words for all pages. In Chapter 5, we study about the user behavior patterns of pages, number of common users in the pages and number of views of pages. In Chapter 6, we compare the vandal and non-vandal edits by studying the content creation and deletion of vandal and non-vandal edits. In Chapter 7, we study vandalism and abuse patterns followed by introduction of a vandalism detection approach called context aware approach. In Chapter 8, we study the effect of occurrence of events to pages. In Chapter 9, we conclude our thesis and show the future work.

CHAPTER 2

Related Work

2.1 Study of Wikipedia

We study Wikipedia by studying the user behavior patterns, abuse patterns, content evolution of pages and content addition and deletion of edits.

Our work takes into account the user information such as whether the user is registered or not and then we study the registered and unregistered user contributions as a whole and also their contribution towards vandalized pages. In the paper [5] by Cristian-Alexandru Dragusanu, the features are grouped into 3 classes: metadata, text and language. Metadata features that are based on general version information are isregistered, comment length, size change, size ratio, previous same author. Text features that are based on basic analysis on text characters are digit ratio, alphanumeric ratio, upper ratio, upper lower ratio, long character sequence, long word, previous length, compression ratio of added words. Language features that are based on advanced analysis over the text content are vulgarity, biased words, sexual words, miscellaneous words, all bad words, good words, comment revert.

We take into account the semantics of a version such as collecting the statistics of purely removed, purely added words, added words and removed words and comparing the content addition and deletion of vandalized and non-vandalized pages. We also collect the most common vandal and non-vandal words. The study [6] by Koen Smets reveals that elementary features which are used by current approaches are not sufficient to fight vandalism. They need to be

accompanied by additional information incorporating the semantics of a version. In their approach for each version they are considering its text, text of previous version, user groups and version comment. They are focusing more on the content of an edit. They perform a diff between current version and previous version. The intuition is that the semantics of offenses, nonsense and spam are likely to differ from the semantics of the revised article and hence are an important feature for classification. Moreover, they believe that the 'text deleted'-feature contains more information than is apparent from the current results, where it appears to be merely a noise factor. They take into account its effect on the semantic level by measuring the text life, i.e. the value of the deleted words. They are among the first to try machine learning techniques to answer the need of improving the recall of current expert systems, which are only capable of identifying 30% of all vandalism. They have demonstrated application of two machine learning algorithms, a straight forward feature representation and using a set of noisy labeled examples, the accuracy of the actual running bots can be improved. There is another paper [4] by Santiago M in which they create fully working antivandalism system and get it working in real world. This paper introduces language independent and language dependent features. Language independent features include comment length, upper lower ratio, upper to all ratio, digit ratio, non-alphanumeric ratio, character diversity, character distribution etc. and language dependent features include vulgarism, pronouns, biased words, sex related words, bad words, good words etc.

On our work we study the spatial and temporal characteristics of vandalism. We find out the most common time frames and regions that are responsible for vandalism. A paper by S. Hahmann[21] investigates on various factors that influence the geospatial characteristics of Wikipedia articles. They found that 4.5% of the documents of a partial web crawl contain a US

zip code and hence are georeferencable. An analysis of geographic entities within newspaper articles came to the result that in average 75% of the investigated newspaper documents contained at least one geographic entity. They have found that 17.5% of all articles on the German Wikipedia are annotated with coordinates. All three figures illustrate that the portion of geospatially referenced information depends on both, the method – search for zip codes, toponyms, coordinates – and the examined corpus – web documents, newspaper documents, Wikipedia articles. They implemented the proposed network based approach with the corpus of the German Wikipedia and combined it with results of a cognitive study. For this approach, they consider Wikipedia a directed graph consisting of articles as vertices and links as edges. An output of this work is the so-called Network Degree of Geospatial Reference (NDGR), which may be seen as a measurement of the 'geospatiality' of information within a network. The NDGR has been computed for all Wikipedia articles.

There have been various criticisms of Wikipedia. We found out that Wikipedia is not efficient for volatile situations and sometimes is affected by the limited knowledge of editors. There is a review of study of Wikipedia in peer reviewed journals by Chitu Okoli [18]. This study concerns how Wikipedia works and why it works successfully. A large body of research is examined to access the reliability of Wikipedia. Applications of Wikipedia are also examined. Some important works in rapidly growing body of research that has focused on the phenomenon of Wikipedia are presented. There are various criticisms discussed in the paper that some scholars have levied against Wikipedia. They contend that in spite of the attempts of some Wikipedians to provide quality control, the lack of formal controls results in the lowest quality contributions prevailing, with unclear standards of accuracy or writing quality. Despite the negative assessments of some scholars, the highest approval of the reliability of Wikipedia's

content might be considered to be those peer-reviewed journal articles that use Wikipedia as a source of data. Another paper by E. Svoboda [19] questions the authenticity of information in Wikipedia. The free online encyclopedia, Wikipedia has generated shared scholarly efforts to rival those of any literary or philosophical movement in history. As such, Wikipedia is vulnerable to user-generated articles that are inaccurate or irrelevant. While a carefully executed and multilayered review process is performed by a team of volunteers, critics believe that the lack of formal gatekeeping procedures ensures that the lowest common denominator will prevail and, since no experts or editors are hired to review the articles, no clear standards exist for accuracy or writing quality. Despite its imperfections, Wikipedia users claim that it works well in practice. Nevertheless, readers are advised to check their online finds against other sources and to be aware of Wikipedia's unique strengths and weaknesses, especially when gathering information for research projects.

2.2 Vandalism Detection

Our study tries to come out with a factor set that can help in vandalism detection. We have proposed two parameters: minimum ratio and mean and have used these parameters in a vandalism detection approach called context aware approach for detection of vandalism. Work by Si-Chi Chin [12] builds statistical models constructing distribution of words from the version history of Wikipedia articles. As vandalism often involves the use of unexpected words to draw attention, the fitness of a new edit when compared with language models built from previous versions may well indicate that an edit is a vandalism instance. Also the paper adopts an active learning model to solve the problem of noisy and incomplete labeling of Wikipedia vandalism. The Wikipedia domain with its version histories offers a novel context in which to explore the potential of language models in characterizing author intention. There is another paper by B.

Thomas Adler [3] makes use of WikiTrust which is a reputation system for Wikipedia authors and content. WikiTrust computes three main quantities i.e. edit quality, author reputation and content reputation. Various features that are extracted and taken into account are author reputation, author is anonymous, time interval to previous version and time interval to next version, hour of the day when version was created, minimum version quality, next version comment length, next comment mentioned in revert, version comment length etc.

We studied the number replacement, topic replacement and status inverse attacks for vandalized edits. The paper [7] by Lakshmish Ramaswamy talks about sophisticated vandal edits called elusive vandal edits. It introduces text stability approach as a measure to quantify the stability of a text block that evaluates the likelihood of a certain text block of an article being modified. They consider the relationship between text-stability and the characteristics of an edit and utilize these factors as features, which are used to drive machine learning-based supervised learning classifiers. Then they evaluate the performance of their approach using the Wikipedia Vandalism PAN corpus. They introduce number replacement, topic replacement and status inverse attacks. Another work by Manoj Harpanali [23] explores more linguistically motivated approaches to vandalism detection. Their hypothesis is that textual vandalism constitutes a unique genre where a group of people share a similar linguistic behavior. Further statistical models give unique language styles in vandalism and deep syntactic patterns based on probabilistic context free grammars discriminate vandalism more effectively than shallow lexico-syntactic patterns based on n-grams.

We study the spatial and temporal characteristics of vandalized and non-vandalized pages and find out the distinguishing time frames and regions that can be distinguishing factors for vandalism detection. We also study the contribution of vandalism by registered and unregistered

users. The paper [1] by Andrew G takes into account the spatial and temporal properties to detect vandalism in Wikipedia. A tool is developed containing server side version processing engine to score edits as they occur and a user facing GUI client which enables quick inspection of likely instances of vandalism. Simple features are exploited like time of the day, version comment length of the edit that operate on the metadata associated with a single edit and aggregate features that combine time decayed behavioral observations to create reputation values for single entities and spatial groupings. Feedback is gathered using administrative form of reversion called rollback. By exploiting these features a lightweight classifier is produced that identifies vandalism. Observation was made that the vandal edits are most prominent between 8pm and 8am. More vandalism occurs on weekdays. Observation was made that often edited articles attract bulk of vandalism and 85% of vandalism is caused by unregistered users. Also it was observed that comment section for vandalized edit is on average 43% of size of those with random edits. This paper also takes into account Wikipedia categories that are topic based or administrative and reputation using topical categories as spatial grouping of articles is calculated. The user reputation by normalizing user reputation by number of user edits and country reputation to find from which country more vandalism can be expected is also taken into account.

Another paper [2] by B. Thomas Adler considers metadata, text, reputation and language features for detecting vandalism. Some of the similarities observed in vandal edits are that highly edited articles are frequent targets of vandalism. Vandalism is most prominent during weekday "school/office" hours. Vandals leave either very short comments or very long ones. The use of first and second person pronouns including slang spellings have high chances of being vandalized edits. User reputation and country reputation are taken into account. Each feature is

categorized as metadata, text, reputation, or language, according to the nature of how they are computed and roughly corresponding to their computational complexity. They discovered that language features only provide an additional 6% of performance over the combined efforts of language-independent features.

CHAPTER 3

Dataset

3.1 PAN Corpus Dataset

The PAN Wikipedia vandalism corpus 2010 (PAN-WVC-10) is a corpus for the evaluation of automatic vandalism detectors for Wikipedia. The corpus compiles 32,452 edits on 28,468 Wikipedia articles, among which 2,391 vandalism edits have been identified. To annotate the corpus Amazon's Mechanical Turk has been used; 753 workers have been recruited who cast more than 1,50,000 votes on the edits, so that each edit was reviewed by at least 3 annotators. The achieved level of agreement was analyzed in order to label an edit as "regular" or "vandalism."

We imported this dataset into a relational database management system. Our dataset contains various tables which store information about Wikipedia pages. **Page** is the current article that is visible in Wikipedia. If we open a Wikipedia article, the text visible to us is the page. **Version** is the previous revision of a page. A page can have one or more versions. If we want to see the versions of an article we can click on the ‘View History’ tab of the Wikipedia page. There are one or more versions of a page visible along with the time and date when the version was created, user who created the version and comments written by users. The dataset contains one table that stores information about the whole article, another table that stores information about the versions of the article and another table storing the content of all the versions of an article.

When a version of a page is created, Wikipedia stores the date and time, user information and comments of that version. **Date and time** represents the date of creation of a version and time of creation of a version respectively. The date is represented in the format 'ddmmyyyy' and time is represented in GMT 24 hour format. If we want to see the date and time when the page was created or when the previous versions of that page were created, we can click on the 'View History' tab of the Wikipedia page and see the time and date of creation for every version and the current page. **ClueBot** is a program which constantly checks Wikipedia for new page edits by users. When it detects a new edit, it analyzes it. If it determines that the edit is vandalism, then it fixes the vandalism by reverting the article to its previous state. It also takes some steps to ensure that vandalism by the same user won't happen again, in accordance with Wikipedia policy. Once ClueBot finishes with a given instance of vandalism, it returns to checking for new edits [22]. **User** is any person or a software application that run automated tasks (Cluebot) that contributes to Wikipedia by adding, removing or modifying information from Wikipedia pages thus creating new versions. User can be of two types: registered or unregistered. Registered users are identified by a unique username while unregistered users are identified by their IP addresses. If we open any Wikipedia article and click on the 'View History' tab, we can see all the versions of that article and besides every version we can see the usernames for registered users and IP addresses for unregistered users who created those versions. **Comments** are the text written by users who add, modify or delete information from pages. Comments help a reader have an idea about the changes made in the previous version. Most of the times whenever a vandalized edit is reverted, users write 'reverted vandalism' comment besides the version. Whenever a Cluebot comes across a vandal edit and reverts it, it definitely places 'reverted vandalism' comment besides the version. If we want to see the comments of various versions of the page, we can

click on the ‘View History’ tab of the Wikipedia page and see the comments for every version besides that version.

Vandalized, non-vandalized and reverted edits are identified by a special field called pathtype. **Pathtype** is the label assigned to each version of every page in PAN Corpus dataset that specifies whether the version is vandalized, non-vandalized, revert or other. If the value of pathtype is 1 it means the version of that page is non-vandalized, if pathtype is 3 or 4 it means the version of the page is vandalized. Value of 2 or 6 means the version of the page was reverted. There are other values of pathtype which are beyond the scope of our work.

CHAPTER 4

Empirical study of Wikipedia

We have used 12 different domains for our experiments. The domains are taken from Dbpedia in which they are represented as classes. Dbpedia is a project that extracts structured content from the information created as a part of Wikipedia project. The following domains are used:

1. Person
2. Work
3. Sports
4. Places
5. Food
6. Currency
7. Disease
8. Chemical substance
9. Planet
10. Color
11. Anatomical Structure
12. Programming Language

Some pages from each domain are selected. We were constrained by the size of PAN Corpus dataset as it has limited number of pages. So we selected 20 pages each from

domains for Person, Places, Work and Food and for the rest of the domains we selected 5 pages each. In total we selected 120 pages.

We have also used Hadoop technology to deal with big data**.** Hadoop MapReduce [24] is a programming model and software framework for writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. It has mainly two phases: map and reduce. Map/reduce is a special form of such a DAG which is applicable in a wide range of use cases. It is organized as a "map" function which transforms a piece of data into some number of key/value pairs. Each of these elements will then be sorted by their key and reach to the same node, where a "reduce" function is use to merge the values (of the same key) into a single result.

HDFS: The distributed file system is designed to handle large files (multi-GB) with sequential read/write operation. Each file is broken into chunks, and stored across multiple data nodes as local OS files. There is a master "NameNode" to keep track of overall file directory structure and the placement of chunks. This NameNode is the central control point and may re-distributed replicas as needed. DataNode reports all its chunks to the NameNode at boot up. Each chunk has a version number which will be increased for all update. **T**herefore, the NameNode know if any of the chunks of a DataNode is stale (e.g. when the DataNode crash for some period of time). Those stale chunks will be garbage collected at a later time.

4.1 Mean Duration between Edits

Duration between two edits means the time difference between one edit and the other edit. In Wikipedia whenever an edit is made, the time and date for that edit is stored. In order to find the mean duration between the edits of a page, firstly we find the time difference between each edit of the page and then we take the mean of all the differences that are calculated for that

page. After calculating the mean duration of all the pages in a domain, we take the mean of the mean durations calculated for all the pages of that domain. So by doing this we can find out the how often the edits occur in a particular domain and which domain has the highest frequency of edits or lowest mean duration between edits.

Figure 1 shows the mean duration between edits. Places domain has the highest frequency of edits or lowest mean duration between edits. Chemical substance has the highest mean duration between edits or lowest frequency of edits. Sports domain has second lowest mean duration between edits. Planet domain has the second highest mean duration between edits.

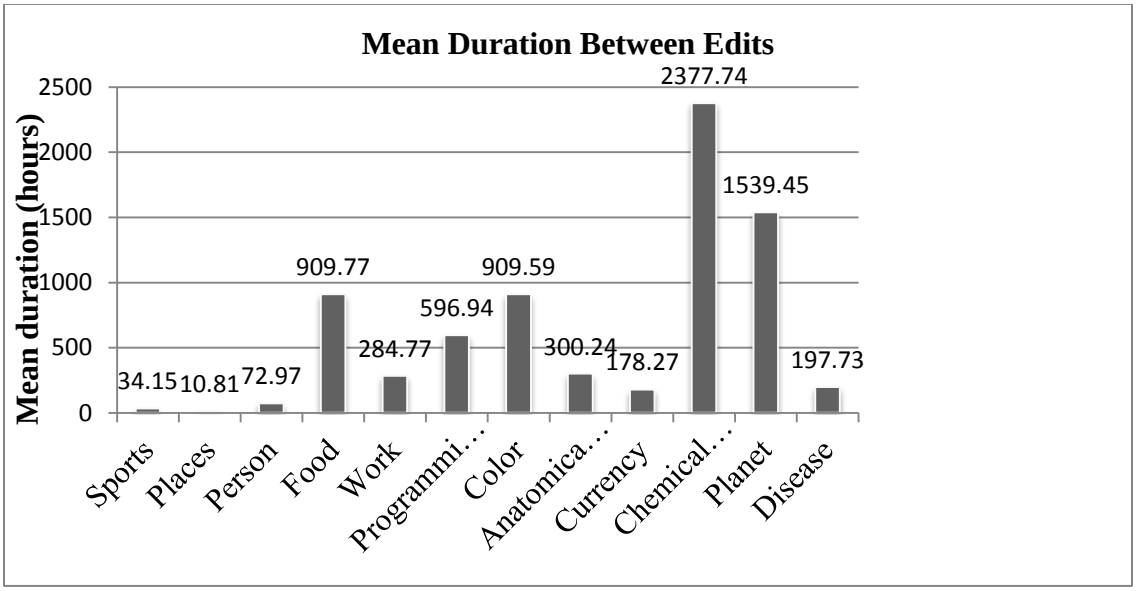

Figure 1: Mean Duration Between Edits

4.2 Mean Number of Edits

Mean number of edits means how many edits took place normalized over all months. To calculate the mean number of edits we take the total number of versions of a page and then divide it by the total number of months the page existed. For calculating the number of months we took into account all the different months of all the years for which the page existed.

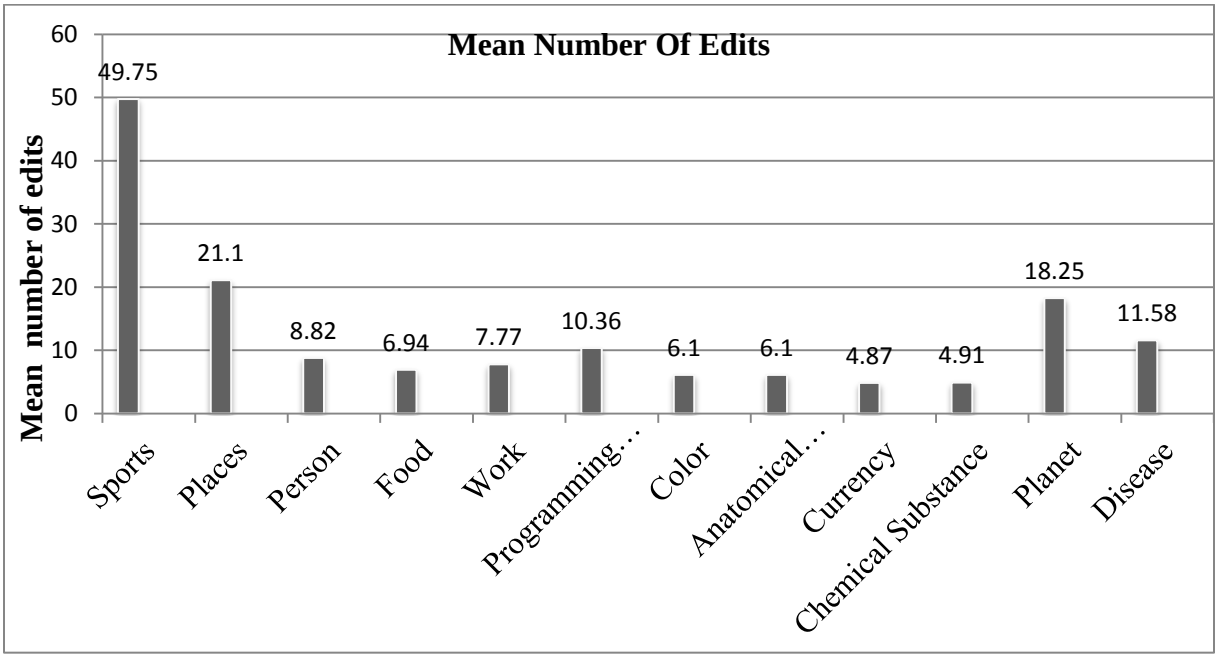

Figure 2: Mean Number Of Edits

Above Figure 2 shows the mean number of edits for various domains. Sports domain has highest mean number of edits and Currency domain has the lowest mean number of edits.

4.3 Evolution of Size of Pages

We studied how the sizes of pages evolve over time. We calculated the size of first 10 versions and the size of last 10 versions of a page. The size is calculated by counting the number of words in the versions. Then we calculated the difference between them. Basically we want to study for which domain size of pages increase over time the most and for which domain size of pages decrease over the least.

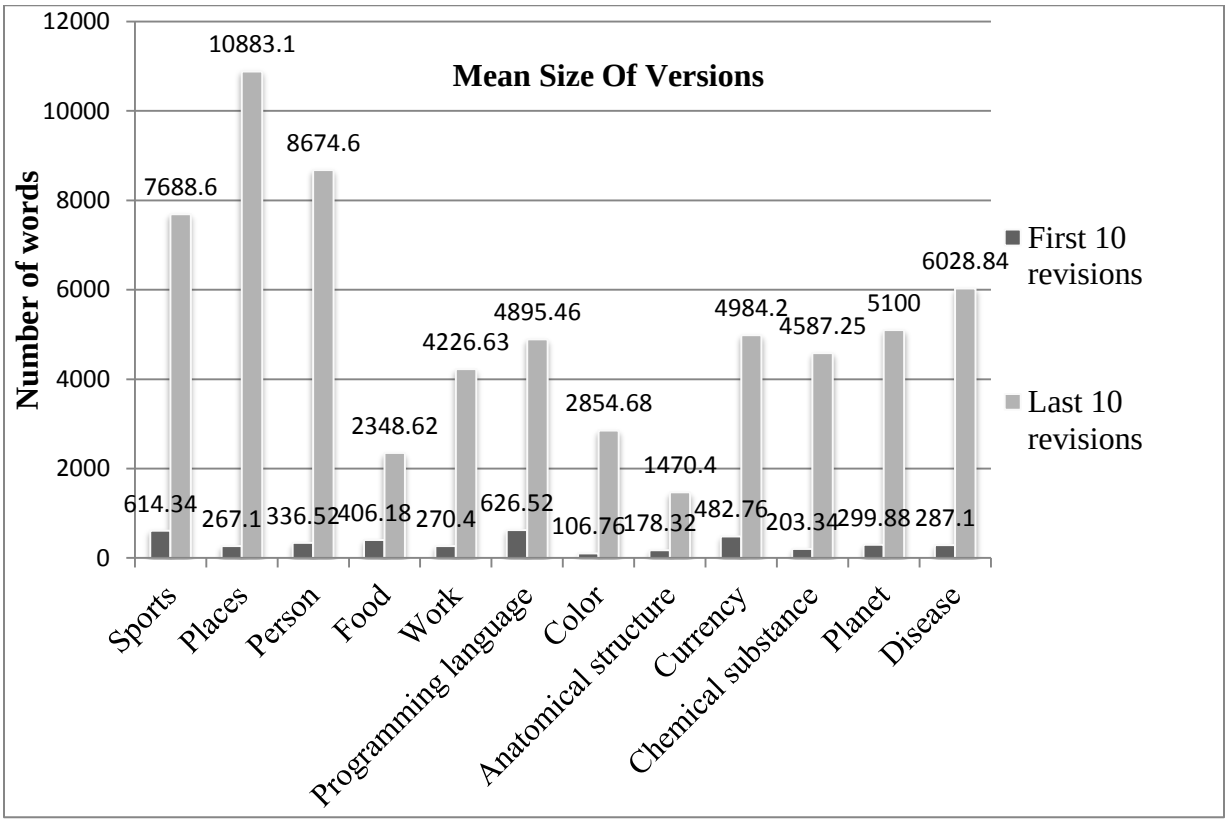

Figure 3: Mean Size Of Versions

Figure 3 shows the mean size of first and last ten versions of pages of a domain. Places domain has the highest increasing size from first ten versions to last ten versions. Also in Person domain mean size increases a lot from first ten to last ten versions. Anatomical Structure domain has the least difference between first ten and last ten versions.

Domains such as Places, Sports, Person are interesting domains in which a lot of users contribute and there are edits occurring frequently because events occur more frequently in the pages of these domains and also they evolve more over time. On the other hand domains such as Chemical substance, Color and Planet have pages in which events occur less frequently and these domains require specialized knowledge and not many people contribute to such domains. Hence these pages evolve less over time and edits are less as compared to other domains.

4.4 Common Words between all Domains

Our goal is to find the most common words that occur in all the pages of all the domains. Firstly all the words that are vandalized and non-vandalized are collected based on the values of pathtype for all pages for each domain. For vandalized versions we take the diff of the previous version with the current version and hence calculate the removed words. On the other hand for non-vandalized versions we take the diff of the current version and the previous version and found out the added words. We classify a word as vandalized based on the version in which the word appears. If pathtype for the version is 1 then the version is non-vandalized and if pathtype of the version is either 3 or4 the version is vandalized. Then we use Hadoop to count how many times a word appeared in all the pages for each domain. Further we use Hadoop to rank the words according to the occurrences. The word that appeared the most is ranked at the top and the word that appears the least is ranked the last.

We collect all the files which contains the word, domain name and category whether it is vandal or non-vandal. Then we use WordCount program in Hadoop to count how many times the word appeared and then we use Ranking program in Hadoop to rank the words according to their occurrence.

List of some commonly occurring vandal words and non-vandal words are shown in Appendix I.

Figure below shows the frequency distribution of first 1000 vandal words.

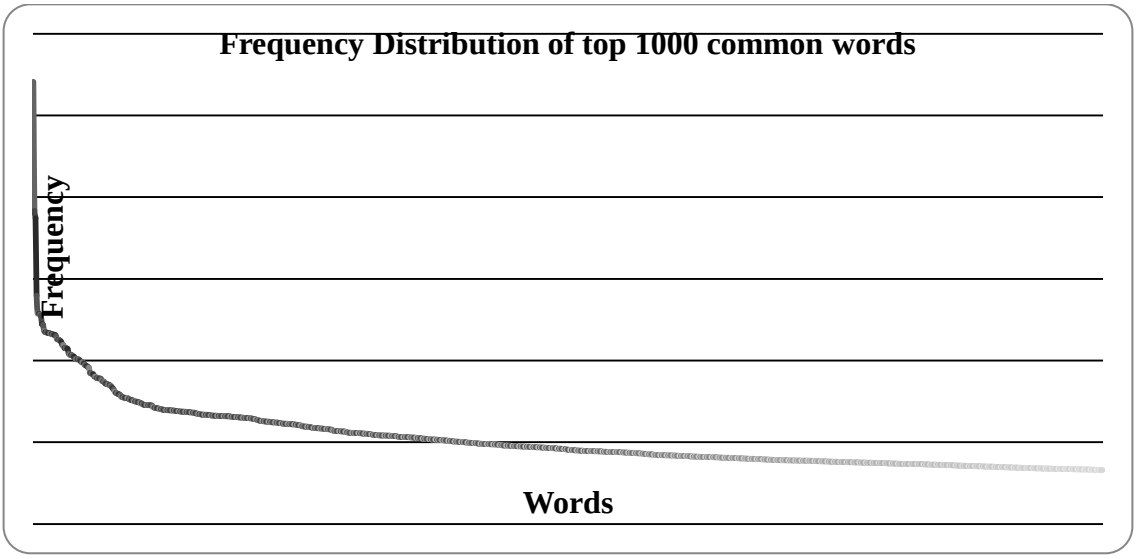

Figure 4: Frequency Distribution Of Top 1000 Common Words

CHAPTER 5

User behavior patterns in Wikipedia

5.1 User Contributions

There are two types of users in Wikipedia- registered users and unregistered users. Every user contributes to Wikipedia, in which some of the contributions are vandalized which means adding incorrect content into the articles.

Firstly mean number of unique users for all domains are taken i.e. first collecting the number of users contributing to all articles in a particular domain and then calculating their mean. Unique user means if there is one user who has contributed more than once he is only counted once.

For every domain, contributions of top user, top 5% users, top 10% users, top 25% users and top 50% users are calculated. Our main motive is to find out how users contribute to various domains. Are the contributions equally distributed or they are unequal and done by single user or a set of users?

Further we study users in terms of registered and unregistered users. We study that what percentage of edits come from registered users and what percentage of edits come from unregistered users. Then we show that how users contribute to an article and what type of users are in general responsible for vandalism. We collect all the users for all vandalized versions and classify them as registered and unregistered users. In general we are trying to show the percentage of distribution of vandal edits of a page among the users.

Then we find how many users contributed to how many different pages of a domain means how many common users are there in the different pages of a single domain. We find out that how many common users contributed to 25% of the pages, 50% of the pages and 75% of the pages of a single domain. Common users can be a distinguishing factor in identifying the reputation of users. If the user has contributed to mostly non-vandalized edits then it is most likely the user will contribute to non-vandalized edits in future and the same holds true for vandalized edits.

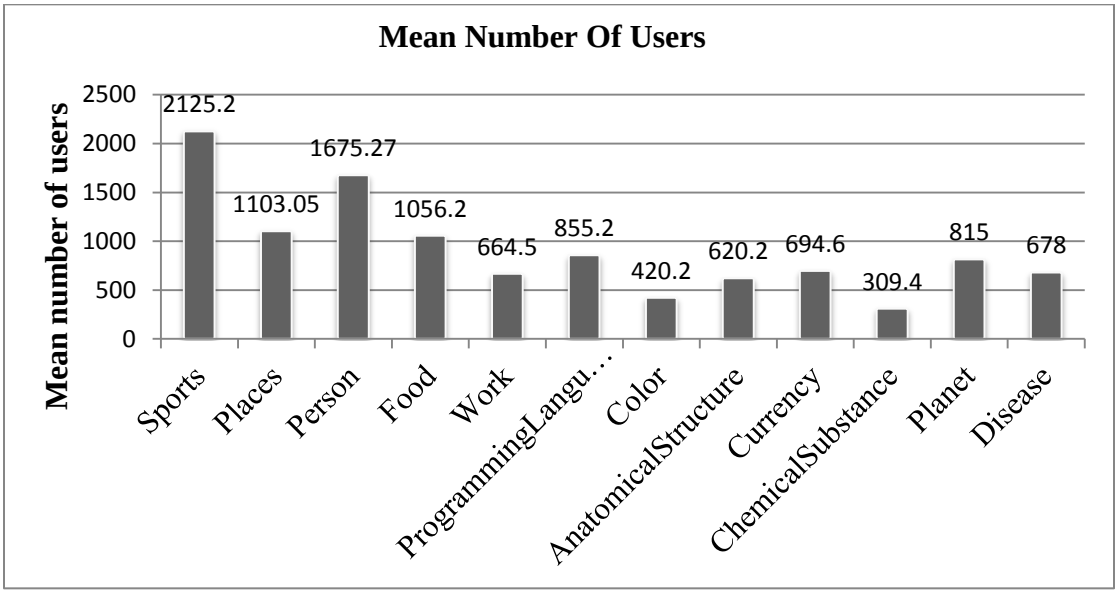

Figure 5: Mean Number Of Users

Figure 5 shows the mean number of users for each domain. Results show that Chemical Substance domain has the lowest mean number of users while Sports domain has the highest mean number of users. The reason may be that Chemical Substance domain requires domain

knowledge which common people do not have, hence resulting in the lowest mean number of users. Sports domain is famous and almost everyone is interested in one or the other kind of sports resulting in highest mean number of users.

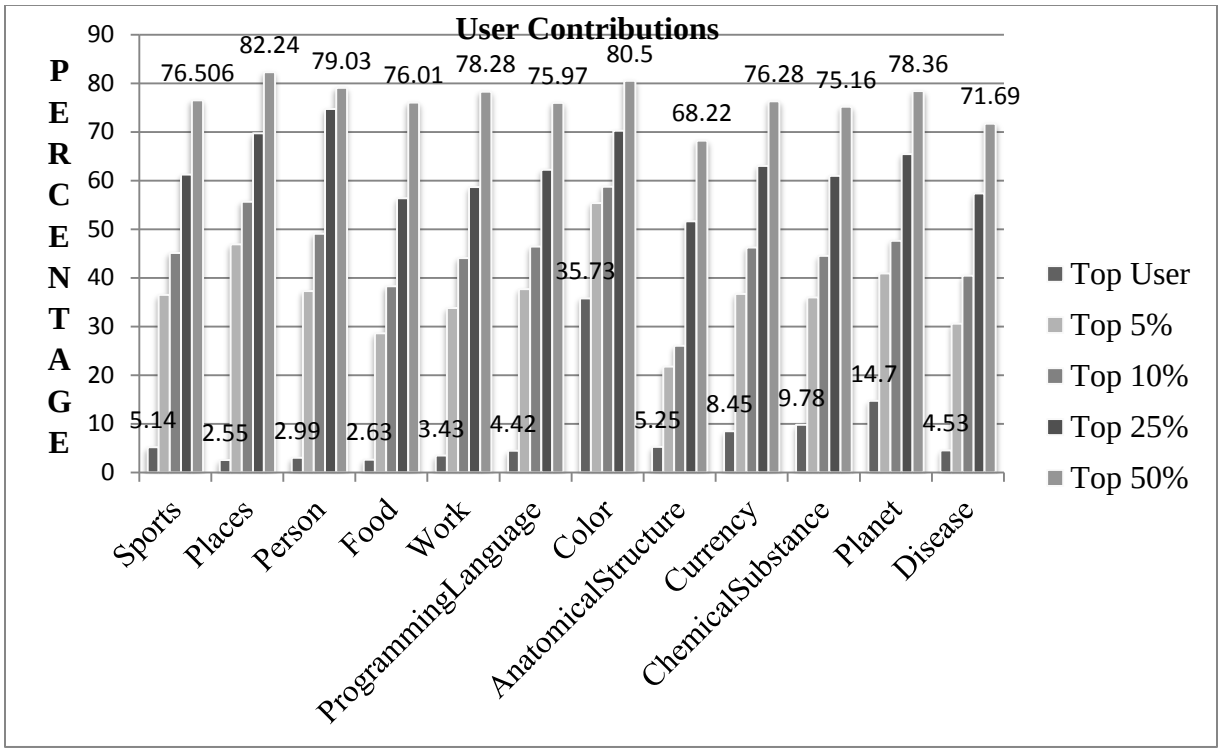

Figure 6: User Contributions

Figure 6 shows the percentage of contributions made by the top user followed by the percentage of contributions by 5% top users, percentage of contributions by top 10% users, percentage of contributions by top 25% users and percentage of contributions by top 50% users for all domains.

Below graphs show the distribution of user contributions to various domains.

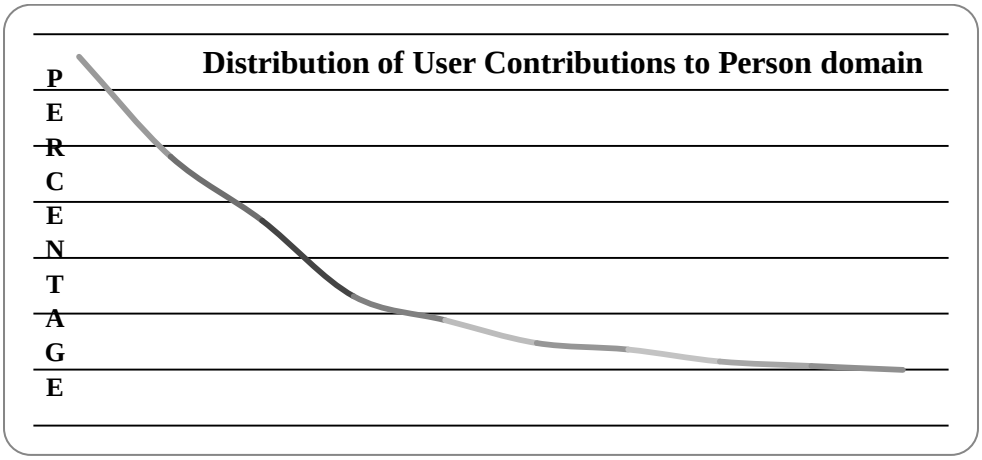

Figure 7: Distribution Of User Contributions To Person Domain

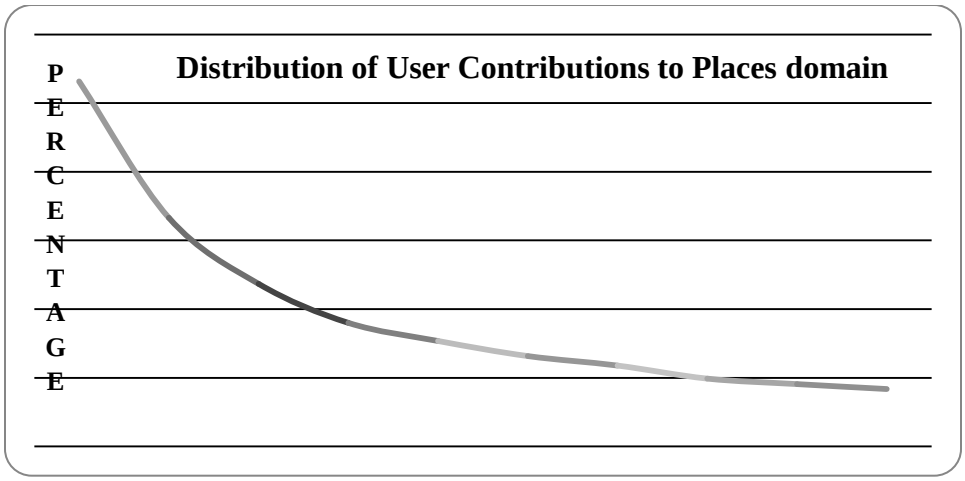

Figure 8: Distribution Of User Contributions To Places Domain

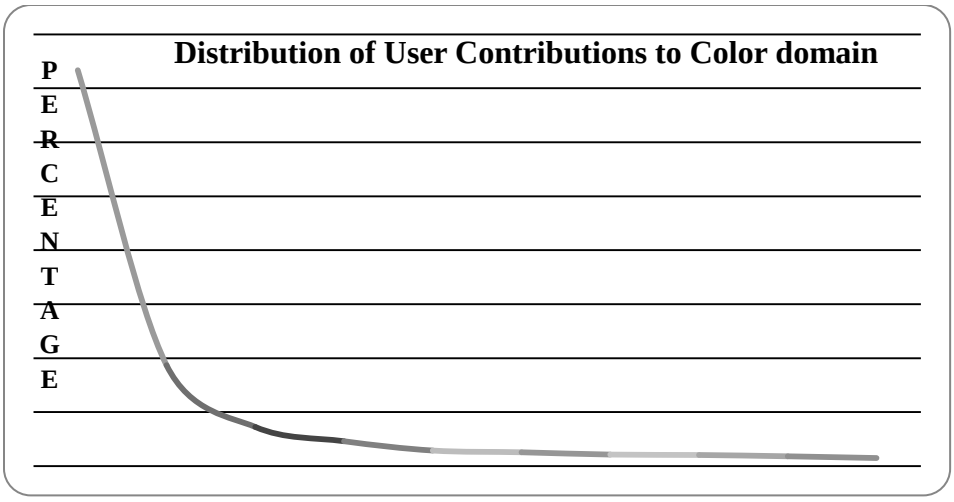

Figure 9: Distribution Of User Contributions To Color Domain

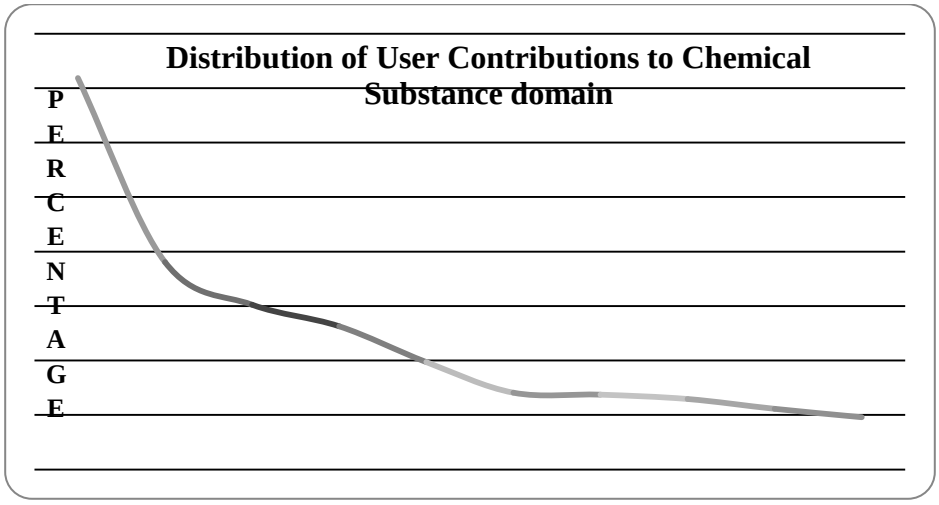

Figure 10: Distribution Of User Contributions To Chemical Substance Domain

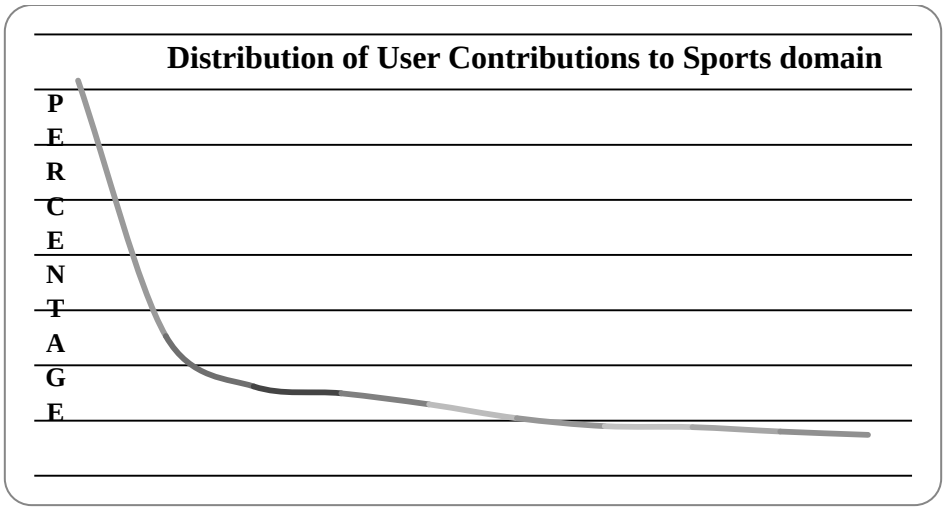

Figure 11: Distribution Of User Contributions To Sports Domain

We observed that top user contributed the most to Color domain and least to Places domain. This may be because Color domain is not a vast and interesting domain so mostly edits are done by single user, maybe the user who created the page while Places domain is wide and the knowledge is more open to people so more users add information in it. If we see the contribution of top 5% users in Figure 6, they contribute the least to Anatomical Structure and highest to Color domain. If we see the contribution of top 10% users, again they contribute the least to Anatomical Structure and the highest to Color domain. Looking at the contribution of top 25% users, they contributed least to Anatomical Structure domain and highest to Person domain. Top 50% users contributed least to Anatomical Structure and highest to Places domain.

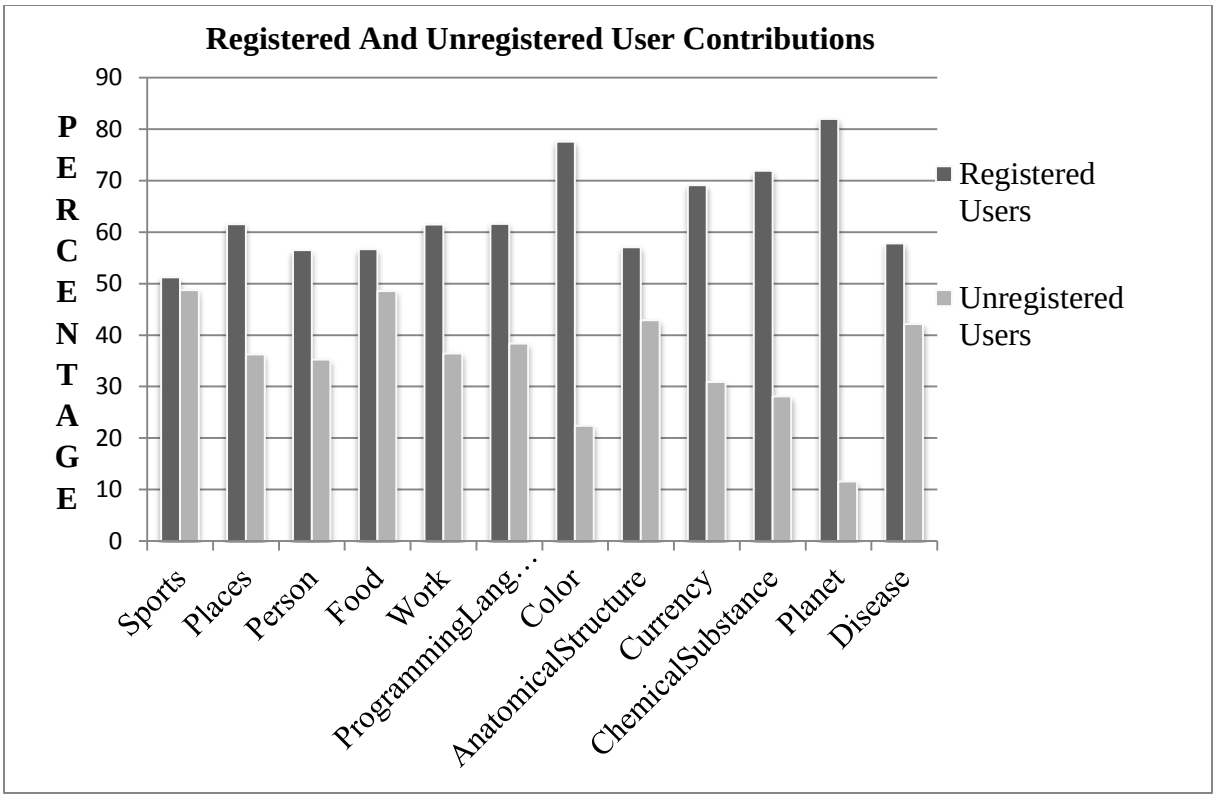

Figure 12: Registered And Unregistered User Contributions

Figure 12 shows the percentage of contributions of registered and unregistered users. Among various domains registered users contributed the most to Planet domain. Secondly Color domain has lot of contributions from registered users. In Sports domain registered and unregistered users almost contributed equally.

Overall registered users contribute to about 65% of edits and unregistered users contribute to about 35% of edits.

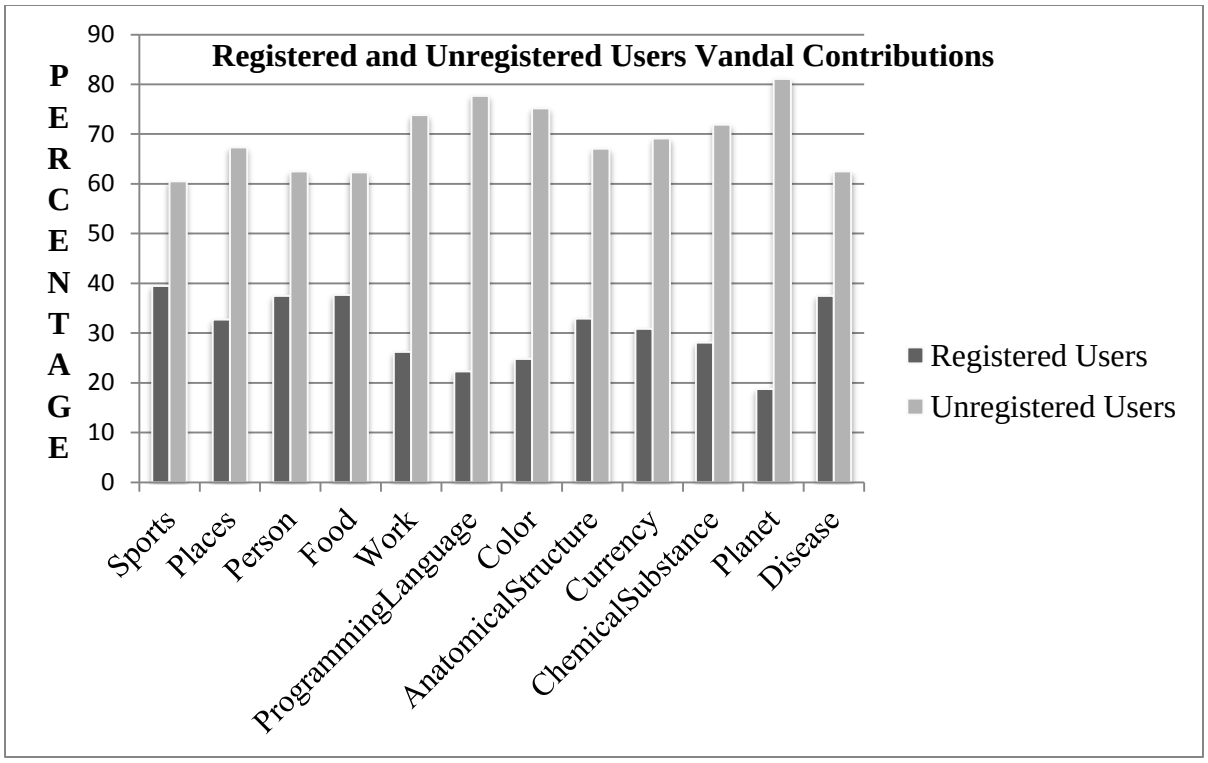

Figure 13: Registered and Unregistered Users Vandal Contributions

Above Figure 13 shows the percentage of contributions of registered and unregistered users for vandal edits. Clearly in every domain unregistered users are the highest contributors of vandalism as compared to registered users. For Planet domain unregistered users highly contributed to vandalism as compared to registered users. For Programming language domain also unregistered users highly contributed to vandalism as compared to registered users.

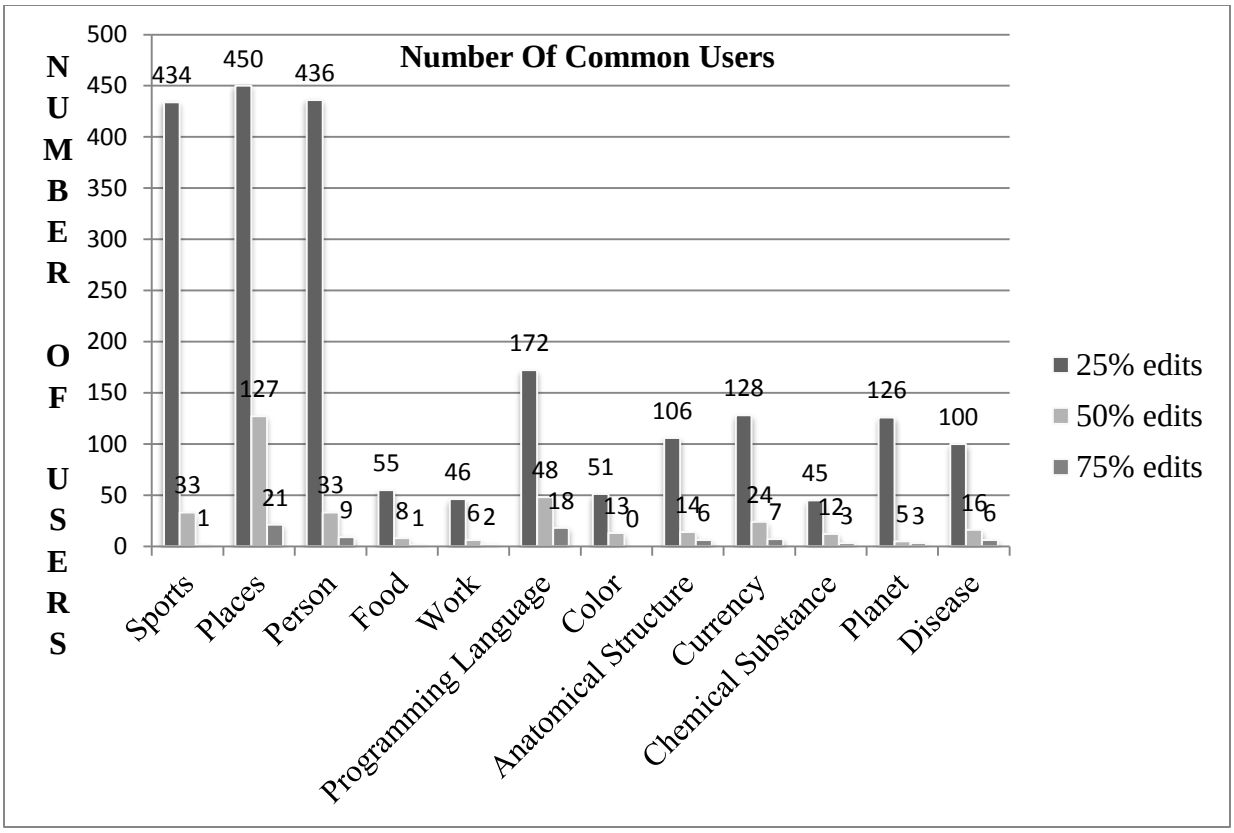

Figure 14: Number Of Common Users

Above Figure shows the common number of users in the pages of a domain. It shows how many common users exist between different pages of a domain. Above figure shows that Places domain has the highest number of common users among 25% of edits, 50% of edits and 75% of edits. Chemical substance domain has lowest number of common users among 25% of edits. Planet domain has lowest number of common users among 50% of edits and Color has lowest number of common users among 75% of edits.

5.2 Number of Views of Pages

We calculate the mean number of views of the pages in the past 30, 60 and 90 days. We calculate the number of views using the statistics on Wikipedia pages. To see the number of views we go to Wikipedia page, click on the 'View History' tab, then click on the link named 'Page View Statistics'. We want to study which domain has the pages having the highest number of views and which domain has the pages having the least number of views. We also study the correlation between the number of views and the number of edits in the pages.

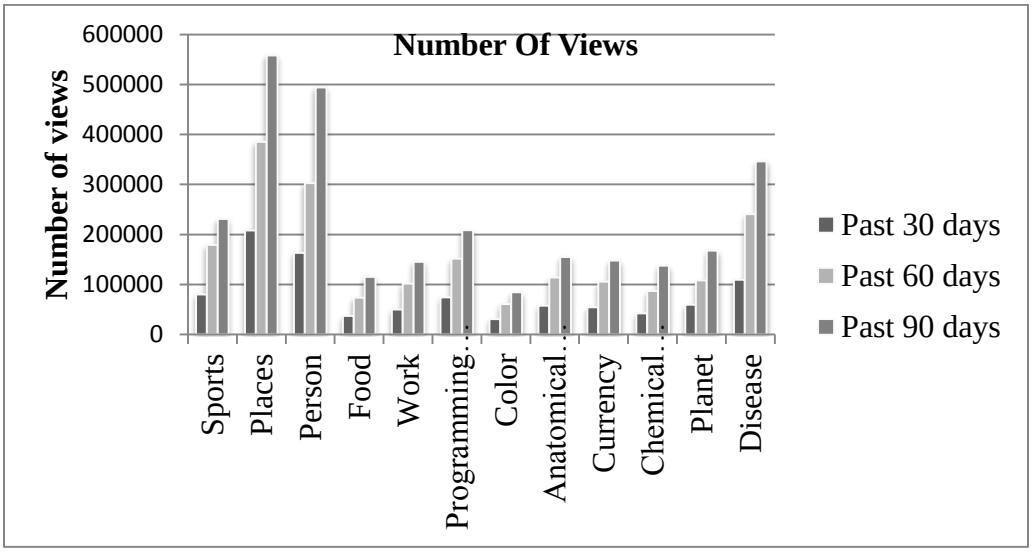

Figure 15: Number Of Views

Above figure shows the maximum number of views among past 30 days, past 60 days and past 90 days. Places domain has the highest number of views in past 30, 60 and 90 days. Color domain has the least number of views in past 30, 60 and 90 days.

If we find the correlation between number of views and number of edits it comes out to be around 0.27. The result 0.27 is not a perfect correlation between fields but not that much negative as it is in the range of 0 to +1.

Hence we are uncertain that more number of views depicts more number of edits.

CHAPTER 6

Content creation/deletion in vandalized and non-vandalized edits

6.1 Comparison between Sizes of Edits

Size of an edit is calculated by counting the total number of words in an edit. A comparison is done between two consecutive edits of a page. We find the added words between two edits by taking the diff between the previous version and the current version. We find the removed words by taking the diff between the current version and the previous version. Diff is a program which returns the words that are different in two pages being compared.

Firstly we calculate how many times the size of a version increased for vandalized and non-vandalized pages. For each version of a page we compare it with the previous version by calculating the diff and then calculating whether the size increased. Then we do it for all the pages of a domain.

Secondly we calculate how many times the size of a version decreased for vandalized and non-vandalized pages. For each version of a page we compare it with the previous version by calculating the diff and then calculating whether the size decreased. Then we do it for all the pages of a domain.

Thirdly we calculate how many versions have purely removed words for both vandalized and non-vandalized pages. For each version of a page we compare it with the previous version by calculating the diff and then calculating whether the words only got removed and no words got added. Then we do it for all the pages of a domain.

Lastly we calculate how many versions have purely added words for both vandalized and non-vandalized pages. For each version of a page we compare it with the previous version by calculating the diff and then calculating whether the words only got added and no words got removed. Then we do it for all the pages of a domain.

Then for vandal and non-vandal versions, we take the percentage of versions containing purely removed words, purely added words, added and removed words.

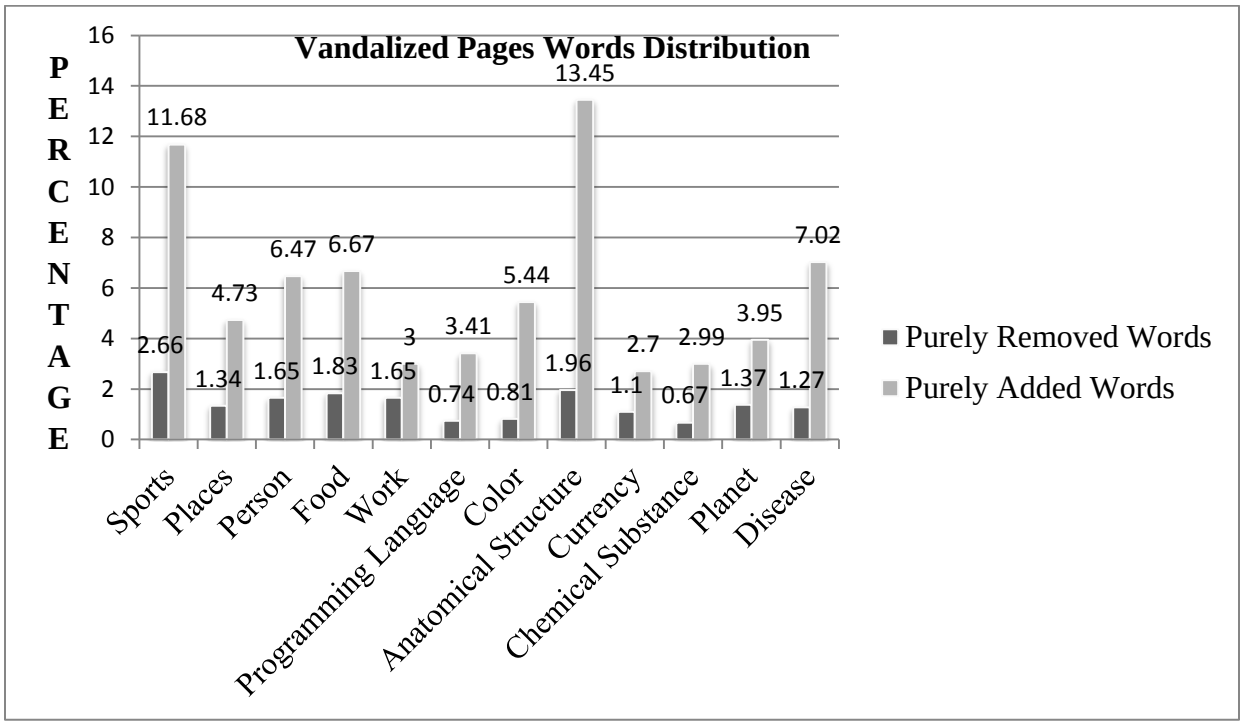

Figure 16: Vandalized Pages Words Distribution

In vandalized pages, versions containing purely added words are lot more than versions containing purely removed words as shown in above figure.

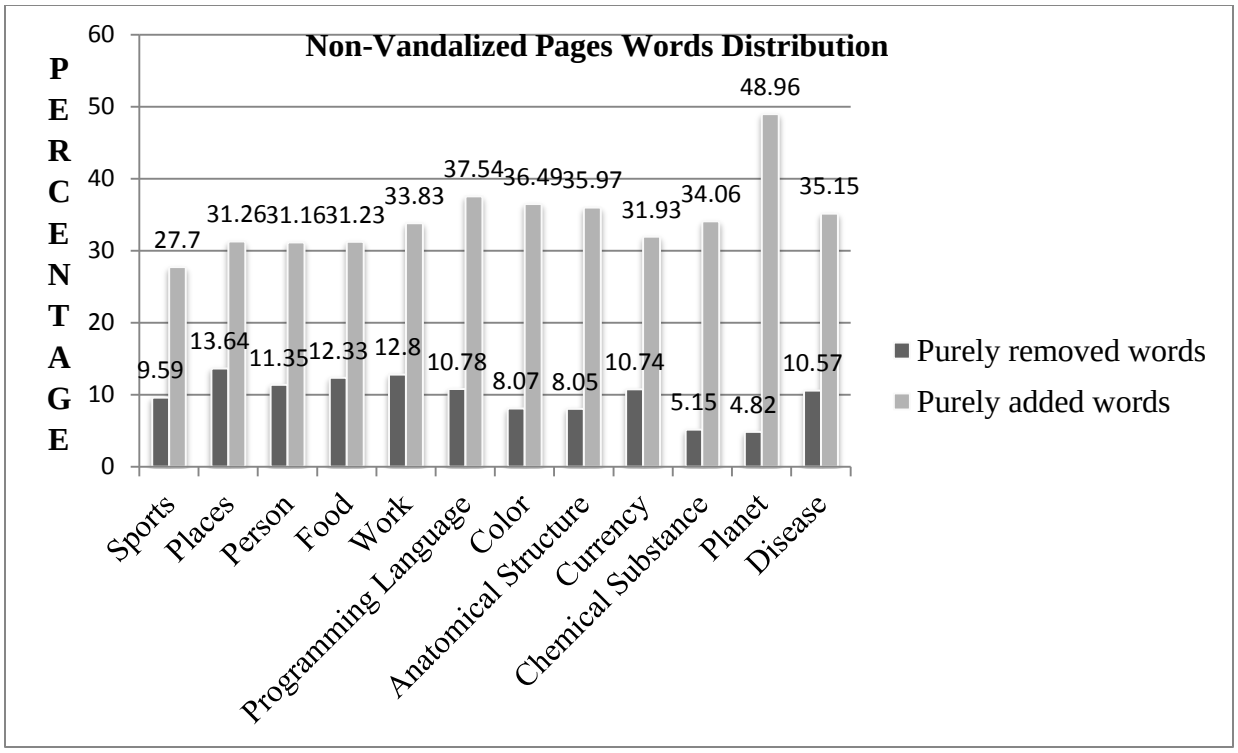

Figure 17: Non-Vandalized Pages Words Distribution

In non-vandalized pages, versions containing purely added words are more but there is high percentage of versions containing purely removed words too and if we see the percentage of versions containing purely removed words in vandalized pages, they are lot more in case of non-vandalized pages as compared to vandalized pages.

Also according to the other results the size of the versions increased more in case of vandalized versions as people are more interested in adding stuff rather than deleting stuff but in case of non-vandalized versions deletion was more than vandalized versions and also addition was more. In case of non-vandalized versions deletion occurs when reverting vandal edits or

inappropriate information and addition occurs by evolution of events and information to be added related to that page.

Hence, addition and removal of words can be a distinguishing factor in identifying the difference between vandal and non-vandal edits as vandalized versions can be distinguished by non-vandalized versions clearly by looking at the purely removed and purely added words.

CHAPTER 7

Study of vandalism and abuse pattern in Wikipedia

7.1 Top Occurring Vandal Words

A list of top occurring vandal words is collected. For each version of a page vandal words are collected based on the pathtype and similarly all the vandal words are collected for all pages for all domains. If the pathtype of a version of a page is either 3 or 4, it means the version of that page is vandalized. Similarly we collect the vandalized words from all versions for all pages. Then all the words are fed into Hadoop to get the count of the words showing how many times a word appeared as a vandalized word and then rank the words accordingly. So the word that appeared the most in vandalized versions is ranked the highest and so on. Top 25, 50 and 100 vandal words are collected.

We collected all the vandal and non-vandal words from various pages of different domains and ranked them according to the number of occurrences and got the top occurring vandal words.

Appendix 2 contains the list of top 25 occurring vandal words followed by the list of next top 25 occurring vandal words.

7.2 Temporal Characteristics of Vandalism

In Wikipedia time is represented in GMT format. We studied three countries having different time zones to study temporal characteristics of vandalism and non-vandalism. The countries are: United States, United Kingdom and India.

The list of pages from United States used was:

George Washington Bush, Jennifer Lopez, Jimmy Carter, Rihanna, Wayne Gretzky

The list of pages from United Kingdom used was:

David Beckham, Charles Dickens, Winston Churchill, Kate Winslet, Britney Spears

The list of pages from India used was:

Shilpa Shetty, Sathya_Sai_Baba , Ranbir_Kapoor ,Saif Ali Khan, John Abraham

For all versions of a page the time on which a particular version occurred is calculated. The time for vandalized versions and non-vandalized versions is calculated separately.

First we classify a version of a page as vandalized or non-vandalized based on path type of version. If pathtype is 1, the version is classified as non-vandalized and if the pathtype is 3 or 4, the version is classified as vandalized. Then we take the time of that version and round it to the nearest digit of hour. Then we calculate what percentage of vandalized versions occurs in a specified time intervals such as between 1.00am to 2am, 2am to 3am and so on. Similarly we calculate the percentage for non-vandalized versions.

Overall we observe the time intervals in which vandalism and non-vandalism occurs the most and time intervals in which vandalism occurs more and can be distinguished from non-vandalism and vice versa.

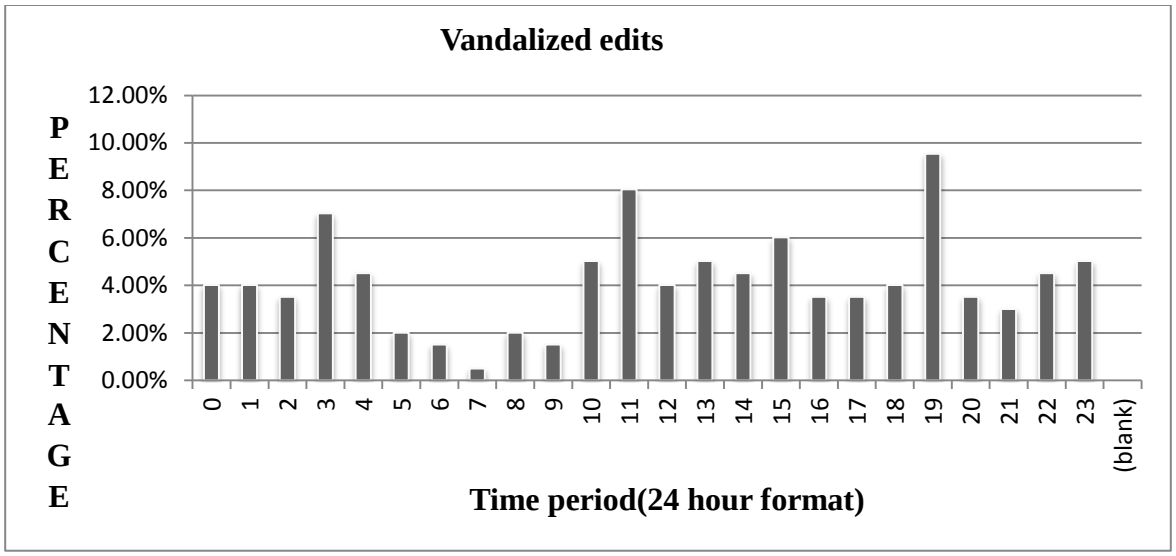

Figure 18: Vandalized Edits

Above figure shows the percentage of vandalized edits that occur during different time periods for one of the pages in United Kingdom. '0' means the time period between 12.00am to 1.00am, similarly '23' means time period between 11pm to 12am.

The lowest number of vandal edits occurs between time periods 7am to 8am and the highest number of vandal edits occurs between time periods 7pm to 8pm for this page.

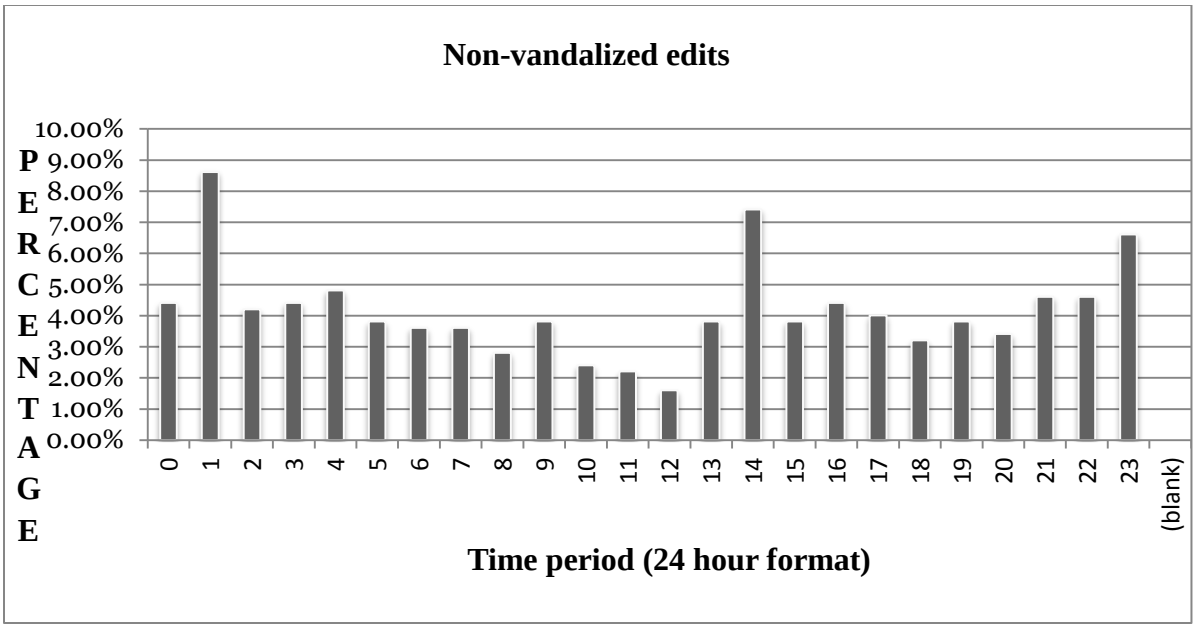

Figure 19: Non-Vandalized Edits

Above figure shows the percentage of non-vandalized edits that occur during different time periods. The lowest number of non-vandal edits occurs between time periods 12pm to 1am and the highest number of non-vandal edits occurs between time periods 1am to 2am for this UK page.

Similarly we have results for 4 more pages belonging to United Kingdom and 5 more pages belonging to India and United States.

Below table shows the country with the time periods in which maximum number of vandal edits occurs.

Table 1: Time Periods Of Vandal Edits

Country	Time periods with maximum no. of edits	Time periods with minimum no. of edits
United Kingdom	7pm-8pm , 11am-12.00pm, 7am-8am	7am-8am ,6am-7am,9am-10am
United Kingdom	2pm-3pm, 6pm-7pm,7pm-8pm	4am-5am, 8am-9am, 2am-3am,
United Kingdom	7pm-8pm, 8pm-9pm, 5pm-6pm	5am-6am, 2am-3am, 3am-4am
United Kingdom	6pm-7pm, 3pm-4pm, 5pm-6pm	5am-6am, 11am-12pm, 4am-5am
United Kingdom	2pm-3pm, 1am-2am, 9pm-10pm	6am-7am, 12am-1am,5am-6am
United States	12pm-1pm, 6pm-7pm, 8pm-9pm	7am-8am,9am-10am, 8am-9am
United States	3am-4am, 2pm-3pm, 3pm-4pm	9am-10am, 12pm-1pm, 6am-7am
United States	4pm-5pm, 9pm-10pm, 1am-2am	6am-7am, 12pm-1pm, 11am-12pm
United States	6pm-7pm, 4pm-5pm, 2am-3am	5am-6am,8am-9am,9am-10am
United States	8pm-9pm, 10pm-11pm, 2am-3am	5am-6am, 4am-5am,3am-4am
India	11am-12am, 8pm-9pm, 12am-1am	5am-6am, 8am-9am, 6pm-7pm
India	3am-4am, 10pm-11pm, 1pm-2pm	11pm-12am, 6am-7am, 11am-12pm
India	4pm-5pm, 6am-7am, 4am-5am	8am-9am, 9am-10am, 1am-2am
India	6pm-7pm,5pm-6pm, 3pm-4pm	6am-7am, 12am-1am, 11am-12pm
India	4pm-5pm, 5pm-6pm, 12pm-1pm	9am-10am, 7am-8am,8am-9am

Below table shows the country with the time periods in which maximum number of non-vandal edits occurs.

Table 2: Time Periods Of Non-Vandal Edits

Country	Time periods with maximum no. of edits	Time periods with minimum no. of edits
United Kingdom	1am-2am , 4pm-5pm, 11pm-12pm	12pm-1pm ,11am-12am,10am-11am
United Kingdom	6am-7am, 4am-5am,5pm-6pm	1pm-2pm, 12am-1am, 11pm-12am,
United Kingdom	7pm-8pm, 8pm-9pm, 5pm-6pm	5am-6am, 2am-3am, 3am-4am
United Kingdom	7pm-8pm, 1pm-2pm, 6pm-7pm	7am-8am, 2am-3pm, 8am-9am
United Kingdom	1pm-2pm, 3pm-4pm, 10pm-11pm	9am-10am, 7am-8am,10am-11am
United States	9pm-10pm, 11pm-12am, 1am-2am	11am-12am,10am-11am, 9am-10am
United States	2pm-3pm, 9am-10am, 3pm-4pm	6am-7am, 7pm-8pm, 12pm-1pm
United States	1pm-2pm, 3pm-4pm, 10pm-11pm	9am-10am, 10am-11am, 7am-8am
United States	3am-4am, 10pm-11pm, 8pm-9pm	6am-7am,10am-11am,12pm-1pm
United States	6pm-7pm, 9pm-10pm, 8am-9am	7am-8am, 10am-11am,11am-12pm
India	11pm-12am, 2pm-3pm, 2am-3am	11am-12pm, 5am-6am, 9am-10am
India	2pm-3pm, 7pm-8pm, 6pm-7pm	5am-6am, 6am-7am, 2am-3pm
India	4pm-5pm, 6am-7am, 4am-5am	8am-9am, 9am-10am, 1am-2am
India	6pm-7pm,5pm-6pm, 3pm-4pm	6am-7am, 12am-1am, 11am-12pm
India	7pm-8pm, 2pm-3pm, 10am-11am	4am-5am, 2am-3am,8am-9am

In the tables above, time is shown in GMT which is different for different regions. On average the maximum number of vandal edits occurs during office hours while on average the

maximum number of non-vandal edits occurs late evenings. The least number of edits occur early in the morning or late at night in both the cases. Trend is almost same for United Kingdom, United States and India, all of these have different time zones, and making it more clear that majority of edits come from the same country a page belongs to.

7.3 Spatial Characteristics of Vandalism

We picked the pages of 26 countries to study vandalism. Our mail goal was to study why vandalism occurs in the pages of the countries and which regions contribute to vandalism the most for these pages.

The list of 26 countries is: Afghanistan, Australia, Bangladesh, Burma, Cambodia, China, Cuba, Germany, Greece, India, Iran, Iraq, Israel, Italy, Japan, Lebanon, Nepal, North Korea, Pakistan, Russia, Saudi Arabia, Sri Lanka, Taiwan, United States, United Kingdom and Zimbabwe.

For getting the regions we extracted the IP Addresses of the users who created the versions of a page and then we mapped those IP Addresses to the countries to which they belong, using Geo IP location service. There might be a possibility that IP may not be the actual IP of the user as shown in Wikipedia, IP might be a turning IP but we consider it as rare and ignore it.When we get the countries we feed them into Hadoop to get the count of how many times the country contributed to vandalism and then rank the countries, meaning the country which contributed to vandalism is ranked one and so on.

United States, United Kingdom, Canada and Australia are the top contributors of vandalism for every country.

Table 3: Countries Contributing To Vandalism

Country	**Other Vandal contributors**	**Country**	**Other Vandal contributors**
Afghanistan	Philippines, Sweden, Germany, Norway, Netherlands, India, Israel, Malaysia, New Zealand, Turkey, Austria	Italy	New Zealand, Hong Kong, Thailand, Argentina, Colombia, Germany, Netherlands, Philippines
Australia	Germany, New Zealand, Norway, Philippines, Sweden, Austria, Singapore, Spain, India, France, Argentina	Japan	New Zealand, Hong Kong, Netherlands, Denmark, Germany, Greece, Ireland, Malaysia, Philippines, Singapore, Japan
Bangladesh	New Zealand, Thailand, Egypt, Pakistan, Serbia, United Arab Emirates	Lebanon	United Arab Emirates, Israel, Colombia, Qatar, Saudi Arabia, Sweden. Lithuania, Dominican republic, Italy, Kuwait, Bahrain
Burma	New Zealand, Myanmar, Singapore, Argentina, Netherlands, Brazil, Romania, Saint Lucia, Slovakia, Finland	Nepal	Nepal, India, Hong Kong, Belgium, Denmark, New Zealand, Spain
Cambodia	France, Brazil, Germany, Italy, Malaysia, Romania, Thailand	North Korea	India, Japan, Mexico, New Zealand, Norway
China	Ireland, Germany,	Pakistan	India, Pakistan,

	Poland, Sweden, Netherlands, Taiwan, New Zealand, Singapore, Colombia, France, Mexico, Estonia		Denmark, Netherlands, Philippines, Singapore
Cuba	Sweden, Belize, Thailand	Russia	New Zealand, Germany, Georgia, Israel, Norway, Philippines, Russian Federation, Mexico, Ukraine, Estonia, Netherlands
Germany	Germany, Norway, Netherlands, Belgium, Egypt, Hong Kong, Poland, Sweden	Saudi Arabia	Saudi Arabia, India, Pakistan, Ireland, New Zealand, Norway, Egypt, Colombia, Malaysia
Greece	Greece, Hungary, Albania, Germany, Ireland, Singapore, Spain, Switzerland, Trinidad and Tobago, Turkey, Poland	Sri Lanka	Sri Lanka, India, New Zealand, Denmark, Germany, Hong Kong, Ireland, France, Norway, Qatar, Seychelles, Egypt, Turkey, Pakistan
India	India, Bahrain, Bangladesh, Pakistan, Chile, France, Germany, Malaysia, New Zealand, Sweden, Hong Kong	Taiwan	Taiwan, China, Germany
Iran	Switzerland, Belgium, Greece, Philippines, Sweden	United Kingdom	Spain, India, Argentina, Chile, Germany, Greece, Norway, Turkey
Iraq	Kuwait, Indonesia,	United States	India, Germany,

	Ireland, India, Malaysia, Pakistan, Sweden, United Arab Emirates, Finland		Greece, France, Poland, Russian Federation, Sweden, United Arab Emirates, Brazil, Norway
Israel	Israel, Canada, Turkey, Oman, United Arab Emirates, South Africa, France, New Zealand, Norway, Egypt, Saudi Arabia, Singapore, Syrian Arab Republic, Taiwan, Denmark, Indonesia	Zimbabwe	New Zealand, Philippines, South Africa, Ireland, Netherlands, Germany, Sweden, Switzerland, Denmark, Norway

By studying about the vandal contributors we observed an interesting fact that hostilities among the countries are a big factor because of which vandalism occurs in the pages of the countries.

For example: In ‘Pakistan’ page ‘India’ is one of the top contributors of vandalism, this is because of hostility between India and Pakistan. In ‘India’ page ‘Pakistan’ is also a vandal contributor.

In ‘Taiwan’ page ‘China’ is one of the top contributors and also in ‘China’ page ‘Taiwan’ is one of the vandal contributors.

‘Israel’ and ‘United Arab Emirates’ do not have diplomatic relationships and the results show that ‘United Arab Emirates’ is one of the top contributors in ‘Israel’ page.

Similar can be observed in the pages of other countries.

Below are some graphs showing vandal contributions of India, China and United Arab Emirates in the pages belonging to other countries and they clearly show that hostilities among the countries is one of the factors that contribute to vandalism.

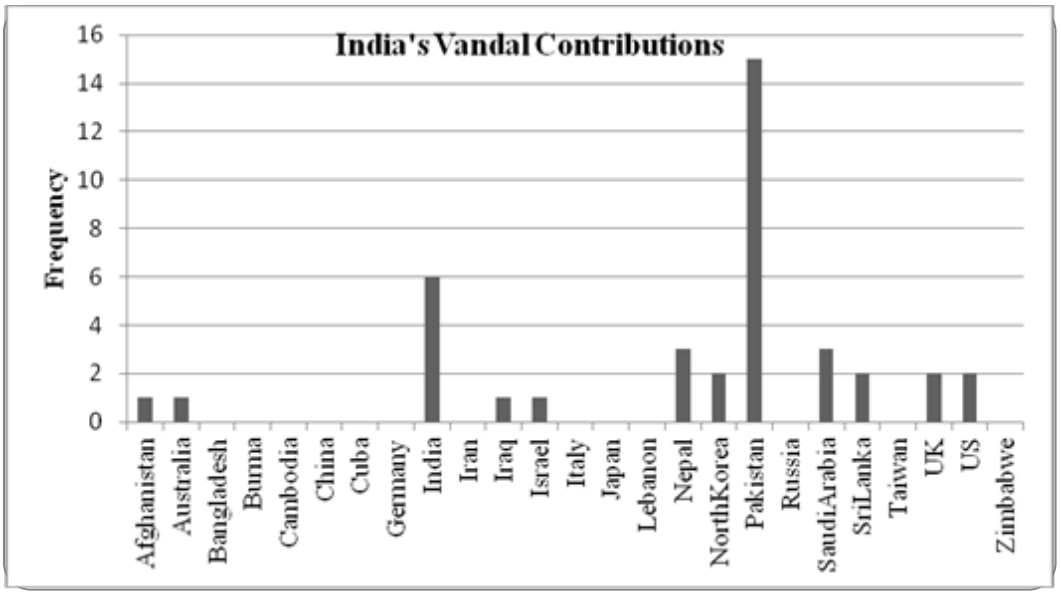

Figure 20: India’s Vandal Contributions

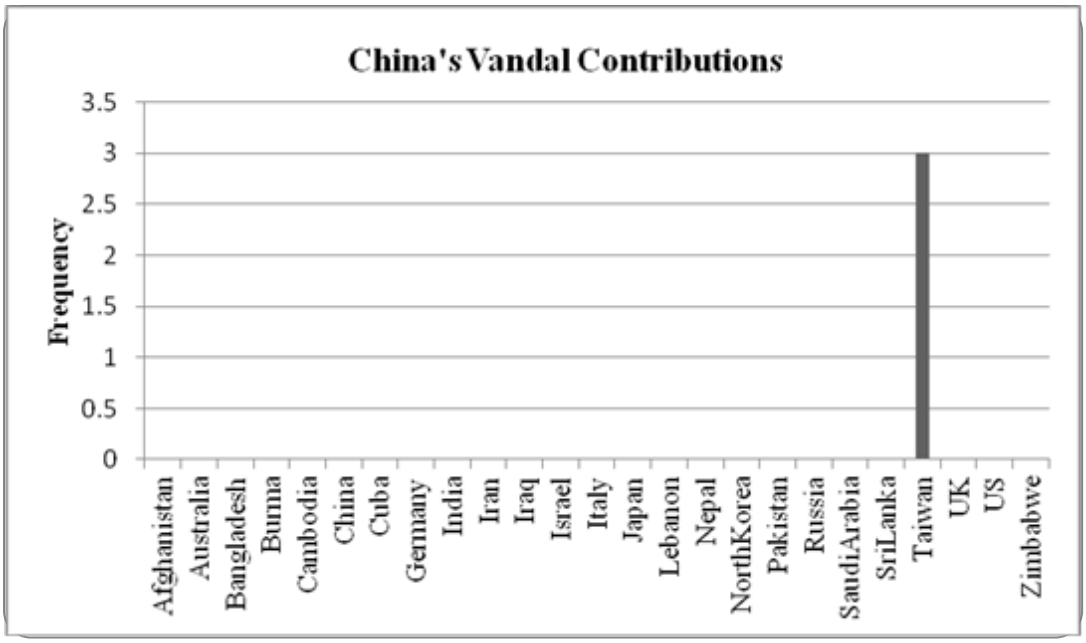

Figure 21: China's Vandal Contributions

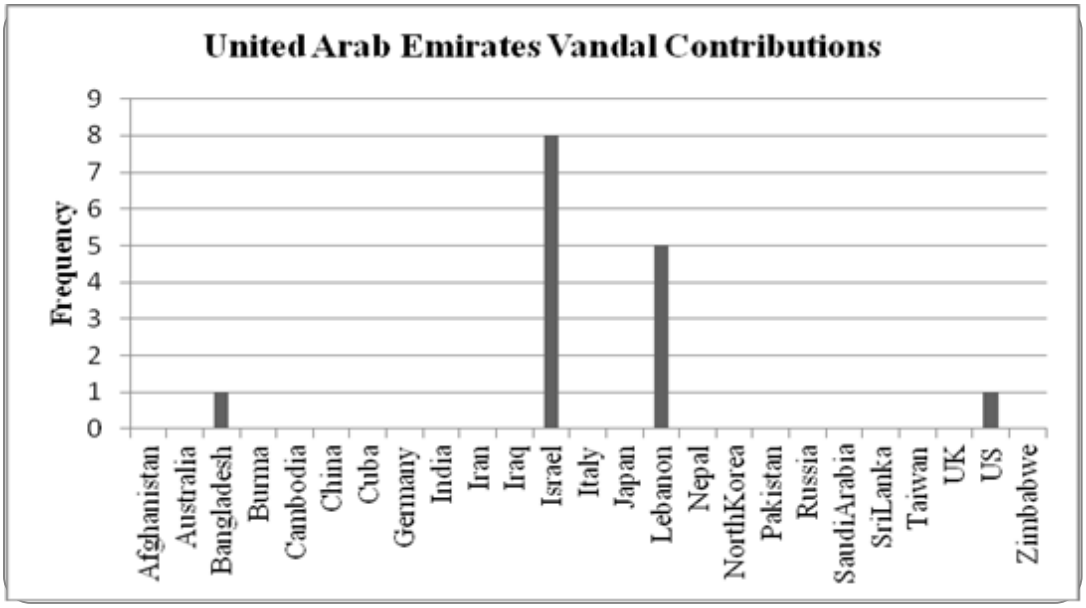

Figure 22: United Arab Emirates Vandal Contributions

7.4 Vandalism Detection using Context Aware Approach

We also devised a vandalism detection method called context aware approach which is based on co-occurrence probability. Co-occurrence probability means the probability of occurrence of the word with the page title. Our hypothesis is that if we calculate the co-occurrence probability of a word with the page title, probability too low means the word doesn't lie in the context of the page, hence may be vandalized.

Firstly we collect all the added words from all the versions of a page by comparing each version with its previous version. Then we query that word with its page title in a trustworthy search engine. We obtain the count by querying the word with its page title. In this way we calculate the co-occurrence probability of the word with its page title.

$$Co - occurence\ Probability$$

$$= (Count\ of\ (pagetitle) + Count\ of(word)/((Count\ of(pagetitle) + Count\ of(word) + (Count\ of(pagetitle) - Count\ of(word))$$

If the co-occurrence probability is too low, it means the word does not lie in the context of the page and hence it may imply vandalism.

If we take an example of 'Geriatrics' page, an added word 'Mongoose' when queried with 'Geriatrics' return very less results as compared to the added word 'Medicine' when queried with 'Geriatrics'.

Performance of words is quantified using three metrics: minimum ratio, mean and maximum ratio.

Minimum Ratio: The least co-occurrence probability of a word among all added words

Mean: The average of co-occurrence probabilities of all added words

Maximum Ratio: The highest co-occurrence probability of a word among all added words

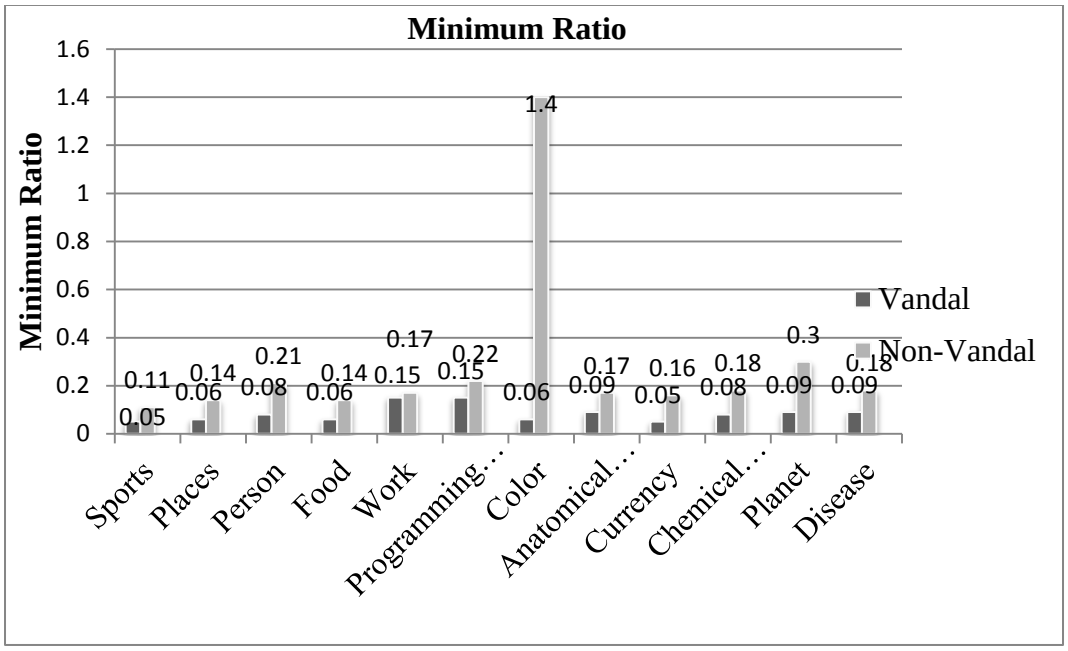

Figure 23: Minimum Ratio

Above figure shows the results of vandal and non-vandal edits with minimum ratio as the metrics. In almost all the domains vandal edits are distinguishable from non-vandal edits as the value of the minimum ratio of vandal edits is almost double than the vandal edits. In Color domain value of minimum ratio for non-vandal edit is about 23 times more than vandal edit. There is no case where minimum ratio of vandal edit is more than non-vandal edit. So if we set up a threshold we can distinguish between a vandal edit and a non-vandal edit using minimum ratio as a threshold value.

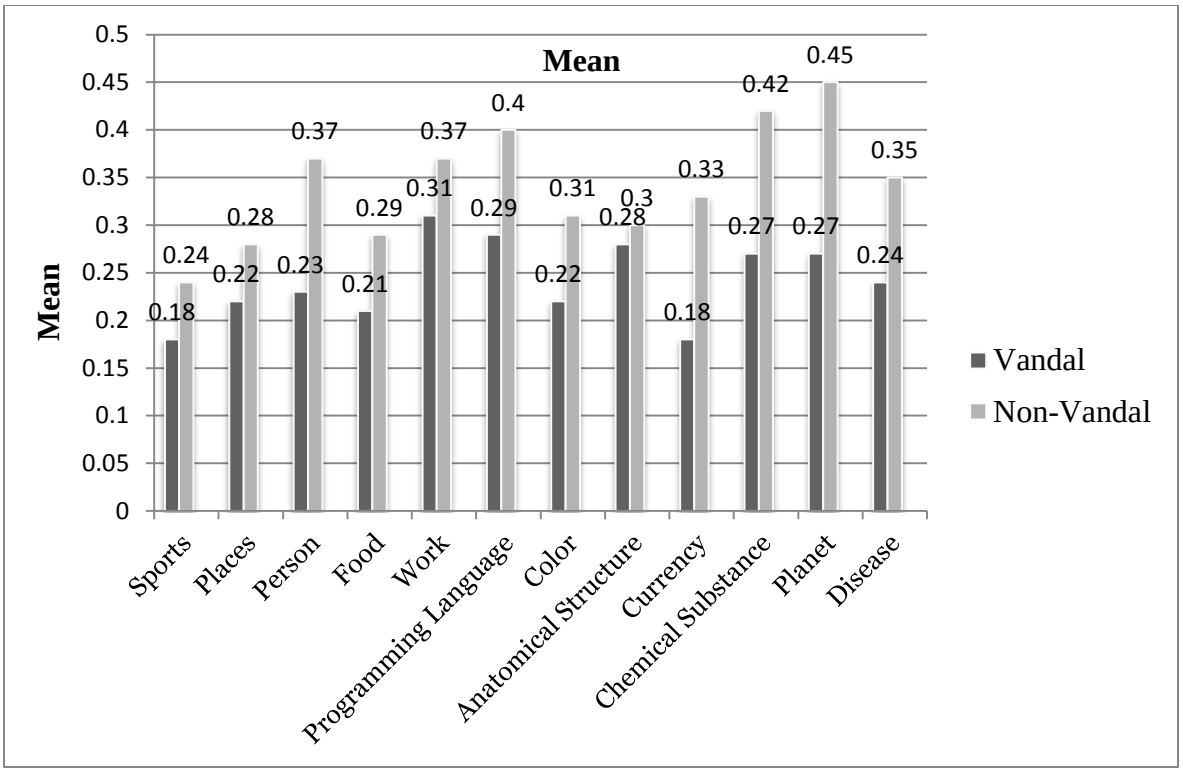

Figure 24: Mean

Above figure shows the results of vandal and non-vandal edits with mean as the metrics. In almost all the domains vandal edits are distinguishable from non-vandal edits as the value of the mean of vandal edits is higher than the vandal edits. There is no case where mean of a vandal edit is more than non-vandal edit. So if we set up a threshold we can distinguish between a vandal edit and a non-vandal edit using mean as a threshold. It is a less effective metric than minimum ratio as minimum ratio distinguishes a vandal edit from a non-vandal edit more efficiently.

Maximum ratio is not a good distinguishing metric according to our experiments; hence the results are not included here. If we combine minimum ratio and mean, we can get a distinguishing factor set to identify the vandal edits.

7.5 Status Inverse, Topic Replacement and Number Replacement attacks

Replacing or changing a number is one form of editing that is prone to vandalism. Regular expression is used to identify numbers in a text and check whether both the deleted text and the inserted text involve different numbers [7]. We calculate the percentage of number replacement attacks that are vandalized in the pages of a particular domain for all the domains.

In status inverse attacks edits inverse the meaning of a sentence. To identify these instances, the content of a new edit contains the words “not”, “none”, or prefixes of “un-”, “dis-” to existing words are checked [7]. We calculate the percentage of statement inverse attacks that are vandalized in the pages of a particular domain for all the domains.

In Topic replacement attacks an edit replaces the link of one Wikipedia topic with another Wikipedia topic. A Wikipedia topic is the title of an article. Experiments show that the majority of hyperlinks in an article between Wikipedia topics are mainly created at the earlier stage of an article. After its content gets stabilized, new edits are less likely changing these hyperlinks [7]. We calculate the percentage of topic replacement attacks that are vandalized in the pages of a particular domain for all the domains.

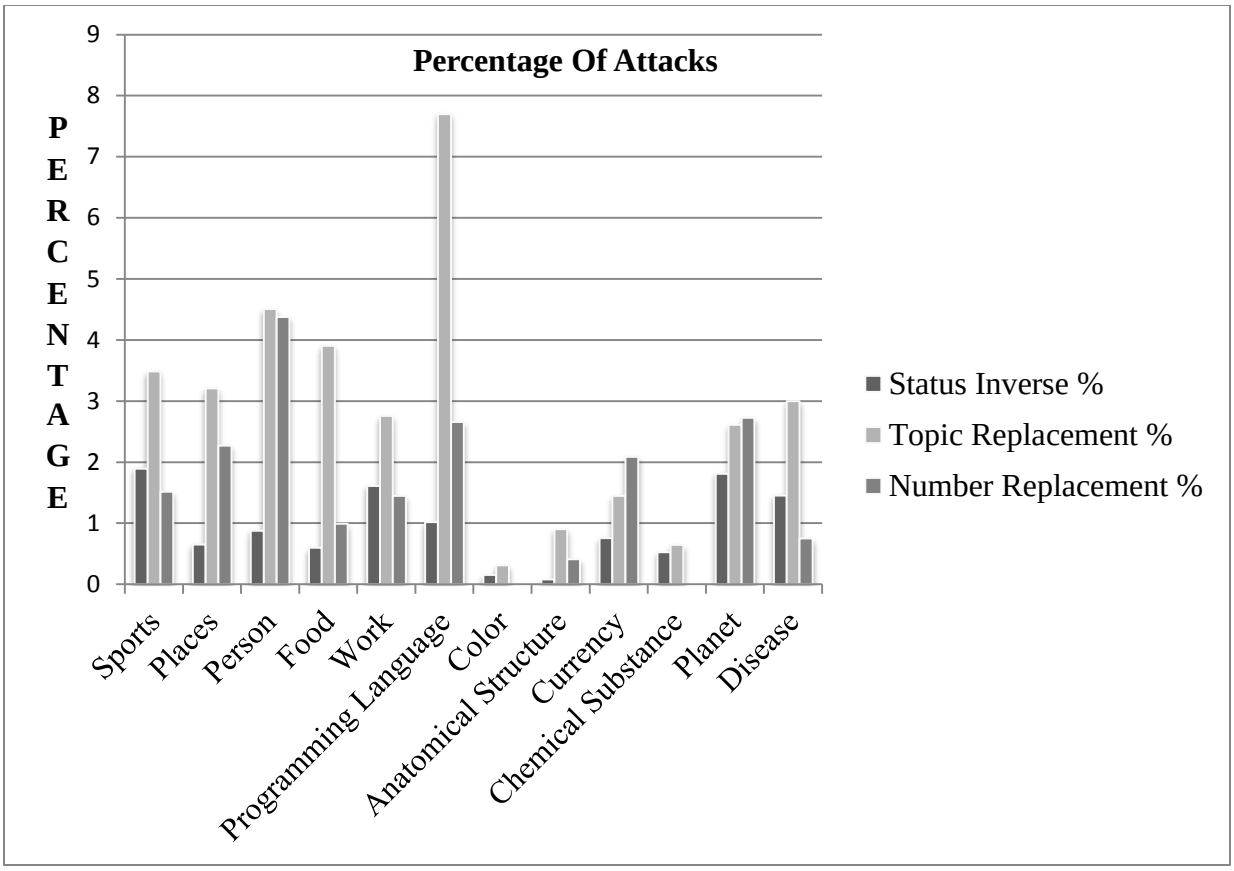

Figure 25: Percentage Of Attacks

Figure 25 shows the percentage of status inverse attacks, topic replacement attacks and number replacement attacks. Status inverse attacks are highest for Planet domain and lowest for Anatomical Structure domain. Topic replacement attacks are highest for Programming Language domain and lowest for Color domain. Number replacement attacks are highest for Person domain and lowest for Color and Chemical substance domain.

7.6 Prevalence of Edit Wars

Edit wars are the consecutive reverts of the versions that occur in Wikipedia pages. Edit wars can be 2RR's means two reverts occurring consecutively, 3RR's means three reverts occurring consecutively and 4RR's means four reverts occurring consecutively. For every page in a domain we calculate how many edit wars were there in each year and then we find the average number of edit wars for each domain.

Reverts are identified based on the pathtype. If pathtype is 2 or 6, it means there is a revert and if there are two consecutive versions whose revert path is 2 or 6 it means it is a 2RR, similarly if there are three consecutive versions having revert path 2 or 6 it means it is a 3RR and with four consecutive versions having revert path 2 or 6 it is a 4RR.

Figure 26 shows the prevalence of edit wars. As we can see in the figure 2RR's, 3RR's and 4RR's occurs the most in Places domain. 2RR's and 3RR's occurs the least in Chemical substance domain and 4RR's occurs the least in Anatomical Structure, Chemical substance and Disease domain.

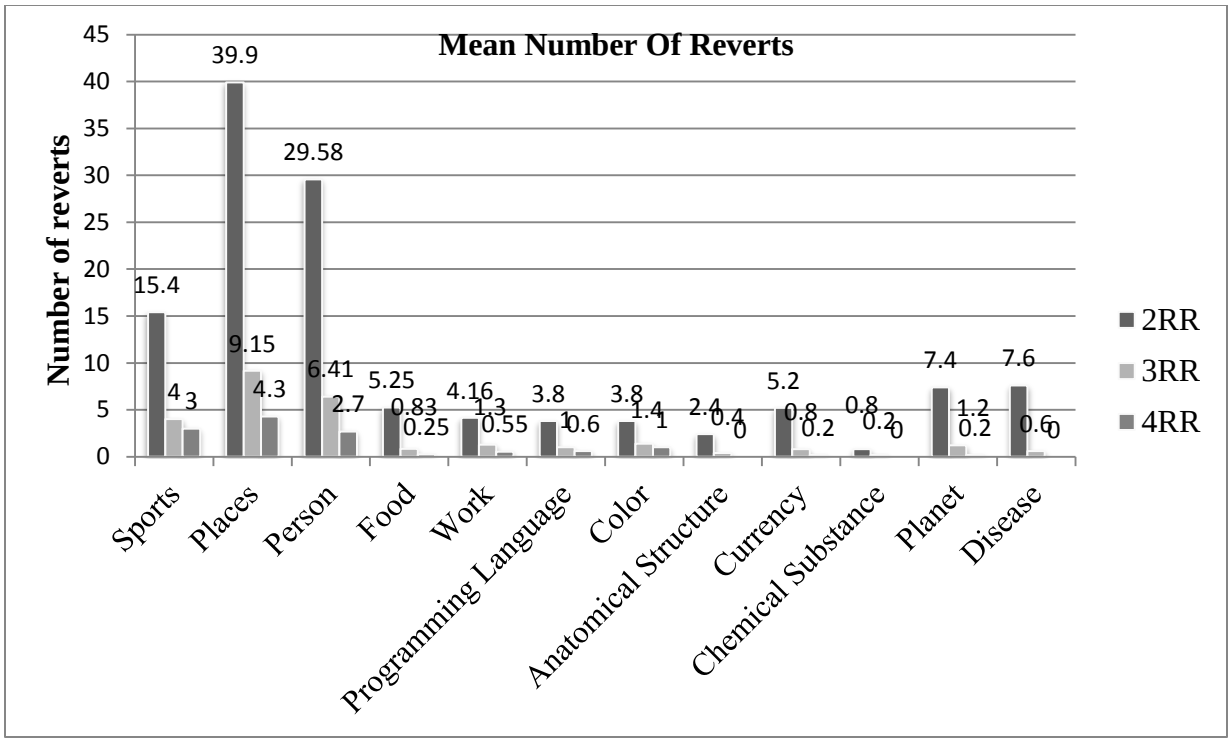

Figure 26: Mean Number Of Reverts

CHAPTER 8

Content evolution of Wikipedia

8.1 Effect of occurrence of events to Wikipedia pages

We study the effect of occurrence of events to Wikipedia pages. We selected 10 Wikipedia pages named 'Egypt', 'Iraq', 'Japan', 'Osama Bin Laden', 'Sachin Tendulkar', 'Steve Jobs', 'Whitney Houston'. Then on each page we took a main event and calculated the mean time duration of 2 months before the event and 2 months after the event occurred.

For example we took the event 'Steve jobs died on October 6 2011'. We calculated the mean time duration of versions for 2 months before October 6 and mean time duration of versions for 2 months after October 6. Further we also studied the effect of occurrence of events on vandalism. We studied how many vandalized versions got introduced in 2 months after the event occurred as compared to 2 months before the event occurred.

Similarly we took other events for various pages listed below:

- 'Egypt's Hosni Mubarak forced from power' on Feb 11th , 2011 for 'Egypt' page
- 'Iraq war ended on Dec 15th 2011' for 'Iraq' page
- 'Japan earthquake and tsunami March 11th 2011' for 'Japan' page
- 'Osama Bin Laden died on May 1st 2011' for 'Osama Bin Laden' page
- 'Sachin Tendulkar scores his 100th international century for India on March 16th 2012' for 'Sachin Tendulkar' page
- 'Whitney Houston died on Feb 11th 2012' for 'Whitney Houston' page

Further we study the effect of occurrence of events on vandalism. We study if the occurrence of events increases vandalism or not.

Occurrence of events has a great impact on Wikipedia. As the events occur edits increase. Mean duration between edits drastically decrease for the two months after the event as compared to the two months before the event.

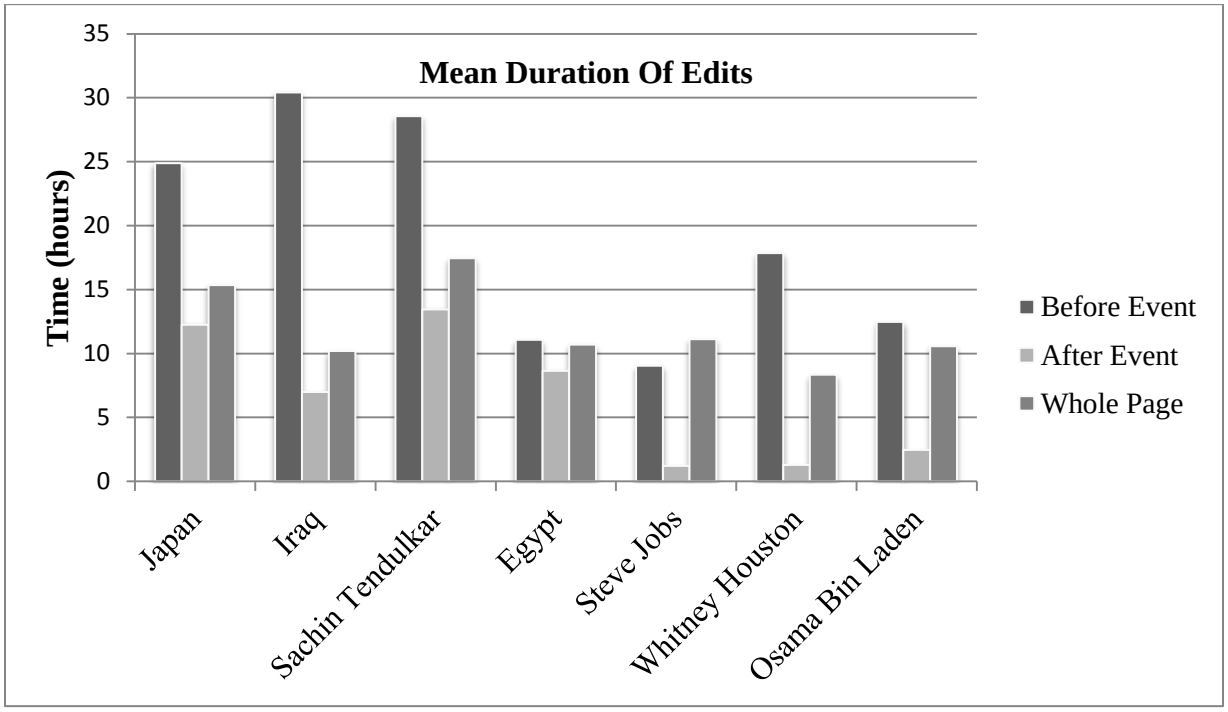

Figure 27: Mean Duration Of Edits

Every page has an impact on mean duration on occurrence of an event and especially for pages Steve Jobs, Iraq and Whitney Houston mean duration decreased drastically on occurrence of the event.

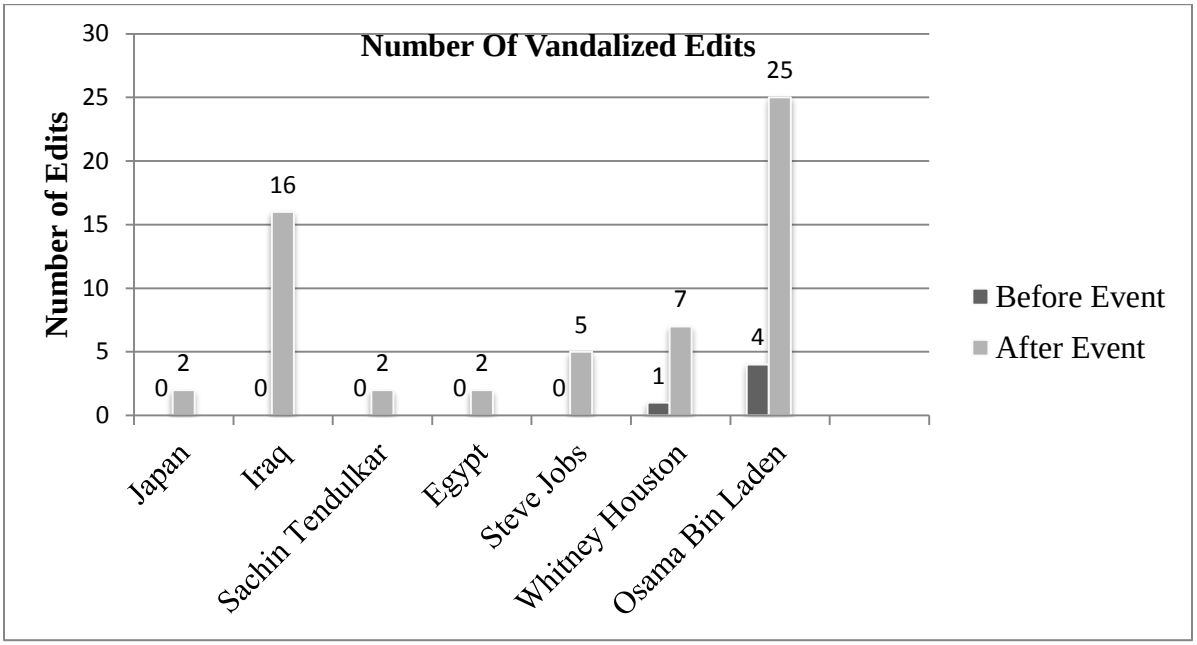

Figure 28: Number Of Vandalized Edits

Above figure clearly shows the increase of vandal edits on occurrence of an event. In some pages vandalism increases drastically, like for pages Iraq and Osama Bin Laden. These pages are more critical and people all over the world are interested in them so mean duration and vandalism is affected greatly when an event occurs.

8.2 Effect of adding real time data in Wikipedia due to limited knowledge

While studying we observed one thing that Wikipedia is not a good source of current information and is more inclined towards providing historical information. Wikipedia is

controlled by people and not everyone has knowledge about everything happening in the world. Whenever an event occurs there are many people who try to add information that is correct but either unreferenced or not in knowledge of other people who revert it stating the information as unreferenced. After various revert finally when the news is officially out, the same information is accepted, making Wikipedia as a source of historical information rather than current information.

But this is not true in case of every article. There are various articles in which not many people edit information. So there are many cases in which the unreferenced edits are reverted and eventually that Wikipedia page does not contain that important part of information.

Hence Wikipedia is not very efficient for volatile situations and sometimes does not contain important information due to limited knowledge of editors.

CHAPTER 9

Conclusion and Future Work

9.1 Conclusion

We used 12 domains to study Wikipedia. We studied the user contributions showing that registered users contribute to about 65% of edits and unregistered users contribute to about 35% of edits. Also unregistered users contribute much more to vandalism as compared to registered users. Hence unregistered users are the main contributors of vandalism.

We list the top occurring common vandal words. We also studied about number replacement, statement inverse and topic replacement attacks. Topic replacement attacks are more prevalent and status inverse attacks are least prevalent in vandalized pages.

Occurrence of events decreases the mean duration and increases vandalism in the articles related to that event. Our study concludes that Wikipedia is a good source of historical information but is not efficient for volatile situations.

We also studied the temporal characteristics of vandalism and non-vandalism. Vandalism occurs the most during office hours while non-vandalism occurs the most during late evenings.

After studying the spatial characteristics we observed that United States, United Kingdom, Canada and Australia are the top contributors of vandalism in every article related to countries.

Hostilities among the countries are one major cause of vandalism.

We proposed parameters: minimum ratio and mean and used them in a vandalism detection method called context aware approach in which we are able to distinguish between vandal and non-vandal edits efficiently.

9.2 Future Work

Data was limited for this study, so in future we can do this study on all the existing 256 domains and more number of pages for more efficient results. Temporal characteristics can be studied on more than 3 time zones, preferably all the time zones that exist. Spatial characteristics can be extended to more number of articles related to the existing countries. Context Aware approach can be more refined to give more efficient results.

There is a lot of research required for studying causes of vandalism in Wikipedia. Wikipedia is written collaboratively by volunteers all over the world and it is open to all, hence it is difficult to stop people from making vandal edits. Vandal edits are made to show hatred, gain recognition, for fun and many other reasons. More strong algorithms need to be devised that can detect and remove vandalism efficiently.

REFERENCES

[1]. Andrew G. West, Sampath Kannan, Insup Lee. Detecting Wikipedia Vandalism via Spatio-Temporal Analysis of Version Metadata. In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, New York, NY, USA, 2010. ACM.

[2]. B. Thomas Adler, Luca de Alfaro, Ian Pye. Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features

[3]. B. Thomas Adler, Luca de Alfaro, Santiago M. Mola-Velasco. Detecting Wikipedia Vandalism using WikiTrust Lab Report for PAN. In WICOW, 2010.

[4]. Santiago M. Mola-Velasco. Wikipedia Vandalism Detection

[5]. Cristian-Alexandru Dragusanu, Marina Cufliuc, Adrian Iftene. Detecting Wikipedia Vandalism using Machine Learning Notebook for PAN. CLEF 2011

[6]. Koen Smets and Bart Goethals and Brigitte Verdonk. Automatic Vandalism Detection in Wikipedia: Towards a Machine Learning Approach. In *Proc. of AAAI workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy*, pages 43–48. AAAI, 2008

[7]. Qinyi Wu, Danesh Irani, Calton Pu, and Lakshmish Ramaswamy. Elusive Vandalism

Detection in Wikipedia: a Text Stability-based Approach. In *CIKM*, 2010.

[8]. Amit Belani. Vandalism Detection in Wikipedia: a Bag-of-Words Classifier Approach. *CoRR*, abs/1001.0700, 2010.

[9]. M. Potthast, B. Stein, and R. Gerling. Automatic vandalism detection in Wikipedia. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, chapter 75, pages 663–668. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

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

[11]. Cluebot. http://en.wikipedia.org/wiki/User:ClueBot,

Version as of 20:29, 22 May 2010.

[12]. Si-Chi Chin, Padmini Srinivasan, W. Nick Street, David Eichmann. Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models. In *Proceedings of 4th Workshop on Information Credibility on the Web*, 2010.

[13]. http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia

[14]. http://en.wikipedia.org/wiki/History_of_Wikipedia

[15]. C. Goldspink. Normative behavior on wikipedia information. *Communications and Society*

[16]. http://en.wikipedia.org/wiki/Wikipedia:Wikipedians

[17]. http://simple.wikipedia.org/wiki/Wikipedia:User_access_levels

[18]. Chitu Okoli. A Brief Review of Studies of Wikipedia in Peer-Reviewed Journals

[19]. E. Svoboda, "One-Click Content, No Guarantees," IEEE Spectrum

[20]. http://en.wikipedia.org/wiki/Criticism_of_Wikipedia

[21]. S. Hahmann, D. Burghardt. Investigation on factors that influence the (geo)spatial characteristics of Wikipedia articles

[22]. Jacobi Carter, Cluebot and Vandalism on Wikipedia

[23]. Manoj Harpalani, Michael Hart, Sandesh Singh, Rob Johnson, and Yejin Choi. Language of Vandalism: Improving Wikipedia Vandalism Detection via Stylometric Analysis

[24]. http://hadoop.apache.org/docs/r0.16.4/hdfs_design.html

[25]. 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.

[26]. 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*, pages 261–270, New York, NY,USA, 2007. ACM.

[27]. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 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, 2009.

[28]. C. Goldspink. Normative behavior on Wikipedia information. *Communications and Society*, 13(3).
[29]. L. Ramaswamy, A. Iyengar, L. Liu, and F. Douglis. Automatic Detection of Fragments in Dynamically Generated Web Pages. In *Proceedings of the* 13th *World Wide Web Conference*, May 2004.

[30]. A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, New York, NY, USA, 2010. ACM.

[31]. A. G. West, S. Kannan, and I. Lee. Detecting Wikipedia vandalism via spatio-temporal analysis of revision metadata? In *EUROSEC '10: Proceedings of the Third European Workshop on System Security*, pages 22–28, New York, NY, USA, 2010. ACM

APPENDIX I

Table 4: List of common vandal words

Word	**Domains in which word appears**
Ball	Sports, Person, Food, Places, Chemical substance, Color, Planet, Disease
Kiss	Work, Person, Food
British	Places, Person, Food, Sports, Currency, Color, Chemical substance, Disease, Planet, Work, Programming language
Hole	Sports, Food, Work, Places
Play	Sports, Person, Places, Work, Disease, Currency, Food,
Course	Sports, Person, Places, Programming language, Chemical substance, Food, Disease
Date	Places, Person, Work, Currency, Sports, Food, Disease, Planet, Programming language, Anatomical structure
Shot	Sports, Person, Food, Work
Parties	Places, Color, Work, Person, Food
Nude	Work, Person
Woman	Sports, Color, Food, Chemical substance, Places, Person, Disease, Work, Anatomical structure
Pork	Food, Person, Places
Dick	Planet, Work, Food, Places, Person, Disease, Anatomical Structure
Sex	Places, Person, Anatomical Structure, Color, Planet, Disease, Work
Laid	Sports, Places, Person, Color
Fuck	Person, Places, Food, Planet, Chemical Substance, Disease, Work
Terrorism	Places, Person
Penal	Person, Places, Sports
English	Places, Person, Food, Sports, Planet, Food, Disease, Chemical substance, Work, Color
Chicken	Food, Place, Person, Work
Meat	Food, Person, Work, Places
Bomb	Places, Person, Food, Programming Language, Work

Death	Places, Person, Disease, Sports, Food, Currency, Anatomical Structure, Programming Language, Planet
Drug	Places, Food, Person, Chemical Structure, Disease
Sick	Person, Places, Chemical Substance, Anatomical Structure

Table 5: List of common non-vandal words

Word	**Domains in which word appears**
Dollar	Currency, Work, Person, Places Food, Sports, Planet, Color
Cup	Person, Places, Sports, Work, Food,
Align	Places, Currency, Person, Work, Food, Chemical Substance, Programming Language, Color
Cite	Person, Planet, Places, Disease, Person, Work, Anatomical Structure, Color, Chemical Substance, Food, Programming Language, Currency
Match	Person, Sports, Places, Work, Planet, Food, Color, Programming Language
Name	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
Real	Person, Places, Work, Food, Planet, Programming Language, Food, Color, Currency, Chemical Substance
Country	Places, Person, Sports, Food, Chemical Substance, Work, Currency, Color, Disease
Ship	Planet, Programming Language, Person, Places, Work, Color
World	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
List	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language,

	Disease, Currency, Sports, Anatomical Structure, Color
Source	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
Left	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
Metal	Places, Person, Work, Chemical Structure, Currency, Sports
Year	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
City	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Color
Language	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Color
Water	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
History	Places, Person, Work, Planet, Chemical Structure, Food, Programming Language, Disease, Currency, Sports, Anatomical Structure, Color
Constitution	Places, Person, Work, Food, Programming Language, Disease, Currency

APPENDIX II

Table 6: List of top 25 vandal words

Word	**Word**	**Word**	**Word**
Ball	Chicken	British	Woman
Hole	Handicap	Meat	Kiss
Play	Old	Love	Death
Course	Kick	American	Bomb
Nude	Hit	Heart	
Voodoo	Party	Mother	
Married	Drug	Romance	

Table 7: List of next top 25 vandal words

Word	**Word**	**Word**	**Word**
Feel	Peg	Duck	Laid
Skins	Black	Fuck	Sex
Hot	Vomit	Ass	Terrorism
Confused	Stupid	Poke	Obscene
Lies	Kill	Animal	
Sick	Affair	Banana	
Pelvic	Penal	Lamb	