THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

MEASURING WIKIPEDIA ARTICLE QUALITY BY REVISION COUNT

Master's Thesis

MUSTAFA UTKU BAYIK

ISTANBUL, 2012

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCE
COMPUTER ENGINEERING

MEASURING WIKIPEDIA ARTICLE QUALITY BY REVISION COUNT

Master's Thesis

MUSTAFA UTKU BAYIK

Supervisor: Yrd. Doç. Dr. TEVFİK AYTEKİN

ISTANBUL, 2012

THE REPUBLIC OF TURKEY
BAHCESEHIR UNIVERSITY

GRADUATE SCHOOL OF NATURAL
AND APPLIED SCIENCE
COMPUTER ENGINEERING

Name of the thesis: Measuring Wikipedia article quality by revision count
Name/Last Name of the Student: MUSTAFA UTKU BAYIK
Date of the Defense of Thesis: 11.06.2012

The thesis has been approved by the Graduate School of Natural and Applied Sciences.

Doç. Dr. Tunç BOZBURA
Graduate School Director
Signature

I certify that this thesis meets all the requirements as a thesis for the degree of Master of Arts.

Yrd. Doç. Dr. Çağrı GÜNGÖR
Program Coordinator
Signature

This is to certify that we have read this thesis and we find it fully adequate in scope, quality and content, as a thesis for the degree of Master of Arts.

Examining Comittee Members	Signature
Thesis Supervisor Yrd. Doç. Dr. Tevfik AYTEKİN	----------------------------------
Member Yrd. Doç. Dr. Egemen ÖZDEN	----------------------------------
Member Yrd. Doç. Dr. Mehmet Alper TUNGA	----------------------------------

ABSTRACT

MEASURING WIKIPEDIA ARTICLE
QUALITY BY REVISION COUNT

Mustafa Utku Bayık

Computer Engineering

Thesis Supervisor: Tevfik Aytekin

June 2012, 55 Pages

Wikipedia is a free encyclopedia which has millions of articles. Since articles are collaboratively edited by many users there is no standard in the quality of articles. Although there are articles which have high quality (such as featured articles), some articles have poor quality or insufficient information.

In this work we propose to use monthly revision histories of articles in order to assess article quality. We use featured articles in Wikipedia as our standard for quality articles. We extract features from revision history of each article and try to classify articles as featured and non-featured using well-known machine learning algorithms. We achieve a satisfactory classification performance using our methodology as the experimental results on a Wikipedia article dataset that we create shows. We think that this performance is open to further improvement by extracting more features.

Keywords: Wikipedia, Revision Count, Measuring Quality

ÖZET

DEĞİŞİKLİK SAYISINA GÖRE WIKIPEDIA KALİTESİNİ ÖLÇME

Mustafa Utku Bayık

Bilgisayar Mühendisliği

Tez Danışmanı: Tevfik Aytekin

Haziran 2012, 55 Sayfa

Wikipedia milyonlarca makale içeren ücretsiz bir ansiklopedidir. Makaleler bir çok kullanıcı tarafından ortaklaşa yazıldığı için makalelerde eşit bir kalite standardı bulunmamaktadır. Çok kaliteli makaleler olduğu gibi (örn., seçkin içerikli makaleler), kalitesiz ya da yeterli bilgi içermeyen makaleler de mevcuttur.

Bu çalışmada makalelerin değişikliklik geçmişlerini inceleyerek kalitelerini değerlendirmeyi öneriyoruz. Kalite standardı olarak Wikipedia'daki seçkin içerikli makaleleri kullandık. Makalelerin değişiklik geçmişlerinden özellikler çıkararak, bilinen makina öğrenmesi yöntemleriyle makaleleri seçkin içerikli ve seçkin içerikli olmayan şeklinde sınıflamaya çalıştık. Wikipedia'daki makalelerden oluşturduğumuz bir veri seti üzerinde elde edilen deneysel sonuçlar göstermektedir ki geliştirdiğimiz yöntemle tatmin edici bir düzeyde sınıflama performansı elde edilebiliyor. Yeni özellikler çıkararak bu performansın daha da artırılabileceğini düşünüyoruz.

Anahtar kelimeler: Wikipedia, Değişiklik Sayısı, Kalite Ölçme

İÇİNDEKİLER

TABLES

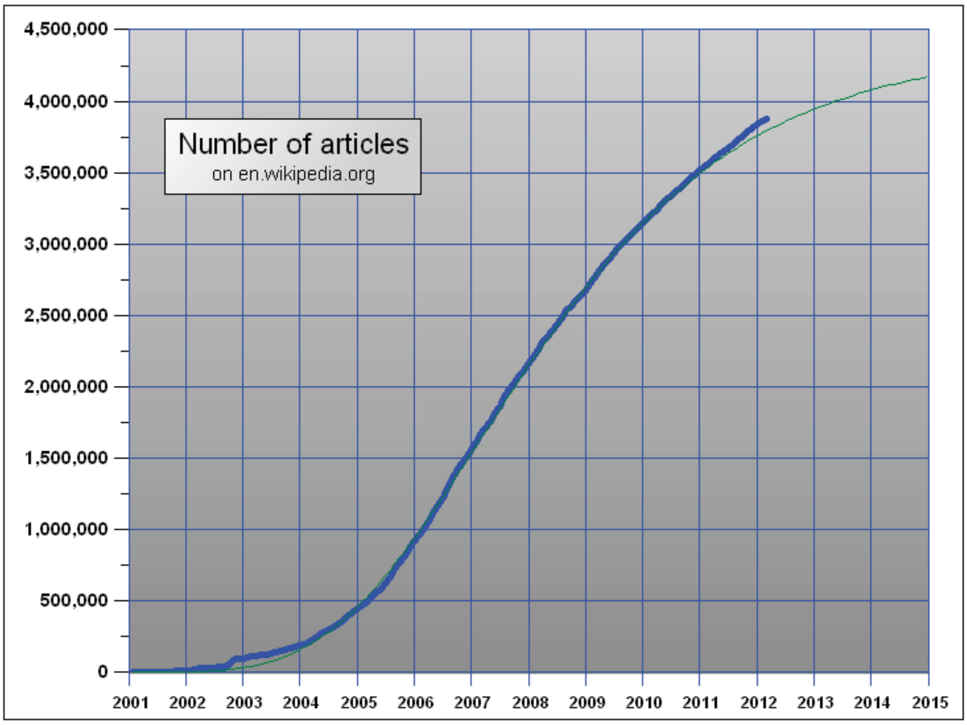
FIGURES

1.INTRODUCTION

1.1 ABOUT WIKIPEDIA

Wikipedia is the most popular free online encyclopedia used by many users to create and revise shared documents for reference in research and daily life. According to Alexa web traffic ranking in 2011, Wikipedia is the most used learners website with over 3.5 million articles in more than 200 languages. Figure 1.1 shows the yearly growth of number of articles in English. It also shows the expected growth for the next few years with a green line. Wikipedia contains approximately one million articles written in English. Since the articles are written by users with or without expertise, Wikipedia is a source of wide information and presentation, which has both reliable and accurate information and untrustable content (McGuinness&Bhaowal 2006, pp.45-67).

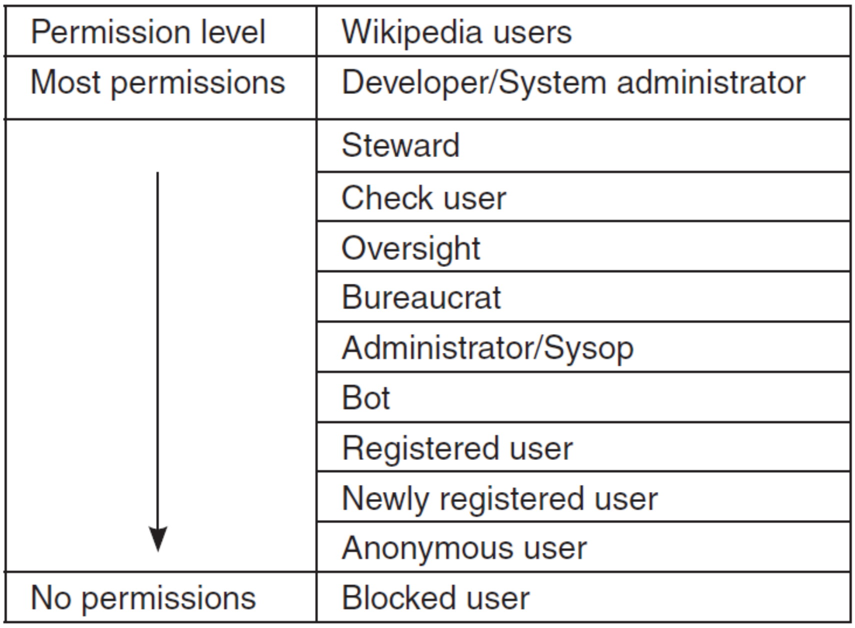
Figure 1.1: Number of articles per month



Recently, due to the expansion, reliability, relevancy and accuracy in its content and due to high ranking of its articles, Wikipedia web engines have provided a wide range of information. The approach taken by the web is different from other encyclopedias paving way for consideration of views from people of diverse backgrounds, knowledge, skills, expertise and experiences. The web is open for critical thinking, analysis, and online research for one to be able to make responsible conclusions and recommendations. (Stvilia &Gasser2005, pp.38-41).

Research has shown that, to reach a high quality and well researched source, comparative analysis adds up value on the quality and reliability of the Wikipedia articles. This is supported from the work published recently by Nature Magazine in 2011, that ranked Wikipedia comparable with Encyclopedia Britannicathe, which is one of the most ancient sources of reference that have been kept up to date with additional information over time. The study conducted showed almost same number of common errors in both,and the kind of information provided from them are almost equal dependable (McGuinness&Bhaowal 2006, pp.122-130). Rigorous mechanisms have been employed for Wikipedia to maintain high quality information on published articles. Before publishing the articles in the Wikipedia, they are supposed to pass through peer-review scrutiny that recommends its publication, which end up with correction or being rejected on non-valueable articles. An article passes through a number of editorial communities and editors measure the articles quality, accuracy and reliability before they got approved for use on the Wikipedia (McGuinness&Bhaowal 2006,pp.81-96). The permission levels of the users are given in Figure 1.2. At that figure can be clearly seen that permission levels are different for users, and so are the trust on the user.

Figure 1.2: Permission levels of users



Permission level	Wikipedia users
Most permissions	Developer/System administrator
↓	Steward
	Check user
	Oversight
	Bureaucrat
	Administrator/Sysop
	Bot
	Registered user
	Newly registered user
	Anonymous user
No permissions	Blocked user

Due to its simple nature to access and understand, the Wikipedia has high web visibility atinformation collection and dissemination among the other web sites. It is based on a varying quality in article presentations. However, Wikipedia has faced a lot of challenges; it has also gained trustworthiness due to the revision counts of articles and reducing erroneous information that can mislead researchers (McGuinness&Bhaowal 2006, pp.231-239). The popularity of the Wikipedia is due to the fact that the articles are written by volunteer users, instead of paid experts. Therefore the kind of information given remains trustworthy and eliminates bias to the paid experts whom may mislead for the shake of getting paid. The Wikipedia.org is open to anyone for access of articles, modification of the already existing one or creating new articles to the site. It is a source that gives one free access to the sum of all human knowledge (Stvilia&Gasser 2005, pp.111-115).

Lack of essential information, accuracy and poor writing of articles poses a great challenge to the quality and application of the Wikipedia articles. As Nicholas Carr put it in 2009, “this is garbage, an incoherent hodge-podge of dubious factoids that adds up to something far less than the sum of its parts”. The good thing, more accurate information may also be found in the same Wikipedia. A lot of criticism has been put forward to challenge the accuracy and reliability.

Vandalism has caused a lot of damage to the articles published in Wikipedia. In order to deal with the challenge, mechanisms have been created, bots and history revisions through the history link at the pages. It organizes the work in two ways. The first way is qualitative level upon which the work uses the total number of edits and unique editor count to measure the article. In this level the color text, font and spacing is determined for the users could immediately understand the quality. The second is quantitative level where machine learning is used by researchers for measuring the quality of the articles and produce algorithmic methods of measurement.

Wikipedia is one of the globally internet encyclopedia accessible to anyone to participate in the online publication and preservation of knowledge. Diversity and experience brings together different kinds of information for users for exposure to a wide diversity of knowledge, opinions, ideas, views, and even reservations. The openness of the web allows the anonymous and unregistered Wikipedia users to play a significant part to the new and existing Wikipedia articles. The Wikipedia's philosophy is that as the content becomes more reliable and accurate over time when community works together on the content. As a result of this, the articles created on the Wikipedia are never "finished" as the addition, correction and collaboration are dynamic (McGuinness&Bhaowal 2006, pp.89-93).

Due to lack of formal peer review, the Wikipedia is subjected to vandalism and access to misleading information to make wrongful judgment in research. Self interested parties on the web can the take the openness advantage also to misinform others through the web (Stvilia &Gasser2005, p.121).

1.2 FEATURED ARTICLES

Many visitors of the Wikipedia find it very difficult to trust the content due to high variance in quality and reliability. For exceptional content quality variance, Wikipedia has taken special attention on articles with exceptional quality through grouping them as "featured" articles. They are the most trusted, reliable and accurate articles for web users. As Wikipedia.org explains, "featured articles contain the best quality that Wikipedia has to offer". These are well researched, prepared and presented articles determined by

Wikipedia editors and their contributions emanates from collaborative works organized in the Wikipedia internet services (Zeng&McGuinness 2006, pp.79-80). The articles are reviewed through diverse method and criteria for determination of accuracy, reliability, neutrality, trustworthiness, completeness and style used in presentation of the article.

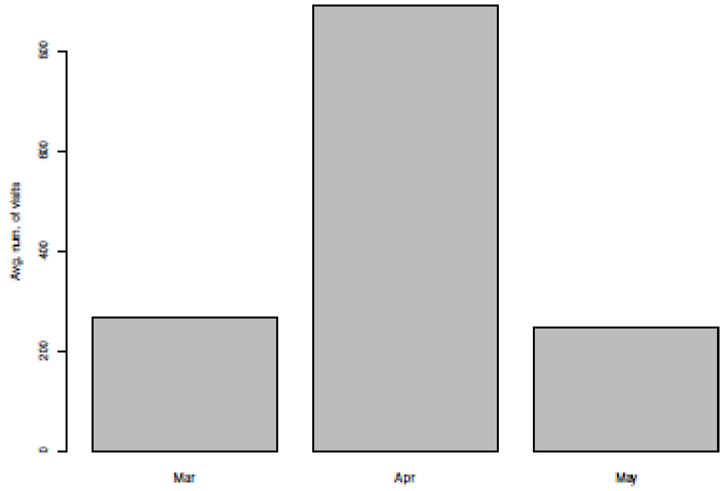
The featured articles are well written, comprehensive and must explain major details and facts concerning the topic. It must be neutral and stable to the fact that it is fair and without any bias without changes as time goes by contents on the featured articles undergo a rigorous and thorough process of review to ensure that high standards are met. The peer-review process involves a group of competitive editors designated for careful scrutiny of every article published on the Wikipedia (McGuinness&Bhaowal 2006, pp.234-236).

The most unfortunate thing is that for one thousand articles published, only one article may be marked as a featured article. This leaves a challenge to the users to decide the quality, reliability and acccuracy of the article. Whilst many articles are indicated with similar metadata to denote low quality articles, majority of the articles found on the Wikipedia does not have such marking for the users to be sensitive and decide how to trust such articles. It is very uncertain for motivated researchers to find mechanisms for determining the quality of an article (Speigelhalter&Thoma 2005, pp.57-61).

The findings from Blumenstock suggestedin 2008, that word count alone can differentiate a featured Wikipedia articles from random Wikipedia articles. From another thinking perspective this conclusion may be intuitive; featured articles should be long for an article to be featured.

The word count and revision count are the qualification for an article to be a featured article. It is tested that revision count outdoes complex techniques in the classification of articles. Long articles are thought to have gone through by several people and therefore have more knowledge and detailed information. Collaborative work of the Wikipedia forces articles to be long articles and of high quality (McGuinness&Bhaowal 2006, pp.78-79).In some occasions, a long article may not be featured and a short article be featured. Therefore not all long Wikipedia articles are high quality and featured. Through collaborative works the quality of an article keeps growing and improves over time. A short article may be of high quality, however not with the Wikipedia context (Speigelhalter&Thoma 2005, pp.94-95).

Figure 1.3: Average visits of featured articles by month

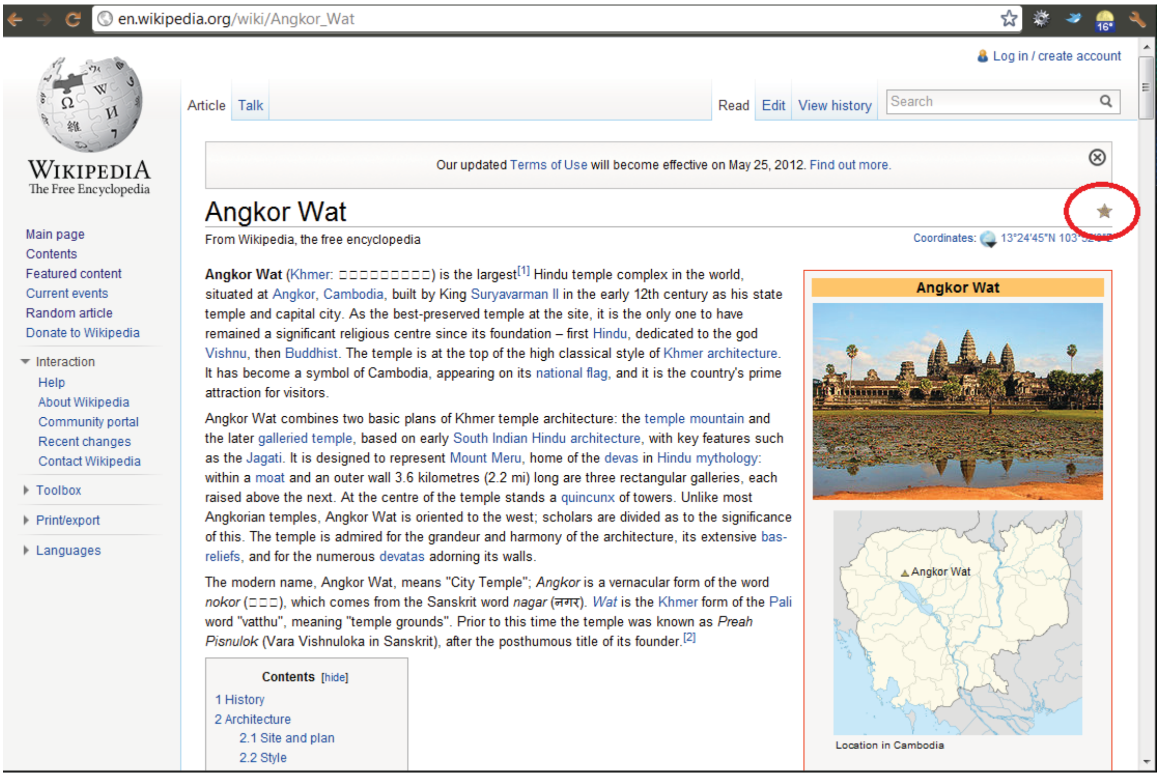


Featured articles are also known to get more visitors. Figure 1.3 shows feature article visits per month. Wikipedia has direct link from its main page to the featured articles, and also it gives additional link to the featured article of the day.

Although using featured articles as a proxy for quality, a higher standard of quality measurement is still required. Organizing human reviewers and editors are very costly and subjective and all articles cannot be reviewed because of their enourmus amount. To come up with quality rating such as Wikiproject Biography or Assessment, and offer great opportunity for future research works for web users, wikipedia bots also play a great role in finding valdalism attempts, and reducing errors.

On the Wikipedia Encyclopedia, featured articles are denoted by a small bronze star icon (★) on the top right corner unless the appropriate preference is set by the user (Wikipedia, 2012). This is meant to show the user have trustworthy on the article in relation to its accuracy and reliability. In figure 1.4 a star can be seen within the red circle.

Figure 1.4: Featured article with a star



Our updated Terms of Use will become effective on May 25, 2012. Find out more.

Angkor Wat

From Wikipedia, the free encyclopedia

Angkor Wat (Khmer: □□□□□□□□□) is the largest[1] Hindu temple complex in the world, situated at Angkor, Cambodia, built by King Suryavarman II in the early 12th century as his state temple and capital city. As the best-preserved temple at the site, it is the only one to have remained a significant religious centre since its foundation – first Hindu, dedicated to the god Vishnu, then Buddhist. The temple is at the top of the high classical style of Khmer architecture. It has become a symbol of Cambodia, appearing on its national flag, and it is the country's prime attraction for visitors.

Angkor Wat combines two basic plans of Khmer temple architecture: the temple mountain and the later galleried temple, based on early South Indian Hindu architecture, with key features such as the Jagati. It is designed to represent Mount Meru, home of the devas in Hindu mythology: within a moat and an outer wall 3.6 kilometres (2.2 mi) long are three rectangular galleries, each raised above the next. At the centre of the temple stands a quincunx of towers. Unlike most Angkorian temples, Angkor Wat is oriented to the west; scholars are divided as to the significance of this. The temple is admired for the grandeur and harmony of the architecture, its extensive bas-reliefs, and for the numerous devatas adorning its walls.

The modern name, Angkor Wat, means "City Temple"; *Angkor* is a vernacular form of the word *nokor* (□□□), which comes from the Sanskrit word *nagar* (नगर). *Wat* is the Khmer form of the Pali word "vatthu", meaning "temple grounds". Prior to this time the temple was known as *Preah Pisnulok* (Vara Vishnuloka in Sanskrit), after the posthumous title of its founder.[2]

Contents [hide]

1 History
2 Architecture
2.1 Site and plan
2.2 Style

1.3 RANDOM ARTICLES

These are Wikipedia articles considered to have very little information or content in that they are short articles. Any Wikipedia article can be viewed as a random article. This means that a random article can come from high quality articles, however, mostly random articles are chosen from other articles of low quality. Whilst a word count at the featured Wikipedia article is around 2700 words, average random Wikipedia article word count is around 200 (McGuinness&Bhaowal 2006, pp.121-123).

Random articles are mostly very short, and this implies that very little revision count work and less collaboration that has been done on them. The work then cannot be relied or accurate for its application in research and study. The revision count was found to be the most correct method to measure the accuracy and reliability of any content from non-featured Wikipedia articles. Commonly articles with more revision count have more words, when more people collaborate, there will be more wisdom. For instance, all articles with more than 2,000 words are classified “featured “and those with less than 2,000 words as “random”, which achievesa very high level of accuracy. Article with less than 2,000 words are considered as below the cutoff threshold accuracy, which is a requirement for the article to be featured (Speigelhalter&Thoma 2005, pp.108-110).

Computing as a classifier method is very costly, as it takes much time in fetching the data to compute. The article considered to be a collection of one author in one revision. One author is regarded to be of very low quality to be trusted. The revision count of very many authors signifies a collective work of many authors with a lot of information put together and hence the level of accuracy and reliability is high. It refers to the action of editing through revision, addition and correction of errors in the work done. Random articles have less revision count, and are considered to be the work of a less authors and thus should not be trusted due to low quality in reliability and accuracy. Random articles are also considered to be not neutral and hence are more biased to the direction of the author, level of experience, techniques and exposure in the society (Hasan&Andr´e Gon 2009, pp.131-133).

2. PREVIOUS WORK

2.1 ANALYSIS BY WORD COUNT

Word count is a much simpler method of measuring the quality of Wikipedia articles as compared to the use of complex quantitative methods. In this method, the length of the article is measured by calculating the number of words in it. As far as there are limitations to this metric, there are substantial reasons to prove that this method will be compared to quality. Figure 2.1 shows word count of randomized and featured articles. Due to Absence of complication by this cadent, they present the following advantages.

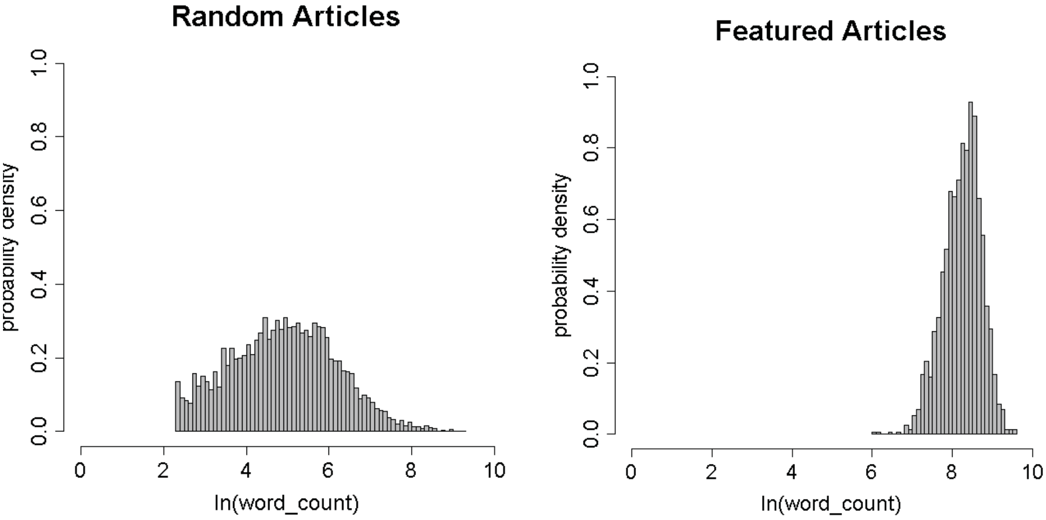
i. Measurement of the article length becomes easy

ii. Length of the article performs significantly much better than the other methods

iii. Most of the approaches require complex information for calculation e.g., history and revision text of article (Speigelhalter&Thoma 2005, pp.94-95).

iv. Other methods mostly operate in an old fashion which constitutes hidden results and parameters that are to be decoded by average Wikipedia visitors.

Figure 2.1: Distribution of word count for featured and random articles



An experiment was conducted to test the quality by article length to separate low and high quality articles by a procedure formulated by Zeng and Stvilia et al. Instead of comparing scalar measure of article quality against metric, it was assumed that random articles are of lower quality than the featured articles. The goal was to maximize precision and recall of non-featured and featured articles (Speigelhalter&Thoma 2005, pp.57-61). To make the conclusion 5,654,236 articles from the 7/28/2007 archives of English Wikipedia were extracted as shown in table 2.1 below.

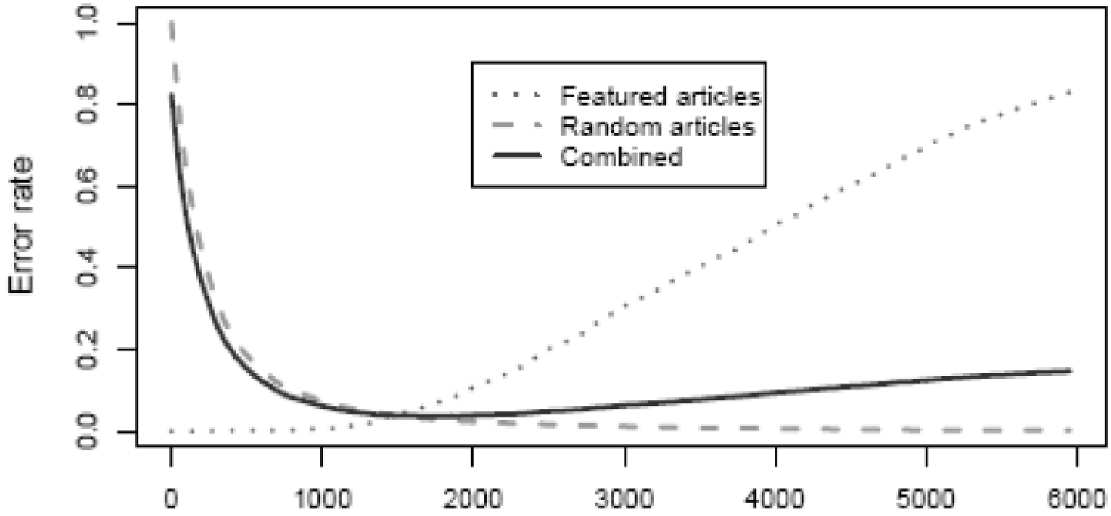
Table 2.1: Word count performance for random vs. featured articles.

Class	n	TP rate	FP rate	PRECISION	RECALL	F-measure
FEATURED	1554	0.936	0.023	0.871	0.936	0.902
RANDOM	9513	0.977	0.064	0.989	0.977	0.983

Specialized files (images and templates) and those articles that contained less than fifty words were removed after stripping the Wikipedia- related mark up. This resulted to cleaning of the data set which contained 1,554 featured articles. Further additional 9,513 cleaned articles which served as non-featured corpus were randomly selected. The total corpus added up to 11,067 articles. To further prove in another experiment 2/3(7378) articles for training were used and 1/3(3689) articles for testing, with a similar ratio of random: featured articles on each set.

The results showed that by classification of articles with more than 2000 words representing featured and those with less than 2000 representing random, 96.31 percent accuracy was achieved in binary in binary classification task. The results were achieved by minimizing the rate of error on the training set. The accuracy reported results from testing on the held out test set. More sophisticated classification techniques could lead to produce of the modest improvements. As example, a multi-layer perception with an overall accuracy of 97.15 percent was archived with an F- measure of .983 for random articles and .902 for measured articles. The *k*-nearest neighbor classifier replicated similar results of 96.94 percent accuracy and a log it model showed 96.74 percent accuracy. A random forest classifier showed 95.80 percent accuracy (Hasan&Andr´e Gon2009,p.118).

Figure 2.2: Error rate by word count



All this techniques shows that word count is a more reliable method of quality measurement over the more complex methods in Zeng et al and Stvilia et al which showed 86 percent and 84 percent accuracy respectively. Error rate can be seen in figure 2.2 above.

Word count matrices have proven to be very accurate which raised curiosity of whether the other simple increased classification accuracy. Features like readability metrics, part of speech tags and *n*-gram bag of words have proven to be moderately successful in other contexts. In Wikipedia quality context, however it was noted that word count was unbeatable. *N*-gram bag of words classification indicated 81percent accuracy for an example, and so did the *n*-1, 2,3 on both Bayesian and sym classifiers. A slightly higher accuracy of 96.46 percent was achieved with a threshold of 1,830 words (Hasan&Andr´e Gon 2009, pp.112-114).

Even with a kitchen sink algorithm with thirtyfeatures it was noted that the classifier could not achieve more than 97.99 percent accuracy. This means, calculating with word count is an improvement against the considerable effort required to build the classifies and produce this metrics (Stvilia&Twidale 2005, pp.13-15).

It has been proven that the article length is a good way of determining whether the article will be featured in Wikipedia. Word count has proven to be a simple method of metric, that is by far more accurate than the other complex methods as proposed in related works done previously. It also performs a well independent classification without parameters and a complex logarithm. We cannot exaggerate the efficiency of this metric by assuming that it features accurate measurement for quality because it is indicated that article length can also be used to determine the article quality. We can conclude that most of the featured articles are long and long articles are mostly featured (Adler&de Alfaro. 2007, pp.67-69).

2.2 QUALITATIVE WORK

Apart from the editorial guidelines in the Wikipedia.org, substantial qualitative work has developed with an aim of helping people to understand quality of Wikipedia particularly and the encyclopedia in general. For example according to Crawford (2001) he presented a thorough framework of assessing the quality of encyclopedia. Further Lih (2003) proposed metrics for online context. He also analyzed the correlation between unique authors of Wikipedia articles and the numbers of revisions.He also analysed the quality of these numbers. He proposed using unique editors and the total number of edits to measure quality of article and later in 2006 he suggested the use of color according to age in order to give visitors some indication of quality (John&Langley 1995, p.90).

2.3 QUANTITATIVE WORK

A more complex system for measuring quality of an article has been designed and developed by researchers. This system basically relies on machine learning techniques with an aim of calculating methods of measurement. Two steps are involved in the standard methodology (John&Langley 1995, pp.49-53). They include;

2.3.1 Feature Extraction

It involves presentation of each article as a combination of various quantifiable metrics. This metrics are called features and might include straight forward information like word count, syllable count, number of references, sentence count, number of links and linguistic information like number of noun phrases, ratio of verbs and adverbs: revision history count, number of unique editors and edit count.etc. (Hasan&Andr´e Gon 2009, pp. 17-19).

2.3.2 Evaluation

The quality predicted is measured against the objective standard of quality. Few studies like the most recent work of de Alfaro and Adler (2007) have included the use of human experts in judging the quality of the predictions. Use of featured articles as a approximations of quality is the most common approach. The algorithm will correctly put into place each article as a not featured or featured article, the accuracy is hence measured by dividing the number of classified articles with the number of correct classification. The main advantage of this method is that it is objective oriented and automatic. When an effective measure of quality is identified it can be applied with any article on Wikipedia. Following this methodology, Stvilia et al. tried to come with a system that would determine the quality of an article based on quality standards described by Crawford (2001). Seven factors were named by Crawford which were important to measuring quality; uniqueness, scope, format, authority, currency, accessibility and accuracy.

These factors then are multiplied by different weight multipliers, and it can be recieved a quality consistency variable, which gives us the quality predicted. The features that are used are the administraot edit share, article age, unique editor count, total edit count, connectivity, revert count, external link count, registered and anonymus edit counts. The admin edit share and the article age gives us the consistency of the article, and the other used fetures gives us the reputation of the article as well.

After computing these factors for each article he then ran a cluster analysis to determine whether each article was featured or not. 86 percent overall accuracy was achieved. Similarly Zeng et al (2006) formulated a method of trying to measure the "trust" of article based on their edit history. In this case the relevant calculations made for the number of deletions, number of revisions and the number of blocked authors who edited each article. In regard to these features he used a dynamic Bayesian network to create evolution of each article. He observed that he could classify featured articles at 84 percent accuracy.

2.4 CLASSIFICATION/QUALITY PREDICTION

Use of these algorithms predicted a quality of an article on the basis of its features. For example; predicting that ages of articles are the most important feature, it is okay to bealive that old articles are better in quality than new articles (John&Langley 1995, pp. 58-60).
Contrary to these complex methods elaborated above, Blumenstock, (2008) formulated how features with more than 97 percent accuracy can be identified. Its potential applications and results are discussed below.

2.5 STABILIZED ARTICLE

A stabilized article is the one that has more or less to do with the total knowledge of the subject matter of topic. This article is considered complete contentwise. Topics in stabilized articles mostly refer to notions, events, people etc. with no chance of changing over time. Changes that happen in this type of articles are mostly related to revision or maintenance such as those made by automatic bots for updating the articles categories and the reverts of random vandalized attack. It is expected that significant accuracy is paramount in stabilized articles content since they are supposed to be complete content wise and to the total topic knowledge.

Stabilized articles can be the articles, that are semi or fully locked. İn this case, the article will be non editable by unauthorized users, and also by blocked users. Although wikipedia lockes articles very rarely, this can be the case that the article is a stabilized article. Lock on the articles can be because of the vandalism attacks, and also because of the edit wars, and the rapidly change of the fast changing articles. In this case, an ongoing event, that has an article edited by many users in the same time, and this changes the stability of the article. When the event finishes, the edit attempts will decrease, and the article will be unlocked, so all users will be able to edit the article again.

Wikipedia's featured articles can serve as benchmarks of quality to model the stabilized articles quality. Some of the better written complete articles are featured in Wikipedia on a rotating basis. A policy of Wikipedia mandates that all the featured articles must be stable. Their content should not be subjected to on going edition wars or do not change significantly from day to day. For this reason stabilized articles aspire to be like the featured article essentially (Hasan&Andr´e Gon 2009, p. 79).

This model of quality uses articles features except for length can be handled like vital building blocks for an article. These are features like citations, images, paragraphs and references which are all essentials of a quality article. Nevertheless excessive use of these building blocks can over or under develop an article.

There are no efforts in determining best features of stabilized articles, because stabilized models are thought to be simple models, and they are parts of more complex article classification schemes. This leads us to choose features which appear more reasonable and simple to extract for stabilized articles.

Featured articles act as quality benchmarks. The model expects that if a stabilized Wikipedia article appears to have exactly same characteristic proportion to the featured article it is possible that it will affect the article length and quality very much. And the same if the articles characteristics differ with those of the featured article, it diminishes the influence on the article length. Samples of featured articles in wikipedia are required in this model. The sample is elaborated as a collection of different components of a mixture model. Six mixture models exist within this mixture model and they are acquired from the featured articles sample set. The components are mostly Gaussian probability density functions for computation of length, internal link density, image count density, citation density, internal link density and section count density. (Hasan&Andr´e Gon 2009, p. 210).

2.6 CONTROVERSIAL ARTICLE

Controversial articles are articles whose content is instable due to different opinions. The policy on Wikipedia editing requires neutral view narratives. Nevertheless editors at Wikipedia are human and prone to biasness that influence how they edit intentionally or unintentionally. Other editors on detection of such biasness may disagree with them making the article a subject of controversy. Some of the articles contain inherited controversy due to their subject content. This may include articles on religion or ancient cultures that are passed down generations ago. Some articles may go through the phase of controversy due to the attention, because they grab at specific times like eye raising current events (McGuinness&Bhaowal 2006, pp.19-21).

Most of the times, controversial articles are a weak target of sabotage and act as a combat zone for reverts events. Historically controversial articles could be identified by how large the number of vandals and revert wars occur, as well as anonymous contributions they attracted. Today we determine the quality of controversial articles by taking into consideration their revision history. The model used to determine the quality of controversial articles is very similar to that used to determine the quality of stabilized articles although it contains different article features (Hasan&Andr´e Gon 2009, pp. 95-97).

The following table is a representation of a controversial model.

Table 2.2: Registered and anonymus user model

Feature Name	**Description**
Avg. Number of Reverts	Average number of reverts in the article's revision history
Revisions Per Registered User	Average revisions per registered authors
Revisions Per Anonymous User	Average revisions per anonymous authors
Percentage of Anonymous Users	Percentage of anonymous authors

2.7 CATEGORIZING ARTICLES

Before applying a quality model for either controversial or stabilized articles to a specific Wikipedia article, it is important to determine first, if the article is controversial or stabilized or it belongs to another different category. This is achieved by using supervised learning techniques of classification. A classifier is developed and trained for specific article category. Finding the category of a Wikipedia article involves a two-step process. First; features of an article are extracted and ran against a battery of classifiers (McGuinness&Bhaowal 2006, pp.49-51).

When the target article is positively classified in the classifier, a quality model that corresponds to that classifier is applied to the article. In case of a target article is classified as positive by more than one classifier, the average of outputs of each applied quality model is considered as the final score of the targeted article. Lastly if the target article is not shown positive classification by any of the classifiers in the series, the stabilized model of that article is applied as the final score. Note that each classifier was qualified from Wikipedia articles dataset which was manually chosen to include a mixture of article type described earlier. The class labels for these data sets were assigned manually. The (SMO) sequential minimal optimization learning algorithm used for training vector machines classifiers was chosen (Stvilia&Smith 2005, pp.56-58).

2.8 EVALUATION OF REVISION HISTORY

The trust values of articles fragments' are used to determine the trust value of the article. It is shown in the previous experiments thatmodels produced strong results on the worthiness of the articles. This also indicates a good performance of the model at the fragment level. (Stvilia&Gasser 2005, pp.111-113).

Many measures have been taken to address challenges of trust. E.g. privileges of many authors to create new articles were recently increased and in the resent past a new feature called article validation is being processed which will enable users to rate an article openly via a restricted form. In addition, Lih formulated a set of metrics to evaluate the Wikipedia articles quality among other factors like number of revisions.

Vi'egas et al presented a tool that visualized revision flow and at the same time which revealed various interesting patterns in Wikipedia. For example it was noticed that half of the mass deletions were being reverted within two minutes.

Theories of trust computation have also been widely studied. For example Kamvar et al introduced a reputation system that helped minimize the effect of malicious peers in peer to peer networks. Propagation of trust and was discussed by Guha in social networks like ePinions.com. All these approaches are targeted on transitivity property of trust. That is if A trusts B and B does the same to C, then it would be automation that A trusts C to a certain level. This model can be improved by development of author trust models that can model complicated author behaviors like letting a blocked author in some cases make trustworthy contributions.

2.9 FRAGMENTED TRUST

2.9.1 Fragment Formulation And Identification

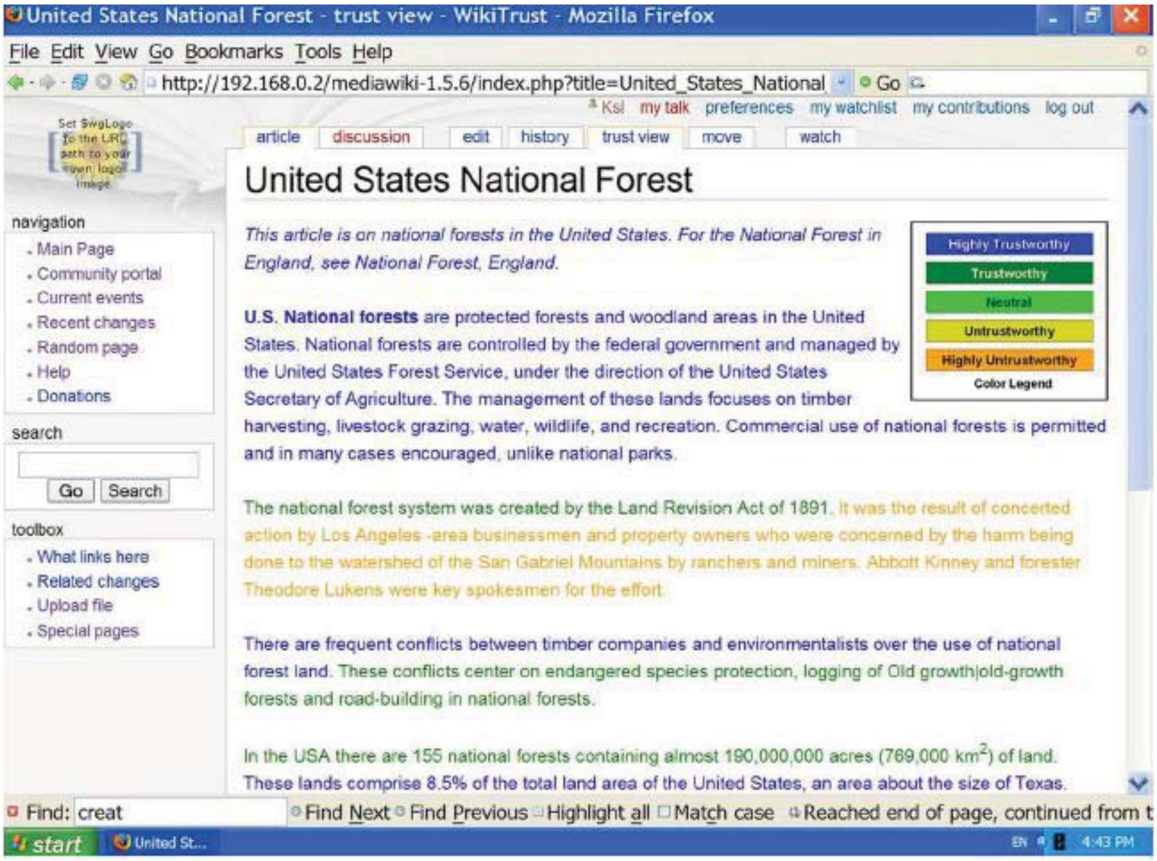
A fragment in an article that is considered to be a compilation of various texts in an article which has been contributed by an author in one of his revisions. If by any chance the author had revised the article severally, different fragments would be formed by each of his revisions. The texts in the fragments are typically continuously located but not for necessarily computation reasons that will be discussed later in the section. Since individual fragments are not stored in most wikis, many interpretations are open for fragment formulation (Zeng&Alhossaini 2006, p.2).

In this research, a revision on an article refers to the authors' action of editing the article. It is considered that a revision is a combination of add and deletion operations. In these operations, user removes a particular content from the article and at the same time zero or more insertion operations have occured, which brings additional increasment of the quality to an article. When a revision on an article is done, a newer version of the article is developed to replace the content that has been revised. Therefore, the *i'th* version of the article is the *i* article after the first revision. The original article is defined to be the 0^{th} version of the article. The revision history of an article is the series of its versions structured by their formation time. This method has an accuracy showed at table 2.3 below.

Table 2.3:Fragmanted model success

	Featured Articles	Clean up articles
Fragment trust model	91%	84%
Article trust model	82%	84%

Figure 2.3: Fragmanted trust with colors


United States National Forest - trust view - WikiTrust - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://192.168.0.2/mediawiki-1.5.6/index.php?title=United_States_National
Go
Ksl my talk preferences my watchlist my contributions log out
Set $wgLogo to the URL path to your own logo image.
article discussion edit history trust view move watch
navigation
Main Page
Community portal
Current events
Recent changes
Random page
Help
Donations
search
Go Search
toolbox
What links here
Related changes
Upload file
Special pages
Highly Trustworthy
Trustworthy
Neutral
Untrustworthy
Highly Untrustworthy
Color Legend
Find: creat
Find Next Find Previous Highlight all □ Match case Reached end of page, continued from t
start United St...
EN 4:43 PM

United States National Forest

This article is on national forests in the United States. For the National Forest in England, see National Forest, England.

U.S. National forests are protected forests and woodland areas in the United States. National forests are controlled by the federal government and managed by the United States Forest Service, under the direction of the United States Secretary of Agriculture. The management of these lands focuses on timber harvesting, livestock grazing, water, wildlife, and recreation. Commercial use of national forests is permitted and in many cases encouraged, unlike national parks.

The national forest system was created by the Land Revision Act of 1891. It was the result of concerted action by Los Angeles -area businessmen and property owners who were concerned by the harm being done to the watershed of the San Gabriel Mountains by ranchers and miners. Abbott Kinney and forester Theodore Lukens were key spokesmen for the effort.

There are frequent conflicts between timber companies and environmentalists over the use of national forest land. These conflicts center on endangered species protection, logging of Old growth|old-growth forests and road-building in national forests.

In the USA there are 155 national forests containing almost 190,000,000 acres (769,000 km^2) of land. These lands comprise 8.5% of the total land area of the United States, an area about the size of Texas.

Fragments of the articles in Wikipedia suggest that user views sentences, which are displayed in different colors based on the trustworthiness, which is shown when a visitor clicks the generated trust view tab. (Figure 2.3) Fragments that have higher trustworthiness are displayed in a vibrant color than fragments with lower trustworthiness to help the users to have an insight on relative trust just by looking at the tab presentation of the article although issues like intuitive mapping from use of color to trustworthiness are still being investigated. The revision trust has a lot of benefits far beyond the trust in Wikipedia. Many applications can be built to fully utilize the trust information that is available. The users may have an option of viewing the most trustworthy versions of an article as well as the most recentone. In addition to this the model,it can be provided an automated method of monitoring changes in trustworthiness therefore providing timely notifications of malicious content modification and vandalism (Stvilia&Gasser2005, pp.456-458).

1.Constitutes of a fragment:

Formulation of a fragment can be done at word level, a paragraph level, sentence level or a word level. At the word level, basically a fragment consists of words collection. If one or more words are modified by an author, a newer fragment comprising of just modified words is developed in the article that is revised. All the other words are retained in the original article fragment. Therefore fragment contribution is limited the number of words that he modifies.

At the sentence level, one or more sentences consists one fragment. If one or more words are modified in a fragment by an author, the whole sentence that contains these words is considered as modified. Therefore the old sentences are removed from the original section and a new fragment is formed by the newly inserted sentences (Hoist&Miksch 2007,p.15).

At the paragraph level, one or more paragraphs make up on fragment. Even when one word or a sentence is modified in a fragment, the whole paragraph that contains that sentence or word is considered to be modified. Paragraph ownership changes from the author of the original fragment to the revision author. Consequently, the old paragraph is erased from the fragment and the new paragraph is inserted an it forms the new fragment. Naturally the formulation of the fragment should depend upon the revision context. e.g. If an obvious spelling error in a fragment is corrected by an author, then in this case the word level fragment is more suitable. However if a sentence containing critical assumptions is removed from a fragment by the author, then the more applicable fragment in that case is the paragraph level. A formulation of a sentence level fragment is chosen and it is assumed that the semantics of a revision can be interpreted. Mostly using a sentence level fragment is an average of word and paragraph levels. In addition, the decisionis based on the consideration of practical implementation. The article fragments may be too fractured for modeling in the word level, while in most cases comparison of articles based on a paragraph may not be so helpful (Giles G 2005, p. 438).

2. Identification of sentence boundaries:

The database of Wikipedia stores articles in raw texts instead of individual sentences. Hence in order to enhance identification of fragments at the sentence level easily, the raw text needto be divided into sentences. Identification of sentence boundaries problem is solved using the (Natural Language Processing) NLP techniques; like Ratnaparkhi and Reynar which present the entropy approach maximally. The same the model that we use can tolerate inaccuracies by dividing sentence; therefore sentence final punctuation marks can be used (e,g., ? ".") for text division and manually setting the decided rules foe handling exceptional cases (Hunt&McIlroy 1975,p.26).

2.10USER COUNT ON FEATURED ARTICLES

As a way of enhancing the quality trustworthiness, we have developed a quality finding system using the revisions user status. To do that, we have fetched all revisions of our data with user id request. Anonymus users have an id of zero. Assuming that authorized users are writing with more quality than the anonymus ones, we compared the user count at the random articles with the featured articles. As we thought, the results show that registered users are writing qualitier content and are assigned higher reputation values. But a significant number of anonymous users also contribute high-quality content. So as we looked, featured articles show more user edits than the random articles.

Table 2.4: User Count and anonymus edit count at featured and random articles

Edit Counts	Featured Article	Random Article
User	3085	1532
Anonymus	2217	2579

As it can be seen on table 2.3, edits made by users at featured articles are clearly higher than the edits made by anonymus viewers. After getting all the values, we decided that user count at featured articles can be used to calculate the quality of the article. And so we knew that user status on wikipedia mattered on the article quality. Higher standards can be achieved through user edits. But only using user count is both not enough, and not efficient. So, we wanted to know when these edits are made, and needed to analyse the user edit times. Thus, we tried to get edit times.

3. DATA AND METHOD

3.1 DATASET

By using our survival analysis and growth modeling insights, we have established a procedure for collection for observations from Wikipedia article quality. We knew all the featured articles, which can be listed at Wikipedia featured article list and are easy to select from, because of their limited count. But when it came to select the random articles, we needed to make sure that our algorithm would pass most of the featured article criterias with ease. So we had to know that the article is satisfactionary with both the word count and the revision count. We took random articles with word count more than 2000 to eliminate all the wrong possibilities, which eliminated almost 97 percent of all random articles. The wiki quality coding, that is before we established the exact amount of revisions would take coding a Wikipedia article on average, we decided to use four major measurement occasions for revisions, which arerevision 100, revision 300, revision 500 as well as revision 1000. We decided to choose revision 100 and 300 to help us in capturing two points as early as possible in the lifestyle of the wiki article, based on the knowledge that a substantial number of edits of Wikipedia articles happen early in the lifestyle of the wiki and by a minimum featured article revision count. We proved that our article lifetime was convenient at 500 revisions which was close to a meaningful marker. Revision 1000 was chosen in order to be used for capturing the bump activities that had been found in Wikipedia articles that had survived for, and that is a maximum number featured articles reached so far. In conclusion, a third measure was added at revision 500 for capturing the quality of wiki article for the revision counts. We decided to choose a closer date to the 300 revisions mark rather than at the maximum which is 1000 revisions because it is known from the featured article that the quality of a wiki decreases during that period of time, and at the 500th revision it would be possible that we will still be measuring all bumps at an articles quality.

Following some of our article coding pilot studies, we concluded that adding more measurement occasions would be unjustifiably time consuming or expensive. During the time of this decision making, we were worried that the sparsely distributed data of the high values of time would initiate difficulty when trying to fix the models of developing article quality with polynomial time specifications. For this reason we have decided to add only article lifetime to the measurements, at approximately average of the featured articles, as the measurement additional coactions. It was also assumed that the quality of article's lifetime would continue to develop and it was relevant to have sufficient data in the whole lifetime of the article in order to model possible complexities in the trajectories of quality growth. After analyzing our complete set of article quality measurements that were in the first sample, it became clear that after calculating revision count, article length and article age, most of the criterias for the featured articles will be fulfilled.

Our dataset rows don't contain any words, or revision differences. So it is very fast to fetch this data. After fetching it, data is also needed to be stored. Storing all the revisions of many articles keep also a lot of space in the memory, especially when working on articles which have a minimum revision count of 500. Also calculating the revision differences is something hard and slow to process. In order to create an algorithm both fast and not memory consuming, we did not get any text from revisions. This both fastened our calculation, and increased our download speed of the articles.

After minding all these, we have taken a dataset of 300 featured articles and 250 random articles with 500 or more revisions, and a word count of 5000 or more. Thus, we fulfilled Wikipedia criterias for promoting articles to featured status for all of the random articles, too. To get all this data, we used Wikipediaapi with a limit time of 15.03.2012 and before.

3.2 FETCHING DATA

In order to fetch data from Wikipedia, we have tried various ways. First we tried to download datasets created by Wikipedia. These backups are easy to get, which is by only downloading a dataset file, and easy to handle. But these backup files failed to give us information about featured articles, and to fetch the detailed data of the articles, we needed to download the full backup file, which includes all the text of the revisions, which is way too big like terabytes of data.

There are also other ways to fetch Wikipedia data. One of them is Wikipedia api. This api can be found at http://en.wikipedia.org/w/api.php, and users can fetch any required data on will. The good thing about this api is that, user can fetch only required data, instead of downloading all data. This api is also easy to learn, and can return values with XML or JSON format. To receive only the required information, we used Wikipedia api to fetch data.

We have created a program in order to fetch data automatically. After receiving all featured article list from the Wikipedia featured article list, which can be found at http://en.wikipedia.org/wiki/Wikipedia:Featured_articles page, we have created a string array to hold all these featured article names. For each item in this array, we have triggered the Wikipedia api with these variables:

i. Format = We have taken this variable as json, this variable can be json or xml.

ii. Action=This parameter can be one of the various actions available like login, logout, review, etc. While we are fetching data, we have taken this variable as "query"

iii. Titles=This parameter is filled with article name.

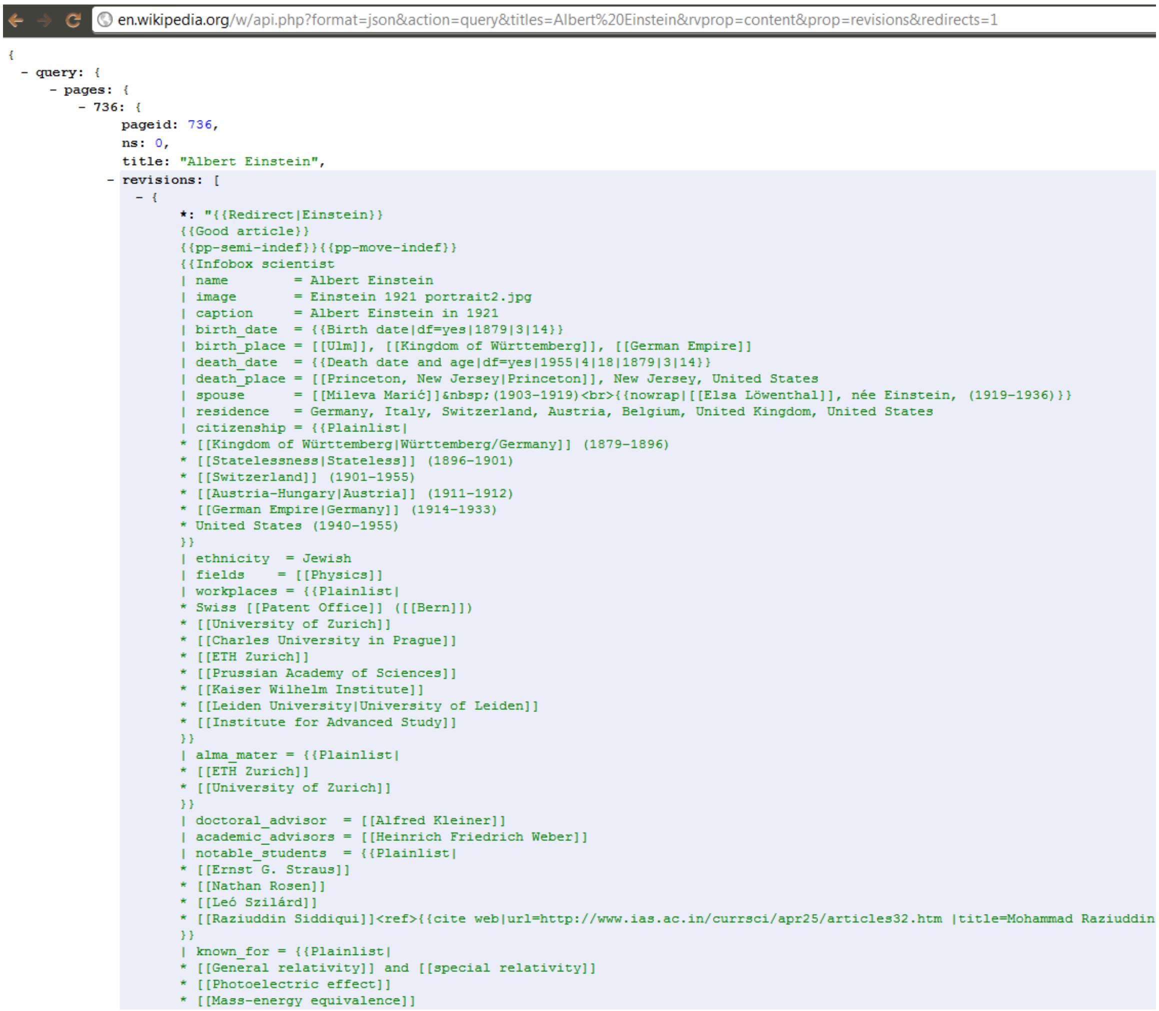
iv. Rvprop=This parameter contains the variables we want as return values. We needed to get various contents, so we have triggered the api with various parameters like content, userid, timestamp etc.

v. Prop = After determining the rvprop parameter, api requests for detailed info about the request it shoud return. This variable can be revisions, pageids, titles etc.

vi. RvStart = This parameter gives us the start revision of the return values list.

Triggering these values gives us a json result which can be seen at figure 3.1

Figure 3.1:JSON result of the api call



en.wikipedia.org/w/api.php?format=json&action=query&titles=Albert%20Einstein&rvprop=content&prop=revisions&redirects=1

```
{
  - query: {
     - pages: {
        - 736: {
              pageid: 736,
              ns: 0,
              title: "Albert Einstein",
           - revisions: [
              - {
                    *: "{{Redirect|Einstein}}
                    {{Good article}}
                    {{pp-semi-indef}}{{pp-move-indef}}
                    {{Infobox scientist
                    | name        = Albert Einstein
                    | image       = Einstein 1921 portrait2.jpg
                    | caption     = Albert Einstein in 1921
                    | birth_date  = {{Birth date|df=yes|1879|3|14}}
                    | birth_place = [[Ulm]], [[Kingdom of Württemberg]], [[German Empire]]
                    | death_date  = {{Death date and age|df=yes|1955|4|18|1879|3|14}}
                    | death_place = [[Princeton, New Jersey|Princeton]], New Jersey, United States
                    | spouse      = [[Mileva Marić]]&nbsp;(1903–1919)<br>{{nowrap|[[Elsa Löwenthal]], née Einstein, (1919–1936)}}
                    | residence   = Germany, Italy, Switzerland, Austria, Belgium, United Kingdom, United States
                    | citizenship = {{Plainlist|
                    * [[Kingdom of Württemberg|Württemberg/Germany]] (1879–1896)
                    * [[Statelessness|Stateless]] (1896–1901)
                    * [[Switzerland]] (1901–1955)
                    * [[Austria-Hungary|Austria]] (1911–1912)
                    * [[German Empire|Germany]] (1914–1933)
                    * United States (1940–1955)
                    }}
                    | ethnicity  = Jewish
                    | fields     = [[Physics]]
                    | workplaces = {{Plainlist|
                    * Swiss [[Patent Office]] ([[Bern]])
                    * [[University of Zurich]]
                    * [[Charles University in Prague]]
                    * [[ETH Zurich]]
                    * [[Prussian Academy of Sciences]]
                    * [[Kaiser Wilhelm Institute]]
                    * [[Leiden University|University of Leiden]]
                    * [[Institute for Advanced Study]]
                    }}
                    | alma_mater = {{Plainlist|
                    * [[ETH Zurich]]
                    * [[University of Zurich]]
                    }}
                    | doctoral_advisor  = [[Alfred Kleiner]]
                    | academic_advisors = [[Heinrich Friedrich Weber]]
                    | notable_students  = {{Plainlist|
                    * [[Ernst G. Straus]]
                    * [[Nathan Rosen]]
                    * [[Leó Szilárd]]
                    * [[Raziuddin Siddiqui]]<ref>{{cite web|url=http://www.ias.ac.in/currsci/apr25/articles32.htm |title=Mohammad Raziuddin
                    }}
                    | known_for = {{Plainlist|
                    * [[General relativity]] and [[special relativity]]
                    * [[Photoelectric effect]]
                    * [[Mass-energy equivalence]]
```

After determining all these values, we have received a fetured article fetching api link as "http://en.wikipedia.org/w/api.php? format = json & action = query & titles={0} & rvprop = content & prop = revisions & redirects = 1". After getting this link, we have replaced the {0} value with the featured article title, so we get the last featured article revision, the word count of that revision and the page title.

To fetch the random articles, we have triggered the list property of the api, so we received an api link as " http : / / en . Wikipedia . org / w / api . php ? format = json & action = query & list = random & rnlimit = 10", where rnlimit gives us the count of the random articles per request.

We tried various algorithms on the dataset. These algorithms are mentioned about at the previous work section. For these algorithms, we needed various properties like revision edit count, user count, anonymus count, etc. To get all these data, we used "http://en.Wikipedia .org/w/api.php?format=json&action=query&prop=revisions&titles={0}&rvprop=ids|user|userid|content&rvlimit=300&rvstart= {1}" api link, which has the rvstart variable, where we specify the start of the list. As it can be seen, rvprop variable requests for page id, content text, user name and user id. Because we fetch all the text data of the article on all revisions, this query is very slow, and very costly. Fetching this data only once returns 300 revisions, but it requires more than 2 minutes to get. When speaking of thousands of revisions at only one article, this is not very time effective, and hard to fetch.

For the algorithm we used requires only the timestamp of the revision. The api link is as follows:"http://en.Wikipedia .org/w/api.php?format=json&action=query&prop=revisions&titles={0}&rvprop=timestamp&rvlimit=500&rvstart = {1}" where {0} is replaced with the title and {1} is replaced with the last revision id fetched. While we are not requesting for any text or complicated data, this query is replied with less than a second. This both reduces the calculation time of the algorithm, and reduces the fetching time of the query.

3.3 FEATURE EXTRACTION

Feature extraction of the process sets of new features are extracted from original features by using some functional mapping. Feature extraction is used for reduction of feature space contrary to feature construction and feature transformation which expands the feature space. Construction and selection are the new methods of feature extraction. Activations in hidden units are interpreted as new features which have been extracted from the dataset that was original. The use of feature transformation as well as subsequent selection depends upon the intended purpose that is for better classification or for simple concept description. All this aims at preservation of the tropical structure of data and at the same time the latter targets which enhance predictive power.

Feature construction is the process of discovering information that is missing on the relationship between augments and features through creation of additional features or by inferring (Michalski&Wnek 1994, pp.139-168).

Feature/ subset selection is the process by which a subset of original features is selected causing the reduction of the feature space. Subset selection simplifies language in the cases which the language is insufficient for problem description. The algorithm for feature selection can be classified as a wrapper or a filter, in dependence on whether it's treated as intertwined or a preprocess with the task of learning. Generally the wrapper approach out performs the filter due its direct application which optimizes the evaluation measure of the task of learning while removing features. Time needed for selection of features is longer than in filter approach.

Feature selection can be defined as the process of finding a minimum subset which satisfy a decisive factor be it wrapper or a filter. The decisive factor can be an error rate,information measure, inconsistency rate, dependence measure or distance measure. The notion of relevance can be can be characterized in a framework that's has a foundation of mathematics using two axioms; the necessity and the sufficiency axiom. Both of them are shown to be equal to maximization of relevance, which is of a subset feature to the class and class relevance to the feature subset. (The relevance degree is measured by positive pairs/negative pairs). This can be used in the identification of irrelevant features using set inclusion relation.

In subset selection, searching plays a very important role. It can be characterized by direction (random, backward, bidirectional and forward), search strategy (exhaustive/ complete, non-deterministic and heuristic), and finally the evaluation measure (classic, consistency and accurate) according to Liu &Motoda 1998. A new way of changing the topology of search space is by creation of dynamic operators which connects directly a node to other nodes in consideration.

Other feature extraction approaches includes fractural encoding, use of mutual information (Petry&Perrin 1998, pp.157-173), binary features conversion to continuous ones (Njiwoua&Nguifo 1998, pp.205-218) and use of wavelet. Fractal encoding is invariant in respect to the objects size. Wavelets are used in extraction of important local features in spectrum of high dimensionality. The numerical taxonomy uses numerical value similarity measure and implicitly measure of distance which assumes that objects representation can be natural in terms of variables which have been continuously valued.

Instead of individually considering feature selection, construction and extraction, researches have seen the importance of both subset selection and feature transformation. The reduction of dimensionality is mostly used in pattern recognition in statistics (Wyse 1980,pp.415-425) and databases. Implicitly, the switching circuit design of electrical engineering addresses the feature transformation issue. By addition of other field approaches of similar problems, it is possible to expand the repertoire of data mining algorithms through equipping them with subset selection and feature transformational tools.

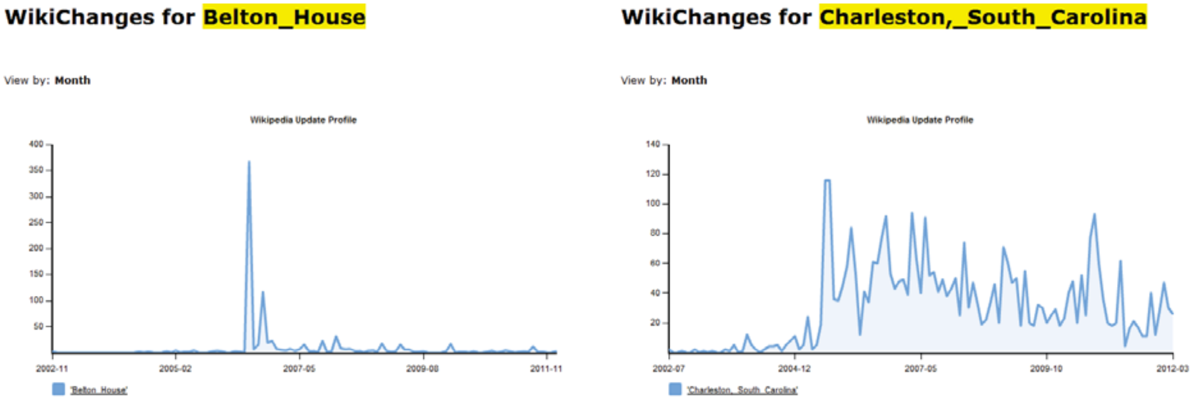
A lot of effort has been directed towards improvements of performance such as estimation of classification accuracy. In the data mining case, however, attention need to be paid to issues like the comprehensibility of the newly constructed of extracted features. People would be interested only in knowing whether or not the information contained in the data is valuable and if the information is within the discovered features and rules.

Complete fragmentary knowledge can contribute greatly to subset selection and feature transformation. To maximize the use of knowledge, conflicts arising from usage of knowledge from various sources need to be handled and to balance domain independence bias and domain specific knowledge.

3.4 BUMPS AT REVISION COUNTS

As we investigated further at the revision times of the articles, we saw some shapes at the featured articles, which rarely exist on random articles. Featured articles mostly get their revisions in a group of months. They recieve extra attention from an user or a group of users, and only for a couple of months they recieve extra attention, and get a revision count above average. So when we look at the revision count of featured articles by monthly interval, we see a bump, an extra ordinary revision count compared to the other months. At the random articles, we saw an ordinary monthly revision count with an average edit count instead as it can be seen at figure 3.2.

Figure3.2: Featured and random revision count by month



As it can be seen at figure 2, featured articles mostly have a “bump”, an anomaly at one to four months, that have more revision count than average edit count for the article. At the random articles there are either multiple bumps, or none.

After recieving all the data we requested, we still need to extract some features to spot the bumps automatically. To extract these features, we first grouped article revisions by month. After grouping, we recieved group counts, and calculated with these values.
Also there are multiple features at the bump, which classifies its shape and style. We also needed to differ it from the shapes at the random articles.

3.3.1 Bump Features

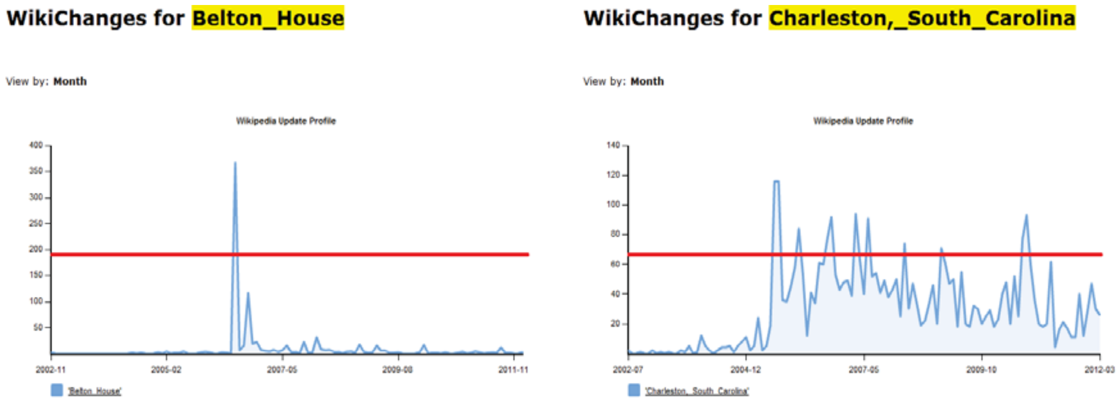
There are multiple bump features, which we needed to calculate in order to find the bumps, and those no other shapes have. Bumps have a very large number of revisions at a month, especially when compared with the other months. So we needed a threshold level, and data above that threshold shall be taken as a bump. We calculated this threshold with various different ways.

In order to find bumps, we used these features:

i. Revision month: This feature gives us how much the article age in months is.
ii. Average value: The monthly average revision count is given at this feature.
iii. Max revision: The maximum revision count in a month.
iv. More than double average: month count, that exceed the double average value.
v. More than triple average: month count, that exceed the triple average value.
vi. More than quatro average: month count, that exceed the quatro average value.
vii. Bump count by average: Gives us the bump count of the monthly revisions by double average threshold.
viii. Bump width by average: Gives us the bump time span in months by average threshold.
ix. Bump month by average: Gives us when the bump occurs by average threshold.
x. Bump count by maximum: Gives us the bump count of the monthly revisions by maximum/2 threshold.
xi. Bump width by maximum: Gives us the bump time span in months by maximum/2 threshold.
xii. Bump month by maximum: Gives us when the bump occurs by maximum/2 threshold.
xiii. Word count: The word count in the article.

The first way we used is getting the threshold by maximum value. We took the maximum revision count of the months, and took it as the bump.(Figure 3.3) Threshold level is the half of this value. This means that all the values above it will be calculated as bumps.

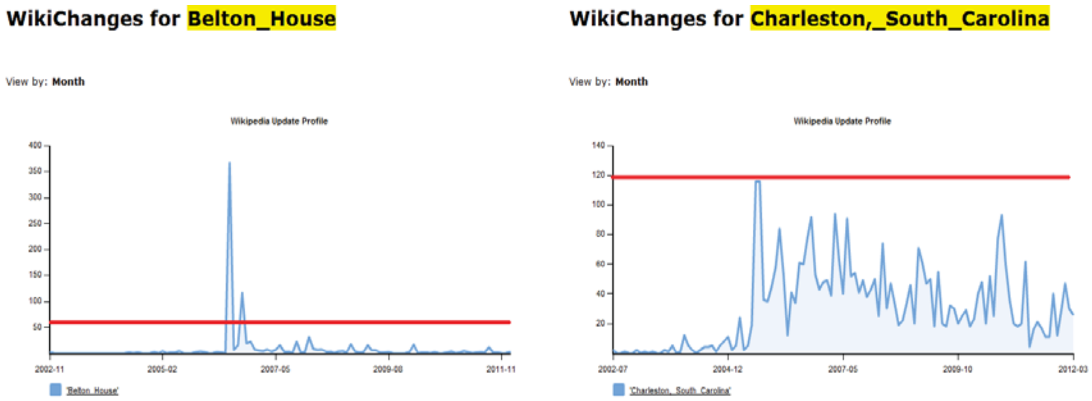
Figure 3.3: Bump through maximum value



At this threshold finding type, we have recieved multiple bumps at the random articles, instead of none. But it gave us a great knowledge of anomalities. If the bump count with this threshold value recieves a high number of results, we can say that, the revision count at the bump is not so high, and it should not be treated as a featured article. Also, this means that the bump width at the featured articles should be a lower value than the random articles.

The second way we used to calculate the bumps is through average.(Figure 3.4) We took the average of the revision counts, and multiplied it by two. All the values above this threshold will be taken as bumps.

Figure 3.4: Bump through average value



After finding the bump threshold, we can calculate further features like count, width and average of the bumps. We have calculated all these values for the both thresholds above. So we achived two of these values all.

Bump count is an important variable when winding featured articles. Since featured articles can have one to four bumps, we can say that, the bump counts are very important when finding featured articles. We calculate bump count by counting the values above threshold.

Bump width is the width of the bump, which specifies the time the revisions proceeded. This is important, because it differs the bump from an edit war. Random articles have commonly have bumps with high bump widths, commonly because of the arguments between groups or people. Bump width is calculated by the consecutive bumps . If two or more bumps are consecutive, they will be counted as only one bump with more width.

Bump average is the number of revisions in a bump, which specifies us how high the bump gets. İf a bump average is way to higher than the average, it means that the bump is very high.

To calculate all these values we took the articles one by one. For every article, we have calculated these values individually. First we have taken every revision of the article, and grouped the revisions for month. After accomplishing that, we recieved monthly revision counts for every article.

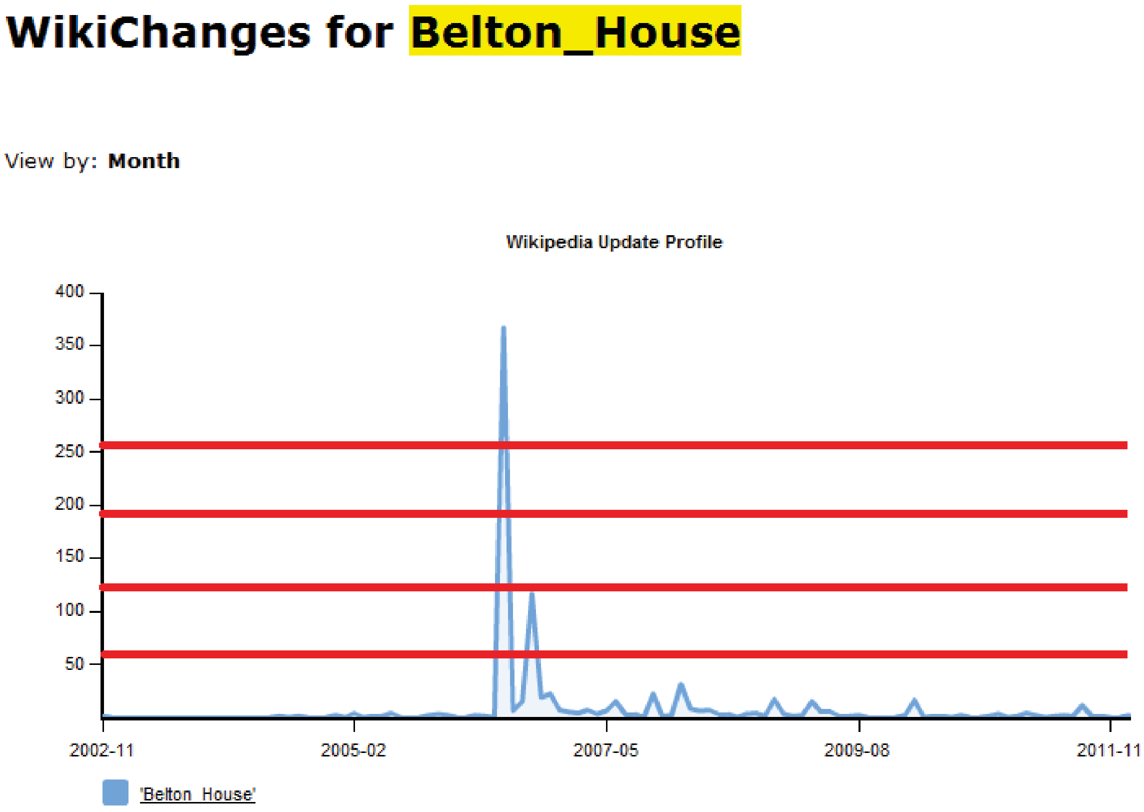
Since most of the articles have different lifetimes, the month count for the articles are different. So we needed to normalise these month counts. In order to do that, we have taken the maximum month count, and assumed that, every article has a revision count of zero from the last revision until that time. So data normalisation has been clarified.

After recieving monthly revision counts, we have found the maximum value, and the index of that value. We took half of this value as threshold. Then we looked at every item at the monthly revision counts, and found every item that is larger than this threshold. For every item, we increased the bump count by one.

After finding all these bump values, we looked at their index values. If the index values are sequential, there is only one bump, and so we reduced the bump count by one, and increased the bump width by one.

Except by these values, we have calculated other values to find featured articles by the revision count. To find the bumps, we found the average of the revision counts, and multiplied it by two, three and four. So we have found bumps higher than double, triple and quatro average. The detailed visualisation can be seen at figure 3.5.

Figure 3.5: Bump through double, triple and quatro average.



WikiChanges for Belton_House

View by: **Month**

This gives us an important knowledge of the bump counts by height. Random articles, that have more revision counts than double average, mostly have less revision count than triple average. And featured articles have mostly higher bumps than triple average, sometimes higher than quatro average.

4.RESULTS

We have analysed all these featured with various algorithms like J48 and SimpleLogistics. After eliminating all the word cuont and revision count criterias, and having the only remaining 4 percent of the random articles, with these algorithms we have still recieved an success rate above 80 percent.

Decision trees are among the most commonly used machine learning algorithms. They are responsible for performing general specific searches of feature space as well as addition of the most informative features to a tree structures as the search progresses. Specific features that are selected during the process of the search are represented by a node in the decision tree which has been learned. Disambiguation of test instances is by addition of a path through the learned decision tree from the root to a leaf node that corresponds with the observed features. The majority classifier assigns the most common sense in the training data to every example in the training data.

When we run this algorithm with SimpleLogistics, we have recieved a model, which uses only five of our features. It uses bump width, bump count by maximum value, bump count by average, more than double average and more than quatro average. The model is given below.

1.23 +[Revision_Month] * -0.01 +
[Word_Count] * 0.003 +
[Average] * 0 +
[More_than_Double_Average] * -0.01 +
[More_than_Triple_Average] * 0.07 +
[More_than_Quatro_Average] * 0.15 +
[BumpCountByAverage] * -0.17 +
[BumpWidthByAverage] * -0.05 +
[BumpAverageMonth] * 0.01 +
[BumpMaxCount] * -0.2 +
[BumpWidthByMaximum] * -0.27

We can clearly see that the more bumps the graph has, the more likely it is a random article. Also it can be seen that the bump height, we can also say that it is the maximum revision count should be more than the quatro average of all the revisions.

When we look at the results, this algorithm gives us a correct classification percentage of 84.6. And if we look at the confusion matrix, we can see that it recieves more errors with random articles.

Confusion Matrix gives us results as:

Table 4.1: Simple Logistics Confusion matrix

	FEATURED CLASSIFIED	RANDOM CLASSIFIED
FEATURED ARTICLE	271	30
RANDOM ARTICLE	56	200

These classifications are a result of values given in table 4.2 below.

Table 4.2: Simple Logistics results

Correctly Classified Instances	471 (84.56%)
Incorrectly Classified Instances	86 (15.44 %)
Kappa statistic	0.6868
Mean absolute error	0.2419
Root mean squared error	0.341
Relative absolute error	48.6996 %
Root relative squared error	68.4253 %
Total Number of Instances	557

At J48 we have run 2 different types of training sets. One of them is cross validation, and the other one is with a training set of data. At cross validation we recieved following results

Table 4.3: J48 cross validation results

Correctly Classified Instances	454 (81.508 %)
Incorrectly Classified Instances	103(18.4919 %)
Kappa statistic	0.6261
Mean absolute error	0.2061
Root mean squared error	0.4073
Relative absolute error	41.4929 %
Root relative squared error	81.7324 %
Total Number of Instances	557

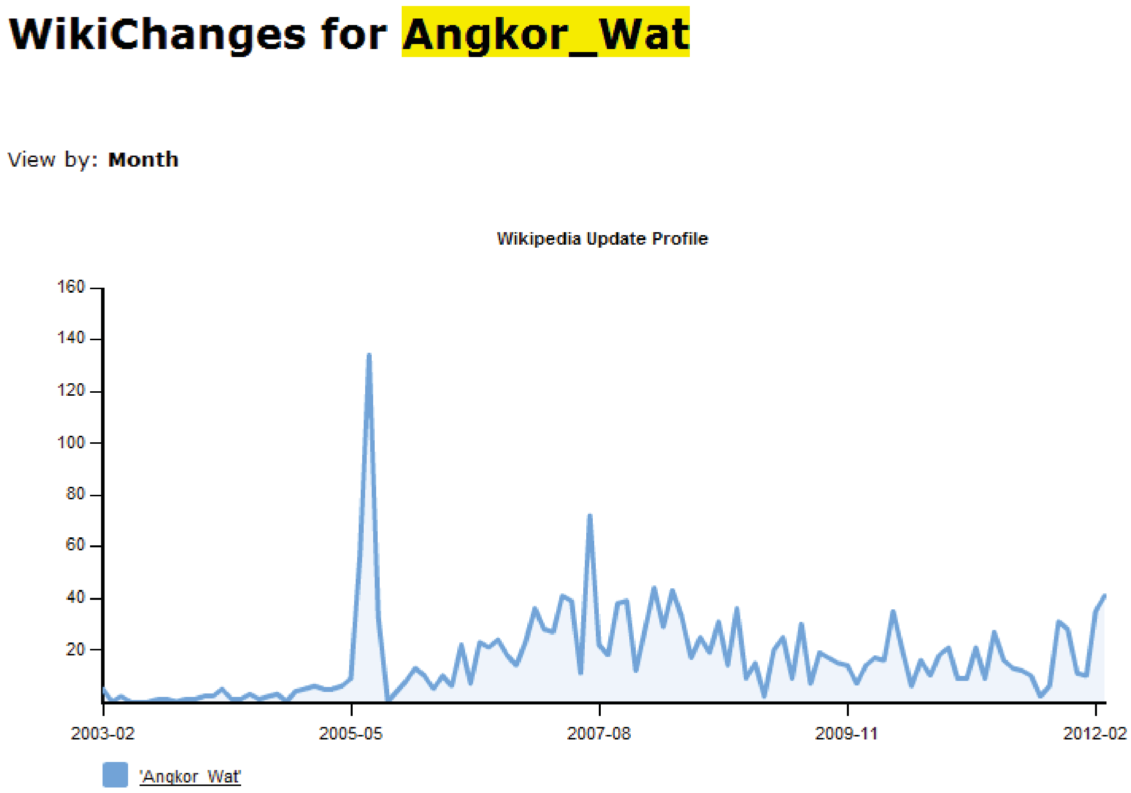
With this algorithm we can see at table 4.3 that we have reached an success rate of 82.4 percent. We have also tried some other algorithms like All results with Precision, Recall and F-Score values tells us, that this algorithm has an average of 80 to 90 percent of success.(Table 4.4)

Table 4.4:Detailed resultsfor various algorithms

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area
SMO	0.904	0.207	0.837	0.904	0.869	0.848
Simple Logistic	0.9	0.219	0.829	0.9	0.863	0.905
J48	0.854	0.23	0.813	0.854	0.833	0.808
M. Perceptron	0.904	0.148	0.877	0.904	0.89	0.921

When we look at the failed results, we can see that, some of the featured articles and some of the random articles do not fit very well at these bump formation. Although it is less than 20 percent of our simplified dataset, we can still see a difference at the graphs. At figure 4.1 can be seen that, at the featured article “angkor wat”, there are many revisions except the bump, and these edits are effecting the algorithm.By only looking at this graph, we can both say that, this graph has a bump, therefore it is a featured article, also there is an ongoing edit war or edits required on the text of this article, and it can’t be selected as a featured article.

Figure4.1:A featured article classified as random

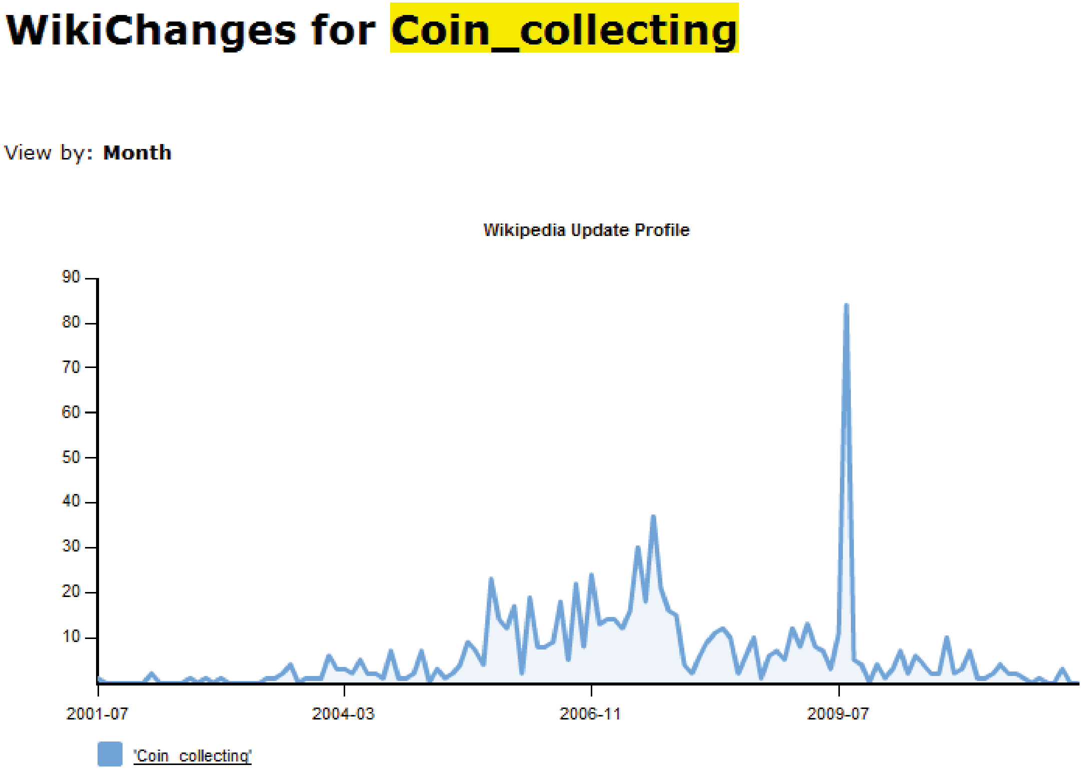


WikiChanges for Angkor_Wat

View by: Month

Also some of the random articles can mislead us if we only look at their revision graph. As example, at figure 4.2 we can see that a random article “coin collecting” can mislead us, because it has a “bump”, but it is a random article. And we should remember that, our dataset only contains data with 500 or more revisions. Random articles with less revisions contain less data, so it they rarely have bumps, because bumps mostly contain more than 100 revisions.

Figure4.2:A random article classified as featured



WikiChanges for Coin_collecting

View by: **Month**

Although it has some exceptions, thisalgorithm gives us very good results. We already know that word count can eliminate 97 percent of all random articles. With this algorithm, we can eliminate 80 percent of the remaining articles, which is an incredible result.

5.CONCLUSION

Wikipedia is a daily changingencyclopedia and the article qualityof it can change over time as they are edited by different users. Research shows us, that the article quality improves through revisions of the articles. To analyze the parameters which affect quality,we compared the behavior of mothly revision count and showed that there is a positive correlation between revision bumps and article quality. Furthermore,we examined the revision history in featuredand non–featured articles.

Quality of a wikipedia article can be calculated throug various ways. Although some of the criterias of the featured articles like word count and revision count can be programatically found easily, most of them are hard to understand. In this work we have found that the evaluation of a featured article relies only on a few months of a hard work.

We hope that other researchers can take away several lessons from this narrative of our dataset finding development of protocols for the application of the Wikipedia quality finding. First, we used feature article finding criterias concerning the length, contribution and timing of our measures to frame our decision-making about when to measure wiki quality. Second, in the absence of existing published research about wiki quality, we were able to use easily obtained data about wiki development to make reasonable assumptions about wiki quality development. Thus, we focused our resources on measuring to the contently satisfying articles of Wikipedia. Finally, we used wiki quality data to refine our protocols in subsequent studies.

Through some research, we found out that most of the editsare made through some groups in a couple of months. This work on the article has a very big role in promoting the article to the featured article status. We can see this change in status clearly at the featured articles with a graph of revision counts per month. At the graph, this change creates a “bump”, that is easy to find out, and it mostly helpes us to find the featured article in a fair stuation.

We eliminated all the criterias which can be programatically found, which includes participation, word count and stabilisation of the article. To do that, we took random articles with only more than 2000 words of length, with a minimum revision count of 500 and with a minimum article age of 12 months, we successfully removedmore than 95 percent of all the random articles. After doing so, and creating a dataset with these values, we still can recieve more than 80 percent of success only by looking at the monthly revision count of the articles.

Bump finding is easy to compute, there is no need for revision fetching, also no need for text manupulation, or difference calculation. Through the api, the results can be calculated immediately with ease. With the fetched data, and a small calculation, which lasts less than a few seconds, we can get the result we want. Through our algorithm can be improved, it is a good start point for a new and fast way of calculating qulity of the random articles.

We presented series of results comparing quality models for Wikipediaarticles with a dataset with articles, that pass all the criterias we could eliminte. In this paper, we were able to confirm the assumption for measuring information quality, that quality of articles depend on one or more months of hard working. We also introduced a new approach to quality finding: monthly revision count should have a bump in order to improve the quality of the article.

This is a very new and improved research for quality finding, and we hope that this method helps other researchers for their algorithms as well. It can be improved through some additional features, as well as it can be used inside some other quality finding algorithms. Eighter way, we think bumps will help researchers finding quality articles easily.

REFERENCES

Books

Bruce R & Wiebe J, 1994. *Word-sense disambiguationusing decomposable models*., pp. 139-146.

Pedersen T, 1996. *Fishing for exactness*. pp.188-200, Austin: TX.

Pedersen T, 2000. *A simple approach to building ensembles of naive bayesian classiers for word sense disambiguation.* pp. 63-69, Seattle: WA.

Washio T &Matoda H, 1997.*Discovering Admissible Models of Complex Systems Based on Scale-Types and Identity Constrains*. pp.810-817.

Periodicals

Barry X.& Miller K. 2006. I want my wikipedia!*Library Journal*, April

Cressie N & Read T, 1984. Multinomial goodnessof _t tests. *Journal of the Royal Statistics Society Series* B, 46 pp.440-464

Giles G, 2005. *Internet encyclopedias go head to head. In Nature*, 438,

Pedersen T &. Bruce R, 1997. A new supervised learning algorithm for word sense disambiguation., pp. 604-609, *Providence*, RI, July.

Kilgarri A _ &. Palmer M, 2000. Special issue on SENSEVAL: Evaluating word sense disambiguation programs. *Computers and the Humanities*, **34**(1-2).

Other Sources

Blumenstock E, 2008. *Size Matters: Word Count as a Measure of Quality on Wikipedia*. Beijing, China.

Miller R, 2004. *Wikipedia founder jimmy wales responds*. Slashdot.

Jimmy Wales. http //en.wikipedia.org/wiki/[Accessed 20 April 2012] Wikipedia

MediaWiki. http //mediawiki.org. [Accessed 20 April 2012]

Hasan M.& Andr´e Gon, 2009. *Automatic quality assessmentof content created collaboratively by web communities:*, NewYork, USA.

Lim A, Sun H & B.-Q. Vuong 2007, *Measuring article quality in wikipedia:* New York, USA.

Zeng H & McGuinness L, 2006. *Computing trust from revision history.* PST.

McGuinness D.& Bhaowal M 2006. *Investigation into trust for collaborative information repositories*: A Wikipedia case study

Stvilia, B & Gasser, L2005,*Assessing information quality of a community-based encyclopedia*.: ICIQ

Stvilia B & Smith L *2005*. *Information Quality Discussions in Wikipedia*. International Conference on Knowledge Management.

Zeng H, & D McGuinness 2006. *Computing trust from revision history*. Intl.Conf. on Privacy, Security and Trust,