Analyzing and Predicting Quality Flaws in User-generated Content: The Case of Wikipedia

by

Maik Anderka

Dissertation to obtain the academic degree of
Dr. rer. nat.

Faculty of Media
Bauhaus-Universität Weimar
Germany

Advisor: Prof. Dr. Benno Stein
Reviewer: Prof. Dr. Michael Granitzer

Date of oral exam: June 28, 2013

Contents

Abstract

Web applications that are based on user-generated content are often criticized for containing low-quality information; a popular example is the online encyclopedia Wikipedia. The major points of criticism pertain to the accuracy, neutrality, and reliability of information. The identification of low-quality information is an important task since for a huge number of people around the world it has become a habit to first visit Wikipedia in case of an information need. Existing research on quality assessment in Wikipedia either investigates only small samples of articles, or else deals with the classification of content into high-quality or low-quality. This thesis goes further, it targets the investigation of *quality flaws*, thus providing specific indications of the respects in which low-quality content needs improvement. The original contributions of this thesis, which relate to the fields of user-generated content analysis, data mining, and machine learning, can be summarized as follows:

(1) We propose the investigation of quality flaws in Wikipedia based on user-defined cleanup tags. Cleanup tags are commonly used in the Wikipedia community to tag content that has some shortcomings. Our approach is based on the hypothesis that each cleanup tag defines a particular quality flaw.

(2) We provide the first comprehensive breakdown of Wikipedia's quality flaw structure. We present a flaw organization schema, and we conduct an extensive exploratory data analysis which reveals (a) the flaws that actually exist, (b) the distribution of flaws in Wikipedia, and, (c) the extent of flawed content.

(3) We present the first breakdown of Wikipedia's quality flaw evolution. We consider the entire history of the English Wikipedia from 2001 to 2012, which comprises more than 508 million page revisions, summing up to 7.9 TB. Our analysis reveals (a) how the incidence and the extent of flaws have evolved, and, (b) how the handling and the perception of flaws have changed over time.

(4) We are the first who operationalize an algorithmic prediction of quality flaws in Wikipedia. We cast quality flaw prediction as a one-class classification problem, develop a tailored quality flaw model, and employ a dedicated one-class machine learning approach. A comprehensive evaluation based on human-labeled Wikipedia articles underlines the practical applicability of our approach.

Chapter 1

Introduction

> *Wikipedia is first and foremost an effort to create and distribute a free encyclopedia of the highest possible quality to every single person on the planet in their own language. (Jimmy Wales, Wikipedia founder)*

The improvement of its information quality is a major task for the free online encyclopedia Wikipedia. Wikipedia is one of the largest and most popular user-generated knowledge sources on the Web: it is available in more than 285 languages, the English version contains about 4 million articles, and the Wikipedia community involves more than 36 million registered editors.[1] Moreover, wikipedia.org is the sixth most visited website, and hence, it is more popular than, for instance, Twitter, Amazon, and LinkedIn.[2] The community of Wikipedia authors is heterogeneous, including people with different levels of education, age, culture, language skills, and expertise [65]. In contrast to printed encyclopedias, the contributions to Wikipedia are not reviewed by experts before publication. Therefore, the most important and probably the most difficult challenge for Wikipedia pertains to its articles' information quality. The improvement of content quality has been stated as one of five strategic priorities that have been established by the Wikimedia Strategy Task Force [61]. Wikipedia founder Jimmy Wales announced in a recent interview [93]: "Our goal is to make Wikipedia as high-quality as possible. [Encyclopædia] Britannica or better quality is the goal."

The size and the dynamic nature of Wikipedia render a comprehensive manual quality assurance unfeasible. This is underlined by the fact that, at the time of writing this thesis, less than 0.1% of the English Wikipedia articles has been

[1] Wikistats 2.0, "List of Wikipedias," last update October 11, 2012, `http://wikistats.wmflabs.org`.

[2] Alexa Internet Inc., "Alexa Top 500 Global Sites," accessed November 12, 2012, `http://www.alexa.com/topsites`.

labeled as *featured*, i.e., are considered to be well-written, comprehensive, well-researched, neutral, stable, and of reasonable length.[3] A variety of approaches for automatic information quality assessment in Wikipedia has been proposed in the relevant literature, whereby a good deal of the existing research targets the classification task "Is an article featured or not?" [22, 46, 71, 77, 99, 100, 126, 143, 157]. Although the developed approaches perform nearly perfectly in distinguishing featured and non-featured articles, or stated generally, high-quality and low-quality articles, they provide virtually no support for quality assurance activities. The classification is based on meta-features that correlate with featured articles in general but cannot (and was not intended to) provide a rationale governing the respects in which an article violates Wikipedia's featured article criteria. This goal, however, is addressed in this thesis. We try to pinpoint the flaws of an article that need to be fixed in order to improve its quality.

There is already a large body of research that addresses the detection of vandalism in Wikipedia articles (see for instance [112, 123, 124]). However, the majority of quality issues in Wikipedia is not caused due to malicious intentions but stem from edits by inexperienced authors; examples include poor writing style, unreferenced statements, or missing neutrality. In this thesis, we do not investigate vandalism but rather target the whole spectrum of quality flaws. We consider a quality flaw as any issue within a Wikipedia article that causes a violation of Wikipedia's quality standards. A few prior studies exist that address quality flaws in Wikipedia, and, in contrast to our research work, they either investigate only small samples of articles [144] or analyze only a restricted set of flaws [63, 129]. This thesis is endeavoring to provide the first comprehensive study of quality flaws in Wikipedia.

The key objectives of this dissertation are twofold:

1. *Analyzing* quality flaws in Wikipedia.
   The objective is to breakdown Wikipedia's quality flaw situation by means of an exploratory data analysis. The focus is on both the current situation and the evolution of quality flaws.

2. *Predicting* quality flaws in Wikipedia.
   The objective is to automatically identify quality flaws in Wikipedia articles using machine learning techniques.

[3] Wikipedia, "Featured articles," last modified November 21, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Featured_articles`.

1.1 Research Questions and Contributions

As mentioned above, the amount of featured articles in the English Wikipedia is less than 0.1%. This begs the question of what is wrong with the remaining 99.9%. We conduct an extensive exploratory analysis of the English Wikipedia to address this issue and to investigate the following three research questions:

RQ 1. *How to compile the set of quality flaws that occur in Wikipedia?* A comprehensive compilation of quality flaws that occur in Wikipedia does not exist so far. We take advantage of the fact that Wikipedia users who encounter a flaw (but who are either not willing or who do not have the knowledge to correct it) can tag the article with a so-called cleanup tag. We implement an automatic mining approach to extract the existing cleanup tags from Wikipedia, which gives us the set of quality flaws that have so far been identified by Wikipedia users.

RQ 2. *What kinds of quality flaws exist in Wikipedia?* We organize the flaws along the three dimensions flaw type, flaw scope, and flaw commonness. This organization reveals Wikipedia's quality flaw structure and provides indications of a flaw's importance as well as of the usefulness of the respective cleanup tag. In particular, the breakdown shows the kinds of flaws a user may encounter when searching information in Wikipedia.

RQ 3. *How to quantify the extent of flawed content?* We quantify the flawed content that has been tagged so far by investigating the incidence of cleanup tags over Wikipedia's namespaces and Wikipedia's main topics. Moreover, we propose a measure for estimating the actual extent of the flawed content, since due to the size and the dynamic nature of Wikipedia, it is more than likely that many flaws have not yet been tagged. Thus we are the first who give empirical evidence for the amount of low-quality content in Wikipedia.

The findings from research questions RQ 1–3 give insights into Wikipedia's current quality flaw situation and support Wikipedia's quality assurance activities by revealing weaknesses with respect to the quality of information. Specifically, our analysis reveals that 26.86% of the English Wikipedia articles contain at least one quality flaw, whereas 50.1% of the tagged quality flaws concern the articles' verifiability, which is one of the most important principles of an encyclopedia. The actual frequency of the flaws is even higher, it is likely that one out of five articles does not cite any references or sources.

The dynamic nature of Wikipedia suggests that the quality flaw situation changes over time. We address this issue by investigating the entire revision history of the English Wikipedia from its launch in January 2001 until January 2012. Our analysis comprises all 508 243 744 revisions of the 25 981 062 pages that existed in January 2012. The contents of all revisions sum up to 7.9 TB (uncompressed), which requires efficient and scalable technology to access and process the relevant data. We use the Wikipedia database backup dumps provided by the Wikimedia Foundation as the basis for our analyses and process the dumps on a Hadoop cluster using Google's MapReduce. In particular, we analyze the occurrence of cleanup tags in the revisions of Wikipedia articles. The benefits of this approach are twofold: First, it shows how the incidence and the extent of (tagged) quality flaws have evolved. Second, it shows how the way the Wikipedia community perceives and handles quality flaws has changed over time. Our analysis targets the following specific research questions:

RQ 4. *When did the first quality flaws emerge, and how have the number and the kind of flaws changed over time?* The age, the number, and the kind of a cleanup tag give some indication of the importance as well as of the scope of the respective quality flaw. In particular, we expect the absolute number of cleanup tags to become stable at some (future) point, when each possible flaw is covered by a respective tag.

RQ 5. *Has the frequency, the type, and the distribution of quality flaws changed over time?* This question relates to the usage of cleanup tags. We expect that certain cleanup tags have been widely used, whereas others have been used infrequently or even not at all. This also gives some indication of the benefit of a certain tag for the Wikipedia community.

RQ 6. *How long does it take until tagged quality flaws are corrected?* It has been shown that certain types of vandalism were repaired faster than others [152]. Similarly, we expect that certain flaws get corrected faster than others; consider for instance a broken link and compare it to an article that is not written from a neutral point of view. In this regard, we quantify the mean correction time of a flaw, which gives some indication of the flaw's complexity.

We provide the first comprehensive breakdown of the evolution of quality flaws in the English Wikipedia. Specifically, our analysis reveals that a number of cleanup tags has been used very infrequently or not at all. We also identify several quality flaws that have never been corrected or that have a very high correction time. Moreover, we show that inline (i.e., within the text) cleanup tags are more effective than tag boxes when it comes to the correction time. Our findings form the basis for valuable conclusions for the Wikipedia community, of which we believe that they will help to make future quality assurance activities more effective.

As motivated earlier, a comprehensive manual tagging of articles that require cleanup is unfeasible due to Wikipedia's size and its constantly changing content. We use the articles that have been tagged so far as a source of human-labeled data, which is then exploited by a machine learning approach to automatically predict quality flaws in untagged articles. In particular, we address the following research questions:

RQ 7. *How to model quality flaws?* The automatic prediction of quality flaws requires a model that captures the flaws' characteristics based on measurable features. We implement 65 features that have been proposed in previous work on automatic quality assessment in Wikipedia, and we introduce 30 new features that directly target certain quality flaws. In particular, we distinguish the two modeling paradigms "intensional" and "extensional". The former allows for an efficient as well as effective prediction of certain flaws that is based on rules, the latter resorts to the realm of machine learning.

RQ 8. *How to predict quality flaws?* An algorithmic prediction of quality flaws in Wikipedia has not been operationalized before. We suggest to cast quality flaw prediction in Wikipedia as a one-class classification problem: Given a sample of articles that have been tagged with flaw f, decide whether or not an unseen article suffers from f. We argue that common binary or multiclass classification approaches are ineffective for flaw detection, and we adapt a dedicated one-class classification machine learning approach to tackle this problem.

RQ 9. *How to assess classifier effectiveness?* The acquisition of significant test data is a difficult undertaking in the Wikipedia setting. A representative sample of Wikipedia articles that have been tagged to *not* contain a particular flaw is not available. To assess the effects of a biased sample selection, we evaluate our classifier on both an optimistic test set, using featured articles as outliers, as well as a pessimistic test set, using random untagged articles as outliers. In addition, since the real-world flaw-specific class probabilities in Wikipedia are unknown, we analyze classifier effectiveness as a function of the flaw distribution.

Our flaw prediction approach is evaluated on the basis of 10 000 English Wikipedia articles that have been tagged with ten important quality flaws. Given the optimistic test set (using featured articles as outliers) and a balanced class distribution, eight flaws can be detected with a precision close to 1. The evaluation results underline the practical applicability of our approach: consider, for instance, a Wikipedia bot that autonomously identifies and tags flawed articles. Our evaluation is based on the corpus of the "1st International Competition on Quality Flaw Prediction in Wikipedia"; a competition that was initiated and co-organized by the author of this thesis and that took place in conjunction with the PAN 2012 lab held at the CLEF 2012 conference.[4] The evaluation corpus comprises 1 592 226 English Wikipedia articles, of which 208 228 have been tagged to contain one of ten important quality flaws.

[4] The "1st International Competition on Quality Flaw Prediction in Wikipedia": `http://www.webis.de/research/events/pan-12/pan12-web/wikipedia-quality.html`.

The analyses in this thesis are performed in the context of the English Wikipedia, which is the largest and most popular language edition. However, we are confident that our research results are relevant to other Wikipedia language editions as well, but we do not demonstrate this relevance here. Finally, we believe that our findings can prove beneficial not only to Wikipedia but also to other wiki-based projects and to user-generated content in general.

1.2 Thesis Organization and Related Publications

This thesis is organized as follows: The remainder of *Chapter 1* gives a brief introduction to Wikipedia, discusses Wikipedia's definition of information quality, and reviews the state of the art in quality assessment.

Chapter 2 investigates research questions RQ 1–3. This chapter breaks down Wikipedia's current quality flaw situation, based on the English Wikipedia snapshot from January 4, 2012.

Chapter 3 investigates research questions RQ 4–6. This chapter breaks down the evolution of quality flaws, taking into account the entire history of the English Wikipedia from its launch in January 2001 until January 2012.

Chapter 4 investigates research questions RQ 7–9. This chapter tackles automated quality flaw prediction and presents a dedicated one-class machine learning approach. This chapter provides also an overview of the "1st International Competition on Quality Flaw Prediction in Wikipedia".

Chapter 5 concludes this thesis. This chapter summarizes main results, discusses them in relation to our research questions, and gives an outlook on future work. Moreover, it draws conclusions regarding the practical suitability of our findings and provides concrete recommendations for the Wikipedia community.

Appendix A provides a complete listing of the quality flaws that have been identified in the English Wikipedia, along with detailed statistics and further information about the individual flaws.

Appendix B provides in-depth descriptions and implementation details of Wikipedia article features, which were either proposed in prior studies on quality assessment in Wikipedia or are newly introduced in this thesis.

This dissertation is based on a number of publications, see Table 1.1. The chapters of this thesis go beyond the content of the respective publications and add significant details as well as new aspects. The most notable innovation is that all analyses that are based on Wikipedia data are carried out anew, using the English Wikipedia snapshot from January 4, 2012. Depending on the respective release date, the publications are based on different Wikipedia snapshots.[5] Using a uniform and well-defined data base makes the individual findings comparable and guarantees their reproducibility. The snapshot is publicly available and was also used in the "1st International Competition on Quality Flaw Prediction in Wikipedia" mentioned above.

The remainder of this section briefly summarizes research by the author that is not covered in this thesis (Table 1.1 overviews the respective publications).

Theory of models for information retrieval. We introduce the idea of collection-relative retrieval models, a paradigm where several important retrieval models fit in, including VSM, GVSM, ESA, and LSI [139]. This unifying view helps to better understand retrieval models, and it can be considered as a step towards a common theoretical framework for text retrieval. Among the retrieval models that have been proposed in the last years, the Explicit Semantic Analysis, ESA, received a lot of attention. We look at the foundations of the ESA from a theoretical point of view and employ a general probabilistic model for term weights, which reveals how the ESA actually works [4, 66]. Based on this model, we provide a theoretical grounding on how properties of the reference collection affect the ESA-based computation of similarities. Moreover, we give evidence that (a) the ESA model is a variation of the generalized vector space model, GVSM, and, (b) that the original conceptual motivation for the ESA model does not hold.

Cross-language information retrieval. We introduce Cross-language Explicit Semantic Analysis, CL-ESA, a multilingual generalization of the ESA model [122]. The model exploits a document-aligned multilingual reference collection, such as Wikipedia, to represent a given text as a language-independent concept vector. The relatedness of two texts in different languages is assessed by the cosine similarity between the corresponding concept vectors. We also propose a formal definition and an alternative interpretation for the CL-ESA model, which is relevant for real-world retrieval applications since it shows how the computational effort of CL-ESA can be shifted from the query phase to a preprocessing phase [7]. This variation of the CL-ESA is evaluated in the TEL@CLEF task of the CLEF 2009 ad-hoc track. Moreover, we contribute to an important variant of cross-language information retrieval, called cross-language high similarity search. Monolingual high similarity search can be tackled in sub-linear time, either by fingerprinting or by "brute force n-gram indexing". We present theoretical and empirical insights that neither of these two approaches can be applied to tackle cross-language high similarity search, and that a linear scan is inevitable [8].

[5] Dates of the Wikipedia snapshots used in previous publications: January 16, 2010 [9]; January 15, 2011 [5, 10, 12, 101]; September 1, 2011 [11]; and January 4, 2012 [6].

Table 1.1: Publications by the author and their usage within this thesis.

Used in	Venue	Pages	Publisher	Year	Reference
Chap. 2	WebQuality	8	ACM	2012	[5]
	M. Anderka and B. Stein. A breakdown of quality flaws in Wikipedia.				
Chap. 2	WWW	2	ACM	2011	[9]
	M. Anderka, B. Stein, N. Lipka. Towards automatic quality assurance in Wikipedia.				
Chap. 3	WPAC	9	online	2012	[11]
	M. Anderka, B. Stein, M. Busse. On the evolution of quality flaws and the effectiveness of cleanup tags in the English Wikipedia.				
Chap. 4	SIGIR	10	ACM	2012	[12]
	M. Anderka, B. Stein, N. Lipka. Predicting quality flaws in user-generated content: the case of Wikipedia.				
Chap. 4	CLEF	7	CLEF	2012	[6]
	M. Anderka, B. Stein. Overview of the 1st international competition on quality flaw prediction in Wikipedia.				
Chap. 4	CIKM	4	ACM	2011	[10]
	M. Anderka, B. Stein, N. Lipka. Detection of text quality flaws as a one-class classification problem.				
–	SIGIR	2	ACM	2012	[101]
	N. Lipka, B. Stein, M. Anderka. Cluster-based one-class ensemble for classification problems in information retrieval.				
–	CIKM	4	ACM	2011	[66]
	T. Gottron, M. Anderka, B. Stein. Insights into explicit semantic analysis.				
–	ECIR	5	Springer	2010	[8]
	M. Anderka, B. Stein, M. Potthast. Cross-language high similarity search: why no sub-linear time bound can be expected.				
–	CLEF	8	Springer	2009	[7]
	M. Anderka, N. Lipka, B. Stein. Evaluating cross-language explicit semantic analysis and cross querying.				
–	TIR	5	IEEE	2009	[139]
	B. Stein, M. Anderka. Collection-relative representations: a unifying view to retrieval models.				
–	SIGIR	2	ACM	2009	[4]
	M. Anderka, B. Stein. The ESA retrieval model revisited.				
–	ECIR	9	Springer	2008	[122]
	M. Potthast, B. Stein, M. Anderka. A Wikipedia-based multilingual retrieval model.				

1.3 A Brief Introduction to Wikipedia

Wikipedia is a free, web-based, collaborative, and multilingual encyclopedia project. It was launched in January 2001 by Jimmy Wales and Larry Sanger.[6] Since 2003, Wikipedia is operated by the Wikimedia Foundation, a nonprofit charitable organization. The Wikipedia project is realized using the free software MediaWiki, which is developed by the Wikimedia Foundation. MediaWiki implements the wiki concept, introduced by Ward Cunningham [96], which provides the environment for collaborative content creation. Apart from very few exceptions, all content in Wikipedia can be edited by anybody, even anonymously. In contrast to printed encyclopedias, the contributions to Wikipedia are not reviewed by experts before publication—contributions to Wikipedia are published immediately.[7]

At first consideration, it may seem absurd that Wikipedia's open editing model can produce a useful knowledge source. One reason why Wikipedia is not run down by vandals is that it is less time consuming to undo vandalism than it is to cause it—with the result that malicious behavior is discouraged [114]. The MediaWiki software maintains the full edit history of each Wikipedia article, which allows for restoring earlier versions of vandalized articles without great effort. It has been shown that vandalism and other types of damage are repaired relatively quickly by the community, and a huge amount even almost immediately [125, 152, 153].

The fundamental principles of the Wikipedia project are summarized in the so-called "five pillars":[8]

1. Wikipedia is an encyclopedia.
2. Wikipedia is written from a neutral point of view.
3. Wikipedia is free content that anyone can edit, use, modify, and distribute.
4. Editors should interact with each other in a respectful and civil manner.
5. Wikipedia does not have firm rules.

[6] Wikipedia, "Wikipedia," last modified October 5, 2011, `http://en.wikipedia.org/wiki/Wikipedia`.

[7] There are some exceptions in particular language editions. In the German Wikipedia, for instance, only the latest "sighted versions“ are presented to common users. A sighted version is a revision of an article that is marked as being free of obvious vandalism.

[8] Wikipedia, "Wikipedia:Five pillars," last modified September 11, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Five_pillars`.

1.3.1 Size and Growth

Wikipedia is often referred to as the largest collaboratively created reference work on the Web. At the time of writing this thesis, the Wikipedia project comprises 285 language editions with a total of nearly 89 million pages, of which more than 34 million are encyclopedic articles. More than 36 million registered users and an unknown number of anonymous users have performed over 1.4 billion edits. Table 1.2 summarizes key statistics of Wikipedia. The largest and most prominent language edition is the English Wikipedia, with a total number of more than 28 million pages and over 4 million encyclopedic articles. The English language edition is also the most popular one, its community comprises more than 17 million registered users, which corresponds to 47.92% of the registered Wikipedia users.

Table 1.2: Key statistics of the ten largest Wikipedia language editions. The bottom row shows the respective values summarized over all 285 language editions. The rows are ordered by the number of articles.

Language	**Pages**	**Articles**	**Users**	**Admins**	**Edits**
English	28 440 532	4 075 519	17 663 842	1 458	561 759 173
German	4 199 003	1 481 083	1 518 124	269	114 362 991
French	5 368 645	1 305 931	1 389 309	186	89 576 384
Dutch	2 387 105	1 123 362	485 508	61	33 581 053
Italian	3 094 092	967 006	802 730	109	57 832 911
Polish	1 803 181	927 222	515 974	155	33 336 799
Spanish	3 907 050	927 240	2 389 795	137	64 866 865
Russian	3 124 683	916 488	928 263	92	55 260 874
Japanese	2 255 700	827 889	668 744	60	45 516 429
Portuguese	3 053 585	757 437	1 058 729	5 245	33 384 652
...	...	...	...	...	...
$\sum$	88 910 183	34 552 293	36 864 055	45 443	1 438 456 759

Source: Data from Wikistats 2.0, “List of Wikipedias,” last update October 11, 2012, `http://wikistats.wmflabs.org`.

Early studies that analyzed the evolution of the English Wikipedia report on an exponential growth in the number of encyclopedic articles [3, 29, 31, 154, 157, 164]. The trend changed in 2007, and since then the growth rate declined, as is shown in Figure 1.1. Up to now, different statistical approaches have been proposed to model Wikipedia's article growth [31, 48, 137, 141], with the result that the article count is expected to plateau in the near future.[9] These

[9] For a comparative survey of approaches that model article growth in Wikipedia, refer to `http://en.wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`.

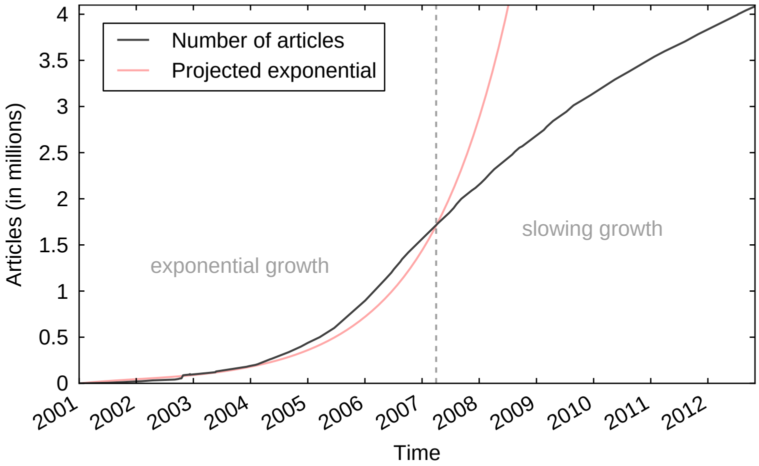


Figure 1.1: Number of encyclopedic articles in the English Wikipedia from January 2001 up to October 2012. The plotted data is taken from `http://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia`.

results led to a rather pessimistic future prospect and it is speculated that the growth of Wikipedia is limited [145]. Several studies explain the slowing growth of Wikipedia by increasing meta activities, such as coordination, content organization, and policy setting [30, 59, 86, 90]. Another explanation is that the opportunities to make novel contributions are limited [145] since Wikipedia already covers a wide range of topics [70, 87, 130]. These findings apply not only to the English Wikipedia, in fact, it is more than likely that a unique growth process underlies all language editions [164]. Despite the slowing growth, one of five critical targets that have been stated in Wikimedia's movement strategic plan is to increase the number of Wikipedia articles to 50 million by 2015.[10]

1.3.2 Community

Almost all Wikipedia content can be edited immediately, by everyone, and with minimum effort. This open editing model begs the question of who actually writes Wikipedia. As mentioned above, there are currently more than 17 million registered English Wikipedia users, and also an unknown number of anonymous contributors. The reasons why people are motivated to contribute to Wikipedia on a voluntary basis have been extensively studied [65, 91, 115, 120, 132, 151, 160]. The main motivations that have been identified are that contributors enjoy a sense of community membership, benevolence, accomplishment, and

[10] Wikimedia Foundation, “Wikimedia Strategic Plan,” February 2011, `http://wikimedia.org/wiki/Wikimedia_Movement_Strategic_Plan_Summary`.

fun. Moreover, many people like the idea of establishing a freely available source of human knowledge and would like to contribute to subject matters in which they have expertise. It has also been shown that cooperation and coordination activities play an important part in fostering the motivation and the activity of individual Wikipedia editors [84, 88, 148, 152, 153]. Cooperation and coordination in Wikipedia are mainly realized via discussion pages and so-called WikiProjects. A WikiProject is composed of a collection of pages and a group of editors with the goal to improve the coverage and quality of a specific topic or family of topics within Wikipedia. At the time of writing this thesis, there are about 2 000 WikiProjects in the English Wikipedia.[11]

A survey [65] conducted in 2010 by the Wikimedia Foundation and the Collaborative Creativity Group at UNU-MERIT[12] shows that the community of Wikipedia editors is heterogeneous, including people with different levels of education, age, and language skills. Moreover, an interesting finding of the survey is that the share of male contributors is substantially larger (86.73%). This gender imbalance has also been observed and further investigated by several recent studies [13, 60, 135]. Despite the huge number of registered Wikipedia users, it has been shown that only a small fraction, so-called elite- or power users, do the majority of work [3, 92, 116, 117, 119, 125, 133, 142]. In particular, less than 10% of the registered users perform more than 90% of the edits in the ten largest language editions of Wikipedia [118]. This fact is in contrast to the "wisdom of the crowd" phenomenon [146]. However, Kittur et al. [85] report on a shift of the workload to the common users, with a corresponding decline in the influence of the elite.

1.3.3 Significance

Wikipedia has become the primary source of knowledge for a huge number of people around the world. In October 2012, Wikipedia received on average 551 million page views per day, of which the vast majority (90.31%) account for the ten largest language editions.[13] At the time of writing this thesis, wikipedia.org is the sixth most visited website according to the Alexa Traffic Rank.[14] Thus, Wikipedia is more popular than for instance Twitter, Amazon,

[11] Wikipedia, "Wikipedia:WikiProject," last modified November 10, 2012, `http://en.wikipedia.org/wiki/Wikipedia:WikiProject`.

[12] UNU-MERIT is a joint research and training center of United Nations University (UNU) and Maastricht University.

[13] Wikimedia Statistics, "Page Views for Wikipedia," accessed November 12, 2012, `http://stats.wikimedia.org/EN/TablesPageViewsMonthly.htm`.

[14] Alexa Internet Inc., "Alexa Top 500 Global Sites," accessed November 12, 2012, `http://www.alexa.com/topsites`.

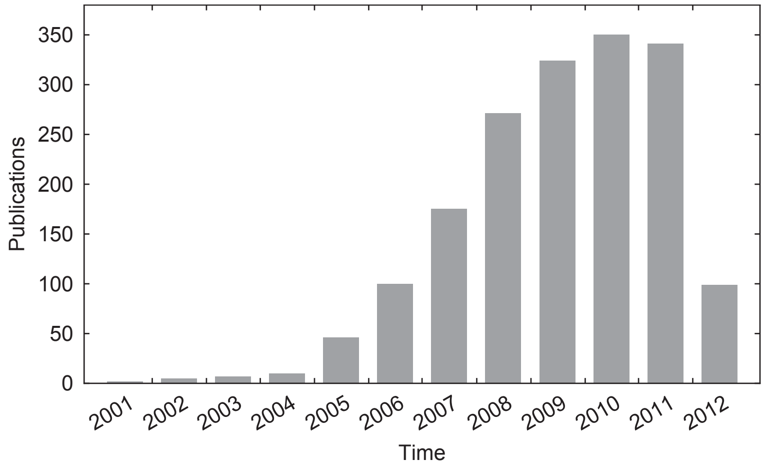


Figure 1.2: Number of Wikipedia-related publications per year. The plotted data is taken from `http://wikipapers.referata.com/wiki/List_of_publications`, last update November 7, 2012.

and LinkedIn. More than half of the visits to wikipedia.org (58.78%) refer to the subdomain en.wikipedia.org, and hence, to the English language edition.[15] Although the reliability of its content might be questionable, for many people it has become a habit to first visit Wikipedia in case of an information need.

The importance of Wikipedia is also witnessed by a large body of relevant scientific publications. The huge amount of publicly available data and the fact that Wikipedia is collaboratively created solely by volunteers have attracted researchers of several academic disciplines. The project WikiPapers, which aims to create the most comprehensive literature compilation for research on wikis, lists currently 1 750 publications.[16] The number of Wikipedia-related publications has grown continuously over the last decade, as shown in Figure 1.2. Wikipedia is often used as an example case in scientific studies to derive and validate general findings; for instance, to analyze the process of content creation by online communities [92], for studying social network aspects [15, 69], and to investigate the effects of cultural values on social media [113]. Moreover, especially in the fields of information retrieval and machine learning, a variety of approaches has been proposed that utilize Wikipedia as a knowledge base; for instance, to tackle the following problems: cluster labeling [32], link prediction [108, 109], multilingual information retrieval [122], text classification [155],

[15] Alexa Internet, Inc., "Wikipedia.org Site Info," accessed November 12, 2012, `http://www.alexa.com/siteinfo/wikipedia.org`.

[16] WikiPapers, "List of publications," last modified November 7, 2012, `http://wikipapers.referata.com/wiki/List_of_publications`.

named entity disambiguation [72], text clustering [18, 76, 78], and measuring semantic relatedness [62, 66].

Wikipedia's popularity renders the detection of quality flaws a relevant task of highest importance since erroneous information affects a huge number of readers as well as a variety of scientific approaches that utilize (possibly flawed) knowledge from Wikipedia.

1.3.4 Technical Aspects

The Wikipedia project is hosted by the Wikimedia Foundation, which operates two colocation centers in Tampa, Florida and another two in Amsterdam, Netherlands. As of November 19, 2012, the Wikimedia Foundation runs 659 servers, representing a total of 7 818 CPUs.[17] The system architecture is realized based on an extended LAMP[18] environment that comprises the following core components:[19]

- Linux operating system (mainly Ubuntu).
- Web server and caching proxy server (Apache and Squid).
- DNS server for geographical request distribution (PowerDNS).
- Load-balancer for incoming and back-end requests (LVS).
- Database and memory object caching system (MySQL and memcached).
- Main Web application (MediaWiki, written in PHP).
- Full-text search engine (Lucene customization).

For a detailed description of the components and further information on Wikipedia's infrastructure, refer to Mituzas [111].

Most relevant for the analyses in this thesis is Wikipedia's database. The relational database management system MySQL is employed to store the data of all Wikimedia projects, including Wikipedia. An individual MySQL database is used for each Wikipedia language edition. The database schema is defined by the MediaWiki software, and in the current MediaWiki version 1.21 it comprises

[17] Wikimedia Foundation, "Wikimedia Grid Report for Mon, 19 Nov 2012," `http://ganglia.wikimedia.org`.

[18] The acronym LAMP refers to the principal components generally used to build Web applications based on open source software: Linux operating system, Apache HTTP Server, MySQL database, and PHP (sometimes Python or Perl respectively). For further information on LAMP, refer to Lee and Ware [94].

[19] Wikimedia Foundation, "Wikimedia servers," last modified October 9, 2012, `http://meta.wikimedia.org/wiki/Wikimedia_servers`.

50 tables.[20] (For further information on the evolution of Wikipedia's database schema, refer to Curino et al. [43].) Many tables relate to Wikipedia's content pages and associated information (e.g., the tables *page*, *text*, *category*, and *revision*). Other tables include information related to users (e.g., *user* and *user_ groups*), media files (e.g., *filearchive* and *image*), statistics and logging (e.g., *site_ stats* and *logging*), caching (e.g., *objectcache* and *querycache*), and software internals (e.g., *job* and *trackbacks*), among others. The Wikipedia database can be accessed directly via the MediaWiki software (e.g., using the Web interface or the API) or via external services provided by the Wikimedia Foundation (e.g., the Wikimedia Toolserver or the database backup dumps). A comparative overview of approaches to access the Wikipedia database is given in Section 3.1.1.

Many features of Wikipedia are provided by the underlying MediaWiki software. MediaWiki organizes the textual content into pages. Each page covers a certain topic and has a unique title. A page belongs to exactly one namespace and should be assigned to at least one category. Namespaces are a means to organize the pages from a technical point of view. At the time of writing this thesis, the English Wikipedia provides 20 namespaces. Categories are used for a topical organization of the pages. The categories themselves are organized in a hierarchical manner. A page is written in wiki markup, also called wikitext, which is a lightweight markup language. Wiki markup was designed with the goal to be easier to learn and to use than HTML. However, a complete and consistent specification of wiki markup is not yet available.[21] Editing a page produces a new revision of the page. The complete revision history of all pages is stored in the Wikipedia database.

1.4 Wikipedia and Quality: Where are We Now?

Despite its size and popularity, Wikipedia is often criticized for containing low-quality information.[22] The major points of criticism pertain to the content's accuracy, completeness, and readability as well as to the neutrality and the reliability of information. During the last decade, great efforts have been made by researchers and practitioners to assess the actual quality of Wikipedia's articles. This section overviews the relevant scientific literature as well as the

[20] MediaWiki, "Manual:Database layout," last modified October 27, 2012, `http://www.mediawiki.org/wiki/Manual:Database_layout`.

[21] MediaWiki, "Markup spec," last modified May 30, 2012, `http://www.mediawiki.org/wiki/Markup_spec`.

[22] For a compilation of critics of Wikipedia in the literature, refer to `http://en.wikipedia.org/wiki/Wikipedia:Criticisms`.

respective guidelines and approaches that have been developed by the Wikipedia community and presents the state of the art in information quality assessment and -assurance activities in Wikipedia. Assessing the quality of Wikipedia articles, however, presumes that there is a consensus on what *quality* means in this context. Before we go on to deal with Wikipedia's definition of information quality it is useful to discuss the general concept of information quality.

1.4.1 General Definition of Information Quality

There is no consensus in the relevant literature about the distinction between data quality and information quality. Madnick et al. [105] mention that there is a tendency to use *data quality* to refer to technical issues (e.g., the integration of records from different databases) and *information quality* to refer to nontechnical issues (e.g., the relevancy of information for a particular user). This fits the distinction between data and information that is known from information theorists: The term *data* refers to the plain facts, which need to be processed and interpreted to become (useful) *information* [24]. Here, this distinction is not made and the term *information quality* is used to refer to all kinds of issues.

Although there is a large body of information quality research, for which Levis et al. [97] give a comprehensive overview, there is no single definition of the quality of information. A widely accepted interpretation of information quality is the "fitness for use in a practical application" [156]. Similar, Juran and Godfrey [82] interpret information "to be of high quality if they are fit for their intended uses in operations, decision making, and planning." I.e., the definition of information quality depends on the particular context and use case.

Information quality is a multi-dimensional concept that combines several quality criteria [156]. A good deal of the existing information quality research focuses on the investigation of quality dimensions and particular classification schemes, for which Madnick et al. [105] give a comprehensive overview. Common information quality dimensions include accuracy, reliability, timeliness, objectivity, completeness, and relevance, for example.

Information quality is often investigated in an organizational context, see e.g., Lee et al. [95], and, especially in early years, many researchers target quality aspects in databases, see e.g., Madnick and Zhu [104]. In recent years, quality assessment of Web documents becomes more and more important, and information quality metrics have been incorporated into information retrieval approaches to improve the search effectiveness of Web search environments [19, 121, 162, 163]. The rise of the Web 2.0 brought an increasing diversity of produced content quality [17], and quality assessment of user-generated content and social media gained particular interest [2, 35, 56, 103, 150]. Wikipedia is solely based on user-generated content, and it is one of the largest and most successful representatives of its kind.

1.4.2 Wikipedia's Definition of Information Quality

As mentioned above, the definition of information quality depends on the particular context. In Wikipedia the context is well-defined, namely by the encyclopedic genre. Crawford [41] defined the quality of an encyclopedic article by seven general criteria: scope, format, uniqueness, authority, accuracy, currency, and accessibility. However, these criteria were defined for traditional printed encyclopedias, and hence, they do not directly fit to Wikipedia. For example, the criteria authority and accessibility are not relevant in case of a free online encyclopedia. The encyclopedic genre forms the ground for Wikipedia's core content policies: neutral point of view, no original research, and verifiability.[23]

The Wikipedia quality task force provides a working definition for the term quality: "the ability of a Wikipedia article to meet the expectations and needs of the article's target audience, i.e. the readers of the article".[24] This definition is compliant with the general "fitness for use" paradigm of Wang and Strong [156] mentioned above. Although the quality task force defines certain requirements that specify the expectations and needs of an article's target audience, the definition is rather unspecific and universal.

The information quality ideal of Wikipedia has been formalized—better: made communicable and quantifiable—within the so-called featured article criteria. Featured articles are considered to be the best articles Wikipedia has to offer. The featured article criteria of the English Wikipedia state that an article should have the following attributes:[25]

1. It is well-written, comprehensive, well-researched, neutral, and stable.
2. It provides a concise lead section, an appropriate structure, and consistent citations.
3. It has images and other media where appropriate.
4. It has a reasonable length and level of detail.

[23] Wikipedia, "Wikipedia:List of policies#Content," last modified June 3, 2012, http://en.wikipedia.org/wiki/Wikipedia:List_of_policies#Content.

[24] Wikimedia Strategy, "Task Force/Wikipedia Quality/Definition of quality," last modified April 27, 2010, http://strategy.wikimedia.org/wiki/Task_Force/Wikipedia_Quality/Definition_of_quality.

[25] Wikipedia, "Featured article criteria," last modified August 9, 2012, http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

To some extent these criteria cover those defined by Crawford [41]. The concept of featured articles exists in many language editions, however, the respective criteria differ slightly. In addition to featured articles, a set of grades for which weaker criteria apply was developed within the Wikipedia Version 1.0 project. Table 1.3 shows the respective grading scheme, which represents a discrete definition of article quality in Wikipedia. (The grading scheme will be discussed in the next subsection.) Additional criteria and guidelines have been defined by certain WikiProjects to cover topical particularities.

Most of the existing criteria and guidelines are intended to assure Wikipedia's quality standards, and hence, they define high-quality information. By contrast, the focus of this thesis is on quality flaws. As already mentioned, we consider a quality flaw as any issue within an article that causes a violation of Wikipedia's quality standards.

1.4.3 Information Quality Assessment

In early years, researchers tried to assess the quality of Wikipedia by comparing samples of articles with the corresponding entries in other well-known encyclopedias. In 2005, the journal *Nature* carried out a study comparing 42 scientific articles from Wikipedia and Encyclopædia Britannica [64]. The articles were reviewed for accuracy by domain experts: the average number of errors per article was four in Wikipedia and three in Encyclopædia Britannica. The study concluded that "Wikipedia comes close to [Encyclopædia] Britannica in terms of the accuracy of its science entries." [64] In 2006, Rosenzweig [128] judged 25 bibliographic Wikipedia articles against comparable entries in Microsoft Encarta and in the American National Biography Online. He found that Wikipedia was as accurate as Microsoft Encarta but not as accurate as American National Biography Online. Another study carried out in 2008 by Rector [127] comparing nine historical Wikipedia articles with respective articles in Encyclopædia Britannica, the Dictionary of American History, and American National Biography Online conclude that Wikipedia was less reliable than the other reference works. Further studies have been conducted by several printed magazines comparing Wikipedia to Bertelsmann Enzyklopädie, World Book Encyclopedia, and Brockhaus Enzyklopädie, among others: in the majority of cases Wikipedia articles were comparable or even better in terms of accuracy, completeness, and currentness.[26] Although the mentioned studies give some indications of Wikipedia's information quality, the findings cannot be generalized since the analyzed samples must be considered as too small.

[26] For an overview of these studies including a summary of the main results, refer to `http://en.wikipedia.org/wiki/Reliability_of_Wikipedia#Comparative_studies`.

A comprehensive quality assessment attempt was initiated by the Wikipedia Version 1.0 project, which was set up in late 2004 with the goal of producing a high-quality offline release version of the English Wikipedia.[27] In order to achieve this goal, an article quality grading scheme was developed, which allows for a quality-based selection of articles to be included in the release version. The grading scheme is shown in Table 1.3. Out of the 4 075 519 English Wikipedia articles (cf. Table 1.2) 3 454 234 have so far been assessed by Wikipedia users, which corresponds to 84.76%. The assessment activities are pushed by a multitude of WikiProjects, which often have a dedicated assessment team that is responsible for evaluating the articles in the project scope. The scheme comprises nine grades ranging from high quality (FA) to low quality (Stub). According to Table 1.3, a relatively small amount of the English Wikipedia articles (0.18%) is considered to be high quality (including the grades FA, A, and FL). On the other hand, more than half of the articles (54.13%) are classified as stubs. The assessment provides the most comprehensive overview of Wikipedia's quality situation so far, however, an enormous manual effort is necessary both to evaluate the large number of articles and to maintain existing assessments after subsequent edits.

A variety of approaches to automatically assess information quality in Wikipedia has been proposed in the relevant literature [22, 44, 46, 49, 71, 77, 80, 98, 99, 100, 126, 143, 157, 159]. The approaches mainly differ in the underlying quality models, i.e., the feature number, the feature complexity, or the rationale to quantify quality. A good deal of the developed approaches employ machine learning techniques to classify Wikipedia articles into the quality grading scheme mentioned above. Especially the grades representing high-quality content are often used, e.g., to determine whether an article is featured or not. The respective approaches perform nearly perfectly in distinguishing featured articles from non-featured ones. However, the practical support for Wikipedia's quality assurance process is marginal because featured articles are not identified, but are made.[28] Moreover, nearly all of the developed approaches are based on meta-features that correlate with featured articles in general; for instance, the number of words in an article. Consequently, these approaches provide no indication of the shortcomings of non-featured articles. There are only a few prior studies that target the identification of specific quality flaws, and, in contrast to this thesis, they either investigate only small samples of articles [144] or analyze only a restricted set of flaws [63, 129].

[27] Wikipedia, "Wikipedia:Version 1.0 Editorial Team," last modified September 29, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team`.

[28] A featured article candidate is rigorously reviewed, discussed, and revised by many users, with the aim to improve it, and therefor, *make* it a featured article. For further information, refer to `http://en.wikipedia.org/wiki/Wikipedia:Featured_article_candidates`.

There is also notable research that relates indirectly to quality assessment in Wikipedia: trust and reliability of articles [42, 161], author reputation [1, 152], automatic vandalism detection [112, 123, 124], and edit quality estimation [51].

Table 1.3: Article quality grading scheme used in the English Wikipedia along with a description of the respective criteria and the respective number of assigned articles. This grading scheme is used by the *Wikipedia:Version 1.0 Editorial Team.* (For a more detailed description of the criteria refer to the source URL stated below the table.)

Grade	Criteria description	Articles
FA	A featured article exemplifies Wikipedia's very best work and is distinguished by professional standards of writing, presentation, and sourcing.	4 271
A	The article is well-organized and essentially complete, having been reviewed by impartial reviewers from this WikiProject or elsewhere.	1 054
GA	A good article is a satisfactory article that has met the good article criteria but may not have met the criteria for featured articles.	16 753
B	The article is mostly complete and without major issues, but requires some further work to reach good article standards.	83 523
C	The article is substantial, but is still missing important content or contains a lot of irrelevant material. The article should have references to reliable sources, but may still have significant issues or require substantial cleanup.	132 444
Start	An article that is developing, but which is quite incomplete and may require further reliable sources.	904 086
Stub	A very basic description of the topic.	2 206 104
FL	A featured list exemplifies Wikipedia's very best work. It covers a topic that lends itself to list format.	1 790
List	Meets the criteria of a stand-alone list, which is an article that contains primarily a list, usually consisting of links to articles in a particular subject area.	104 209
	$\sum$	3 454 234

Source: Wikipedia, "Wikipedia:Version 1.0 Editorial Team/Assessment," last modified October 26, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment`.

Chapter 2

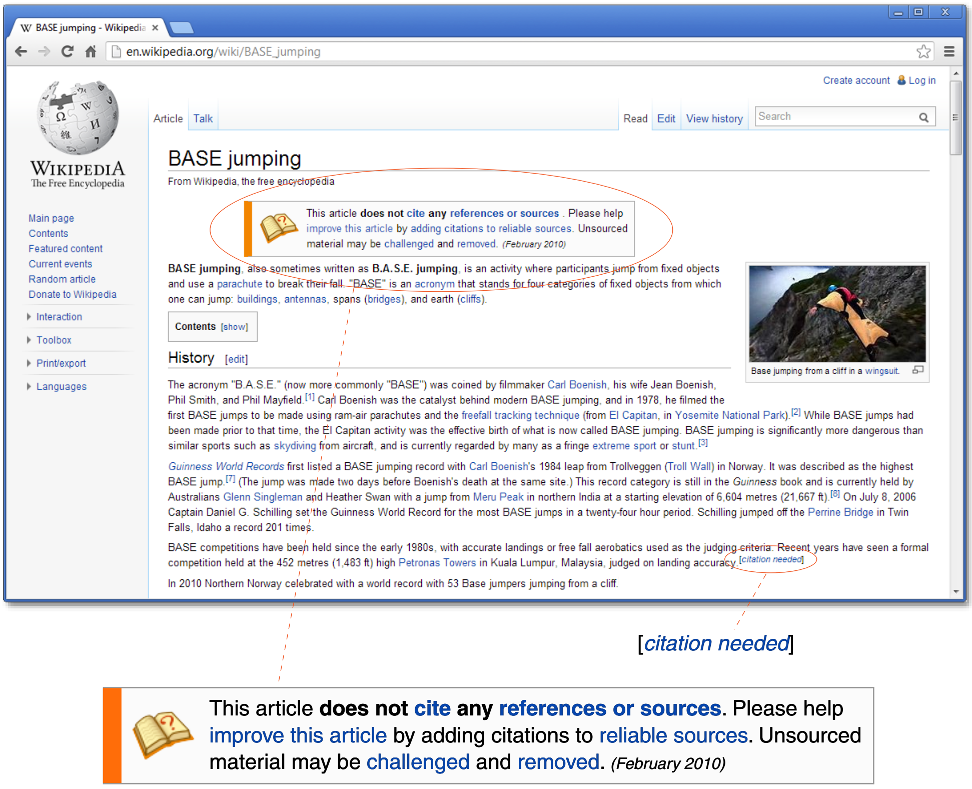
A Breakdown of Quality Flaws

Cleanup tags are a means to tag quality flaws in Wikipedia. As shown in Figure 2.1, cleanup tags are used to inform readers and editors about specific problems in articles, sections, or certain text fragments. The tags are defined by the Wikipedia community and can be placed by every user. Cleanup tags have been utilized previously for different purposes, however, in contrast to this thesis, each of the prior studies targets only a single tag. Gaio et al. [63, 129] investigate the usage and the effectiveness of the cleanup tags *Complex* and *NPOV* (neutral point of view) in SimpleWiki[1] and in the English Wikipedia respectively. Kittur et al. [86] use the cleanup tag *Controversial* to characterize conflict in Wikipedia. Apic et al. [14] propose an indicator of a country's geopolitical instability based on the number of Wikipedia articles that have a link to the Wikipedia article of the respective country and that have been tagged with the cleanup tag *NPOV disputes*. Here, we investigate all existing cleanup tags to breakdown the quality flaws that have so far been identified by Wikipedia users.

Chapter organization. Section 2.1 describes the data underlying our analyses as well as the data preprocessing. Section 2.2 describes our cleanup tag mining approach. Section 2.3 presents the resulting set of quality flaws organized along the three dimensions flaw type, flaw scope, and flaw commonness. Section 2.4 presents the distribution of tagged flaws over Wikipedia's namespaces and main topics; furthermore, this section describes our approach to estimate the actual extent of the flawed content.

Key contributions. We present the first comprehensive breakdown of Wikipedia's quality flaw structure. Our analysis reveals the kinds of flaws that actually exist, the distribution of flaws in Wikipedia, and the extent of flawed content.

[1] SimpleWiki is a relatively small language edition of Wikipedia that is written in basic English: `http://simple.wikipedia.org`.



BASE jumping

From Wikipedia, the free encyclopedia

This article **does not cite any references or sources** . Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(February 2010)*

BASE jumping, also sometimes written as **B.A.S.E. jumping**, is an activity where participants jump from fixed objects and use a parachute to break their fall. "BASE" is an acronym that stands for four categories of fixed objects from which one can jump: buildings, antennas, spans (bridges), and earth (cliffs).

Base jumping from a cliff in a wingsuit.

Contents [show]

History [edit]

The acronym "B.A.S.E." (now more commonly "BASE") was coined by filmmaker Carl Boenish, his wife Jean Boenish, Phil Smith, and Phil Mayfield.[1] Carl Boenish was the catalyst behind modern BASE jumping, and in 1978, he filmed the first BASE jumps to be made using ram-air parachutes and the freefall tracking technique (from El Capitan, in Yosemite National Park).[2] While BASE jumps had been made prior to that time, the El Capitan activity was the effective birth of what is now called BASE jumping. BASE jumping is significantly more dangerous than similar sports such as skydiving from aircraft, and is currently regarded by many as a fringe extreme sport or stunt.[3]

Guinness World Records first listed a BASE jumping record with Carl Boenish's 1984 leap from Trollveggen (Troll Wall) in Norway. It was described as the highest BASE jump.[7] (The jump was made two days before Boenish's death at the same site.) This record category is still in the *Guinness* book and is currently held by Australians Glenn Singleman and Heather Swan with a jump from Meru Peak in northern India at a starting elevation of 6,604 metres (21,667 ft).[8] On July 8, 2006 Captain Daniel G. Schilling set the Guinness World Record for the most BASE jumps in a twenty-four hour period. Schilling jumped off the Perrine Bridge in Twin Falls, Idaho a record 201 times.

BASE competitions have been held since the early 1980s, with accurate landings or free fall aerobatics used as the judging criteria. Recent years have seen a formal competition held at the 452 metres (1,483 ft) high Petronas Towers in Kuala Lumpur, Malaysia, judged on landing accuracy.[*citation needed*]

In 2010 Northern Norway celebrated with a world record with 53 Base jumpers jumping from a cliff.

[*citation needed*]

This article **does not cite any references or sources**. Please help improve this article by adding citations to reliable sources. Unsourced material may be challenged and removed. *(February 2010)*

Figure 2.1: The Wikipedia article "BASE jumping" with two cleanup tags. The tag box *Unreferenced* refers to the whole article while the inline tag *Citation needed* refers to a particular claim.

2.1 Data Base and Preprocessing

The analyses in this chapter are based on a snapshot instead of investigating Wikipedia up-to-the-minute. This guarantees reproducibility and makes our results comparable. The Wikimedia Foundation periodically compiles and provides Wikipedia snapshots in the form of database backup dumps. An automated backup process goes through the set of Wikipedia databases and dumps the data to files. The aim is to have a complete backup for each Wikipedia language edition every two weeks; except for the English Wikipedia, which has the largest database and should has a complete backup once a month.[2] The parts of the backups that comprise content-related data are publicly disclosed.[3]

[2] For further information about the dump process, including technical details, refer to `http://wikitech.wikimedia.org/view/Dumps`.

[3] Wikimedia downloads: `http://download.wikimedia.org`.

Table 2.1: Statistics of the English Wikipedia snapshot from January 4, 2012.

Number of pages	25 981 062
Number of encyclopedic articles	3 865 587
Number of featured articles	3 438
Number of images	812 059
Number of registered users	15 993 675
Number of active users (users who performed an action in the last 30 days)	133 219
Total number of edits	508 243 744
Average number of edits per page	19.56

Source: The statistics are based on the data provided by the *site_ stats* table, which is included in the database backup dumps of the Wikimedia snapshot.

User-related data are withheld for privacy reasons as well as data that relate to MediaWiki software internals. Therefore, the backups provided by the Wikimedia Foundation can be considered as partial snapshots of the Wikipedia databases at a certain date.[4]

As already mentioned in the introduction of this thesis, we use a snapshot of the English Wikipedia from January 4, 2012. The English language edition is most appropriate because it is the largest and most popular one (cf. Section 1.3.1). Table 2.1 summarizes key statistics of the snapshot. Nearly 26 million English Wikipedia pages existed at the time of the snapshot. More than 3.8 million pages are encyclopedic articles, and a subset of 3 438 articles (0.09%) are labeled as featured. The respective backup provided by the Wikimedia Foundation includes 18 so-called SQL dumps of which each comprises a single table of the English Wikipedia database in the form of raw SQL statements, totaling about 40 GB. Table 2.2 lists the 18 database tables.[5] The tables comprise meta data about pages, categories, images, templates, users, and links. We import the 18 SQL dumps into a MySQL database, which gives us a local partial copy of the English Wikipedia database. The local database allows for efficient analyses, without causing traffic on the Wikimedia servers. Note that all of our analyses can be performed on the original Wikipedia database as well.

[4] The snapshot date corresponds to the backup dump's completion time. As the dump process takes some time, especially, for larger databases, there may be inconsistencies in the data of a single snapshot. Here, we neglect possible inconsistencies since it concerns typically only a few days. In case of the English Wikipedia, it takes eight to nine days for a backup run to complete.

[5] For further information about the database schema and a description of all tables, refer to `http://www.mediawiki.org/wiki/Manual:Database_layout`.

Table 2.2: The 18 tables of the Wikipedia database that are included in the English Wikipedia snapshot from January 4, 2012 in the form of SQL dumps.

Table	Description
category	Stores the existing Wikipedia categories.
categorylinks	Category membership of pages.
externallinks	URLs and referring pages of external links.
image	Meta data about images and other media files.
imagelinks	Inclusions of images and other media files into pages.
interwiki	Prefixes and URLs of other projects, e.g., Commons, Wikibooks etc.
iwlinks	URLs and referring pages of interwiki links (to other projects).
langlinks	Links to related pages in other languages.
oldimage	Meta data about old revisions of images and other media files.
page	Meta data about each page, including id, title, and namespace.
page_ props	Properties of pages (used by the MediaWiki software).
page_ restrictions	Protection levels of pages (used to disable editing for specific users).
pagelinks	Referring pages and target pages of (Wikipedia) internal links.
protected_ titles	Protection levels of non-existent (unwanted) pages.
redirect	Pages that are redirects, along with respective target pages.
site_ stats	A number of statistics, including page, article, and edit counts.
templatelinks	Inclusions of pages, especially templates, into other pages.
user_ groups	Group membership of users.

2.2 Cleanup Tag Mining

There is no single strategy to spot the entire set of all cleanup tags. Cleanup tags are realized based on templates, which are special Wikipedia pages that can be included into other pages. Although templates can be separated from other pages by their namespace (the prefix “Template:” in the page title), there is no dedicated qualifier to separate templates that are used to implement cleanup tags from other templates. A complete manual inspection is unfeasible as Wikipedia contains more than 450 000 different templates.[6] We hence employ a two-step approach to compile the set of cleanup tags automatically:

1. an initial set of cleanup tags is extracted from two meta sources within Wikipedia, and

2. the initial set is further refined by applying several filtering substeps.

[6] Wikistats: Wikimedia Statistics, “Database records per namespace,” last modified September 21, 2012, `http://stats.wikimedia.org/EN/TablesWikipediaEN.htm#namespaces`.

Step 1: Extraction

We exploit two sources within Wikipedia containing meta information about cleanup tags. The first source that we employ is the Wikipedia administration category ”Category:Cleanup templates”, which comprises templates that are used for tagging pages as requiring cleanup. The category also has several subcategories to further organize the cleanup tags by their usage, e.g., inline cleanup templates or cleanup templates for WikiProjects. The page titles of those templates linking to the category or some subcategory are obtained from the local Wikipedia database, using the tables *categorylinks* and *page* (cf. Table 2.2), which results in 437 different cleanup tags.

The second source is the meta page ”Wikipedia:Template messages/Cleanup”, which comprises a manually maintained listing of templates that may be used to tag pages as needing cleanup. From a technical point of view, the page is a composition of several pages (transclusion principle). For each of these pages, the content of the revision from the snapshot time is retrieved using the MediaWiki API[7]. A total of 286 different cleanup tags are extracted from the wiki markup of the retrieved pages using regular expressions. Merging the findings from both sources gives 530 different cleanup tags.

Step 2: Refinement

We apply the following filtering substeps to the initial set of cleanup tags:

Redirect resolving. A cleanup tag may have several alternative titles linking to it through redirects. For example, the tag *Unreferenced* has the redirects *Unref*, *Noreferences*, and *No refs* among others. We resolve all redirects using the tables *redirect* and *page* of the local Wikipedia database (cf. Table 2.2).

Subtemplate removal. We discard particular subtemplates, namely experimental pages and documentation pages. Experimental pages are identified by the suffixes “/sandbox” and “/testcases” in the page title and are used for testing purposes only. Documentation pages are identified by the suffix “/doc” and provide a template description.

Meta-template removal. We discard templates that are solely used as building blocks, for instance, to instantiate other templates with a particular parameterization. The two Wikipedia categories ”Category:Wikipedia metatemplates” and ”Category:Wikipedia substituted templates” are used to identify these templates. Moreover, we discard templates that implement technical features (categories ”Category:Search templates” and ”Category:Maintenance navigation”) as well

[7] The MediaWiki Web service API provides direct access to the Wikipedia databases. For further information, refer to Section 3.1.1 or to `http://www.mediawiki.org/wiki/API`.

as templates that are used for documentation and testing purposes (category ”Category:Template namespace templates”).

Altogether we collect a set of 458 cleanup tags.

Discussion

To evaluate our mining approach, we manually inspected the documentation pages of the 458 templates that have been identified as cleanup tags. A documentation page gives information about purpose, usage, and scope of a template. Consider for example the tag *Unreferenced* (shown in Figure 2.1). The respective documentation page states that a tagged article “does not cite any references or sources” and that citations to reliable sources should be added in order to improve the article. This tag indeed relates to a particular cleanup task, and, since the verifiability of information is one of Wikipedia’s core content policies (cf. Section 1.4.2), it defines a quality flaw. Our evaluation reveals that of the 458 templates 445 are actually cleanup tags, and hence, define a particular quality flaw. The analyses in this thesis are based on the 445 quality flaws, which are listed in Appendix A.

The remaining twelve templates are listed in Table 2.3. None of them can be considered as a proper cleanup tag. The first ten templates in Table 2.3 are specific meta-templates that implement technical features. The template *Geodata-check* does not produce any output, and hence, it cannot be considered as a tag. The last template in Table 2.3 is a kind of placeholder that need to be replaced by the respective English cleanup tags. The twelve templates are identified by our mining approach because they are assigned to the category ”Category:Cleanup templates” (or to some subcategory respectively; see Step 1). However, the category’s documentation page states that: “This is a category of templates used for marking articles as requiring cleanup.”[8] The twelve templates are no cleanup tags, and hence, their assignment to this category is incorrect. We initiated discussions on the respective talk pages of the twelve templates in order to correct the wrong assignments in Wikipedia.[9] Note that after the miscategorizations are corrected our approach is able to identify the set of cleanup tags without any manual intervention.

Our mining approach does not guarantee completeness though, since the true set of cleanup tags is unknown in general. However, from a quantitative point of view we are confident that we identify the most common cleanup tags, and hence, the most important quality flaws.

[8] Wikipedia, “Category:Cleanup templates,” last modified November 7, 2012, `http://en.wikipedia.org/wiki/Category:Cleanup_templates`.

[9] The template *Moveoptions* has already been deleted on October 30, 2012.

Table 2.3: The twelve templates that are identified by our mining approach but which are actually no cleanup tags. This issue is due to the incorrect assignment of the templates to the Wikipedia category ”Category:Cleanup templates”.

Template name	Template description
Reflist-talk	Shows a reference section for a talk page discussion within a bordered box.
Cleanup template documentation see also section generic list	Produces a section containing a list of certain cleanup tags.
Edit	Creates an “edit” link.
Editlink	Creates an “edit this section” link.
Editlink-right	Creates a right-aligned “edit” link.
Tagged	Creates a tag for the user page of an user who tagged a page but left no justification on the respective talk page.
Introduction cleanup maintenance templates	Produces a navbox containing a listing of certain cleanup tags.
Details removed	Produces an inline statement stating that personal information has been removed to protect the user's privacy.
Postchronicle	Produces an inline statement stating that a link to a Post Chronicle has been removed.
Moveoptions	Is used to substitute other content.
Geodata-check	Tagging of geodata which needs further checking or correction. No output other than adding the tagged page to the category ”Category:Pages requiring geodata verification”.
...	Originates from the French Wikipedia and is replaced by *Empty section* or *Expand section* in the English language edition. (The template name is composed of three dots.)

2.3 Wikipedia's Quality Flaw Structure

To better understand quality flaws in Wikipedia we organize the 445 flaws along three dimensions: flaw type, flaw scope, and flaw commonness.

2.3.1 Flaw Type

Several quality flaws relate to the same type. For instance the flaws *Unreferenced* and *Citation needed* (shown in Figure 2.1) both concern article verifiability. No organization scheme has been proposed before that covers the complete set of quality flaws in Wikipedia. We consider the problem types identified by Stvilia et al. [144] as inappropriate in this respect, as they result from the manual analysis of only 60 Wikipedia talk pages and hence are very specific. Similarly, the set of ten flaw types proposed in [9] is too specific, as these types target a particular subset of only 70 flaws. At first sight, the featured article criteria (described in Section 1.4.2) may appear as a set of relevant flaw types. However, there are several drawbacks related to this idea: the featured article criteria are not stable, they do not consider technical aspects, and they focus on particularities of high-quality articles. Another idea may be to utilize the organization of the cleanup tags on the meta page ”Wikipedia:Template messages/Cleanup” and consider the section headings as flaw types. However, this organization is mainly intended to serve navigational purposes, and hence, the respective sections are very specific. Ultimately, we organize the quality flaws along a newly formed set of twelve general flaw types, which are described in Table 2.4. Our set is an extension of the ten types proposed in [9], and to some extent it covers the featured article criteria, the problem types of Stvilia et al. [144], and the organization of the above mentioned meta page. The listing in Appendix A organizes the 445 quality flaws along our flaw types. The labeling is exclusive, i.e., each flaw belongs to exactly one type.

The flaw type *Verifiability* is of particular interest as the verifiability of information is one of the most important principles of an encyclopedia. The flaws that belong to this type refer to articles that cite no references at all (e.g., *Unreferenced*, *No footnotes*, and *Unreferenced section*), to articles with inadequate and invalid references (e.g., *Refimprove*, *Primary sources*, and *Dead link*), and to unsourced statements within articles (e.g., *Citation needed*, *Who*, and *By whom*). The flaw type *Style of writing* targets stylistic issues related to grammar, style, cohesion, tone, and spelling. Most of these issues are described in Wikipedia's manual of style.[10] The flaw type *Miscellaneous* comprises flaws that are very specific and that occur relatively infrequently. Flaws that focus on a particular topic are organized under the flaw type *Specific subjects*. For instance, the flaw *Plot* states that an article's plot summary may be too long or excessively detailed, which may only apply to certain articles describing films or novels for instance. The flaw type *Unwanted content* comprises flaws that refer to content that is either not appropriate for an encyclopedia (e.g., *Notability*, *Advert*, and *Original research*) or that is better suited for a different project of the Wikimedia Foundation (e.g., *Copy to Wikiquote* and *Copy to Wikibooks*).[11]

A fundamental principle of Wikipedia is that articles should be written from a neutral point of view, i.e., representing all significant views unbiased and

[10] Wikipedia style guide: `http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style`.

[11] For further information on Wikipedia's inclusion criteria, refer to: `http://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not`.

Table 2.4: The twelve general flaw types along with a description and the respective number of quality flaws that belong to a particular type.

Flaw type	Description	Flaws
Verifiability	Missing or inadequate references and sources.	98
Style of writing	Does not conform with Wikipedia's manual of style.	74
Miscellaneous	Very specific and infrequent flaws.	54
Specific subjects	Issues that occur exclusively in certain subjects.	44
Unwanted content	Does not conform with Wikipedia's inclusion criteria.	42
Neutrality	Not written from a neutral point of view.	40
Wiki tech	Issues concerning markup, links, and categorization.	24
General cleanup	Unspecific or general issues.	17
Expand	Missing or insufficient information.	17
Structure	Inadequate content organization.	14
Time-sensitive	Outdated or unclear temporal information.	13
Merge	Similar content that should be combined.	8
		$\sum$ 445

without opinions.[12] The respective flaws are organized under the flaw type *Neutrality*. The flaw type *Wiki tech* targets technical aspects of an article, including categorization issues (e.g., *Uncategorized*, *Uncategorized stub*, and *Cat improve*), syntactical problems (e.g., *Cleanup-HTML*), connectivity in terms of Wikipedia-internal links (e.g., *Orphan*, *Wikify*, and *Dead end*), and ambiguity of links (e.g., *Dn* and *Dablinks*). The flaw type *General cleanup* groups those cleanup tags that either list several flaws in a single tag (e.g., *Multiple issues* and *Expertsubject-multiple*) or merely state that some cleanup is required at all but provide no further information (e.g., *Cleanup* and *Expert subject*). The flaws that belong to the flaw type *Expand* state that particular sections are under-represented or that certain information is missing. The flaw type *Structure* addresses the articles' organization into sections as well as the length of the sections. For example, an article is expected to have a lead section that summarizes its content. Flaws that address the currency and the lifespan of information are organized under the flaw type *Time-sensitive*. The flaws under the type *Merge* refer to articles that deal with a similar subject and hence should be combined.

The number of quality flaws that are covered by each flaw type varies widely, see Table 2.4 (rightmost column). The flaw types *Verifiability*, *Style of writing*, and *Miscellaneous* comprise more than half of the 445 flaws. However, the number of flaws provides no indication of the seriousness of a certain flaw type (we will

[12] Wikipedia, "Neutral point of view," last modified November 29, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view`.

come to this in Section 2.3.3). Types that comprise a relatively large number of flaws can be considered as better elaborated than others. For example, the flaws in the type *Verifiability* characterize a variety of specific verifiability issues (e.g., *By whom*, *Volume needed*, and *Self-published*; see Appendix A), whereas, the flaws in the type *Expand* are rather general (e.g., *Expand section*, *Incomplete*, and *Missing information*).

2.3.2 Flaw Scope

The quality flaws differ by their scope. Some flaws refer to the whole page (e.g., *Unreferenced*), others to a certain section (e.g., *Expand section*), and still others to particular claims (e.g., *Citation needed*) or links (e.g., *Dead link*). Here, we distinguish two scopes, which are described in Table 2.5. The scope of a flaw is quantified by the kind of cleanup tag that defines the flaw. A cleanup tag is either a tag box or an inline tag, see Figure 2.1. Tag boxes are placed at the top of a page (or at the top of a section) and hence their scope refers to the whole page (or section). By contrast, inline tags are placed within the text after the sentence, claim, or word they refer to. From a technical point of view both kinds of cleanup tags can be distinguished by the respective meta-templates that are used to implement the tags (e.g., *Ambox* and *Fix* respectively). The listing in Appendix A shows the scope for each of the 445 quality flaws.

Table 2.5: The two flaw scopes along with a description and the respective number of quality flaws that belong to a particular scope.

Scope	**Description**	**Flaws**
Page flaw	The flaw refers to the whole page (or to a certain section).	318
Inline flaw	The flaw refers to a certain text fragment within an article.	127
		$\sum$ 445

From the 445 quality flaws, 318 are page flaws and 127 are inline flaws. A possible explanation for this imbalance is the fact that the concept of inline tags has been introduced only after several tag boxes already existed (we will come to this in Section 3.2). In general, inline flaws are more specific than page flaws. For example, the page flaw *Unreferenced* states that the page does not cite any references or sources. By contrast, the inline flaw *Citation needed* gives a direct indication about a claim that needs to be referenced. Consequently, it is easier for a human corrector to repair inline flaws. However, some flaws refer to the whole page per definition and hence it is not appropriate to use inline tags for these flaws. This applies for instance to the flaws that belong to the flaw type *Structure*, which addresses the organization of a page (e.g., *Lead missing* and

Table 2.6: Breakdown of the 445 quality flaws that exist in the English Wikipedia. Each row contains the number of flaws that belong to the respective flaw type along with the flaw's frequency, i.e., the number of times the respective cleanup tag occurs. The values are given separately for the two scopes: page flaws (first multicolumn) and inline flaws (second multicolumn); the third multicolumn summarizes the values.

Flaw type	**Page flaws**		**Inline flaws**		$\sum$	
	Flaws	Frequency	Flaws	Frequency	Flaws	Frequency
Verifiability	42	547 033	56	378 847	98	925 880
Style of writing	47	26 444	27	24 658	74	51 102
Miscellaneous	47	128 305	7	195	54	128 500
Specific subjects	38	10 975	6	2 591	44	13 566
Unwanted content	35	321 640	7	3 151	42	324 791
Neutrality	28	20 413	12	2 697	40	23 110
Wiki tech	22	177 276	2	16 003	24	193 279
General cleanup	17	80 319	0	0	17	80 319
Expand	13	69 777	4	355	17	70 132
Structure	14	7 709	0	0	14	7 709
Time-sensitive	7	6 460	6	4 312	13	10 772
Merge	8	18 921	0	0	8	18 921
$\sum$	318	1 415 272	127	432 809	445	1 848 081

Very long). Table 2.6 breaks down the 445 quality flaws by flaw type and flaw scope (the table also comprises frequency values, which will be discussed in the next section). Analog to the flaw type *Structure*, the flaw types *General cleanup* and *Merge* refer to the whole page per definition, and hence, they comprise solely page flaws. Except for the type *Verifiability*, all flaw types comprise more page flaws than inline flaws. The preponderance of inline flaws in the flaw type *Verifiability* indicates that this type is better elaborated than others because the respective flaws are defined in a more specific manner. Moreover, for some flaws exist a page version and an inline version. Consider for instance the page flaw *Disputed* and the inline flaw *Disputed-inline*. The former states that the article's factual accuracy is disputed, whereas the latter refers to a disputed statement or alleged fact.

Although the analyses in this chapter target the English Wikipedia, it is worth mentioning that the distinction of tag boxes and inline tags applies to the major Wikipedia language editions. An exception is the German Wikipedia, where the application of inline tags is still a subject of ongoing discussions.[13]

[13] Discussion in the German Wikipedia about the usefulness of inline tags and their general applicability to tag quality flaws: `http://de.wikipedia.org/wiki/Wikipedia_Diskussion:Meinungsbilder/Vorlage_zur_Markierung_von_Belegmängeln` (in German).

2.3.3 Flaw Commonness

Certain quality flaws are more common than others. We quantify the commonness of a flaw by the number of times the respective cleanup tag occurs in Wikipedia. We therefore investigate the incidences of the 445 cleanup tags in the 25 981 062 pages of the Wikipedia snapshot. Our local Wikipedia database provides all necessary information; in particular, we use the database tables *templatelinks* and *page* (cf. Section 2.1).

The Wikipedia snapshot contains 1 848 081 instances of the 445 cleanup tags. This means, each quality flaw has been tagged on average 4 152.99 times. The actual distribution of instances per flaw, however, is very skewed. Table 2.7 shows the distribution of the flaws over five discrete commonness classes. The commonness of a flaw gives some indication of its importance and also of the usefulness of the respective cleanup tag. Of the 445 flaws, 16 have not been tagged at all. The majority of the flaws (78.20%) have been tagged more than once but less than 1 000 times, and hence, these flaws are considered as relatively uncommon. Moreover, 60 flaws belong to the third category, their frequency is one order of magnitude higher compared to the flaws in the second category. We consider 14 flaws as very common, the respective cleanup tags occur between 10 000 and 100 000 times. Finally, there are seven highly used flaws which are tagged more than 100 000 times. Table 2.8 describes the seven highly used flaws. The table shows the frequency values for each flaw, i.e., the number of times the respective cleanup tag occurs in Wikipedia. (The listing in Appendix A shows the frequency values for each of the 445 flaws.) The most tagged flaw is *Copy to Wikimedia Commons*, the respective cleanup tag occurs 262 753 times, which corresponds to 14.22% of all tagged flaws. The facts that this is a page flaw and that it belongs to the flaw type *Unwanted content* mean that there are 262 753 pages that are not appropriate for Wikipedia. Altogether, the seven most common flaws account for 67.72% of all tagged flaws.

Table 2.7: Distribution of the 445 quality flaws over five commonness classes. A flaw is assigned to a class if its frequency, i.e., the number of times the respective cleanup tag occurs in Wikipedia, conforms to the values shown in parentheses.

Not used	**Uncommon**	**Common**	**Very common**	**Highly used**
(0)	(1–999)	(1 000–9 999)	(10 000–99 999)	(> 100 000)
16	348	60	14	7

In the previous section, we already discussed Table 2.6, which breaks down the frequencies of the 445 flaws by flaw type and flaw scope. Page flaws are more common than inline flaws, of the 1 848 081 tagged flaws 1 415 272 (76.58%)

are tagged with a page flaw. The ratio between tagged page flaws and tagged inline flaws is about 3 : 1, which roughly matches the ratio between existing page flaws and existing inline flaws. Looking at the corresponding values for each flaw type shows that the respective ratios do not match in general. Consider for instance the flaw type *Style of writing*: the frequency of tagged page flaws and tagged inline flaws is roughly the same, but there are nearly twice as much page flaws than inline flaws. Another example is the flaw type *Verifiability*: although it comprises more inline flaws than page flaws, the frequency of tagged inline flaws is smaller compared to the frequency of tagged page flaws. Table 2.6 (rightmost column) shows that the flaw type *Verifiability* is the most common one, the 98 flaws belonging to this type have been tagged 925 880 times, which corresponds to 50.10% of all tagged flaws. Remember in this respect that the *Verifiability* type comprises five of the seven most common flaws (cf. Table 2.8). The *Unwanted content* type is the second common one, followed by *Wiki tech*.

Table 2.8: The seven most common quality flaws in the English Wikipedia, along with a description and the respective flaw type, flaw scope, and frequency.

Flaw name	**Description**	**Type** **Scope**	**Frequency**
Copy to Wikimedia Commons	The article is a candidate to be copied to Wikimedia Commons.	*Unwanted content* Page flaw	262 753
Unreferenced	The article does not cite any references or sources.	*Verifiability* Page flaw	253 153
Citation needed	The claim is doubtful and lacks a citation to a reliable source.	*Verifiability* Inline flaw	236 589
Orphan	The article has fewer than three incoming links.	*Wiki tech* Page flaw	157 727
Refimprove	The article needs additional citations for verification.	*Verifiability* Page flaw	128 998
Image requested	The article needs an image or photograph to improve its quality.	*Miscellaneous* Page flaw	112 009
Dead link	The external link has become irrelevant or broken.	*Verifiability* Inline flaw	100 313

2.4 Extent of Flawed Content

So far we have analyzed the nature of the existing quality flaws. This section focuses on the content that has been tagged with these flaws. To quantify the extent of the flawed content, we investigate the incidence of cleanup tags over Wikipedia's namespaces and over Wikipedia's main topics. Moreover, we propose a measure to estimate the actual frequency of a flaw.

2.4.1 Flaw Distribution over Namespaces

The MediaWiki software provides the concept of namespaces as a means to organize pages from a technical point of view. At the time of writing this thesis, the English Wikipedia provides ten basic namespaces, which come along with an associated talk namespace for user discussion pages.[14] Figure 2.2 shows the 20 namespaces grouped into three content types: encyclopedia, organization, and community.

The namespace "Main" contains the encyclopedic content of Wikipedia. An important subset of this namespace is called *articles*, also known as proper content pages. A page is considered as an article if it fulfills the following three conditions: 1. it belongs to the namespace "Main", 2. it is not a redirect page, and, 3. it contains at least one wiki link.[15] A redirect page does not contain any content at all, but is used to provide an alternative title for another page and to forward the reader to this page. Pages without wiki links are usually pretty short and simple pages. The content type organization provides namespaces that contain meta pages. Meta pages are used to organize the encyclopedic content ("Portal", "Category", and "Book"), to provide policies and support information ("Wikipedia" and "Help"), to describe non-textual content ("File"), and to handle technical and administrative stuff ("Template" and "MediaWiki"). The content type community provides the talk namespaces and the namespace "User". These namespaces contain pages that relate to the community of Wikipedia authors. Each basic namespace has an associated talk namespace, indicated by the suffix "talk". An exception is the talk namespace "Talk" that is associated with the basic namespace "Main". A talk namespace contains the discussion pages for the regular pages of the respective basic namespace. Registered users are allowed to maintain personal pages, which in turn belong to the namespace "User". The namespace membership of a page is encoded as prefix in the page title (followed by a colon); for example, ”Talk:Main Page" or ”Category:History". Each page belongs to exactly one namespace. A page without prefix belongs to the namespace "Main".

[14] There are also two virtual namespaces: "Media" and "Special". "Media" refers to non-textual content, like images, videos, or audio files. "Special" refers to pages that are created on demand; an example is the page " Special:WhatLinksHere". These two namespaces are not considered here because they do not relate to pages stored in the Wikipedia database.

[15] We use the "automatic definition" of an article, which is also used in the MediaWiki software: Wikipedia, "Wikipedia:What is an article?," last modified October 30, 2012, `http://en.wikipedia.org/wiki/Wikipedia:What_is_an_article?`.

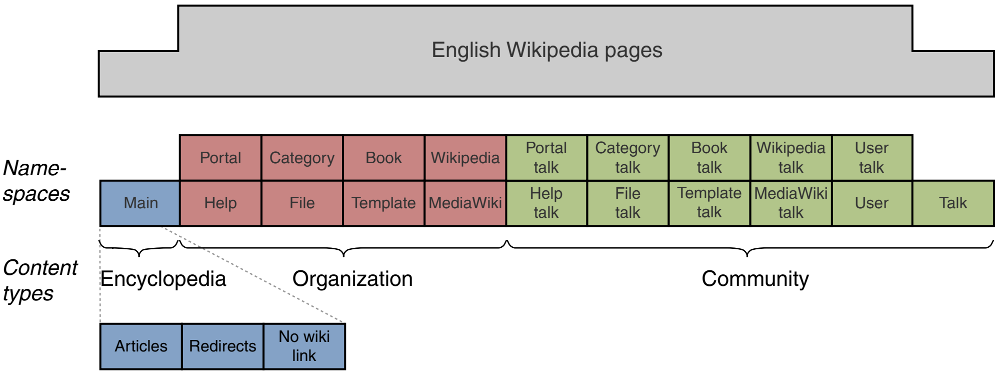


Figure 2.2: The English Wikipedia pages are organized into 20 namespaces. The namespace "Main" is subdivided into articles, redirects, and pages that do not contain any wiki link. In this thesis, we group the namespaces into the three content types encyclopedia, organization, and community.

Table 2.9 shows the distribution of pages and tagged pages over the namespaces. The Wikipedia snapshot comprises a total of 25 981 062 pages, of which 1 472 442 have been tagged with at least one of the 445 cleanup tags, which corresponds to 5.67%. The largest namespace is "Main", which comprises 9 081 332 pages (34.95%), including 3 865 587 articles. The namespaces in the content types organization and community comprise 10.97% and 54.08% of the pages respectively. The majority (70.52%) of the tagged pages belong to the namespace "Main", and nearly all tagged pages in this namespace are articles. Altogether, 26.86% of the articles have been tagged. Another 9.87% of the tagged pages belong to the namespace "Talk". The majority (94.75%) of these pages are associated with articles, and 92.55% of the articles have a talk page. Note that it is unclear whether a cleanup tag on a talk page refers to the content of the talk page itself or to the content of the associated page. There is no general policy that specifies whether cleanup tags should be placed on articles or on the respective talk pages. The same applies for the namespace "File": file description pages account for 18.17% of the tagged pages and the respective cleanup tags might refer to the files themselves or to the descriptions. Furthermore, consider the flaw *Copy to Wikimedia Commons*, which we already discussed above, and, which is the most common one of the 445 flaws (cf. Table 2.8): 98.17% of the tagged pages in the "File" namespace have been tagged with this flaw, and it is more than likely that these tagged flaws refer to the files themselves since Wikimedia Commons is a database for media files.[16] The "File" namespace also shows the largest tagged page ratio compared to the other namespaces, 32.19% of the pages in this namespace are tagged.

[16] Wikimedia Commons: `http://commons.wikimedia.org`.

Table 2.9: For Wikipedia's namespaces: the total number of pages, the number of pages that have been tagged with at least one cleanup tag, and the tagged page ratio.

Namespace		**Pages**	**Tagged pages**	**Ratio in %**
Encyclopedia				
Main	Articles	3 865 587	1 038 338	26.86
	Redirects	5 210 312	22	< 0.01
	No wiki links	5 433	1 140	20.98
Organization				
File		831 288	267 564	32.19
Wikipedia		692 308	2 909	0.42
Template		406 985	617	0.15
Portal		107 048	278	0.26
Category		806 473	104	0.01
Help		908	21	2.31
Book		2 784	14	0.50
MediaWiki		1 572	0	0.00
Community				
Talk		4 359 295	145 379	3.33
User		1 362 204	10 839	0.08
User talk		7 420 828	3 852	0.05
Wikipedia talk		119 330	1 085	0.91
Template talk		137 555	180	0.13
Category talk		493 944	43	0.01
File talk		132 120	31	0.02
Portal talk		21 117	15	0.07
MediaWiki talk		913	5	0.55
Help talk		447	6	1.34
Book talk		2 611	0	0.00
		∑ 25 981 062	∑ 1 472 442	Ø 5.67

The remaining namespaces, beside the mentioned namespaces "Main", "Talk", and "File", contain relatively few tagged pages, the total amount is less than 1.5%. Finally, it can be said that cleanup tagging work in the English Wikipedia mostly targets the encyclopedic content and articles in particular. We restrict the remaining analyses in this thesis to the set of articles because we are particularly interested in quality flaws that occur in proper content pages. These are also the pages that are mostly viewed by typical Wikipedia readers, so that quality flaws in articles must be considered as more serious because they affect a large number of users.

2.4.2 Flaw Distribution over Topics

We utilize the category system of Wikipedia to derive a set of main topics. The category system is a directed cyclic graph with the category “Category:Contents” as the (virtual) root. The root category contains both articles and non-content pages. A subcategory of the root category is “Category:Articles”, which is intended as a starting point for browsing categories that contain only articles. This category has two relevant subcategories: 1. ”Category:Fundamental categories”, whose subcategories represent fundamental areas of human knowledge, and, 2. ”Category:Main topic classifications”, whose subcategories are organized thematically and reflect more detailed fields of knowledge. It is guaranteed that any article can be reached through either of these two categories. We use the 24 direct subcategories of ”Category:Main topic classifications” as main topics; see the leftmost column of Table 2.10.

To identify the main topics of an article, we traverse the category graph bottom-up starting from the categories that are associated with the article. The traversal ends if one path reaches a main topic. Note that several paths may reach different main topics by traversing the same number of categories. In this case, a single article has multiple main topics. For example, the article “Algorithm” belongs to the topics “Science” and “Mathematics”. The majority (64.9%) of the articles has exactly one main topic.

Table 2.10 shows the distribution of articles as well as of tagged articles over the topics. A good deal of the articles belong to the topic “Chronology”. These articles describe time periods such as years (e.g., “1895 BC”), days of the year (e.g., “July 4”), and decades (e.g., “70s”). A large number of articles also belong to the topics “People” and “Geography”. These articles mainly describe individuals (e.g., “Albert Einstein”) and places (e.g., “Badwater, California”) respectively. The ratio of tagged articles varies widely among the topics. This is in line with the findings of Ehmann et al. [52], who report on variability in article quality across diverse disciplines. The largest proportions of tagged articles are in the topics “Computers” (50.07%) and “Belief” (46.37%). One possible explanation for the unequal proportions of tagged articles among the topics is that certain topics, such as “Computers”, are more popular than others, and hence, the respective articles are more likely to be evaluated with respect to flaws. Another explanation is that the articles in controversial topics such as “Belief” are more likely to be challenged than those with more agreement such as “Geography”.

Table 2.10: For Wikipedia's main topics: the total number of articles, the number of articles that have been tagged with at least one cleanup tag, and the ratio of tagged articles. The rows are ordered by the number of tagged articles.

Topic	Articles	Tagged articles	Ratio in %
Chronology	1 005 168	260 182	25.88
People	837 471	243 579	29.09
Geography	808 437	155 674	19.26
Society	402 858	143 558	35.63
Culture	354 952	124 738	35.14
Arts	303 027	99 904	32.97
Technology	244 084	87 088	35.68
Humanities	219 287	77 272	35.24
History	233 273	70 215	30.10
Politics	216 453	63 513	29.34
Business	154 196	61 820	40.09
Life	209 203	57 001	27.25
Nature	208 555	48 870	23.43
Education	120 252	45 696	38.00
Applied sciences	113 473	39 655	34.95
Health	91 291	32 324	35.41
Science	127 268	31 022	24.38
Environment	128 067	28 930	22.59
Language	70 298	23 057	32.80
Agriculture	101 580	22 083	21.74
Law	56 135	19 593	34.90
Computers	22 645	11 339	50.07
Belief	24 265	11 252	46.37
Mathematics	27 893	8 051	28.86

2.4.3 Estimating the Actual Extent of Flawed Content

So far we have analyzed the extent of flawed content by means of tagged flaws. However, the number of tagged flaws underestimates the actual extent of flawed content since it is more than likely that many flaws are not yet identified. As already mentioned, the size and the dynamic nature of Wikipedia render a comprehensive manual quality assurance unfeasible. Stated formally: Let D be the set of the 3 865 587 Wikipedia articles and let $D^- \subset D$ be the 1 038 338 tagged articles, see Figure 2.3. We have no information about the remaining articles $D \setminus D^-$, these articles are either flawless or have not yet been evaluated. The same applies to each single flaw $f_i \in F$, where F denotes the 445 flaws. It is unclear whether the articles in $D^- \setminus D_i^-$ either do not contain f_i, or if they have not been evaluated yet with respect to f_i, where

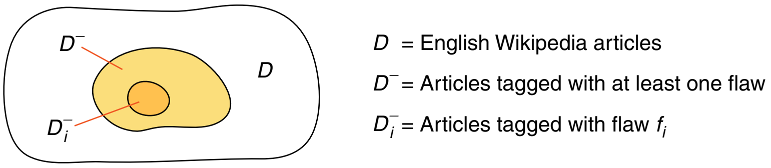

D^-
D
D_i^-
D = English Wikipedia articles
D^- = Articles tagged with at least one flaw
D_i^- = Articles tagged with flaw f_i

Figure 2.3: Sets of Wikipedia articles that we distinguish in this thesis.

$D_i^- \subset D^-$ denotes the articles that have been tagged with f_i. The number of tagged articles $|D_i^-|$, however, can be considered as a lower bound of the actual frequency of a flaw f_i. Consequently, the statistics reported above quantify the minimum extent of flawed content, whereas the actual extent is expected to be even higher.

In order to estimate the actual frequency of a flaw f_i we make two assumptions:

1. each article in D^- is tagged completely, i.e., with all flaws that it contains (Closed World Assumption), and
2. the distribution of f_i in D^- is identical to the distribution of f_i in D.

The first assumption ensures for each flaw $f_i \in F$ that D_i^- comprises all articles $d \in D^-$ that suffer from f_i. Put another way, all articles in $D^- \setminus D_i^-$ are flawless with respect to f_i. The second assumption ensures that D^- is a representative sample of D, and thus, the flaw distribution in D^- can be generalized to the whole Wikipedia. Based on these assumptions, we estimate the actual frequency of a quality flaw f_i by the ratio of articles in D_i^- and articles in D^-, which corresponds to the ratio of flawed and flawless articles.

Table 2.11 (rightmost column) lists the estimated flaw ratios for the ten flaws that have the highest number of tagged articles. (The listing in Appendix A shows the flaw ratios for each of the 445 flaws.) The page flaw *Unreferenced* is the most frequent one according to our estimate. Of the 1 038 338 tagged articles, 251 447 have been tagged with the *Unreferenced* flaw, which corresponds to a ratio of about 1 : 4 (251 447 : 1 038 338). In other words, about every fifth English Wikipedia article is expected to contain this flaw, and hence, does not cite any references or sources. Six of the ten flaws that are listed in Table 2.11 belong to the flaw type *Verifiability*, i.e., the tagged articles have missing or inadequate references and sources.

Our estimate is based on the number of tagged articles, and hence, it does not reflect the fact that an individual flaw may be tagged several times in a single article. However, this only applies for inline flaws since page flaws refer to the whole article per definition. Among the ten flaws that are listed in Table 2.11 are two inline flaw, *Citation needed* and *Dead link*. Both flaws can occur several times within a single article as they refer to an individual claim or link respectively. Because of this, the flaw ratio underestimates the actual frequency of inline flaws. Consider for instance the inline flaw *Citation needed*: its estimated ratio is 1 : 5, which actually means that every sixth article is expected to contain *at least one* instance of this flaw.

Table 2.11: The ten quality flaws with the highest estimated flaw ratios in the English Wikipedia, along with the respective description, flaw type, flaw scope, number of tagged articles, and flaw ratio.

Flaw name	**Description**	**Type** **Scope**	**Tagged articles**	**Flaw ratio**
Unreferenced	The article does not cite any references or sources.	*Verifiability* Page flaw	251 447	1 : 4
Citation needed	The claim is doubtful and lacks a citation to a reliable source.	*Verifiability* Inline flaw	218 612	1 : 5
Orphan	The article has fewer than three incoming links.	*Wiki tech* Page flaw	157 066	1 : 7
Refimprove	The article needs additional citations for verification.	*Verifiability* Page flaw	127 667	1 : 8
Dead link	The external link has become irrelevant or broken.	*Verifiability* Inline flaw	99 012	1 : 10
BLP sources	The biographical article needs additional citations for verification.	*Verifiability* Page flaw	47 817	1 : 22
Multiple issues	The article has multiple maintenance issues.	*General cleanup* Page flaw	43 029	1 : 24
Empty section	The article has at least one section that is empty.	*Expand* Page flaw	43 016	1 : 24
Notability	The article does not meet the general notability guideline.	*Unwanted content* Page flaw	35 183	1 : 30
No footnotes	The article's sources remain unclear because of its inline citations.	*Verifiability* Page flaw	30 773	1 : 34

Chapter 3

Quality Flaw Evolution

The previous chapter investigates quality flaws in current Wikipedia articles, based on a snapshot representing the state of Wikipedia at a certain time. This chapter goes further, and investigates the evolution of quality flaws in the entire history of Wikipedia. In particular, we utilize the 445 cleanup tags that have been identified in Chapter 2 to analyze the quality flaws that have been tagged by the Wikipedia community in the English Wikipedia from its launch in January 2001 until January 2012. During the analysis period, the English Wikipedia has received more than 508 million edits (cf. Table 2.1). Each edit produces a new revision of the edited page, and the contents of all revisions sum up to 7.9 TB. The complete revision history is stored in the Wikipedia database, however, the huge amount of data and the fact that a direct unrestricted database access is in general not available poses particular challenges for a comprehensive and efficient analysis.

Chapter organization. Section 3.1 gives a comparative overview of different approaches to access the Wikipedia database and describes the methods we used to process the database dumps and to identify tagged revisions of Wikipedia articles. Section 3.2 reveals how the number and the kind of cleanup tags have evolved. Section 3.3 addresses the usage of cleanup tags and gives insights into how the extent of (tagged) quality flaws has evolved. Section 3.4 investigates how quickly tagged quality flaws have been corrected, which gives some indication of the flaws' complexity.

Key contributions. We present the first comprehensive breakdown of the evolution of quality flaws in English Wikipedia articles. Our findings yield interesting insights regarding 1. the development of Wikipedia's quality flaw structure and 2. the usage and the effectiveness of cleanup tags.

3.1 Identifying Tagged Revisions

The 445 cleanup tags that existed at January 4, 2012 provide the basis for our analysis (cf. Section 2.1 and Section 2.2). We do not consider cleanup tags that may have existed at some point in the past and that have been deleted. Similarly, we consider only those pages that existed at January 4, 2012, and we investigate their revisions up to this date.

3.1.1 Accessing the Wikipedia Database

There are different possibilities to access the Wikipedia database, see Figure 3.1. The MediaWiki software provides three basic approaches: First, the standard Web interface, which is implemented in the main script of the MediaWiki software `index.php`.[1] Second, the MediaWiki Web service API, which is implemented in `api.php` and provides access for scripts and bots.[2] Third, the page "Special:Export", which uses `dumpBackup.php` to export pages to XML.[3] The database backup dumps, which we already used in Section 2.1, provide another (indirect) way to access the Wikipedia database. As already mentioned, backup dumps are compiled in regular time intervals and correspond to a snapshot of the Wikipedia database at a certain time. Further, the Wikimedia Toolserver provides access to a replication of the Wikipedia database.[4]

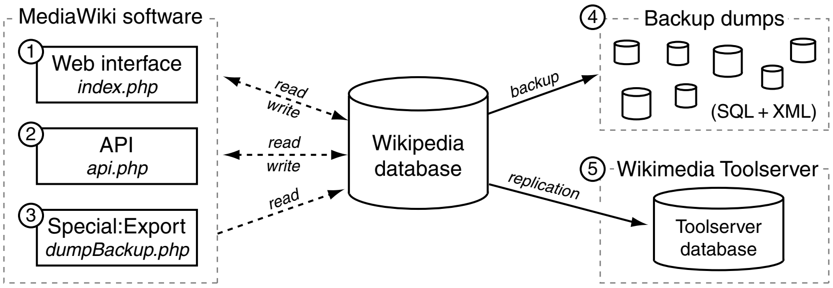


Figure 3.1: Five approaches to access the Wikipedia database: 1. the Web interface, 2. the MediaWiki API, 3. the page "Special:Export", 4. the backup dumps provided by the Wikimedia Foundation, and, 5. the Wikimedia Toolserver.

[1] For further information on the `index.php` script, refer to the MediaWiki manual: `http://www.mediawiki.org/wiki/Manual:Index.php`.

[2] For an introductory overview of the MediaWiki API, refer to `http://www.mediawiki.org/wiki/API:Main_page`.

[3] A description of the XML format as well as of the export procedure is available at: `http://meta.wikimedia.org/wiki/Help:Export`.

[4] The Toolserver is operated by Wikimedia Deutschland, which is the German chapter of the Wikimedia Foundation. For further information, refer to `http://www.toolserver.org`.

Table 3.1: Comparison of the five approaches to access the Wikipedia database that are depicted in Figure 3.1 using seven criteria, which are quantified as follows: the criteria applies (✓), it applies only partially (○), and it does not apply (✕).

Criterion	**Database access methods**				
	Web interface	API	Special: Export	Backup dumps	Wikimedia Toolserver
Read access	✓	✓	✓	✓	✓
Write access	✓	✓	✕	✕	✕
Scalable	✕	✕	✕	✓	○
Page content	✓	✓	✓	✓	✕
Open access	✓	✓	✓	✓	○
Different formats	✕	✓	✓	✕	✕
Current data	✓	✓	✓	✕	○

The mentioned database access approaches have advantages and disadvantages. We have compiled seven criteria to assess the existing approaches and to identify the one that is most appropriate for the task at hand, see Table 3.1. All approaches provide read access to the Wikipedia database, and only the Web interface as well as the API also provide write access. The three MediaWiki methods are suitable to access smaller amounts of data; for instance, to get meta information for a certain set of pages. However, these methods do not scale when a huge amount of data is required; for instance, the revision history of all pages. The Wikimedia servers are not designed to deliver large data volumes, and therefore, the three MediaWiki methods are subject to certain limitations. For instance, the maximum number of page revisions that are returned by the MediaWiki API is limited to 500 for a single query. The database dumps are most suitable for comprehensive analyses as they can be processed locally using any scalable technology. The scalability of the Toolserver is given only partially because the resources must be shared with all users.

Except for the Toolserver, all approaches provide access to the content of the Wikipedia pages, i.e., the pages' wiki markup sources. The Toolserver's replication process does not include the page contents of all revisions. The MediaWiki methods and the database dumps are publicly available, whereas using the Toolserver requires a personal account. Toolserver accounts must be requested providing a justification for usage, and the requests need to be approved by Toolserver staff. The MediaWiki API and the "Special:Export" page provide various output formats, including JSON, serialized PHP, and XML, whereas the other approaches are limited to predefined output formats. The MediaWiki methods provide access to current Wikipedia data. By contrast, the database dumps represent a snapshot of the Wikipedia database at a certain time. The Toolserver database cannot be considered a real-time copy of the Wikipedia database because of replication lags (typically only a few seconds), so that the Toolserver database may represent a state at some point in the past.

Here, we need to analyze the content of each page revision in the English Wikipedia, and hence, scalability as well as the availability of the page content are the most determining criteria. Open access is also important as it ensures the reproducibility of our results. A specific output format is not required, and we need neither access to the latest Wikipedia data nor write access to the database. Consequently, the database backup dumps are most appropriate for the analyses in this chapter.

3.1.2 Processing the Database Dumps

The Wikimedia Foundation provides two kinds of database backup dumps: SQL dumps and XML dumps. We already used the SQL dumps in Chapter 2 to create a local partial copy of the Wikipedia database (cf. Section 2.1), which is used to analyze the state of Wikipedia at a certain time. However, the SQL dumps contain no revision history information because the relevant Wikipedia database tables *revision* and *text* are not dumped directly.[5] The data of the two tables is provided as XML dumps, which use the same XML wrapper format that the aforementioned "Special:Export" produces for individual pages.[6] In order to augment our local copy of the Wikipedia database, we retrieve the relevant information to create the table *revision* from the XML dump *stub-meta-history*. This dump contains various meta information about each revision, including the page identifier, the size of the revision in byte, the user who performed the edit, and an optional editor comment. The size of the uncompressed dump amounts to 127 GB. We process the dump on a 44-node Apache Hadoop cluster using MapReduce.[7] In particular, we implement a tailored XML parser in the map phase, which extracts the relevant revision information from each `<page>` element.

In order to identify those revisions that have been tagged with some cleanup tag, we need to process the revisions' actual content. The raw wiki markup of every revision is comprised in the XML dump *pages-meta-history*. The size of the uncompressed dump amounts to 7.9 TB. We process the dump on the

[5] Wikimedia, "Data dumps," last modified October 23, 2012, `http://meta.wikimedia.org/wiki/Data_dumps#What_happened_to_the_SQL_dumps.3F`.

[6] The XML schema of MediaWiki's page export format is available at: `http://www.mediawiki.org/xml/export-0.3.xsd`.

[7] For details on Apache Hadoop, refer to `http://hadoop.apache.org`. For a description of the MapReduce programming model, refer to [50].

Hadoop cluster as well. However, we had to adapt our parsing approach because there are pages with more than 40 000 revisions, which cannot be handled efficiently by a single map job.[8] We therefore implemented a tailored Wikipedia revision input format, so that one revision is processed by a single mapper which parses its wiki markup using regular expressions to identify cleanup tags. Cleanup tags are included in a page using the template syntax[9]:

```
{{template name | parameter 1 | parameter 2 | ...}}
```

Whereas the template name corresponds to the page title of a cleanup tag. As already mentioned, a cleanup tag may have several alternative titles linking to it through redirects. We resolve all redirects using the approach described in Section 2.2. Parsing the 7.9 TB lasts about five hours. Altogether, we identified 103 575 646 revisions that have been tagged with at least one of the 445 cleanup tags, which corresponds to 23.49% of all revisions. As motivated previously, this thesis targets quality flaws in articles: 92.2% of the tagged revisions corresponds to articles, and of the article revisions 32.91% has been tagged.

3.1.3 Filtering Vandalized Revisions

Vandalism edits interfere the identification of tagged revisions. This is especially the case when analyzing sequences of tagged revisions, for instance, to determine the correction time of a flaw. We identify such sequences by traversing the revisions of a single article in chronological order using the timestamps provided in the *revision* table of the local Wikipedia database. The duration between the date when an article has been tagged with a certain cleanup tag and the date when the tag has been removed is considered as the correction time of this particular flaw. Although vandalism is repaired relatively quickly by the Wikipedia community [152], the vandalized revisions distort our analyses. For example, a common type of vandalism is the deletion of article content. If an existing cleanup tag has been deleted due to a vandalism edit, our approach incorrectly assumes that the respective flaw has been fixed—even if the vandalism is repaired in a subsequent revision. To prevent such issues, we do not consider empty revisions in our analyses. Furthermore, we discard revisions that stem from the following types of edits:

- *Anonymous edits.* It has been shown that the majority of vandalism edits is caused by anonymous users (editing without prior registration), whereas the amount of serious anonymous edits is relatively small [16, 136]. The

[8] As of January 4, 2012, "George W. Bush" is the most edited article, with 44 655 revisions.

[9] For further information on templates, refer to the respective Wikipedia help page: `http://en.wikipedia.org/wiki/Help:Template`.

information whether an edit has been made by an anonymous user or by a registered user is stored in the *revision* table of the Wikipedia database, which provides either the IP address of the user's host machine (anonymous) or the Wikipedia username (registered).

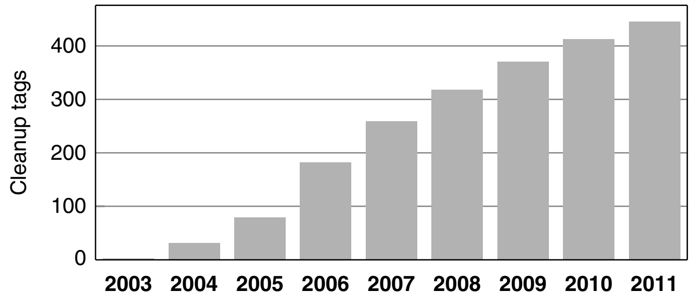
- *Vandalism edits identified by users.* It has become common practice in the Wikipedia community to use dedicated keywords in the editor comment if a vandalism edit has been reverted. We compiled the following list of keywords: "revert", "rv" (revert), "rvv" (revert due to vandalism), "vandalism", "spam", "undid", and "rollback". We consider an edit as vandalism if the subsequent revision's editor comment contains one of the keywords from our list or any combination respectively. A similar approach is used by Suh et al. [145]. Editor comments are stored in the *revision* table of the Wikipedia database.
- *Vandalism edits identified by anti-vandalism bots.* Anti-vandalism bots are programs that operate under a common Wikipedia user account and repair certain types of vandalism autonomously. We consider an edit as vandalism if it has been reverted by one of the 13 bots listed in the Wikipedia category "Category:Wikipedia anti-vandal bots".[10] We identify such edits by the bots' user names.

We are confident that we are able to identify the majority of vandalized revisions using the mentioned heuristics. Note in this respect that vandalism detection in Wikipedia is a separate research area, see for instance Potthast et al. [123, 124], and that the detection of vandalism in general is beyond to scope of this thesis. Without considering vandalized revisions, we identified 53 715 900 article revisions that have been tagged with at least one of the 445 cleanup tags. These revisions provide the basis for the analyses in this chapter.

3.2 Evolution of Cleanup Tags

This section investigates the question of when did the first cleanup tags emerge, and how have the number and the kind of tags changed over time. We will not show individual results for each of the 445 cleanup tags since this would be less insightful. Instead, we use the organization of the cleanup tags into flaw scope and flaw type, which we proposed in Section 2.3. Nevertheless, we will discuss interesting particularities of individual cleanup tags where it is appropriate.

[10] Wikipedia, "Category:Wikipedia anti-vandal bots," last modified May 14, 2011, `http://en.wikipedia.org/wiki/Category:Wikipedia_anti-vandal_bots`.



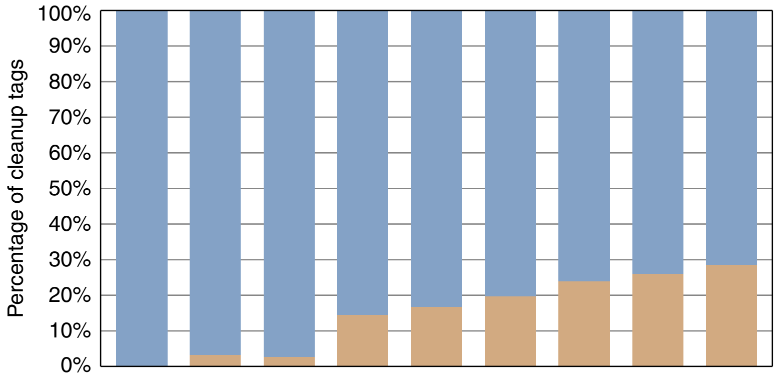
	2003	2004	2005	2006	2007	2008	2009	2010	2011
Absolute	2	31	79	182	259	318	370	412	445
Increase	2	29	48	103	77	59	52	42	33

Figure 3.2: Number of existing cleanup tags per year in the English Wikipedia. The plot shows the absolute values; the table below lists the absolute values and the increase compared to the respective previous year (in gray).

The first cleanup tags were *Disputed* and *POV*, which have been created in December 2003. The former refers to an article's factual accuracy, the later stands for "point of view" and refers to missing neutrality. Since 2003, the number of cleanup tags grows steadily, see Figure 3.2. In 2006, 103 new tags have been created, so that the absolute number of tags more than doubled to 182. A reason for the significant increase in 2006 might be the increasing awareness of cleanup tags in the Wikipedia community due to the creation of the page "Template messages/Cleanup" in the previous year. This page comprises a manually maintained listing of cleanup tags.[11] More than half of the 445 cleanup tags already existed in 2007. Since 2007, the number of newly created tags per year declines. In 2011, only 33 new tags were created. If this trend continues, a stable number of cleanup tags is to be expected within the next few years. An explanation for this development is that the set of relevant quality flaws occurring in Wikipedia is at some point covered by the existing cleanup tags.

In the previous chapter we have organized quality flaws along the two scopes page flaw and inline flaw (cf. Section 2.3.2). Remember in this regard that a cleanup tag is either a tag box that refers to the whole article or an inline tag that refers to a particular text fragment. Tag boxes define page flaws and inline tags define inline flaws. Figure 3.3 shows the number of cleanup tags per year broken down into the two flaw scopes. The two cleanup tags that have been created in 2003 are tag boxes. The first inline tag was *Dubious*, it has been created in July 2004. *Dubious* is the inline version of the tag box

[11] We already used this page as a source in our cleanup tag mining approach, described in Section 2.2 (step 1).

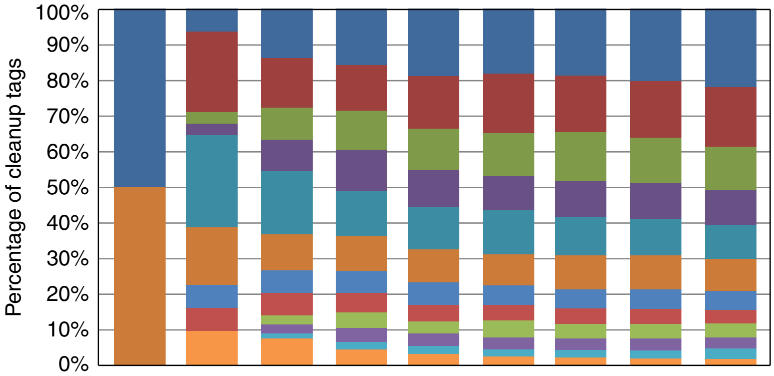


Flaw scope	2003	2004	2005	2006	2007	2008	2009	2010	2011
▬ Page flaw	2	30	77	156	216	256	284	305	318
	2	28	47	79	60	40	28	21	13
▬ Inline flaw	0	1	2	26	43	62	86	107	127
	0	1	1	24	17	19	24	21	20

Figure 3.3: Number of existing cleanup tags per year in the English Wikipedia broken down into the two flaw scopes page flaw and inline flaw. The plot shows the percentages; the table below lists the absolute values along with the increase compared to the respective previous year (in gray).

Disputed and refers to a specific statement or alleged fact that is subject to dispute. *Citation needed* was the second inline tag, it has been created in June 2005. At the end of 2005, there were already 77 tag boxes and only 2 inline tags. In 2006, the number of inline tags increased substantially, and also the number of tag boxes experienced its largest increase. Since 2006, the number of newly created inline tags per year was nearly constant, whereas the number of newly created tag boxes declined. Consequently, the percentage of inline tags has increased over the last years, and at the end of 2011 it corresponded to 28.54%. A large part of the newly created inline tags are inline versions of the existing tag boxes. We therefore expect the number of newly created inline tags per year to stagnate as well in the near future, when for each relevant quality flaw both a page version and an inline version exist. However, some flaws refer to the whole page per definition (e.g., *Lead missing*), and hence, it is not possible to create an inline version for respective tag boxes.

Figure 3.4 shows the development of cleanup tags broken down into the twelve flaw types that we introduced in the previous chapter (cf. Section 2.3.1). The earliest two cleanup tags *Disputed* and *POV* belong to the flaw types *Verifiability* and *Neutrality* respectively. The majority (64.52%) of those cleanup tags that existed at the end of 2004 belong to the flaw types *Style of writing*, *Unwanted content*, and *Neutrality*. This indicates that the Wikipedia community perceived



Flaw type	2003	2004	2005	2006	2007	2008	2009	2010	2011
Verifiability	1	2	11	29	49	58	69	84	98
	1	1	9	18	20	9	11	15	15
Style of writing	0	7	11	23	38	53	59	65	74
	0	7	4	12	15	15	6	6	9
Miscellaneous	0	1	7	20	30	38	51	52	54
	0	1	6	13	10	8	13	1	2
Specific subjects	0	1	7	21	27	31	37	42	44
	0	1	6	14	6	4	6	5	2
Unwanted content	0	8	14	23	31	39	40	42	42
	0	8	6	9	8	8	1	2	0
Neutrality	1	5	8	18	24	28	35	39	40
	1	4	3	10	6	4	7	4	1
Wiki tech	0	2	5	11	16	17	20	23	24
	0	2	3	6	5	1	3	3	1
General cleanup	0	2	5	10	12	14	16	17	17
	0	2	3	5	2	2	2	1	0
Expand	0	0	2	8	9	15	15	17	17
	0	0	2	6	1	6	0	2	0
Structure	0	0	2	7	9	11	12	14	14
	0	0	2	5	2	2	1	2	0
Time-sensitive	0	0	1	4	6	6	8	9	13
	0	0	1	3	2	0	2	1	4
Merge	0	3	6	8	8	8	8	8	8
	0	3	1	2	0	0	0	0	0

Figure 3.4: Number of existing cleanup tags per year in the English Wikipedia broken down into the twelve flaw types. The plot shows the percentages; the table below lists the absolute values along with the increase compared to the respective previous year (in gray). The order of the flaw types is the same in the plot and in the table.

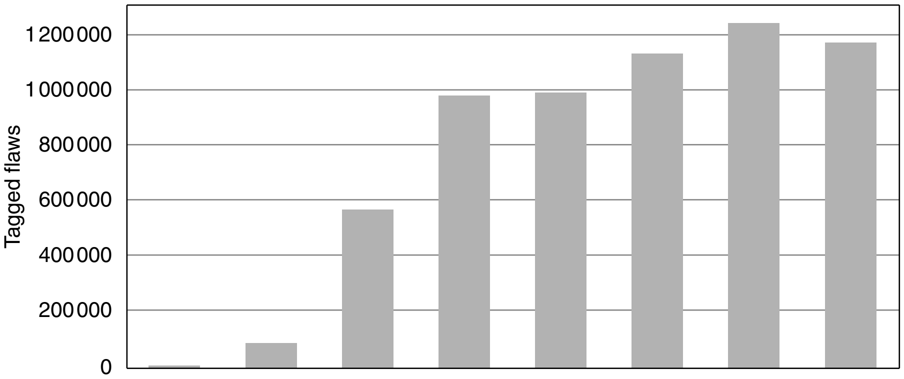
these flaw types as relatively serious in the initial phase of the Wikipedia project, and it also suggests that early articles suffered from low writing style, lacking notability, and missing neutrality. In 2005, at least one cleanup tag existed for each flaw type. The *Unwanted content* type still comprised the largest number of flaws at the end of 2005. This suggests that the implementation of Wikipedia's inclusion criteria, which have been defined already in September 2001, was still one of the primary issues.[12] Also in 2005, the flaw type *Verifiability* gained particular importance: the number of cleanup tags that belong to this flaw type increased by 18 and 20 in the years 2006 and 2007 respectively, which correspond to the largest yearly increases over all types and years. Since 2007, the *Verifiability* type comprised the largest number of cleanup tags of all flaw types. Already in 2006, certain flaw types emerged that comprised more cleanup tags than other, including *Verifiability*, *Style of writing*, *Miscellaneous*, *Specific subject*, *Unwanted content*, and *Neutrality*. Especially the types *Verifiability*, *Unwanted content*, and *Neutrality* are of particular interest as they directly refer to Wikipedia's core content policies, which have been defined in January 2006 as follows: verifiability, no original research, and neutral point of view (cf. Section 1.4.2).

Although, the total number of newly created cleanup tags per year decreased since 2006 (cf. Figure 3.2), certain flaw types experienced an interim increase. For instance, the number of newly created cleanup tags that belong to the *Verifiability* type increased in the years 2009 and 2010; see Figure 3.4. Despite the interim increases, each flaw type showed a significant decline in the number of newly created cleanup tags since 2006, which is another indication that a stable set of cleanup tags is within reach. The set of cleanup tags that belong to the flaw type *Merge*, for instance, is already stable since 2006.

3.3 Evolution of Tagged Quality Flaws

So far we have analyzed the number and the kind of cleanup tags. Now we analyze the usage of cleanup tags and thus the extent of (tagged) quality flaws. The first flaw has been tagged in May 2004 in the article "Stage name" using the cleanup tag *Merge*. Since 2004, the number of tagged flaws per year grew steadily until 2011, see Figure 3.5. Whereas only 11 122 flaws have been tagged in 2004, the number of tagged flaws per year increased rapidly to 91 346 and 578 977 in 2005 and 2006 respectively. The significant increase in 2006 is in line with the development of cleanup tags (cf. Figure 3.2), and, as already mentioned, a possible explanation is the increasing awareness of cleanup tags due to the creation of the overview page "Template messages/Cleanup". In 2008, the number of tagged flaws per year came close to 1 million, however, this year showed the lowest annual increase since 2004. 2011 was the first year in which, compared to the previous year, fewer flaws have been tagged. The fact that, since 2007, about 1 million flaws have been tagged annually indicates that the tagging of quality flaws has become a relevant issue for the Wikipedia

[12] The page "Wikipedia:What Wikipedia is not", which defines Wikipedia's inclusion criteria, was created at September 24, 2001: `http://en.wikipedia.org/wiki/Wikipedia:What_Wikipedia_is_not`.

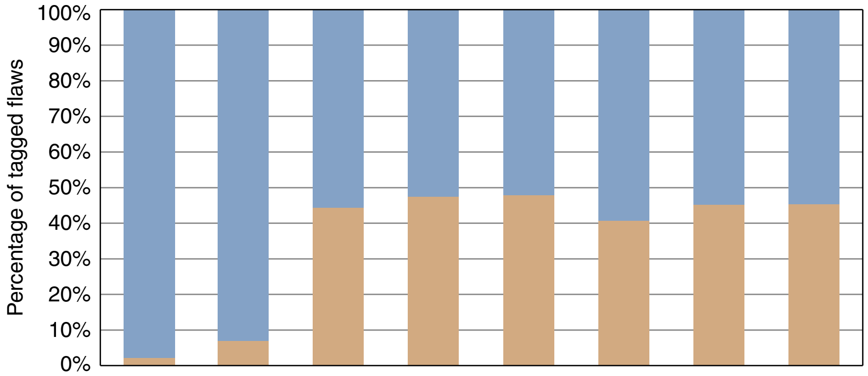


	2004	2005	2006	2007	2008	2009	2010	2011	∑
Absolute	11 122	91 346	578 977	982 682	994 450	1 135 951	1 246 869	1 179 788	6 235 192
Difference	+11 122	+80 224	+487 631	+403 705	+11 768	+141 501	+110 918	-67 081	

Figure 3.5: Number of quality flaws that have been tagged in English Wikipedia articles per year. The plot shows the absolute values; the table below lists the absolute values and the difference compared to the respective previous year (in gray). The rightmost column summarizes the annual absolute values.

community. Altogether, 6 235 192 quality flaws have been tagged from May 2004 until January 2012, which corresponds to 2.15% of all edits that have been made to English Wikipedia articles.

Figure 3.6 breaks down the number of tagged quality flaws per year into the two flaw scopes (described in Section 2.3.2). Of the 6 235 192 tagged quality flaws, 3 465 571 (55.58%) are page flaws and 2 769 621 (44.42%) are inline flaws, i.e., the flaws have been tagged using either a tag box or an inline tag respectively. Only two inline tags existed in 2005 (cf. Figure 3.3), and therefore, it is not surprising that, in 2004 and 2005, page flaws comprised the vast majority of tagged flaws. In 2006, the percentage of tagged inline flaws significantly increased to 44.3%, although, there was still a relatively small number of 26 inline tags compared to the 156 tag boxes that existed in this year (cf. Figure 3.3). Since 2006, the percentage of inline flaws remained relatively constant between 40% and 48%, which means that tag boxes have been used more frequently than inline tags. Note in this respect that placing a tag box at the top of an article is generally easier for a Wikipedia user than identifying and tagging the specific flaws within the text; consider, for instance, the page flaw *Unreferenced* in contrast to the inline flaw *Citation needed*, which we already discussed earlier (cf. Figure 2.1). The table in Figure 3.6 shows that the absolute number of tagged flaws per year grew steadily in both flaw scopes until 2011; except for the year 2009, where fewer inline flaws have been tagged than in the previous year. The annual

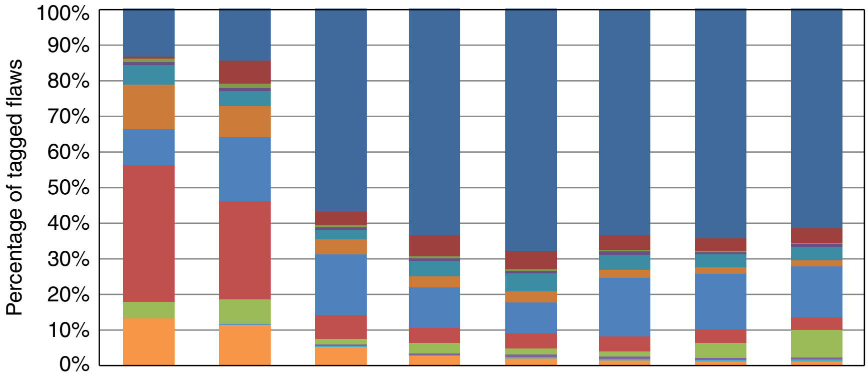


Flaw scope	2004	2005	2006	2007	2008	2009	2010	2011	∑
▬ Page flaw	10912	85592	322497	516536	518541	674942	684190	645507	3465571
	+10912	+74680	+236905	+194039	+2005	+156401	+9248	-38683	
▬ Inline flaw	210	5754	256480	466146	475909	461009	562679	534281	2769621
	+210	+5544	+250726	+209666	+9763	-14900	+101670	-28398	

Figure 3.6: Number of quality flaws that have been tagged in English Wikipedia articles per year broken down into the two flaw scopes page flaw and inline flaw. The plot shows the percentages; the table below lists the absolute values along with the difference compared to the respective previous year (in gray). The rightmost column summarizes the annual absolute values.

numbers of tagged page flaws and tagged inline flaws significantly increased in 2009 and 2010 respectively. In 2011, the annual number of tagged flaws declined in both flaw scopes. Finally, it can be said that tag boxes and inline tags have become a common means to tag flaws in Wikipedia and that both kinds of tags have become almost equally important for the Wikipedia community.

Analog to the previous section, we also breakdown the development of tagged quality flaws into the twelve flaw types (described in Section 2.3.1), see Figure 3.7. In 2004 and 2005, a considerable amount of tagged flaws (45.78% and 27.34% respectively) belonged to the flaw type *General cleanup*, although, the number of respective cleanup tags that existed at this time was relatively small (cf. Figure 3.4). The cleanup tags under the *General cleanup* type merely state that some cleanup is required at all but provide no further information. This shows that the tagging behavior was rather unspecific at the beginning of the Wikipedia project. Since 2006, the specific flaw types gained more and more importance. Most notably is the flaw type *Verifiability*: the respective amount of tagged flaws increased rapidly in 2006 to 55.86%, and, since 2007, the *Verifiability* type comprised between 61% and 68% of the annually tagged flaws. Moreover, it can be seen from the table in Figure 3.7 that the continuously growing number of tagged quality flaws per year (cf. Figure 3.5) was mainly



Flaw type	2004	2005	2006	2007	2008	2009	2010	2011	∑
▬ Verifiability	657	13 274	323 392	639 350	674 225	721 038	**802 242**	725 556	3 899 734
	+657	+12 617	+310 118	+315 958	+34 875	+46 813	+81 204	-76 686	
▬ Style of writing	55	5 890	21 023	**57 046**	50 784	46 164	44 487	47 881	273 330
	+55	+5 835	+15 133	+36 023	-6 262	-4 620	-1 677	+3 392	
▬ Miscellaneous	124	1 111	4 499	**6 582**	6 181	5 481	5 096	4 795	33 869
	+124	+987	+3 388	+2 083	-401	-700	-385	-301	
▬ Specific subjects	91	886	3 483	6 096	6 096	**11 274**	7 436	8 976	44 300
	+91	+795	+2 597	+2 575	+38	+5 178	-3 838	+1 540	
▬ Unwanted content	608	3 763	16 197	42 199	**51 617**	46 758	44 605	44 164	249 911
	+608	+3 155	+12 434	+26 002	+9 418	-4 859	-2 153	-441	
▬ Neutrality	1 388	7 991	23 721	**31 301**	29 382	26 449	23 114	20 916	164 264
	+1 388	+6 603	+15 730	+7 580	-1 917	-2 935	-3 335	-2 198	
▬ Wiki tech	1 126	16 438	98 267	110 890	87 260	187 071	**193 939**	169 018	864 009
	+1 126	+15 312	+81 829	+12 623	-23 630	+99 811	+6 868	-24 921	
▬ General cleanup	5 092	25 019	39 495	41 151	41 901	**46 652**	46 652	40 344	286 371
	+5 092	+19 927	+14 476	+1 656	750	+4 816	-65	-6 308	
▬ Expand	510	6 371	15 275	15 610	16 963	16 399	52 444	**91 843**	215 424
	+510	+5 861	+8 904	+344	+1 344	-564	+36 045	+39 399	
▬ Structure	1	66	1 776	3 820	7 106	**8 347**	7 680	6 882	35 678
	+1	5	+1 710	+2 044	+3 286	+1 241	-667	-798	
▬ Time-sensitive	0	220	1 733	2 666	4 336	4 572	5 546	**6 922**	25 995
	0	+220	+1 513	+933	+1 670	+236	+974	+1 376	
▬ Merge	1 470	10 317	**30 116**	26 000	18 597	15 681	13 628	12 491	128 300
	+1 470	+8 847	+19 799	-4 116	-7 403	-2 916	-2 053	-1 137	

Figure 3.7: Number of quality flaws that have been tagged in English Wikipedia articles per year broken down into the twelve flaw types. The plot shows the percentages; the table below lists the absolute values along with the difference compared to the respective previous year (in gray). The rightmost column summarizes the annual absolute values. Bold numbers indicate the maximum number of annually tagged flaws per type. The order of the flaw types is the same in the plot and in the table.

driven by the *Verifiability* type. The only flaw type that showed a continuous annually increasing number of tagged flaws was *Time-sensitive*, however, the absolute values were relatively small. The development of the remaining flaw types followed a different trend, which is characterized by an annually growing number of tagged flaws up to a certain peak, followed by a continuously decline. For instance, the annual number of tagged flaws that belong to the flaw type *Style of writing* peaked in 2007 and decreased in the following years. The year in which the respective peak has been reached varies for the different flaw types. This indicates that the observed development was not driven by external factors, like, for example, a decline of tagging work or of editing activity in general—such factors would have affected all flaw types equally. Instead, the flaw-type-specific development was governed by particularities of the respective quality flaws, such as frequency, community awareness, and seriousness.

On the basis of what was said before, we can conclude from the tagging behavior that for nine flaw types the quality flaw situation is in the process of improvement. For the types *Verifiability*, *Style of writing*, *Miscellaneous*, *Specific subjects*, *Unwanted content*, *Neutrality*, *General cleanup*, *Structure*, and *Merge* the annual number of tagged quality flaws already peaked and was continuously declining afterwards, see the table in Figure 3.7. Consider for instance the *Style of writing* type: the fact that, since 2008, the annual number of tagged flaws was declining suggests that the overall writing style has been improved during the last years. Similarly, article organization has been improved since 2007, which is indicated by a continuously decreasing annual number of tagged flaws that belong to the flaw type *Merge*. On the other hand, the quality flaw situation was worsening for the *Time-sensitive* type as the respective number of tagged flaws per year was increasing during the whole analysis period. This indicates that this type is still a relevant issue for the Wikipedia community. The two flaw types *Wiki tech* and *Expand* showed a different trend, which is characterized by a recently increasing annually number of tagged flaws after the decline had already happened. The number of annually tagged flaws that belong to the flaw type *Wiki tech* peaked in 2007 and increased significantly again in 2009. The flaw *Expand* also showed an interesting development: after the peak in 2008, the number of tagged flaws per year has more than tripled in 2010, although, at this time, only a relatively small number of 17 respective cleanup tags existed (cf. Figure 3.4).

Of the 6 235 192 tagged quality flaws, 3 899 734 (62.54%) accounted for the flaw type *Verifiability*, see the table in Figure 3.7 (rightmost column). The second most tagged flaw type was *Wiki tech* with 13.56% of the tagged flaws, followed by the *General cleanup* type with 4.6%. The most frequent individual quality flaw was *Citation needed*, see Table 3.2. The *Citation needed* flaw has been tagged 2 072 397 times, which corresponds to 33.24% of the overall tagged flaws. On average, one in 245 revisions has been tagged with this flaw. Moreover, the inline flaw *Citation needed* accounted for the majority (74.83%) of tagged inline flaws as well as for the majority (53.14%) of tagged flaws in the *Verifiability* type. The usage statistics shown in Table 3.2 also incorporate the age of the respective cleanup tags. The cleanup tag *Citation needed* has been created on June 5, 2005, and thus, it exists since 2 394 days (related to our analysis period, which ends on January 4, 2012). That a cleanup tag existed for a long time does not necessarily mean that it has been used more frequently compared to a younger tag. Consider for instance the cleanup tags *Refimprove* and *Uncategorized*: the former existed since nearly five years and has been used on average 155.91 times per day, whereas the latter existed since more than seven years and has been used only 121.01 times per day. Nevertheless, the 33 cleanup tags with the highest number of usages per day were all older than three years, and those cleanup tags that have been created before 2008 accounted for 95.64% of the ever tagged quality flaws. The lower part of Table 3.2 shows the other extreme. There are five cleanup tags that have not been used at all. Moreover, 24 cleanup tags have been used less than once per year, and another 156 cleanup tags have been used less than once per month (not listed in the table).

Table 3.2: The five most widely used and the five least used cleanup tags in the English Wikipedia, along with the respective number of instances, the age of the tag in days (from its creation until January 4, 2012), and the ratio of instances per day. The rows are ordered by the ratio.

Cleanup tag	Instances	Age	Ratio
Citation needed	2 072 397	2 394	865.66
Unreferenced	632 075	2 532	249.63
Refimprove	272 995	1 751	155.91
Dead link	271 615	1 802	150.73
Uncategorized	308 948	2 553	121.01
...	...	...	...
Cleanup-statistics	0	1 974	0.0
Convert to SVG and copy to Wikimedia Commons	0	1 609	0.0
Diagram requested	0	2 244	0.0
ShadowsCommons	0	2 009	0.0
Time references needed	0	228	0.0

3.4 Correction Time of Quality Flaws

This far, it was sufficient to investigate those revisions where a cleanup tag occurred for the first time, i.e., the date when a quality flaw has been tagged. For the analyses in this section, we also investigate those revisions where a cleanup tag has been removed, i.e., the date where a quality flaw has been corrected. The bottom row of Table 3.3 shows correction statistics for the 445 cleanup tags. Of the 6 235 192 tagged quality flaws, 4 367 356 have been corrected as of January 4, 2012, which corresponds to 70.04%. The duration between the tagging date and the correction date is considered as a flaw's correction time. The average correction time over all 445 quality flaws is 155 days. However, the correction time of the individual flaws varied widely, and 39.38% of the corrected flaws have been corrected within the first week after tagging. Page flaws account for the majority (54.22%) of the corrected flaws, see Table 3.3. However, a larger amount (72.18%) of the tagged inline flaws has been corrected, compared to the amount of the corrected page flaws (68.33%). Furthermore, inline flaws have been corrected faster than page flaws. The average correction time of page flaws is 176 days, whereas inline flaws have been corrected on average after 130 days. Likewise, the first week ratio for inline flaws (42.80%) is larger than for page flaws (36.49%). As already mentioned, inline cleanup tags are more specific, and hence, they provide a concrete indication of what needs to be done to correct the respective flaw. It is, for instance, easier to find a reference for a tagged statement than for a latent idea mentioned somewhere in an article that is tagged with a page flaw.

Table 3.3: For the two scopes page flaw and inline flaw: the number of tagged quality flaws, the number of corrected flaws, the percentage of corrected flaws, the average correction time in days, and the ratio of corrected flaws that have been corrected within the first week.

Flaw scope	**Tagged flaws**	**Corrected flaws**	**Ratio in %**	**∅ correction time in days**	**First week ratio in %**
Page flaw	3 465 571	2 368 165	68.33	176	36.49
Inline flaw	2 769 621	1 999 191	72.18	130	42.80
$\sum$	6 235 192	4 367 356	70.04	155	39.38

The majority (60.39%) of the 4 367 356 corrected flaws belonged to the flaw type *Verifiability*, see Table 3.4. The second most corrected flaw type was *Wiki tech* (15.01%), followed by *Style of writing* (4.95%) and *General cleanup* (4.69%). The ratio between tagged flaws and corrected flaws varied among the flaw types. For instance, the corrected flaw ratios of 93.23% and 88.03% for the types *Miscellaneous* and *Merge* respectively are relatively high. On the other hand, only approximately half of the tagged flaws that belonged to the types *Neutrality* and *Expand* have been corrected. In the case of the *Expand* type, the low ratio can be explained by the large number of recently tagged flaws in 2010 and 2011 (cf. Figure 3.7). Nevertheless, flaws that belong to the type *Expand* showed the highest average correction time and a relatively low first week ratio, which might be an indication that these flaws are too complex. The same applies to the *General cleanup* type, whose average correction time was relatively high, although, 71.56% of the tagged *General cleanup* flaws have been corrected.

Table 3.4: For the twelve flaw types: the number of tagged quality flaws, the number of corrected flaws, the percentage of corrected flaws, the average correction time in days, and the ratio of corrected flaws that have been corrected within the first week.

Flaw scope	**Tagged flaws**	**Corrected flaws**	**Ratio in %**	**∅ correction time in days**	**First week ratio in %**
Verifiability	3 899 734	2 637 252	67.63	164	32.92
Style of writing	273 330	216 262	79.12	173	28.64
Miscellaneous	33 869	31 508	93.23	101	38.44
Specific subject	44 300	28 601	64.56	187	24.15
Unwanted content	249 911	184 713	73.91	131	36.95
Neutrality	264 264	142 757	54.02	114	43.25
Wiki tech	864 009	655 485	75.87	106	38.49
General cleanup	286 371	204 934	71.56	216	26.38
Expand	215 424	108 438	50.34	223	24.78
Structure	35 678	27 645	77.49	152	31.99
Time-sensitive	25 995	16 814	64.68	162	27.90
Merge	128 300	112 947	88.03	119	30.91
$\sum$	6 235 192	4 367 356	70.04	155	39.38

Moreover, Table 3.4 identifies several flaw types that have been corrected relatively quick. The *Miscellaneous* and *Wiki tech* types showed the lowest average correction time and nearly 40% of the tagged flaws have been corrected within the first week. Furthermore, although the ratio of corrected flaws that belonged to the *Neutrality* type was relatively low, these flaws showed the highest first week ratio and also a low average correction time. Finally, it can be said that certain flaw types have been corrected faster than others, and that there is a backlog of long-tagged flaws that need to be corrected.

Table 3.5 shows correction statistics for individual quality flaws. The tag box *Hoax* states that an article contains false facts. The fact that this flaw has been corrected after 7 days on average is an indication that hoaxes are considered quite serious by the Wikipedia community. On the other hand, a high average correction time does not necessarily mean that the respective flaw was regarded as unimportant. The average correction time can be considered as a measure for a flaw's complexity. Consider for instance the article flaws *Uncategorized* and *Orphan*: It is relatively easy to identify a proper category for an uncategorized article, but finding related articles for an orphaned article and defining reasonable links is much more complicated, also because in many cases related articles are yet to be created. As a consequence, the *Orphan* flaw showed a relatively high average correction time. However, despite its high complexity, the *Orphan* flaw is considered as important, which is witnessed by the large number of

Table 3.5: The five quality flaws with the shortest and the longest average correction time respectively, along with a description, the total number of corrections, and the average correction time in days. Only flaws with more than 1 000 corrections are listed.

Flaw name	**Description**	**Corrected flaws**	**∅ correction time in days**
Hoax	The truthfulness of the article has been questioned.	1 573	7
Not English	The article needs translation into English.	3 243	9
Image requested	An image or photograph should be included in the article.	2 346	19
Uncategorized	The article has not been added to any categories.	306 969	22
Disambiguation cleanup	The disambiguation page lists articles associated with the same title.	8 273	24
...	...	...	...
No footnotes	The article's sources remain unclear because of its inline citations.	29 389	264
Unreferenced	The article does not cite any references or sources.	366 780	279
Orphan	The article has fewer than three incoming links.	125 375	295
Cleanup FJ biography	Import from the Biographical Directory of Federal Judges requires rewriting and/or reformatting.	1 616	335
Cleanup-school	Cleanup of school-related article.	1 108	388

125 375 corrected flaws. The same applies to the flaw *Unreferenced*, which has been corrected many times but with a relatively high average correction time of 279 days. A possible way to increase the correction times of such complex but still important flaws is the development of respective tools that support potential human correctors. In the case of the *Orphan* flaw, this might be a tool that retrieves articles with similar content.

There are also 17 flaws that have never been corrected: *Author incomplete*, *Title incomplete*, *Define?*, *Check quotation*, *Clarify-span*, *Idetail*, *Citation needed cheap*, *List years*, *Cleanup-statistics*, *Video game cleanup*, *Outdated as of*, *Time references needed*, *ShadowsCommons*, *RJL*, *Convert to SVG and copy to Wikimedia Commons*, *Diagram requested*, and *Map requested.* An explanation is that either these flaws are too complex or the respective cleanup tags are too unspecific.

Chapter 4

Quality Flaw Prediction

This chapter deals with the prediction of quality flaws in Wikipedia articles. We target a subset of the 445 quality flaws that have been identified in Chapter 2. We do not consider the 17 flaws that belong to the flaw type *General Cleanup* (cf. Section 2.3.1) as these flaws are too unspecific. Moreover, we consider only page flaws (cf. Section 2.3.2), i.e., those flaws that refer to the article as a whole.[1] This focus does not reduce the impact of our research, since only 7.48% of the articles that have been tagged so far by the Wikipedia community are tagged with one of the 17 unspecific flaws and since the majority (78.49%) of the tagged articles suffer from a flaw that refers to the whole article. In this way, 301 specific article flaws remain and the ten most frequent are listed in Table 4.1. Altogether, 760 882 articles are tagged with the 301 quality flaws, whereas 81.82% are tagged with the ten flaws shown in Table 4.1. We aim to use the tagged articles as a source of human-labeled data, which is then exploited by a machine learning approach to predict flaws of untagged articles.

Chapter organization. Section 4.1 argues that the prediction of quality flaws in Wikipedia articles is essentially a one-class classification problem and presents a respective problem definition. Section 4.2 gives a comprehensive overview of Wikipedia article features. Section 4.3 describes two paradigms to model quality flaws and devises a tailored one-class machine learning approach to address the problem. The effectiveness of the learning approach in predicting the ten most frequent quality flaws is evaluated and discussed in Section 4.4. Section 4.5 overviews the "1st International Competition on Quality Flaw Prediction in Wikipedia" held in conjunction with the PAN 2012 lab.

Key contributions. We operationalize an algorithmic prediction of quality flaws in Wikipedia articles, which includes a formal problem definition, a tailored flaw model, a dedicated machine learning approach, and an in-depth evaluation.

[1] Note that our prediction approach can be applied to inline flaws as well, by breaking articles into paragraphs or sentences; but we do not demonstrate this here.

Table 4.1: The ten most frequent quality flaws from the 301 specific article flaws that we target in this chapter, along with a description and the number of articles that have been tagged with the respective cleanup tag in the English Wikipedia snapshot from January 4, 2012.

Flaw name	**Description**	**Tagged articles**
Unreferenced	The article does not cite any references or sources.	251 447
Orphan	The article has fewer than three incoming links.	157 066
Refimprove	The article needs additional citations for verification.	127 667
Empty section	The article has at least one section that is empty.	43 016
Notability	The article does not meet the general notability guideline.	35 183
No footnotes	The article's sources remain unclear because of its inline citations.	30 773
Primary sources	The article relies on references to primary sources.	23 126
Wikify	The article needs to be wikified (links and layout).	13 101
Advert	The article is written like an advertisement.	8 009
Original research	The article contains original research.	6 951

4.1 Problem Statement

Let D be the set of Wikipedia articles and let F be a set of quality flaws. A document $d \in D$ can contain up to $|F|$ flaws, where, without loss of generality, the flaws in F are considered as being uncorrelated. A classifier c hence has to solve the following multi-labeling problem:[2]

$$c : \mathbf{D} \to 2^F,$$

where 2^F denotes the power set of F. An article $d \in D$ is modeled as a feature vector $\mathbf{d}$, called document model, and for the set D, $\mathbf{D}$ denotes the set of associated document models.

Basically, there are two strategies to tackle multi-labeling problems:

1. by multiclass classification, where a single classifier is learned on the power set of all classes, and

2. by multiple binary classification, where a specific classifier $c_i : \mathbf{D} \to \{1, 0\}$ is learned for each class $f_i \in F$.

[2] Possibly existing correlations among the flaws in F will not affect the nature of the multi-labeling problem.

The second strategy is favorable since the high number of classes under a multiclass classification strategy entails a very large number of training examples.

In most classification problems training data is available for all classes that can occur at prediction time, and hence, it is appropriate to train a classifier c_i with (positive) examples of the target class f_i and (negative) examples from the classes $F \setminus f_i$. When spotting quality flaws, an unseen document can either belong to the target class f_i or to some unknown class that was not available during training. This means that the standard discrimination-based classification approaches (binary or multiclass) are not applicable to learn a class-separating decision boundary: given a flaw f_i, its target class is formed by those documents that contain (among others) flaw f_i—but it is impossible to model the "co-class" with documents *not* containing f_i. Even if many counterexamples were available, they could not be exploited to properly characterize the universe of possible counterexamples. As a consequence, we model the classification $c_i(\mathbf{d})$ of an document $d \in D$ with respect to a quality flaw f_i as the following one-class classification problem: Decide whether or not d contains f_i, whereas a sample of documents containing f_i is given.

As an additional illustration consider the flaw *Refimprove*, which is described in Table 4.1. An even large sample of articles that suffer from this flaw can be compiled without problems (127 667 articles have been tagged with this flaw). However, it is impossible to compile a representative sample of articles that have a reasonable number of proper citations for verification. Although many articles with sufficient citations exist (e.g., featured articles), they cannot be considered as a *representative* sample. The fact that featured articles are not representative for typical Wikipedia articles becomes clear when looking at Figure 4.1, which shows a sample of Wikipedia articles represented under the first two principle components. Figure 4.1 also shows that quality flaw prediction is a significantly harder problem than discriminating featured articles. Training a binary classifier using featured articles and flawed articles would lead to a biased classifier that is not able to predict flaws on the entire Wikipedia. Also, using random articles and flawed articles to train a binary classifier is unacceptable because, as motivated earlier, it is more than likely that many flawed articles are not yet identified (cf. Section 2.4.3). Stated another way, quality flaw prediction is a one-class classification problem.

One-class classification is also called outlier detection, unary classification, or single-class classification. For an in-depth discussion of one-class classification and a survey of respective methodologies, refer to Tax [147] and Hodge and Austin [75] respectively. Typical one-class classification problems include typist recognition [73], authorship verification [89], plagiarism analysis [140], anomaly detection [36], and novelty detection [106].

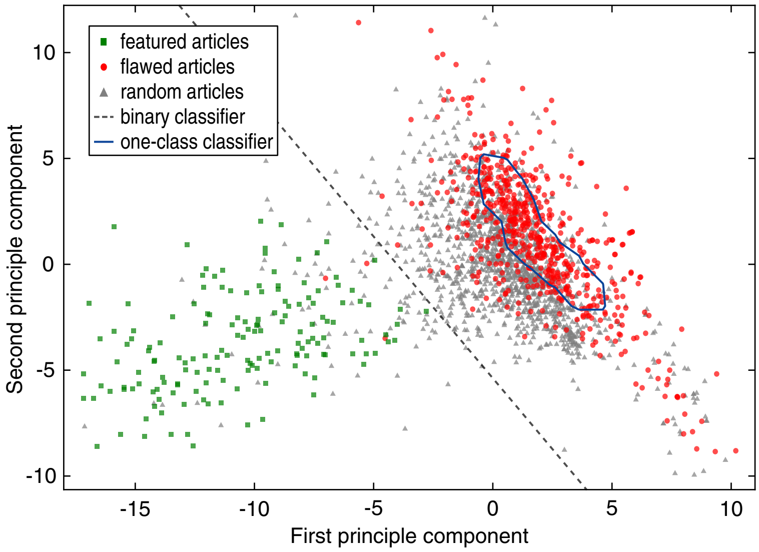


Figure 4.1: Distribution of featured articles, articles that are tagged with the flaw *Refimprove*, and random articles in the English Wikipedia. The articles are represented under the first two principle components, computed based on the model described in Section 4.2. The binary classifier (linear SVM) is trained using featured articles and flawed articles, the one-class classifier (one-class SVM with RBF kernel) is trained solely on the set of flawed articles. (For details on principal component analysis and on one-class SVMs, refer to Jolliffe [81] and to Schölkopf et al. [131] respectively.)

4.2 Article Features

The prediction of quality flaws requires a model that captures the flaws' characteristics based on measurable features. It has been stated earlier that encyclopedia quality cannot be assessed using a single metric [41]. To this day a large body of features—allegedly quality predicting—has been proposed, of which we have compiled a comprehensive breakdown. We implemented the major part of the previously proposed features, i.e., our quality flaw model captures the state-of-the-art with respect to feature expressiveness and information quality research. We organize the features along four dimensions: content, structure, network, and edit history. The source of information that is required for feature computation as well as the computational complexity differs for each dimension. Our model can be adjusted with respect to its transferability to other text documents than Wikipedia articles as well as to its computational complexity, by restricting to the features from a subset of the four dimensions. Some of the previously proposed features are omitted since their implementation effort exceeds the expected benefit by far. In addition, we devise several new features that directly target flaw-specific aspects. Altogether, our model comprises 95 Wikipedia article features. The remainder of this subsection discusses the four dimensions and overviews the respective features. A detailed description of the individual features along with implementation details is given in Appendix B.

4.2.1 Content Features

Content features rely on the plain text representation of an article and are intended to quantify aspects like writing style and readability. Basically, these features target quality flaws that belong to the flaw type *Style of writing* (cf. Section 2.3.1). Table 4.2 lists the content features, organized into the following four subcategories: text statistics, part of speech, readability formulas, and closed-class word sets. The first subcategory comprises features that are based on basic text statistics, such as the frequency of characters, words, sentences, and paragraphs. The features in the second subcategory quantify parts of speech, and therefore, the usage of words and phrases. The third subcategory comprises so-called readability formulas, which measure the complexity and the understandability of text based on basic text statistics. More complex approaches (which use statistical language modeling for instance [39]) to predict the reading difficulty of text have been proposed recently; however, for large scale data analysis, easy measures perform best [79]. The fourth subcategory comprises new features that measure the presence of certain words from predefined closed-class word sets. Besides writing style issues, these features address also the flaw types *Unwanted content* and *Neutrality*. For instance, peacock words, such as legendary, great, and brilliant, may be an indicator of advertising or promotional content; similarly, the presence of sentiment-bearing words can be considered as an indicator of missing neutrality.

Given the plain text representation of an article d, content features can be computed with a complexity of $O(|d|)$, where $|d|$ denotes the length of d in characters. The computational effort is relatively low compared to the features from the other dimensions, such as network and edit history: both of which require a processing of Wikipedia's link graph and revision history respectively. Moreover, content features can be computed for every text document, and hence, they are applicable in other contexts than Wikipedia.

4.2.2 Structure Features

Structure features rely on the wiki markup of an article and are intended to quantify the article's organization. These features target the flaw types *Structure* and *Wiki tech* in the first place (cf. Section 2.3.1). Table 4.3 lists the structure features. We employ features which measure quantity, length, and nesting of

Table 4.2: Content features. Overview of features that are derived from an article's plain text. The references refer to prior studies on quality assessment in Wikipedia that exploit the respective feature. New features that are used for the first time here are marked with an asterisk (∗); features that are not employed within our document model are shown in gray. For a detailed feature description, refer to Section B.1.

Feature	Reference
Text statistics	
Character count	[23, 47, 143]
Character trigrams	[100]
Complex word rate	*
Information-to-noise ratio	[143]
Long sentence rate	[47]
Long word rate	*
Longest sentence length	[47]
One-syllable word count	[23]
One-syllable word rate	*
Paragraph count	*
Paragraph length	[47]
Question count	*
Question rate	[47]
Sentence count	[23, 47]
Sentence length	*
Short sentence rate	[47]
Shortest sentence length	*
Syllable count	[23]
Word count	[23, 47, 77, 100]
Word length	*
Word syllables	*
Part of speech	
Article sentence rate	[47]
Auxiliary verb rate	[47]
Conjunction rate	[47]
Conjunction sentence rate	[47]
Interrog. pronoun sentence rate	[47]
Nominalization rate	[47]
Passive sentence rate	[47]
Preposition rate	[47]
Preposition sentence rate	[47]
Pronoun sentence rate	[47]
Pronoun rate	[47]
Subord. conjunction sentence rate	[47]
"To be" verb rate	[47]
Readability formulas	
Automated Readability Index	[23, 47]
Bormuth Index	*
Coleman-Liau Index	[23, 47]
FORCAST Readability	[23]
Flesch Reading Ease	[23, 47, 143]
Flesch-Kincaid	[23, 47, 143]
Gunning Fog Index	[23, 47]
Läsbarhetsindex	[47]
Miyazaki EFL Readability Index	*
New Dale-Chall	*
SMOG Grading	[23]
Closed-class word sets	
Common word rate	*
Difficult word rate	*
Peacock word rate	*
Stop word rate	*
Weasel word rate	*

(continued on the right)

sections as well as of subsections. Special attention is paid to the lead section, also called intro, as several quality flaws directly refer to it. Moreover, the usage of images, tables, and files is quantified, as well as the categories an article belongs to. Note that the usage of templates and lists is quantified for the first time. Other features measure the usage of references, including citations and footnotes, and shall target flaws that belong to the flaw type *Verifiability*. We introduce new features that check the presence of special sections that are either mandatory, such as "References", or that should be avoided, such as "Trivia".

The computational effort for structure features is governed by the complexity of parsing an article's wiki markup, which is in $O(|d|)$. Structure features and content features are comparable with respect to their computational effort. Regarding their transferability to other contexts, structure features are meaningful only for text documents that contain some kind of markup, like HTML for instance. Therefore, these features are applicable to Web documents in general. However, some structure features are dedicated to wiki markup, such as *template count*, while others are tailored to Wikipedia, such as *reference sections count*.

Table 4.3: Structure features. Overview of features that are derived from an article's wiki markup. The references refer to prior studies on quality assessment in Wikipedia that exploit the respective feature. New features that are used for the first time here are marked with an asterisk ($*$); features that are not employed within our document model are shown in gray. For a detailed feature description, refer to Section B.2.

Feature	Reference
Category count	[22]
File count	[22]
Heading count	*
Image count	[22, 46, 143]
Images per section	[46]
Infobox count	[22]
Lead length	[46]
Lead rate	*
List ratio	*
Reference count	[22, 46]
Reference sections count	*
References per section	[46]
References per text length	[46]
Section count	[22, 46]
Section length	[46]
Section length deviation	[46]
Section nesting	[46]
Shortest section length	[46]
Shortest subsection length	*
Shortest subsubsection length	*
Subsection count	[46]
Subsection length	*
Subsection nesting	*
Subsubsection count	*
Subsubsection length	*
Longest section length	[46]
Longest subsection length	*
Longest subsubsection length	*
Table count	[22]
Template count	*
Trivia sections count	*

(continued on the right)

4.2.3 Network Features

Network features quantify an article's integration by means of hyperlinks and mainly target the flaw type *Wiki tech* (cf. Section 2.3.1). Table 4.4 lists the network features. We distinguish between the following types of outgoing links:

- Internal links, which point to articles in the same language.
- Inter-language links, which point to the same article in a different language.
- External links, which point to sources outside of Wikipedia.

The features count the number of outgoing links as well as their frequency relative to the article size. An internal link is considered as broken if it refers to a non-existing article. There is also a special feature that quantifies incoming internal links, where the origin has to be an article (i.e., links from disambiguation, redirect, and discussion pages are excluded). Several features specifically address individual quality flaws, for instance, the incoming link count targets the flaw *Orphan*, and the outgoing link counts target the flaw *Wikify* (cf. Table 4.1). Other features give some indication of an article's popularity (*assortativity*), relatedness to other articles (*clustering coefficient*), importance (*PageRank*), and link quality (*reciprocity*).

The computation of network features is based on Wikipedia's link graph and is in $O(|d|\cdot|D|)$, where D denotes the set of all Wikipedia articles. In the graph articles represent nodes and internal links represent edges. The computational effort is higher than for content and structure features. Network features can be computed for all kinds of document collections that are connected via hyperlinks.

Table 4.4: Network features. Overview of Wikipedia article features that quantify an article's integration by means of hyperlinks. The references refer to prior studies on quality assessment in Wikipedia that exploit the respective feature. Features that are not employed within our document model are shown in gray. For a detailed feature description, refer to Section B.3.

Feature	**Reference**
Assortativity	[46]
Broken internal link count	[143]
Clustering coefficient	[46]
External link count	[22, 46, 143]
External links per section	[46]
Incoming internal link count	[46]
Internal link count	[22, 46, 143]
Internal links per text length	[46]
Inter-language link count	[46]
PageRank	[46, 157]
Reciprocity	[46]

(continued on the right)

4.2.4 Edit History Features

Edit history features rely on an article's revision history and model article evolution, which pertains to the frequency and the timing of edits as well as to the community of editors. Table 4.5 lists the edit history features. These features quantify article stability (e.g., *modified lines rate* and *revert count*), up-to-dateness (e.g., *currency* and *edit currency rate*), and cooperation (e.g., *connectivity* and *discussion edit count*). Edit history features have been proven a valuable means to classify featured articles [46, 77, 99, 143, 157]. However, the expected benefit of these features for flaw prediction is low when being compared to the implementation effort: hence, we omitted several features of this type.

The computational effort for edit history features is in $O(|d| \cdot r_d)$, where r_d denotes the number of revisions of an article d. This effort is high compared to the other feature dimensions: The average number of revisions per article is 75.07, however, the most-edited article, which is "George W. Bush", counts 44 655 revisions. As already mentioned, the contents of all page revisions in the English Wikipedia sum up to 7.9 TB (uncompressed), which poses a challenge for an efficient feature computation. Regarding their transferability to other contexts, edit history features can be computed for all kinds of document collections that maintain a revision history, which, of course, is the case for all wiki-based projects.

Table 4.5: Edit history features. Overview of Wikipedia article features that are derived from an article's revision history. The references refer to prior studies on quality assessment in Wikipedia that exploit the respective feature. Features that are not employed within our document model are shown in gray. For a detailed feature description, refer to Section B.4.

Feature	**Reference**
Active editor rate	[46]
Admin editor rate	[143]
Age	[46, 143]
Age per edit	[46]
Anonymous editor rate	[46, 143]
Connectivity	[143]
Currency	[143]
Discussion edit count	[46, 157]
Edit count	[46, 99, 143, 157]
Edit currency rate	[46]
Edits per editor	[46, 157]
Edits per editor deviation	[46]
Editor count	[99, 143, 157]
Editor rate	[143]
Modified lines rate	[46]
Occasional editor rate	[46]
ProbReview	[46, 77]
"Quick-turnaround" edit rate	[157]
Registered editor rate	[46, 143]
Revert count	[143]
Revert time	[143]

(continued on the right)

4.3 Modeling Quality Flaws

The modeling of quality flaws can happen intensionally or extensionally, depending on a flaw's nature and the knowledge that is at our disposal.[3] An intensional model of a flaw f can be understood as a set of rules, which define, in a closed-class manner, the set of articles that contain f. An extensional model is given by a set of positive examples, and modeling means learning a classifier that discriminates positive instances (containing the flaw) from all other instances.

[3] For special cases also a hybrid model is conceivable, where a filtering step (intensional) precedes a learning step (extensional).

4.3.1 Intensional Modeling

The descriptions in Table 4.1 show that three flaws from the set of the ten most frequent quality flaws, namely *Unreferenced*, *Orphan*, and *Empty section*, can be modeled with rules based on the afore-mentioned Wikipedia article features.

An article suffers from the flaw *Unreferenced* if it does not cite any references or sources. Wikipedia provides different ways of citing sources, including inline citations, footnotes, and parenthetical referencing.[4] Here, we summarize all types of citations under the term "references". Using the structure features *reference count* and *reference sections count* we define the predicate *unreferenced*(d):

$$unreferenced(d) = \begin{cases} 1, & \text{if } reference\text{-}count(d) = 0 \\ & \text{and } reference\text{-}sections\text{-}count(d) = 0 \\ 0, & \text{else} \end{cases}$$

An evaluation on $D^-_{Unreferenced}$, the set of articles that have been tagged to be unreferenced, reveals that the *unreferenced*-predicate is fulfilled for 85.3% of the articles. We analyzed the remaining 14.7% and found that they actually provide references, and hence, are mistagged. This observation shows a well-known problem in the Wikipedia community, and there is a WikiProject dedicated to cleanup mistagged unreferenced articles.[5] The fact that there is no such WikiProject for other quality flaws suggests that this problem is not considered to be serious for other flaws.

The *Orphan* flaw is well-defined: an article is called orphan if it has fewer than three incoming links. In this regard, the following page types are not counted: disambiguation pages, redirects, soft redirects, discussion pages, and pages outside of the article namespace.[6] Using the network feature *incoming internal link count* we define the predicate *orphan*(d):

$$orphan(d) = \begin{cases} 1, & \text{if } incoming\text{-}internal\text{-}link\text{-}count(d) < 3 \\ 0, & \text{else} \end{cases}$$

An evaluation on D^-_{orphan} reveals that the *orphan*-predicate is fulfilled for 98.4% of the articles.

[4] For detailed information about citing sources in Wikipedia, refer to: `http://en.wikipedia.org/wiki/Wikipedia:Citing_sources`.

[5] WikiProject "Mistagged unreferenced articles cleanup": `http://en.wikipedia.org/wiki/Wikipedia:Mistagged_unreferenced_articles_cleanup`.

[6] Wikipedia, "Wikipedia:Orphan," last modified October 27, 2012, `http://en.wikipedia.org/wiki/Wikipedia:Orphan#Criteria`.

An article suffers from the flaw *Empty section* if it has a section that does not contain content at all. Using the structure feature *shortest section length* we define the predicate *empty-section*(d):

$$\textit{empty-section}(d) = \begin{cases} 1, & \text{if } \textit{shortest-section-length}(d) = 0 \\ 0, & \text{else} \end{cases}$$

An evaluation on $D^-_{\textit{empty-section}}$ reveals that the *empty-section*-predicate is fulfilled for 99.43% of the articles.

The intensional modeling paradigm is efficient since no training data is required and since the computation relies on few basic features. Moreover, as the above evaluations show, it is effective at the same time. However, if the definition of a quality flaw changes, an explicit model needs to be adapted as well.

4.3.2 Extensional Modeling

The majority of quality flaws is defined informally and cannot be modeled by means of explicit rules (cf. Table 4.1); the knowledge is given in the form of examples instead. For an article $d \in D$ we model these flaws as a vector $\mathbf{d}$, called document model. The dimensions of $\mathbf{d}$ quantify the features listed in Table 4.2, 4.3, 4.4, and 4.5. For a set D of Wikipedia articles, $\mathbf{D}$ denotes the set of associated document models. By means of machine learning a mathematical decision rule is computed from $\mathbf{D}$ that discriminates between elements from D^- and $D \setminus D^-$ (see Figure 4.2).

Method

As motivated earlier, quality flaw prediction is essentially a one-class problem. Following Tax [147] three principles to construct a one-class classifier can be distinguished: density estimation methods, boundary methods, and reconstruction methods. Here, we resort to a one-class classification approach as proposed by Hempstalk et al. [73], which combines density estimation with class probability estimation. There are two reasons for using this approach: 1. Hempstalk et al. show that it is able to outperform state-of-the-art approaches, including a one-class SVM, and 2. it can be used with arbitrary density estimators and class probability estimators. Instead of employing an out-of-the-box classifier we apply dedicated density estimation and class probability estimation techniques to address the problem defined in Section 4.1.

The idea is to use a reference distribution to model the probability $P(\mathbf{d} \mid f'_i)$ of an artificial class f'_i, and to generate (artificial) data governed by the distribution characteristic of f'_i. For a flaw f_i let $P(f_i)$ and $P(f_i \mid \mathbf{d})$ denote the a-priori probability and the class probability function respectively. According to Bayes' theorem the class-conditional probability for f_i is given as follows:

$$P(\mathbf{d} \mid f_i) = \frac{(1 - P(f_i)) \cdot P(f_i \mid \mathbf{d})}{P(f_i) \cdot (1 - P(f_i \mid \mathbf{d}))} \cdot P(\mathbf{d} \mid f_i') \tag{4.1}$$

$P(f_i \mid \mathbf{d})$ is estimated by a class probability estimator, i.e., a classifier whose output is interpreted as probability. Since we are in a one-class situation we have to rely on the face value of $P(\mathbf{d} \mid f_i)$. More specifically, $P(\mathbf{d} \mid f_i)$ cannot be used to determine a maximum a-posterior (MAP) hypothesis among the $f_i \in F$. As a consequence, given $P(\mathbf{d} \mid f_i) < \tau$ with $\tau = 0.5$, the hypothesis that d suffers from f_i could be rejected. However, because of the approximative nature of $P(f_i \mid \mathbf{d})$ and $P(f_i)$ the estimation for $P(\mathbf{d} \mid f_i)$ is not a true probability, and the threshold τ has to be chosen empirically. In practice, the threshold τ is derived from a user-defined target rejection rate, trr, which is the rejection rate of the target class training data.

The one-class classifier is built as follows: at first, a class with artificial examples is generated, whereas the feature values obey a Gaussian distribution with $\mu = 0$ and $\sigma^2 = 1$. We employ the Gaussian distribution in favor of a more complex reference distribution to underline the robustness of the approach. The proportion of the generated data is 0.5 compared to the target class. As class probability estimators we apply bagged random forest classifiers with 1 000 decision trees and ten bagging iterations. A random forest is a collection of decision trees where a voting over all trees is run in order to obtain a classification decision [27, 74]. The decision trees of a forest differ with respect to their features. Here, each tree is build with a subset of $log_2(|\mathit{features}|) + 1$ randomly chosen features, i.e., no tree minimization strategy is followed at training time. The learning algorithm stops if either all leaves contain only examples of one class or if no further splitting is possible. Each decision tree perfectly classifies the training data—but, because of its low bias the obtained generalization capability is poor [110, 158]. However, the combination of several classifiers in a voting scheme reduces the variance and introduces a stronger bias. While the bias of a random forest results from several feature sets, the bias of the bagging approach results from the employment of several training sets, and it is considered as being even stronger [26].

4.4 Evaluation and Discussion

We report on experiments to assess the effectiveness of our modeling and classification approach in predicting the ten quality flaws shown in Table 4.1. The evaluation treats the following issues:

1. Since a bias may not be ruled out when collecting outlier examples for a classifier's test set, we investigate the consequences of two extreme settings: overly optimistic and overly pessimistic (Section 4.4.1).
2. Since users (Wikipedia editors) have different expectations regarding the classification effectiveness given different flaws, we analyze the optimal operating point for each flaw-specific classifier within the controlled setting of a balanced class distribution (Section 4.4.2).
3. Since the true flaw-specific class imbalances in Wikipedia can only be hypothesized, we illustrate the effectiveness of the classifiers in different settings, this way enabling users (Wikipedia editors) to assume an optimistic or a pessimistic position (Section 4.4.3).

4.4.1 Outlier Selection

Recall that no articles are available that have been tagged to *not* contain a quality flaw $f_i \in F$ (cf. Section 4.1). Thus a classifier c_i can be evaluated only with respect to its recall, whereas a recall of 1 can be achieved easily by classifying all examples into the target class of f_i. In order to evaluate c_i with respect to its precision one needs a representative sample of examples from outside the target class, so-called outliers. As motivated earlier, in a one-class situation it is not possible to compile a representative sample, and a way out of the dilemma is the generation of uniformly distributed outlier examples [147]. Here, we pursue two strategies to derive examples from outside the target class, which result in the following settings:

1. *Optimistic setting.* Use of featured articles as outliers. This approach is based on the hypothesis that featured articles do not contain a quality flaw at all, see Figure 4.2.[7] Under this setting one introduces some bias, since featured articles cannot be considered as a representative sample of Wikipedia articles (see Figure 4.1).
2. *Pessimistic setting.* Use of a random sample from $D \setminus D_i^-$ as outliers for each f_i. This approach may introduce considerable noise since the set $D \setminus D_i^-$ is expected to contain untagged articles that suffer from f_i.

The above settings address two extremes: classification under laboratory conditions (overly optimistic) versus classification in the wild (overly pessimistic). The experiment design is owing to the facts that "no-flaw features" cannot be

[7] The hypothesis may hold in many cases but not always: the Wikipedia snapshot comprises 19 featured articles that have been tagged with at least one of the ten flaws listed in Table 4.1. We discarded these articles in our experiments.

stated and that the number of false positives as well as the number of false negatives in the set D^- of tagged articles are unknown.

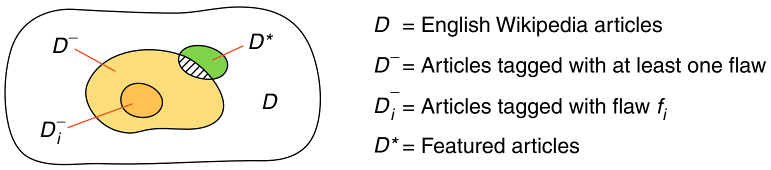

D^-
D^*
D
D_i^-
D = English Wikipedia articles
D^- = Articles tagged with at least one flaw
D_i^- = Articles tagged with flaw f_i
D^* = Featured articles

Figure 4.2: Sets of Wikipedia articles that we distinguish in this thesis. Without loss of generality, we assume in our experiments that the hashed area $D^- \cap D^*$ is empty, i.e., featured articles are flawless.

4.4.2 Effectiveness of Flaw Prediction

The evaluation relies on the same data basis that underlies the exploratory data analyses in Chapter 2 and Chapter 3, the English Wikipedia snapshot from January 4, 2012 (cf. Section 2.1).

Experiment Design

The evaluation is performed for the set $F' \subset F$ of the ten flaws shown in Table 4.1. In the optimistic setting 1 000 outliers are randomly selected from the 3 419 (untagged) featured articles in the Wikipedia snapshot. In the pessimistic setting 1 000 outliers are randomly selected for each flaw $f_i \in F'$ from $D \backslash D_i^-$. We evaluate our approach under both settings by applying the following procedure: For each flaw $f_i \in F'$ the one-class classifier c_i is evaluated with 1 000 articles randomly sampled from D_i^- and the respective 1 000 outliers, applying tenfold cross-validation. Within each cross-validation run the classifier is trained with 900 articles from D_i^-, whereas testing is performed with the remaining 100 articles from D_i^- plus 100 outliers. Note that c_i is trained exclusively with the examples of the respective target class, i.e., the articles in D_i^-. This means that the training of c_i is neither affected by the class distribution nor by the outlier selection strategy that is used in the respective setting.

Operating Point Analysis

For the major part of the relevant use cases precision is the determining measure of effectiveness. Consider for instance a Wikipedia bot that autonomously tags flawed articles. In this use case, false positives cause mistagged articles, which is worse for two reasons: First, Wikipedia is brought into disrepute because regular content is tagged to be flawed. Second, mistagged articles upset users who regularly correct flaws on a voluntary basis, which results in the fact that fewer of the actual flaws are corrected. On the other hand, false negatives do not worsen the initial situation. We therefore aim to optimize the classifiers to achieve the highest possible precision.

The precision of the one-class classifiers is controlled by the hyperparameter "target rejection rate". We empirically determine the optimal operating point for each of the ten flaws under the optimistic and the pessimistic setting, see Figure 4.3. The recall is the same in both settings since it solely depends on the target class training data. With increasing target rejection rate the recall values decrease while the precision values increase. As expected, the precision values under the optimistic setting are still high for small target rejection rates, which is due to the fact that featured articles and tagged articles can be separated easy (cf. Figure 4.1). Although precision is the determining measure in this use case, the analysis shows that it is advisable to consider the recall as well. Consider for example the flaw *No footnotes*: the target rejection rate of the maximum precision classifier under the optimistic setting is 0.6, which corresponds to precision and recall values of 0.93 and 0.44 respectively. However, the classifier with a target rejection rate of 0.2 achieves precision and recall values of 0.92 and 0.82 respectively. The example shows that accepting a slightly lower precision can result in a significant higher recall. We therefore use the F_β-measure as an optimization criterion [149]:

$$F_\beta = (1+\beta^2) \cdot \frac{\mathit{precision} \cdot \mathit{recall}}{\beta^2 \cdot \mathit{precision} + \mathit{recall}} \tag{4.2}$$

The measure incorporates both precision and recall, whereas recall is considered β times as much important as precision. We empirically determined $\beta = 0.2$, i.e., precision is considered five times as important as recall. Hence, the optimal operating point corresponds to the target rejection rate of the classifier that maximizes the $F_{0.2}$-measure. Figure 4.3 shows the optimal operating points for the ten flaws. Consider for example the flaw *Unreferenced*: its optimal operating points under the optimistic and the pessimistic setting are at target rejection rates of 0.15 and 0.65 respectively (with precision values of 1.0 and 0.88).

Note that the precision of a one-class classifier cannot be adjusted arbitrarily since the target rejection rate controls only the probability threshold τ for the classification decision (cf. Section 4.3.2). For instance, a target rejection rate of 0.1 means that a τ is chosen such that 10% of the target class training data will be rejected, which results in a classifier that performs with an almost stable recall of 0.9. Increasing the target rejection rate entails an increase of τ. However, if τ achieves its maximum no further examples can be rejected, and hence, both the precision and the recall remain constant beyond a certain target rejection rate; for the flaw *Unreferenced*, for instance, this is 0.7 (see Figure 4.3).

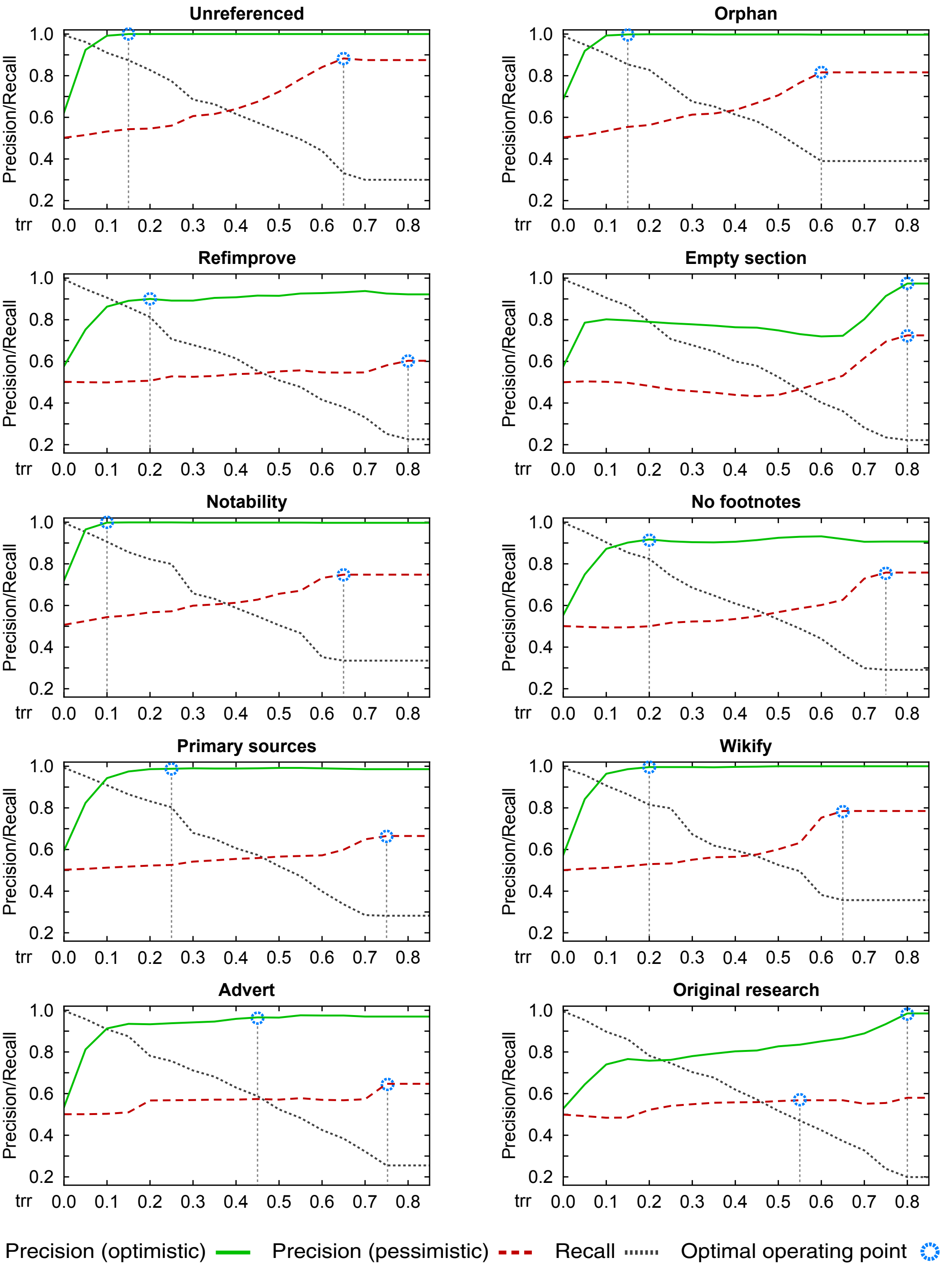


Figure 4.3: Precision and recall over target rejection rate, *trr*, for the ten quality flaws, using featured articles as outliers (optimistic setting) and using random articles as outliers (pessimistic setting). The recall is the same under both settings. The optimal operating point corresponds to the *trr* that maximizes the $F_{0.2}$-measure.

Results

Table 4.6 shows the performance values for each of the ten quality flaws. The values correspond to the performance at the respective optimal operating points, which are shown in Figure 4.3. The performance is quantified in terms of precision and recall. We also report the area under ROC curves (AUC) [53], which is important to assess the tradeoff between specificity and sensitivity of a classifier. An AUC value of 0.5 means that all specificity-sensitivity-combinations are equivalent, which in turn means that the classifier is random guessing.

Table 4.6: Individual performance values for each of the ten quality flaws at the optimal operating point (cf. Figure 4.3), using featured articles as outliers (optimistic setting) and using random articles as outliers (pessimistic setting). The class distribution is balanced under both settings. The flaw ratio 1:n (flawed articles : flawless articles) corresponds to the estimated actual frequency of a flaw.

Flaw name	**Optimistic setting**			**Pessimistic setting**			**Flaw ratio**
	precision	recall	AUC	precision	recall	AUC	
Unreferenced	1.00	0.87	0.94	0.88	0.33	0.64	1:4
Orphan	1.00	0.85	0.93	0.82	0.39	0.65	1:6
Refimprove	0.90	0.81	0.86	0.60	0.23	0.54	1:8
Empty section	0.97	0.22	0.61	0.73	0.22	0.56	1:24
Notability	1.00	0.91	0.95	0.75	0.34	0.61	1:30
No footnotes	0.92	0.82	0.88	0.76	0.29	0.60	1:34
Primary sources	0.99	0.80	0.90	0.67	0.28	0.57	1:44
Wikify	1.00	0.82	0.91	0.79	0.36	0.63	1:79
Advert	0.97	0.59	0.78	0.65	0.26	0.56	1:129
Original research	0.99	0.20	0.60	0.57	0.56	0.59	1:149

Under the optimistic setting four flaws can be detected with a perfect precision and the precision values for another four flaws are very close to 1. Except for the flaws *Empty section*, *Advert*, and *Original research*, the recall values are greater than or equal to 0.8. Note in this respect that the recall can be increased by slightly adjusting the optimization criterion; for example, in the case of the *Advert* flaw using $\beta = 0.25$ results in an operating point at a target rejection rate of 0.15, with precision and recall values of 0.94 and 0.88 respectively (see Figure 4.3). For the flaws *Empty section* and *Original research* an acceptable precision can only be achieved at a low recall, indicating that the respective classifiers constitute an insufficient model of the flaws, which is also witnessed by the relatively low AUC values. For the flaw *Notability* even the achieved recall value is very high, which means that this flaw can be detected exceptionally well.

As expected, the effectiveness of the one-class classifiers deteriorates under the pessimistic setting. However, the classifiers still achieve reasonable precision values, and even in the noisy test set the flaws *Unreferenced* and *Orphan* can be detected with a good precision. Notice that the expected performance in the wild lies in between the two extremes. For some flaws the effectiveness of the one-class classifiers, in terms of both recall and precision, is pretty low under both settings, including *Empty section* and *Original research*. We explain this behavior as follows: 1. Either the document model is inadequate to capture certain flaw characteristics, or 2. the hypothesis class of the one-class classification approach is too simple to capture the flaw distributions.

4.4.3 Flaw-specific Class Imbalances

The performance values shown in Table 4.6 presume a balanced class distribution, i.e., the classifiers are evaluated with the same number of flawed articles and outliers. The real distribution of flaws in Wikipedia is unknown, and we hence report precision values as a function of the class imbalance. Given the recall and the false positive rate (*fpr*) of a classifier for the balanced setting, its precision for a class size ratio of 1:n (flawed articles : flawless articles) computes as follows:

$$\mathit{precision} = \frac{\mathit{recall}}{\mathit{recall} + n \cdot \mathit{fpr}} \tag{4.3}$$

The false positive rate is the ratio between the detected negative examples and all negative examples, and hence, it is independent from the class size ratio. The same argument applies to the recall, which is the ratio between the detected positive examples and all positive examples. Figure 4.4 shows the precision values for the ten quality flaws as a function of the flaw distribution under the optimistic and the pessimistic setting. The precision values of the 1:1 ratio correspond to the values listed in Table 4.6.

In Section 2.4.3, we proposed a measure to estimate the actual frequency of a flaw. Table 4.6 lists the estimated flaw ratios for the ten flaws. For example, the ratio of the flaw *Unreferenced* is 1:4, i.e., every fifth article is expected to contain this flaw. Figure 4.4 identifies the precision values that can be expected in the wild based on the estimated flaw ratios. Under the optimistic setting the flaws *Unreferenced*, *Orphan*, *Notability*, and *Wikify* can be detected with a precision of 1, i.e., the false positive rate is 0, and hence, the prediction performance is independent of the class imbalance. This shows that the respective one-class classifiers capture the characteristics of the flaws exceptionally well. Moreover, the expected precision values for the three flaws *Refimprove*, *Empty section*, and *Primary sources* are still good (0.62, 0.7, and 0.7 respectively); although, the estimated flaw ratio of the *Primary sources* flaw is 1:44.

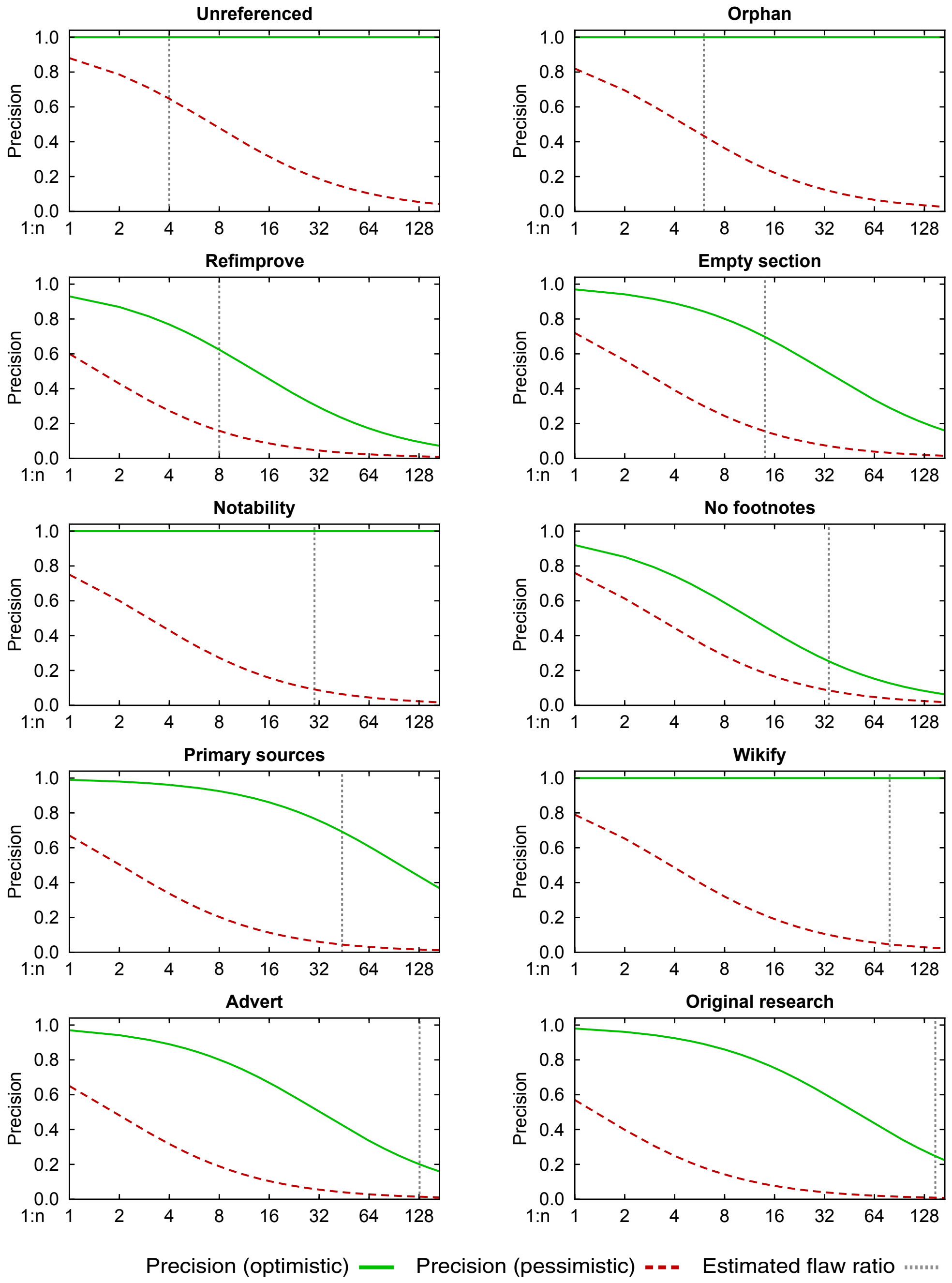


Figure 4.4: Precision under the optimistic and the pessimistic setting over flaw ratio for the ten quality flaws: 1:n (flawed articles : flawless articles) with n ∈ [1; 170]. The figure puts the classification performances reported in Table 4.6 into perspective, since it considers imbalances in the test sets that might occur in the wild.

Under the pessimistic setting the expected precision deteriorates significantly, and only the flaw *Unreferenced* can be detected with a reasonable precision of 0.65. The expected precision values for those flaws with an estimated flaw ratio 1:n where $n > 8$ are lower than 0.2. Aside from conceptual weaknesses regarding the employed document model, the weak performance indicates also that the training set of the one-class classifiers may be too small.

4.5 Competition on Quality Flaw Prediction

The "1st International Competition on Quality Flaw Prediction in Wikipedia" was initiated and co-organized by the author of this thesis, with the goals to: 1. compile and provide a uniform evaluation corpus for quality flaw prediction in Wikipedia, 2. push the development of advanced flaw prediction methods, and, 3. build up a community of researchers and practitioners working on related topics. The competition took place in conjunction with the PAN 2012 lab, held at the CLEF 2012 conference in Rome, Italy.[8] PAN is a series of evaluation labs on uncovering plagiarism, authorship, and social software misuse.[9] Wikimedia Deutschland, the German chapter of the Wikimedia Foundation, supported the competition by sponsoring the price for the winning team.

The competition addressed the problem defined in Section 4.1: Given a set of Wikipedia articles that are tagged with a particular quality flaw, decide whether an untagged article suffers from this flaw.

Evaluation Corpus

The evaluation corpus is based on the English Wikipedia snapshot from January 4, 2012. The corpus contains for each of the ten quality flaws listed in Table 4.1 Wikipedia articles that are exclusively tagged with the respective cleanup tag. The corpus contains also untagged articles, which have not been tagged with any cleanup tag. Altogether, 1 592 226 articles are provided of which 208 228 are tagged and 1 383 998 are untagged.[10] The cleanup tags are removed from the articles' wiki markup because this information is not allowed as a feature in the classification task—such features are unusable in practice.

[8] CLEF 2012, Conference and Labs of the Evaluation Forum (formerly known as Cross-Language Evaluation Forum): `http://clef2012.org`.

[9] For further information about PAN, refer to the website of the current evaluation lab: `http://pan.webis.de`.

[10] The corpus is available at `http://www.webis.de/research/corpora`.

For the PAN competition, the corpus is divided into a training corpus and a test corpus.[11] The training corpus contains tagged articles for each of the ten quality flaws plus additional 50 000 untagged articles. In the training corpus the respective class labels are given, i.e., it is known which article is tagged with what flaw. In particular, tagged articles may be considered as "positive" training examples while untagged articles may be considered as outlier examples to evaluate and tune flaw predictors (cf. Section 4.4.1, "pessimistic setting"). In case of a semi-supervised learning approach, the untagged articles serve as additional training examples.[12] The test corpus contains a balanced number of tagged articles and untagged articles for each of the ten quality flaws. In the test corpus the class labels are omitted. Moreover, it is ensured that 10% of the untagged articles are featured articles, in order to address both the optimistic and the pessimistic setting (cf. Section 4.4.1).

Flaw Prediction Approaches

Of 21 registered teams three submitted runs for the competition, see Table 4.7. Ferretti et al. [54] and Ferschke et al. [55] submitted a report describing their quality flaw classifiers, while Pistol and Iftene provided a brief description.

Ferretti et al. apply *PU learning*, a semi-supervised learning paradigm proposed by Liu et al. [102]. The algorithm is implemented as a two-step strategy: 1. a set of so-called "reliable negatives" is identified from the set of untagged articles, and 2. the reliable negatives and the tagged articles are used to train a binary classifier. Ferretti et al. employ a Naive Bayes classifier within the first step and a Support Vector Machine within the second step. Their document model is based on 73 features, which form a subset of the features that are used in this theses.[13] For each of the ten flaws the same document model is used.

Ferschke et al. regard the problem as a binary classification task, using the tagged articles as positive instances and the untagged articles as negative instances. They employ two machine learning approaches, namely a Naive Bayes classifier and C4.5 decision trees. Their document model is based on 31 feature types from the seven categories: structural features, reference features, network features, named entity features, revision-based features, lexical features, and other features. Information Gain is used for feature selection to determine a dedicated document model for each flaw.

[11] The evaluation corpus of the PAN competition is available at: `http://www.webis.de/research/events/pan-12/pan12-web/wikipedia-quality.html`.

[12] Learning from both labeled and unlabeled data is called semi-supervised learning. For further information about semi-supervised learning, refer to Chapelle et al. [37].

[13] In particular, the document model of Ferretti et al. [54] is based on a subset of the features proposed by Anderka et al. [12].

Table 4.7: Participating teams of the "1st International Competition on Quality Flaw Prediction in Wikipedia".

Team name	Participants and affiliations
Ferretti et al.	Edgardo Ferretti[⋆], Donato Hernández Fusilier[°], Rafael Guzmán Cabrera[°], Manuel Montes-y-Gómez[†], Marcelo Errecalde[⋆], and Paolo Rosso[‡] [⋆] Universidad Nacional de San Luis, Argentina [°] Universidad de Guanajuato, Mexico [†] Óptica y Electrónica (INAOE), Mexico [‡] Universidad Politécnica de Valencia, Spain
Ferschke et al.	Oliver Ferschke, Iryna Gurevych, and Marc Rittberger Technische Universität Darmstadt, Germany
Pistol and Iftene	Ionut Cristian Pistol and Adrian Iftene "Alexandru Ioan Cuza" University of Iasi, Romania

Instead of using machine learning, Pistol and Iftene resort to a rule-based approach. They define a particular set of rules for each flaw and classify an article as flawed if it fulfills the formulated requirements.

Overview of Results

The quality flaw classifiers are evaluated for each of the ten flaws individually. To determine the winning classifier, the prediction performance is judged by averaging precision, recall, and F_1-measure over all ten quality flaws. Table 4.8 shows the prediction performance of the three quality flaw classifiers. The classifier of Ferretti et al. performs best in terms of the averaged F_1-measure and the averaged recall. The classifier of Ferschke et al. achieves a slightly higher averaged precision, but a much lower averaged recall. The third classifier of Pistol and Iftene falls far behind because of a very low averaged precision. The situation is nearly the same for the individual flaws: except for the flaw *Wikify*, Ferretti et al. achieve in general a higher recall than Ferschke et al. For seven of the ten quality flaws Ferschke et al. achieve the highest precision. However, in terms of the F_1-measure the classifier of Ferretti et al. performs best for seven of the ten quality flaws.

The results of the competition can be summarized as follows: three quality flaw classifiers have been developed, which employ a total of 105 features to quantify ten important quality flaws in the English Wikipedia. Two classifiers achieve promising performance for particular flaws. An important "by-product" of the competition is the first corpus of flawed Wikipedia articles, the PAN Wikipedia quality flaw corpus 2012 (PAN-WQF-12).

Table 4.8: Evaluation results of the "1st International Competition on Quality Flaw Prediction in Wikipedia". The prediction performance for the ten quality flaws is given for each of the three teams in terms of precision, recall, and F_1-measure. The maximum F_1-measure for each flaw is shown in bold. The averaged values over all flaws are given at the bottom of the table.

Flaw name	Team name	Precision	Recall	F_1-measure
Unreferenced	Ferretti et al.	0.744731	0.954000	**0.836475**
	Ferschke et al.	0.780229	0.884000	0.828880
	Pistol and Iftene	0.056462	1.000000	0.106889
Orphan	Ferretti et al.	0.830365	0.979000	0.**898577**
	Ferschke et al.	0.862873	0.925000	0.892857
	Pistol and Iftene	0.016669	0.241000	0.031181
Refimprove	Ferretti et al.	0.734848	0.970000	**0.836207**
	Ferschke et al.	0.614566	0.751000	0.675968
	Pistol and Iftene	0.034962	0.357000	0.063687
Empty section	Ferretti et al.	0.741546	0.921000	0.821588
	Ferschke et al.	0.876081	0.912000	**0.893680**
	Pistol and Iftene	0.056462	1.000000	0.106889
Notability	Ferretti et al.	0.739655	0.858000	**0.794444**
	Ferschke et al.	0.661491	0.852000	0.744755
	Pistol and Iftene	0.055024	0.477000	0.098666
No footnotes	Ferretti et al.	0.720446	0.969000	**0.826439**
	Ferschke et al.	0.730364	0.902000	0.807159
	Pistol and Iftene	0.034518	0.170000	0.057384
Primary sources	Ferretti et al.	0.716615	0.923000	**0.806818**
	Ferschke et al.	0.735769	0.866000	0.795590
	Pistol and Iftene	0.052055	0.423000	0.092702
Wikify	Ferretti et al.	0.742195	0.737000	0.739589
	Ferschke et al.	0.677912	0.844000	**0.751893**
	Pistol and Iftene	0.056462	1.000000	0.106889
Advert	Ferretti et al.	0.736133	0.929000	0.821397
	Ferschke et al.	0.853306	0.826000	**0.839431**
	Pistol and Iftene	0.046575	0.582000	0.086248
Original research	Ferretti et al.	0.647462	0.930966	**0.763754**
	Ferschke et al.	0.739544	0.767258	0.753146
	Pistol and Iftene	0.022903	0.542406	0.043951
Averaged over all flaws	Ferretti et al.	0.735400	0.917097	**0.814529**
	Ferschke et al.	0.753213	0.852926	0.798336
	Pistol and Iftene	0.043209	0.579241	0.079449

Chapter 5

Conclusions

The machine-based assessment of text quality is becoming a topic of enormous interest. This fact is rooted, among other things, in the increasing popularity of user-generated Web content and the broad range of the delivered content's quality. A comprehensive manual quality assurance is unfeasible in this context because of the large data volumes and the constantly evolving contents. We believe that the identification of specific quality flaws is the right course of action to effectively support quality assurance activities and to improve the quality of user-generated content.

This thesis deals with the analysis and the prediction of quality flaws based on cleanup tags, taking the example of the largest and most popular user-generated knowledge source on the Web: the online encyclopedia Wikipedia. Our findings are relevant for all people who use Wikipedia, including authors, readers, and researchers. Moreover, our findings can be beneficial for a variety of information retrieval and machine learning approaches that utilize (possibly flawed) knowledge from Wikipedia [18, 32, 62, 66, 72, 76, 78, 108, 109, 122, 155].

The analyses in this thesis are limited in that they target the English Wikipedia. Nevertheless, we are confident that our research is relevant for other Wikipedia language editions as well, because all of them rely on the same fundamental principles; and the relevant concepts, such as cleanup tagging, are gaining more and more importance. It is in the nature of things that our findings cannot be applied directly to user-generated content in general: The definition of information quality varies depending on the particular context and use case [82, 156]. However, the methodology and the algorithms that are developed in this thesis can be applied to other contexts than Wikipedia as well.

In what follows, we summarize the main contributions of this thesis, discuss them in relation to our research questions RQ 1–9 that are stated in Section 1.1, draw relevant conclusions regarding the practical suitability of our findings, and formulate concrete recommendations for the Wikipedia community (Section 5.1). Finally, we give an outlook on future research directions based on the findings from this dissertation (Section 5.2).

5.1 Research Answers and Contributions

RQ 1. *How to compile the set of quality flaws that occur in Wikipedia?*

Cleanup tags provide an effective means to identify and analyze quality flaws in Wikipedia. We propose an automatic mining approach to compile an overview of cleanup tags that actually exist, which gives us the set of quality flaws that have been tagged to date. Wikipedia users currently spend a lot of their time trying to compile such an overview manually, however, with limited success. Our mining approach automates this task to save users time. Furthermore, our approach can be used to generate an up-to-date overview of cleanup tags whenever a Wikipedia user encounters some quality flaw and needs to find the respective cleanup tag. Altogether, we identify a set of 445 cleanup tags in the English Wikipedia snapshot from January 4, 2012, each of which defines a particular quality flaw.

RQ 2. *What kinds of quality flaws exist in Wikipedia?*

Wikipedia is often criticized for containing low-quality information, but until now there is no comprehensive analysis that gives empirical evidence. We close this gap: our quality flaw breakdown and the organization along flaw type, flaw scope, and flaw commonness reveals the quality flaw structure of Wikipedia. Our compilation will be useful for the development of future quality assurance strategies, such as mechanisms and policies to help editors to avoid certain types of flaws. In particular, the breakdown shows the types of flaws a user may encounter when searching information in Wikipedia.

RQ 3. *How to quantify the extent of flawed content?*

We are the first who give empirical evidence for the amount of low-quality content in Wikipedia. We quantify the flawed content that has been tagged so far by analyzing the incidence of cleanup tags. An important finding is that 26.86% of the English Wikipedia articles are tagged to contain at least one quality flaw, whereas 23.42% of the tagged flaws refer to a certain text fragment and 76.58% to the whole article. Furthermore, 50.1% of the tagged quality flaws concern article verifiability, which is one of the most important principles of an encyclopedia. The actual extent of (untagged) quality flaws is expected to be still higher.

We analyze the distribution of cleanup tags in Wikipedia's namespaces as well as in 24 top-level topics. The benefits for the Wikipedia community are twofold: First, the distribution over namespaces reveals the potential for making cleanup tagging more efficient, e.g., by adjusting the tagging policies to allow cleanup tags either in articles or in associated talk pages. Second, the analysis shows what topics are likely to contain flaws, which is valuable information for a large number of WikiProjects that are associated with the respective topic. In this respect, it has been found that tagging work in Wikipedia mostly targets the encyclopedic content and that the tagging behavior varies in different topical domains.

RQ 4. *When did the first quality flaws emerge, and how have the number and the kind of flaws changed over time?*

We investigate the evolution of quality flaws in the English Wikipedia, whereas, for the first time, the entire revision history of all articles is considered. We analyze the time period from January 2001 until January 2012, which comprises 508 243 744 revisions, whose content sum up to 7.9 TB (uncompressed). To process the large amount of data, we employ state-of-the-art data mining technology in form of MapReduce and Apache Hadoop. Our analysis reveals that the first cleanup tags emerged at the end of 2003. The early tags mainly target quality flaws that belong to the flaw types *Style of writing*, *Unwanted content*, and *Neutrality*, whereas since 2005 the flaw type *Verifiability* gained more and more importance. Our analysis also shows that the number of newly created cleanup tags per year declined since 2006, which indicates that a stable set of cleanup tags that covers all relevant quality flaws is to be expected within the next few years. Another important finding is that inline tags have become more and more popular; their percentage in relation to tag boxes increased steadily since 2006.

RQ 5. *Has the frequency, the type, and the distribution of quality flaws changed over time?*

Altogether, 6 235 192 quality flaws have been tagged in English Wikipedia articles in the period from May 2004 until January 2012, using the 445 cleanup tags. Verifiability constitutes the biggest issue, 62.45% of the tagged flaws relate to this flaw type. Tagging of quality flaws is a relevant issue for the Wikipedia community: since 2007, about 1 million flaws have been tagged annually. However, the tagging behavior has changed over time, from flagging whole articles with general cleanup statements in the initial phase of the Wikipedia project to tagging specific inline flaws in recent years. The most frequent cleanup tag is the inline tag *Citation needed*, which was created in June 2005 and has been used nearly 2 million times. This indicates that this is a fundamental quality flaw that should be particularly considered in future quality assurance activities. We also identify several cleanup tags that have seldom or never been used and that should be deleted because they rather distract than support editors.

RQ 6. *How long does it take until tagged quality flaws are corrected?*

Of the 6 235 192 tagged quality flaws, 70.04% have been corrected as of January 4, 2012. The average correction time is 155 days, where 39.38% of the corrected flaws have been corrected within the first week after tagging. We identify several quality flaws that have been tagged but that have never been corrected, and hence, the respective cleanup tags should be redefined because they are either too unspecific or too complex. Moreover, several cleanup tags show a large correction time but still address relevant flaws. Dedicated tools should be developed that support potential correctors so that these flaws can be corrected earlier. If possible, users should use inline tags rather than tag boxes to tag quality flaws as inline flaws are corrected earlier. The average correction time of inline flaws is 26.14% faster compared to a page flaw. According to this, it should be investigated whether existing tag boxes can be redefined as inline tags.

RQ 7. *How to model quality flaws?*

We compiled a set of 113 Wikipedia article features. Our feature organization relies on four dimensions: content, structure, network, and edit history. The organization reflects the computational complexity of the features as well as the source of information that is required for feature computation. The feature breakdown may serve as a reference for researchers and practitioners to help them building dedicated models for different requirements as well as for different contexts than Wikipedia. The document model that we implemented combines 65 state-of-the-art quality assessment features, which have been proposed in prior work, with 30 new features that directly target certain quality flaws. By distinguishing between intensional and extensional modeling we show that particular flaws can be modeled based on rules, which is both efficient and effective. However, most of the flaws are defined somewhat informally and hence need to be modeled in an extensional manner, based on examples that are exploited by a machine learning approach.

RQ 8. *How to predict quality flaws?*

We treat quality flaw prediction as a process where for each known flaw an expert is asked whether or not a given Wikipedia article suffers from it; the experts in turn are operationalized by one-class classifiers. The tagged articles provide a source of human-labeled data, which is used to train the classifiers. To demonstrate the applicability of this approach we design a tailored one-class classifier that is based on a combination of density estimation and class probability estimation. We are convinced that the presented or similar approaches will help to simplify Wikipedia's quality assurance process by spotting weaknesses within articles without human interaction. The automated prediction of quality flaws also supports the principle of *intelligent task routing* [40], which addresses the automatic delegation of flaws to appropriate human correctors.

RQ 9. *How to assess classifier effectiveness?*

Our evaluation takes into account the facts that representative outlier examples are not available and that the flaw-specific class distribution is unknown. We report classifier effectiveness for different settings, which enables users (Wikipedia editors, researchers, or flaw prediction tools) to assume an optimistic or a pessimistic position. Specifically, we propose the usage of featured articles as outlier examples, and we demonstrate the effects of a biased sample selection. Moreover, we illustrate classifier effectiveness as a function of the flaw distribution to consider class imbalances that might occur in the wild.

We use the evaluation corpus of the "1st International Competition on Quality Flaw Prediction in Wikipedia" to assess our flaw prediction approach. The corpus was compiled by the author of this thesis and comprises 208 228 English Wikipedia articles that are tagged with ten important quality flaws. Our flaw prediction approach achieves precision values close to 1 for eight flaws—presuming an optimistic test set with little noise and a balanced flaw distribution. Even for a class size ratio of 1:16, five flaws can still be detected with a precision of more than 0.9.

5.2 Future Research Directions

This thesis shows that an automated quality flaw prediction in Wikipedia is within reach. A possible next step could be the operationalization of our classification technology in the form of a Wikipedia bot that autonomously identifies and tags flawed articles. We demonstrate that certain flaws can be predicted with a near-perfect precision, so that the respective classifiers can directly be applied in the wild. Further research needs to be done in order to improve the prediction effectiveness for the remaining flaws. In this regard, we suggest the development of knowledge-based predictors for the individual flaws. Instead of resorting to a single document model, a flaw-specific view should be developed that combines feature selection, expert rules, and multi-level filtering. In this regard, it should be analyzed in detail which features prove essential for the prediction of a certain quality flaw. Moreover, instead of resorting to a single learning approach, the amenability of the different one-class classification approaches (density estimation, boundary identification, and reconstruction analysis [147]) with respect to the different flaws should be investigated.

The large number of untagged articles shows room also for improving the flaw prediction performance. Especially when only a few tagged articles are available for a particular flaw, untagged articles can play an important role. The work of Ferretti et al. [54] is a first step in this direction. We suggest to investigate in detail whether and how partially supervised classification techniques, such as PU-learning [102], can be applied to predict quality flaws.

An open issue is to investigate possible correlations between the 445 quality flaws. In particular, it would be interesting to analyze whether certain flaws occur relatively often in the same articles or in articles that belong to the same topic. Moreover, the flaws that have been tagged in non-content pages should be analyzed in detail. For instance, for file description pages and talk pages it is unclear whether a tag refers to the page itself or to an associated article. Another interesting question is whether and how the editing behavior changed after an article has been tagged. This question relates to the effectiveness of cleanup tagging, which is relevant not only for Wikipedia but for user-generated content in general. Further investigations are also required regarding the question how vandalism effects our analysis. We have applied simple heuristics to discard vandalism edits—nevertheless, a more sophisticated vandalism detection approach would be desirable [123, 124].

Appendix A

Overview of Quality Flaws

The following listing shows the 445 quality flaws that have been identified in the English Wikipedia snapshot from January 4, 2012 (cf. Section 2.2). The flaws are organized into twelve general flaw types (cf. Section 2.3.1). The number of flaws that belong to a particular type is given in parentheses after the type name. For each flaw type the absolute and relative frequency of tagged flaws is given, i.e., the number of times the respective cleanup tags defining the flaws that belong to this type occur in Wikipedia (cf. Section 2.3.3). The percentage relates to the whole number of tagged flaws, which is 1 903 442.

For each individual flaw its frequency, scope, and ratio is given (in parentheses after the flaw name). The scope distinguishes page flaws that refer to the whole page and inline flaws that refer to a certain text fragment, indicated by "p" and "i" respectively (cf. Section 2.3.2). The ratio 1:n (flawed articles : flawless articles) corresponds to the estimated actual frequency of a flaw (cf. Section 2.4.3). We do not provide ratios for $n > 1\,000$ since this would be less insightful.[1]

Further information about the individual quality flaws is available at the documentation pages of the respective cleanup tags. The documentation page of a flaw can be accessed using the following URL:

`http://en.wikipedia.org/wiki/Template/NAME`,

whereas `NAME` must be replaced by the flaw name.

[1] The flaw ratio is based on the number of tagged articles, whereas the flaw frequency refers to tagged flaws in all pages. Therefore, the ratio cannot be computed for flaws that are tagged solely in non-article pages, like *Copy to Wikimedia Commons* for example.

Flaw type: Verifiability (98) **Tagged flaws**: 770 416, 41.69%

Unreferenced (253 153, p, 1:4), Citation needed (236 589, i, 1:4), Refimprove (128 998, p, 1:8), Dead link (100 313, i, 1:10), BLP sources (48 089, p, 1:21), No footnotes (31 052, p, 1:33), Primary sources (23 484, p, 1:44), Cleanup-link rot (13 399, p, 1:78), Unreferenced section (12 591, p, 1:84), Who (9 692, i, 1:119), One source (8 132, p, 1:128), More footnotes (7 964, p, 1:131), Refimprove section (4 842, p, 1:219), Dubious (4 461, i, 1:280), By whom (3 856, i, 1:286), BLP IMDb refimprove (3 616, p, 1:288), Verify source (3 302, i, 1:351), Verify credibility (3 003, i, 1:389), Page needed (2 554, i, 1:434), Failed verification (2 517, i, 1:487), Citation style (2 508, p, 1:432), Ibid (1 901, p, 1:552), Subscription required (1 874, i, 1:610), Which? (1 659, i, 1:691), Volume needed (1 606, i, 1:665), Disputed (1 350, p, 1:931), BLP unsourced section (1 118, p, 1:932), Specify (1 024, i), Where (842, i), Cite quote (788, i), Full (754, i), Unreliable sources (747, p), Self-published (665, p), Section OR (615, p), Reference necessary (490, i), Self-published inline (449, i), Request quotation (368, i), Whom? (348, i), ISBN (346, p), Disputed-section (323, p), ISSN-needed (318, p), Third-party (316, p), Citation broken (297, i), Cite check (295, p), Better source (282, i), Registration required (273, i), Attribution needed (254, i), Religious text primary (226, p), Nonspecific (226, i), Page numbers needed (181, p), BLP unsourced (163, p), Primary source-inline (155, i), Hoax (133, p), BLP sources section (133, p), Unreliable medical source (112, i), Crystal (102, p), Citation needed (lead) (101, i), Chronology citation needed (95, i), Citations broken (90, p), Disputed-inline (75, i), Third-party-inline (74, i), Author missing (73, i), List fact (61, i), Copyvio link (58, i), Speculation (53, p), Year missing (48, i), Unreferenced-law (45, p), Biblio (39, p), BLP primary sources (27, p), Medical citation needed (21, i), Date missing (20, i), Better citation (18, i), Whosequote (17, i), Citation not found (16, i), ISBN missing (15, i), Title missing (12, i), Additional citation needed (12, i), Tertiary (10, i), Self-reference (7, p), Speculation-inline (6, i), Film IMDb refimprove (6, p), Circular-ref (5, i), Citation needed by (4, i), Imagefact (4, i), Page numbers improve (4, p), SCIRS (3, i), COI source (2, i), SCICN (2, i), Author incomplete (2, i), Title incomplete (2, i), Cite plot points (1, p), Citation needed cheap (1, p), Publisher missing (1, i), Check cite (1, i), Who else (1, i), BLP selfpublished (0, p), Verify sources (0, p), Unreferenced2 (0, p)

Flaw type: Style of writing (74) **Tagged flaws**: 48 174, 2.61%

Clarify (14 713, i, 1:74), When (7 337, i, 1:145), Tone (4 828, p, 1:221), Context (4 666, p, 1:230), Copy edit (2 750, p, 1:400), In-universe (2 443, p, 1:439), Essay-like (1 839, p, 1:584), Prose (1 836, p, 1:585), Technical (1 811, p), Vague (1 537, i, 1:741), Confusing (1 407, p, 1:793), Overly detailed (503, p), Rough translation (389, p), Example farm (326, p), Quantify (294, i), Off-topic (269, p), Review (261, p), Over-quotation (255, p), Trivia (253, p), Story (222, p), Elucidate (215, i), In popular culture (214, p), Contradict (186, p), Cleanup-tense (178, p), Buzzword (168, p), Travel guide (153, p), Inappropriate person (153, p), Contradict-other (144, p), Magazine (144, p), Copy edit-section (126, p), Time-context (110, p), Abbreviations (109, p), Incoherent (104, p), Examples (103, i), Confusing section (87, p), Ambiguous (86, i), Misleading (85, p), Format footnotes (71, p), Contradiction-inline (70, i), Technical-statement (65, i), Example needed (61, i), Repetition (54, p), Too many see alsos (49, p), Manual (47, p), Definition (43, p), Expand acronym (42, i), Off-topic-inline (30, i), Textbook (30, p), Capitalization (25, p), Inconsistent (24, i), Debate (21, p), Tone-inline (19, i), Inline relevance (19, i), Contradict-other-multiple (19, p), Pro and con list (17, p), Directory (16, p), Context-inline (15, i), Awkward (14, i), Term paper (10, p), Clarify-section (7, p), List missing criteria (5, p), Lacking overview (5, p), Context needed (4, i), Check quotation (3, i), Colloquial (3, p), Too abstract (3, p), Off topic sentence (3, i), Off topic paragraph (1, i), Specific time (1, i), Clarify-span (1, i), Define? (1, i), Buzz (0, i), Necessary? (0, i), Clarifyref (0, i)

(continued on the next page)

Flaw type: Miscellaneous (54) **Tagged flaws**: 127 938, 6.92%

Image requested (112 009, p), Translated page (12 276, p), Infobox requested (589, p), Cleanup-laundry (479, p), Convert to SVG and copy to Wikimedia Commons (456, p), Diagram requested (413, p), Cleanup-gallery (245, p), Map requested (214, p), Split (204, p), Copypaste (203, p), Split section (193, p), Translation WIP (155, p), Close paraphrasing (124, p), MOSLOW (92, p), Need-IPA (76, i), Duplication (74, p), TBD (73, i), Sync (69, p), Split-apart (68, p), Cleanup-translation (55, p), ORList (46, p), Cleanup-IPA (40, p), Pronunciation needed (35, i), Screenshot requested (33, p), List to table (33, p), Cleanup-images (33, p), Disputed-list (23, p), List dispute (23, p), Too many photos (19, p), Summarize (17, p), Metricate (14, p), Create-list (12, p), Cleanup split (12, p), Overcolored (12, p), Formula missing descriptions (11, p), Split sections (9, p), Overcoloured (7, p), Bad summary (7, p), Icon-issues (7, p), Cleanup-lang (7, p), Not English-inline (7, i), Split dab (6, p), Cleanup-colors (4, p), Bad unit conversions (4, p), TranslatePassage (3, p), Dubious conversion (2, i), NFimageoveruse (2, p), Romanization needed (2, i), Integrate (2, p), RJL (1, p), Whose translation (0, i), Cv? (0, p), Cleanup-list-sort (0, p), Repair coord (0, p)

Flaw type: Specific subjects (44) **Tagged flaws**: 12 929, 0.7%

Plot (2 953, p, 1:357), Ship infobox request (1 546, p), Issue (1 521, i, 1:692), Cleanup FJ biography (1 318, p), All plot (953, p), NRIS dead link (900, i), Like resume (833, p), Single infobox request (465, p), Famous (374, p), Famous players (356, p), Alumni (297, p), Mileposts (273, p), Video game cleanup (266, p), Cleanup-school (233, p), Cleanup-biography (199, p), Game guide (135, p), CIA (97, p), No plot (94, p), Episode (93, i), Fiction (89, p), USRD-wrongdir (84, p), Local (80, p), Cleanup-university (59, p), Cleanup Congress Bio (57, p), Where is it (49, p), Cleanup-tracklist (35, p), Fictionrefs (28, p), ME-fact (27, i), Season needed (25, i), Include-eb (25, i), More plot (23, p), ToLCleanup (11, p), Cleanup-comics (9, p), Kmposts (8, p), Animals cleanup (7, p), Cleanup-ICHD (7, p), Symbolism (7, p), Cleanup-book (7, p), Religion primary (6, p), Cleanup-GM (5, p), NCBI taxonomy (4, p), Nonfiction (4, p), Cleanup-chartable (4, p), Aero-table (0, p)

Flaw type: Unwanted content (42) **Tagged flaws**: 323 067, 17.48%

Copy to Wikimedia Commons (262 753, p), Notability (35 809, p, 1:29), Advert (8 487, p, 1:129), Original research (7 326, p, 1:149), Or (2 830, i, 1:437), Now Commons (2 573, p), External links (2 096, p, 1:521), Howto (409, p), NOT (368, p), Synthesis (328, p), Dicdef (257, p), Importance-section (222, p), Syn (214, i), Movenotice (185, p), Copy to Wikiquote (160, p), Non-free (130, p), Neologism (127, p), TWCleanup (95, p), Obituary (81, p), Relevance note (39, i), Copy to Wiktionary (38, p), Importance-inline (36, i), Copy to Wikisource (31, p), Copy to Wikibooks (30, p), ShadowsCommons (28, p), TWCleanup2 (21, p), Not English (18, p), Spam link (14, i), Cleanup-articletitle (14, p), Contact information (13, i), Copy to Wikiversity (10, p), Copied to Wikibooks (10, p), Schedule (9, p), Copy to Wikibooks Cookbook (7, p), Science review (6, p), Neologism inline (5, i), Copy to Meta (5, p), Move to userspace (4, p), Copied to Wikibooks Cookbook (2, p), Almanac (1, p), Copied section to Wikisource (0, p), Copied howto (0, p)

Flaw type: Neutrality (40) **Tagged flaws**: 20 865, 1.13%

POV (5 732, p, 1:200), COI (3 449, p, 1:319), Globalize (3 229, p, 1:346), Peacock (1 545, p, 1:711), POV-check (1 133, p), POV-section (1 104, p), Weasel-inline (930, i), Weasel (854, p), Says who (724, i), News release (526, p), Autobiography (465, p), Fanpov (443, p), Why? (443, i), POV-statement (370, i), Unbalanced (347, p), Peacock term (325, p), Recentism (297, p), Criticism section (296, p), Undue (218, p), NPOV language (90, p), Editorial (83, p),

(continued on the next page)

Lopsided (75, i), Geographical imbalance (55, p), Coat rack (55, p), POV-lead (52, p), Opinion (47, i), POV-title (47, p), Undue-inline (43, i), Editorializing (40, i), News release section (23, p), Cleanup-weighted (20, p), POV tag (16, i), Cherry picked (14, p), ASF (6, i), Booster (6, p), Mission (4, p), Compared to? (2, i), Strawman (1, p), Howoften (1, i), Criticism title (0, p)

Flaw type: Wiki tech (24) **Tagged flaws**: 189 466, 10.25%

Orphan (157 727, p, 1:6), Disambiguation needed (15 929, i, 1:67), Wikify (13 533, p, 1:79), Uncategorized (2 408, p, 1:701), Improve categories (1 223, p, 1:895), New infobox (957, p), Incoming links (378, p), Disambiguation cleanup (276, p), Dead end (225, p), Uncategorized stub (134, p), Overlinked (77, p), MisleadingNameLink (74, i), Missing fields (56, p), Dablinks (54, p), Category unsourced (53, p), Cleanup red links (51, p), Cleanup-infobox (48, p), Dead link header (40, p), Cleanup-HTML (10, p), Unlinked references (8, p), Recategorize (7, p), Broken (5, p), Category relevant? (4, p), More-specific-links (2, p)

Flaw type: General cleanup (17) **Tagged flaws**: 79 317, 4.29%

Multiple issues (43 486, p, 1:24), Cleanup (27 159, p, 1:39), Expert-subject (4 619, p, 1:236), Cleanup-rewrite (2 639, p, 1:408), Lead rewrite (736, p), Expert-subject-multiple (529, p), Cleanup-reorganize (519, p), Expert-talk (198, p), Cleanup-list (125, p), Further reading cleanup (87, p), Cleanup AfD (59, p), Cleanup-remainder (49, p), MOS (48, p), Prune (38, p), Spacing (20, p), Checkcategory (5, p), Refactor (3, p)

Flaw type: Expand (17) **Tagged flaws**: 68 213, 3.69%

Empty section (43 148, p, 1:24), Expand section (19 624, p, 1:54), Expand Spanish (4 299, p, 1:269), Incomplete (1 664, p, 1:759), Expand further (576, p), Missing information (255, p), Data missing (194, i), Year needed (159, i), Generalize (135, p), Incomplete table (32, p), Specific (18, p), Alphabetize (13, p), List years (7, p), Generalize-section (6, p), Called (1, i), Idetail (1, i), Cleanup-statistics (0, p)

Flaw type: Structure (14) **Tagged flaws**: 7 632, 0.41%

Lead too short (4 570, p, 1:232), Lead missing (1 657, p, 1:638), Very long (456, p), Sections (396, p), Lead too long (267, p), Inadequate lead (191, p), Condense (70, p), Cleanup-combine (32, p), Section-diffuse (22, p), Summarize section (20, p), Sub-sections (18, p), Summary style (7, p), Section-sort (2, p), Too-many-boxes (1, p)

Flaw type: Time-sensitive (13) **Tagged flaws**: 10 629, 0.58%

Update (5 102, p, 1:211), Update after (4 197, i, 1:509), Out of date (1 273, p, 1:839), As of? (92, i), Recently revised (56, p), Unclear date (16, p), Clarify timeframe (12, i), Anachronism (9, p), Time needed (6, i), Outdated as of (4, p), Oldfact (3, i), Currentevent-inline (2, i), Time references needed (0, p)

Flaw type: Merge (8) **Tagged flaws**: 18 800, 1.02%

Merge to (7 052, p, 1:153), Merge (4 289, p, 1:258), Merge from (3 633, p, 1:300), Merged-to (1 638, p), Merged-from (1 620, p), Afd-merge from (341, p), Afd-merge to (338, p), Merging (10, p)

Appendix B

Wikipedia Article Features

This chapter provides detailed descriptions of Wikipedia article features that were either proposed in prior studies on quality assessment in Wikipedia or are newly introduced in this thesis. We organize the features along four dimensions: content, structure, network, and edit history. Section 4.2 discusses the dimensions and the respective features in the context of automated quality flaw prediction. Table 4.2, 4.3, 4.4, and 4.5 list the individual features and provide references to prior work. We implemented the major part of the previously proposed features; implementation details will be provided where it is appropriate.

B.1 Content Features

The computation of content features requires a conversion of an article's wiki markup into plain text. We employ an extended version of *Wikipedia Extractor*, which is a light-weight tool written in Python that uses a rule-based approach to convert wiki markup into plain text.[1] This section describes the content features listed in Table 4.2, organized into four subcategories: text statistics, part of speech, readability formulas, and closed-class word sets.

Text Statistics

The following features quantify basic text statistics. To compute these features, the articles' plain texts are preprocessed using the publicly available Natural Language Toolkit, NLTK. NLTK is an open source library written in Python that comprises a variety of natural language processing and computational linguistic tools.[2]

[1] Wikipedia Extractor: `http://medialab.di.unipi.it/wiki/Wikipedia_Extractor`.

[2] For further information about the Natural Language Toolkit, NLTK, see `http://nltk.org` or refer to Bird et al. [20].

- *Character count.* Number of characters, without spaces, digits, and punctuation.
- *Character trigrams.* Character trigram distribution of the article. Character trigrams were originally applied for writing style analysis [138]. Lipka and Stein [100] were the first who used this feature for quality assessment of Wikipedia articles.
- *Complex word rate.* Percentage of words with three or more syllables, excluding proper nouns.
- *Information-to-noise ratio.* Ratio between the number of individual index terms (bag-of-word stems without stop words) and the word count. This measure was originally proposed by Zhu and Gauch [163], as "the proportion of useful information contained in a Web page of a given size."
- *Long sentence rate.* Percentage of long sentences. A sentence is defined as long if it contains at least 30 words.
- *Long word rate.* Percentage of words with more than six characters.
- *Longest sentence length.* Number of words in the longest sentence.
- *One-syllable word count.* Number of one-syllable words.
- *One-syllable word rate.* Percentage of one-syllable words.
- *Paragraph count.* Number of paragraphs. Paragraphs are separated by two consecutive newline characters.
- *Paragraph length.* Average paragraph length in sentences. (Dalip et al. [46] use the character count to compute the mean paragraph size.)
- *Question count.* Number of sentences that are questions. We identify questions as sentences that end with a question mark.
- *Question rate.* Percentage of sentences that are questions.
- *Sentence count.* Number of sentences. We use NLTK's *Punkt* module to tokenize the text into sentences.
- *Sentence length.* Average sentence length in words.
- *Short sentence rate.* Percentage of short sentences. A sentence is defined as short if it contains at most 15 words.
- *Shortest sentence length.* Number of words in the shortest sentence.

- *Syllable count.* Number of syllables. To estimate the number of syllables for each word, we used the Perl module *Lingua::EN::Syllable.*[3]
- *Word count.* Number of words, without digits. We use NLTK's *Punkt* module to tokenize the text into words.
- *Word length.* Average word length in characters.
- *Word syllables.* Average number of syllables per word.

Part of Speech

The following features quantify parts of speech, and therefore, the usage of words. We employ a part-of-speech tagger that is developed against NLTK and that is trained from the Brown Corpus to identify word classes.[4]

- *Article sentence rate.* Percentage of sentences beginning with an article.
- *Auxiliary verb rate.* Percentage of auxiliary verbs.
- *Conjunction rate.* Percentage of coordinating conjunctions and subordinating conjunctions.
- *Conjunction sentence rate.* Percentage of sentences beginning with a conjunction.
- *Interrogative pronoun sentence rate.* Percentage of sentences beginning with an interrogative pronoun.
- *Nominalization rate.* Percentage of nominalizations. A word is a nominalization if its suffix is equal to either "tion", "ment", "ence", or "ance".[5]
- *Passive sentence rate.* Percentage of passive voice sentences. A sentence is identified as passive voice if it contains a "to be" verb and then later on a non-gerund.
- *Preposition rate.* Percentage of prepositions.
- *Preposition sentence rate.* Percentage of sentences beginning with a preposition.
- *Pronoun sentence rate.* Percentage of sentences beginning with a pronoun.

[3] *Lingua::EN::Syllable* – Routine for estimating syllable count in words by Greg Fast: `http://search.cpan.org/dist/Lingua-EN-Syllable`.

[4] The part-of-speech tagger is available at: `http://code.google.com/p/narorumo/source/browse/trunk/passive`.

[5] This definition of a nominalization is used in the GNU Style software: `http://www.gnu.org/software/diction`.

- *Pronoun rate.* Percentage of personal, interrogative, relative, indefinite, and demonstrative pronouns.
- *Subordinate conjunction sentence rate.* Percentage of sentences beginning with a subordinate conjunction.
- *"To be" verb rate.* Percentage of "to be" verbs.

Readability Formulas

The following features measure readability of text based on basic text statistics, like character, word, syllable, and sentence counts. Several formulas provide a score indicating the grade level that readers need to comprehend a text.

- *Automated Readability Index, ARI.* Approximates the age needed to understand a text. Proposed by Senter and Smith [134].

$$ari = 4.71 \cdot \frac{characterCount}{wordCount} + 0.5 \cdot \frac{wordCount}{sentenceCount} - 21.43$$

- *Bormuth Index.* Estimates the reading grade level required to read a text. Proposed by Bormuth [25].

$$\begin{aligned} bormuth = & 0.886593 - 0.03640 \cdot \frac{characterCount}{wordCount} + \\ & 0.161911 \cdot \frac{difficultWordCount}{wordCount} - 0.21401 \cdot \frac{wordCount}{sentenceCount} - \\ & 0.000577 \cdot \frac{wordCount}{sentenceCount} - 0.000005 \cdot \frac{wordCount}{sentenceCount} \end{aligned}$$

A word is considered as difficult if it is not included in the Dale-Chall list of 3 000 common words (see *New Dale-Chall*).

- *Coleman-Liau Index.* Approximates the U.S. grade level needed to understand a text. Proposed by Coleman and Liau [38].

$$coleman\text{-}liau = 5.89 \cdot \frac{characterCount}{wordCount} - 30 \cdot \frac{sentenceCount}{wordCount} - 15.8$$

- *FORCAST Readability.* Estimates the years of education required to understand a text. Proposed by Caylor and Sticht [34].

$$forcast = 20 - \frac{oneSyllablesCount}{10}$$

- *Flesch Reading Ease.* Estimates the ease of reading and understanding a text. The formula computes a value between 0 and 100; the higher the number, the easier the text is to read. Proposed by Flesch [57].

$$\mathit{flesch} = 206.835 - 1.015 \cdot \frac{\mathit{wordCount}}{\mathit{sentenceCount}} - 84.6 \cdot \frac{\mathit{syllablesCount}}{\mathit{wordCount}}$$

- *Flesch-Kincaid.* Modification of *Flesch Reading Ease* to produce a grade-level score. Proposed by Kincaid et al. [83].

$$\mathit{flesch\text{-}kincaid} = 0.39 \cdot \frac{\mathit{wordCount}}{\mathit{sentenceCount}} + 11.8 \cdot \frac{\mathit{syllablesCount}}{\mathit{wordCount}} - 15.59$$

- *Gunning Fog Index.* Measures reading ease. The score can also be mapped to grade levels. Proposed by Gunning [68].

$$\mathit{gunning\text{-}fog} = 0.4 \cdot \left(\frac{\mathit{wordCount}}{\mathit{sentenceCount}} + 100 \cdot \frac{\mathit{complexWordCount}}{\mathit{wordCount}} \right)$$

  Whereas words with three or more syllables are considered as complex words, excluding proper nouns.

- *Läsbarhedsindex, LIX.* Assesses text difficulty. Proposed by Björnsson [21]. Depending on the language of the text the score is interpreted differently.

$$\mathit{lix} = \frac{\mathit{wordCount}}{\mathit{sentenceCount}} + 100 \cdot \frac{\mathit{longWordCount}}{\mathit{wordCount}}$$

  Whereas long words are words with more than six characters.

- *Miyazaki EFL Readability Index.* Computes a reading ease score between 0 and 100; the higher the number, the easier the text is to read. Proposed by Greenfield [67].

$$\mathit{miyazaki\text{-}efl} = 164.935 - 18.792 \cdot \frac{\mathit{characterCount}}{\mathit{wordCount}} - 1.916 \cdot \frac{\mathit{wordCount}}{\mathit{sentenceCount}}$$

- *New Dale-Chall.* The new Dale-Chall readability formula provides a score that measures the difficulty to comprehend a text. The formula was originally proposed by Edgar Dale and Jeanne Chall in 1948 [45]. Here, we used the revised version.

$$\mathit{new\text{-}dale\text{-}chall} = 0.1579 \cdot \frac{\mathit{difficultWordCount}}{\mathit{wordCount}} + 0.0496 \cdot \frac{\mathit{wordCount}}{\mathit{sentenceCount}}$$

  A word is considered as difficult if it is not included in a list of 3 000 common words. If the difficult words account for more than 5%, 3.6365 is added to the score (adjusted score).

- *SMOG Grading.* Estimates the reading grade a person must have reached to understand a text. Proposed by McLaughlin [107].

  $$\mathit{smog\text{-}grade} = \sqrt{30 \cdot \frac{\mathit{complexWordCount}}{\mathit{sentenceCount}}} + 3$$

  Whereas words with three or more syllables, excluding proper nouns, are considered as complex words.

Closed-class Word Sets

The following features measure the presence of certain words from predefined closed-class word sets.

- *Common word rate.* Percentage of words that are included in the list of 3 000 common words used in the *New Dale-Chall* readability formula.
- *Difficult word rate.* Percentage of words that are not included in the list of 3 000 common words used in the *New Dale-Chall* readability formula.
- *Peacock word rate.* Percentage of words and phrases that promote the subject of an article by making unprovable proclamations about its importance; see Table B.1.
- *Stop word rate.* Percentage of words that occur very frequently and that have little lexical meaning, such as function words, articles, and conjunctions. We use the English stop word list provided by NLTK.
- *Weasel word rate.* Percentage of words and phrases that are intended to create an impression of authority and relevance without giving the reader the possibility to verify the respective statement (also called anonymous authority); see Table B.1.

B.2 Structure Features

The computation of structure features is based on an article's wiki markup. For a description of the wiki markup syntax and the general article layout, refer to the respective MediaWiki help page.[6] This section describes the structure features listed in Table 4.3.

[6] MediaWiki help page "Help:Wiki markup": `http://en.wikipedia.org/wiki/Help:Wiki_markup`.

Table B.1: Words and phrases used to compute the two closed-class word set features *Peacock word rate* and *Weasel word rate*.

Feature	Words / phrases
Peacock word rate	acclaimed, amazing, astonishing, authoritative, beautiful, best, brilliant, canonical, celebrated, charismatic, classic, cutting-edge, defining, definitive, eminent, enigma, exciting, extraordinary, fabulous, famous, infamous, fantastic, fully, genius, global, great, greatest, iconic, immensely, impactful, incendiary, indisputable, influential, innovative, inspired, intriguing, leader, leading, legendary, major, masterly, mature, memorable, notable, outstanding, pioneer, popular, prestigious, really good, remarkable, renowned, respected, seminal, significant, skillful, solution, single-handedly, staunch, talented, the most, top, transcendent, undoubtedly, unique, visionary, virtually, virtuoso, well-known, well-established, world-class, worst
Weasel word rate	about, adequate, and/or, appropriate, approximately, are a number, as applicable, as circumstances dictate, as much as possible, as needed, as required, as soon as possible, at your earliest convenience, basically, clearly, completely, critics say, depending on, exceedingly, excellent, experts declare, extremely, fairly, few, frequently, good, huge, if appropriate, if required, if warranted, is a number, in a timely manner, in general, in most cases, in our opinion, in some cases, in most instances, indicated, interestingly, it is believed, it is often reported, it is our understanding, it is widely thought, it may, it was proven, largely, major, make an effort to, many, many are of the opinion, many people think, maybe, more or less, most feel, mostly, normally, often, on occasion, perhaps, primary, quite, relatively, relevant, remarkably, research has shown, roughly, science says, significantly, several, should be, some people say, sometimes, striving for, substantially, suitable, surprisingly, tentatively, tiny, try, typically, usually, valid, various, vast, very, we intend to, when necessary, when possible

- *File count.* Number of files including images and other media files, identified by file links: `[[file:...]]`.
- *Category count.* Number of Wikipedia categories an article belongs to, identified by category links: `[[category:...]]`.
- *Heading count.* Total number of headings, including section, subsection, and subsubsection headings.
- *Image count.* Number of images, identified by image links: `[[image:...]]`.
- *Images per section.* Ratio between the image count and the section count.
- *Infobox count.* Number of infoboxes. Infoboxes are fixed-format tables used to summarize relevant information in a unified and structured manner (typically in the top right-hand corner of an article).

- *Lead length.* Number of words in the lead section. (Dalip et al. [46] use the character count instead of word count.) A lead section is defined as the text before the first heading.
- *Lead rate.* Percentage of words in the lead section.
- *List ratio.* Percentage of words in lists. A list can be either an itemization, an enumeration, or a definition; identified by lines starting with an asterisk (`*`), a number sign (`#`), or a semicolon (`;`) respectively.
- *Reference count.* Number of references and citations used in an article, identified by the tags: `<ref>...</ref>`.
- *Reference sections count.* Number of reference sections, identified by the section heading. We use the headings listed in Table B.2.
- *References per section.* Ratio between the reverence count and the section count.
- *Reference per text length.* Ratio between the reference count and the word count. (Dalip et al. [46] use the character count instead of the word count.)
- *Section count, subsection count, subsubsection count.* Number of sections, subsections, and subsubsections.
- *Section length, subsection length, subsubsection length.* Average section, subsection, and subsubsection length in words. (Dalip et al. [46] use the character count to compute the mean section size.)
- *Section length deviation.* Standard deviation of section length.
- *Section nesting, subsection nesting.* Average number of subsections per section and average number of subsubsections per subsection.
- *Shortest section length, shortest subsection length, shortest subsubsection length.* Number of words in the shortest section, subsection, and subsubsection.
- *Longest section length, longest subsection length, longest subsubsection length.* Number of words in the longest section, subsection, and subsubsection.
- *Table count.* Number of tables, identified by: `{|...|}`.
- *Template count.* Number of templates that are used in an article.
- *Trivia sections count.* Number of trivia sections, identified by the section heading. We use the following headings: “facts”, “miscellanea”, “other facts”, “other information”, and “trivia”.

Table B.2: Common headings of reference sections in English Wikipedia articles, used to compute the *Reference sections count* feature.

“references”, “notes”, “footnotes”, “sources”, “citations”, “bibliography”, “works cited”, “external references”, “reference notes”, “references cited”, “bibliographical references”, “cited references”, “notes, references”, “sources, references, external links”, “sources, references, external links, quotations”, “notes & references”, “references & notes”, “external links & references”, “references & external links”, “references & footnotes”, “footnotes & references”, “citations & notes”, “notes & sources”, “sources & notes”, “notes & citations”, “footnotes & citations”, “citations & footnotes”, “reference & notes”, “footnotes & sources”, “note & references”, “notes & reference”, “sources & footnotes”, “notes & external links”, “references & further reading”, “sources & references”, “references & sources”, “references & links”, “links & references”, “references & bibliography”, “references & resources”, “bibliography & references”, “external articles & references”, “references & citations”, “citations & references”, “references & external link”, “external link & references”, “further reading & references”, “notes, sources & references”, “sources, references & external links”, “references/notes”, “notes/references”, “notes/further reading”, “references/links”, “external links/references”, “references/external links”, “references/sources”, “external links / references”, “references / sources”, “references / external links”

B.3 Network Features

The computation of network features is based on Wikipedia’s link graph. In the graph articles represent nodes and internal links (links between articles) represent edges. We use our local Wikipedia database, described in Section 2.1, for feature computation. The link graph is build based on the *pagelinks* table (cf. Table 2.2). This section describes the network features listed in Table 4.4.

- *Assortativity.* Ratio between an article’s degree and the average degree of its neighbors. The degree of an article (node) is the sum of the incoming internal link count and the internal link count (incoming and outgoing edges). This measure was originally used for Web spam detection [33].

- *Broken internal link count.* Number of internal links that refer to non-existing articles.

- *Clustering coefficient.* Ratio between the number of existing edges and the number of all possible edges between an article and its nearest neighbors.

- *External link count.* Number of links that point to sources outside of Wikipedia. External links are identified using the *externallinks* table of the Wikipedia database.

- *External links per section.* Ratio between the external link count and the section count.

- *Incoming internal link count.* Number of incoming internal links.
- *Internal link count.* Number of outgoing internal links.
- *Internal links per text length.* Ratio between the internal link count and the word count.
- *Inter-language link count.* Number of links that point to the same article in a different language. Inter-language links are identified using the *langlinks* table of the Wikipedia database.
- *PageRank.* The article's PageRank value, computed according to Brin and Page [28] based on Wikipedia's link graph.
- *Reciprocity.* Ratio between the incoming internal link count and the internal link count.

B.4 Edit History Features

Edit history features rely on an article's revision history. We use our local Wikipedia database for feature computation; in particular, we use the table *revision*, which comprises meta data for all edits ever made (cf. Section 3.1.2). In what follows, the features listed in Table 4.5 are described.

- *Active editor rate.* Percentage of edits made by the top 5% of most active editors of an article.
- *Admin editor rate.* Percentage of edits made by administrator users.
- *Age.* Days between article creation and now (date of the snapshot).
- *Age per edit.* Ratio between the age and the edit count.
- *Anonymous editor rate.* Percentage of edits made by anonymous users, identified by the editor's username, which is an IP address in case of an anonymous user.
- *Connectivity.* For a given article, the number of articles with common editors. Two articles have common editors when at least one of their revisions has been made by the same user.
- *Currency.* Days between the last edit and now (date of the snapshot).
- *Discussion edit count.* Edit count of an article's discussion page.
- *Edit count.* Number of edits (or number of revisions respectively).

- *Edit currency rate.* Percentage of edits made in the last three months.
- *Edits per editor.* Ratio between the edit count and the editor count.
- *Edits per editor deviation.* Standard deviation of edits per editor.
- *Editor count.* Number of distinct editors (all users who edited an article).
- *Editor rate.* Ratio between the editor count and the edit count.
- *Modified lines rate.* Number of lines that have been modified when comparing the current revision to a revision three-months old.
- *Occasional editor rate.* Percentage of edits made by users who edited an article less than four times.
- *ProbReview.* A measure proposed by Hu et al. [77] that quantifies article quality based on the quality of its editors.
- *"Quick-turnaround" edit rate.* Percentage of edits that followed within 30 minutes on a previous edit, whereas both edits were made by different (human) editors.
- *Registered editor rate.* Percentage of edits made by registered users, identified by the editor's username.
- *Revert count.* Number of reversions, i.e., the number of times a previous revision has been restored. For more information on reverts and respective detection approaches, refer to Flöck et al. [58].
- *Revert time.* Average duration in minutes between an edit and its reversion.

References

[1] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the Wikipedia. In *Proceedings of the 16th international conference on World Wide Web (WWW'07)*, pages 261–270. ACM, 2007. ISBN 978-1-59593-654-7. doi: 10.1145/1242572.1242608. 20

[2] Eugene Agichtein, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. Finding high-quality content in social media. In *Proceedings of the 1st ACM international conference on Web search and Web data mining (WSDM'08)*, pages 183–194. ACM, 2008. ISBN 978-1-59593-927-2. doi: 10.1145/1341531.1341557. 17

[3] Rodrigo B. Almeida, Barzan Mozafari, and Junghoo Cho. On the evolution of Wikipedia. In *Proceedings of the 1st international conference on Weblogs and social media (ICWSM'07)*, 2007. 10, 12

[4] Maik Anderka and Benno Stein. The ESA retrieval model revisited. In *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'09)*, pages 670–671. ACM, 2009. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1572070. 7, 8

[5] Maik Anderka and Benno Stein. A breakdown of quality flaws in Wikipedia. In *Proceedings of the 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12)*, pages 11–18. ACM, 2012. ISBN 978-1-4503-1237-0. doi: 10.1145/2184305.2184309. 7, 8

[6] Maik Anderka and Benno Stein. Overview of the 1st international competition on quality flaw prediction in Wikipedia. In *Notebook Papers of CLEF'12 Labs and Workshops*, 2012. ISBN 978-88-904810-3-1. 7, 8

[7] Maik Anderka, Nedim Lipka, and Benno Stein. Evaluating cross-language explicit semantic analysis and cross querying. In *Multilingual information access evaluation I: text retrieval experiments. Selected papers of the 10th cross-language evaluation forum (CLEF'09)*, pages 50–57. Springer, 2009. ISBN 978-3-642-15753-0. doi: 10.1007/978-3-642-15754-7_4. 7, 8

[8] Maik Anderka, Benno Stein, and Martin Potthast. Cross-language high similarity search: why no sub-linear time bound can be expected. In *Advances in information retrieval. 32nd european conference on information retrieval research (ECIR'10)*, pages 640–644. Springer, 2010. ISBN 978-3-642-12274-3. doi: 10.1007/978-3-642-12275-0_66. 7, 8

[9] Maik Anderka, Benno Stein, and Nedim Lipka. Towards automatic quality assurance in Wikipedia. In *Proceedings of the 20th international conference on World Wide Web (WWW'11)*, pages 5–6. ACM, 2011. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963196. 7, 8, 28

[10] Maik Anderka, Benno Stein, and Nedim Lipka. Detection of text quality flaws as a one-class classification problem. In *Proceedings of the 20th ACM international conference on information and knowledge management (CIKM'11)*, pages 2313–2316. ACM, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063954. 7, 8

[11] Maik Anderka, Benno Stein, and Matthias Busse. On the evolution of quality flaws and the effectiveness of cleanup tags in the English Wikipedia. In *Wikipedia Academy (WPAC'12)*, 2012. 7, 8

[12] Maik Anderka, Benno Stein, and Nedim Lipka. Predicting quality flaws in user-generated content: the case of Wikipedia. In *Proceedings of the 35th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'12)*, pages 981–990. ACM, 2012. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348413. 7, 8, 79

[13] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. Gender differences in Wikipedia editing. In *Proceedings of the 7th international symposium on wikis and open collaboration (WikiSym'11)*, pages 11–14. ACM, 2011. ISBN 978-1-4503-0909-7. doi: 10.1145/2038558.2038561. 12

[14] Gordana Apic, Matthew J. Betts, and Robert B. Russell. Content disputes in Wikipedia reflect geopolitical instability. *PLoS ONE*, 6(6), 2011. doi: 10.1371/journal.pone.0020902. 21

[15] Pablo Aragón, Andreas Kaltenbrunner, David Laniado, and Yana Volkovich. Biographical social networks on Wikipedia: a cross-cultural study of links that made history. In *Proceedings of the 8th international symposium on wikis and open collaboration (WikiSym'07)*. ACM, 2012. ISBN 978-1-4503-1605-7. 13

[16] B. Thomas Adler, Luca de Alfaro, and Ian Pye. Detecting Wikipedia vandalism using WikiTrust: lab report for PAN at CLEF'10. In *Notebook Papers of CLEF'10 Labs and Workshops*, 2010. ISBN 978-88-904810-2-4. 45

[17] Ricardo Baeza-Yates. User generated content: how good is it? In *Proceedings of the 3rd workshop on information credibility on the Web (WICOW'09)*, pages 1–2. ACM, 2009. ISBN 978-1-60558-488-1. doi: 10.1145/1526993.1526995. 16

[18] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using Wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'07)*, pages 787–788. ACM, 2007. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277909. 14, 83

[19] Michael Bendersky, W. Bruce Croft, and Yanlei Diao. Quality-biased ranking of web documents. In *Proceedings of the 4th ACM international conference on Web search and Web data mining (WSDM'11)*, pages 95–104. ACM, 2011. ISBN 978-1-4503-0493-1. doi: 10.1145/1935826.1935849. 16

[20] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python*. O'Reilly, 2009. 93

[21] Carl-Hugo Björnsson. Läsbarhet. Stockholm: Liber, 1968. 97

[22] Joshua E. Blumenstock. Size matters: word count as a measure of quality on Wikipedia. In *Proceedings of the 17th international conference on World Wide Web (WWW'08)*, pages 1095–1096. ACM, 2008. ISBN 978-1-60558-085-2. doi: 10.1145/1367497.1367673. 2, 19, 65, 66

[23] Joshua E. Blumenstock. Automatically assessing the quality of Wikipedia articles. Technical report, Recent work, school of information, UC Berkeley, 2008. 64

[24] Max Boisot and Agustí Canals. Data, information and knowledge: have we got it right? *Journal of Evolutionary Economics*, 14(1):43–67, 2004. ISSN 0936-9937. doi: 10.1007/s00191-003-0181-9. 16

[25] John R. Bormuth. Readability: a new approach. *Reading Research Quarterly*, 1(3):79–132, 1966. ISSN 00340553. doi: 10.2307/747021. 96

[26] Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. ISSN 0885-6125. doi: 10.1023/A:1018054314350. 70

[27] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. 70

[28] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998. ISSN 0169-7552. doi: 10.1016/S0169-7552(98)00110-X. 102

[29] Luciana S. Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. Temporal analysis of the Wikigraph. In *Proceedings of the 2006 IEEE/WIC/ACM international conference on Web intelligence (WI'06)*, pages 45–51. IEEE Computer Society, 2006. ISBN 0-7695-2747-7. doi: 10.1109/WI.2006.164. 10

[30] Brian Butler, Elisabeth Joyce, and Jacqueline Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In *Proceeding of the 26th annual SIGCHI conference on human factors in computing systems (CHI'08)*, pages 1101–1110. ACM, 2008. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357227. 11

[31] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74(3):036116, 2006. doi: 10.1103/PhysRevE.74.036116. 10

[32] David Carmel, Haggai Roitman, and Naama Zwerdling. Enhancing cluster labeling using Wikipedia. In *Proceedings of the 32nd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'09)*, pages 139–146. ACM, 2009. ISBN 978-1-60558-483-6. doi: 10.1145/1571941.1571967. 13, 83

[33] Carlos Castillo, Debora Donato, Aristides Gionis, Vanessa Murdock, and Fabrizio Silvestri. Know your neighbors: Web spam detection using the Web topology. In *Proceedings of the 30th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'07)*, pages 423–430. ACM, 2007. ISBN 978-1-59593-597-7. doi: 10.1145/1277741.1277814. 101

[34] John S. Caylor and Thomas G. Sticht. Development of a simple readability index for job reading material. Technical report, Human Resources Research Organization, Monterey, CA. Div. 3., 1973. 96

[35] Kevin Chai, Vidyasagar Potdar, and Tharam Dillon. Content quality assessment related frameworks for social media. In *Proceedings of the international conference on computational science and its applications: part II (ICCSA'09)*, pages 791–805. Springer, 2009. ISBN 978-3-642-02456-6. doi: 10.1007/978-3-642-02457-3_65. 17

[36] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: a survey. *ACM Computing Surveys (CSUR)*, 41(3):15:1–15:58, 2009. ISSN 0360-0300. doi: 10.1145/1541880.1541882. 61

[37] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning.* MIT Press, 2006. 79

[38] Meri Coleman and T. L. Liau. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284, 1975. doi: 10.1037/h0076540. 96

[39] Kevyn Collins-Thompson and Jamie Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of the human language technology conference / North American chapter of the association for computational linguistics annual meeting (HLT/NAACL'04)*, pages 193–200, 2004. 63

[40] Dan Cosley, Dan Frankowski, Loren Terveen, and John Riedl. Using intelligent task routing and contribution review to help communities build artifacts of lasting value. In *Proceedings of the 24th SIGCHI conference on human factors in computing systems (CHI'06)*, pages 1037–1046. ACM, 2006. ISBN 1-59593-372-7. doi: 10.1145/1124772.1124928. 87

[41] Holly Crawford. Encyclopedias. In Richard E. Bopp and Linda C. Smith, editors, *Reference and information services: an introduction*, pages 433–459, Englewood, CO, 2001. Libraries Unlimited. 17, 18, 62

[42] T. Cross. Puppy smoothies: improving the reliability of open, collaborative wikis. *First Monday*, 11(9), 2006. URL `http://131.193.153.231/www/issues/issue11_9/cross/index.html`. 20

[43] Carlo Curino, Hyun J. Moon, Letizia Tanca, and Carlo Zaniolo. Schema evolution in Wikipedia: toward a web information system benchmark. In *Proceedings of the 10th international conference on enterprise information systems (ICEIS'08), Volume DISI*, pages 323–332, 2008. ISBN 978-989-8111-36-4. 15

[44] Alberto Cusinato, Vincenzo Della Mea, Francesco Di Salvatore, and Stefano Mizzaro. QuWi: quality control in Wikipedia. In *Proceedings of the 3rd workshop on information credibility on the Web (WICOW'09)*, pages 27–34. ACM, 2009. ISBN 978-1-60558-488-1. doi: 10.1145/1526993.1527001. 19

[45] Edgar Dale and Jeanne S. Chall. A formula for predicting readability: instructions. *Educational Research Bulletin*, 27(2), 1948. ISSN 15554023. URL `http://www.jstor.org/stable/1473669`. 97

[46] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic quality assessment of content created collaboratively by Web communities: a case study of Wikipedia. In *Proceedings of the 9th ACM/IEEE-CS joint conference on digital libraries (JCDL'09)*, pages 295–304. ACM, 2009. ISBN 978-1-60558-322-8. doi: 10.1145/1555400.1555449. 2, 19, 65, 66, 67, 94, 100

[47] Daniel Hasan Dalip, Marcos André Gonçalves, Marco Cristo, and Pável Calado. Automatic assessment of document quality in Web collaborative digital libraries. *Journal of Data and Information Quality (JDIQ)*, 2(3): 14:1–14:30, 2011. ISSN 1936-1955. doi: 10.1145/2063504.2063507. 64

[48] Sanmay Das and Malik Magdon-Ismail. Collective wisdom: information growth in wikis and blogs. In *Proceedings of the 11th ACM conference on electronic commerce (EC'10)*, pages 231–240. ACM, 2010. ISBN 978-1-60558-822-3. doi: 10.1145/1807342.1807380. 10

[49] Gabriel De la Calzada and Alex Dekhtyar. On measuring the quality of Wikipedia articles. In *Proceedings of the 4th workshop on information credibility on the Web (WICOW'10)*, pages 11–18. ACM, 2010. ISBN 978-1-60558-940-4. doi: 10.1145/1772938.1772943. 19

[50] Jeffrey Dean and Sanjay Ghemawat. MapReduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008. ISSN 0001-0782. doi: 10.1145/1327452.1327492. 44

[51] Gregory Druck, Gerome Miklau, and Andrew McCallum. Learning to predict the quality of contributions to Wikipedia. In *Proceedings of the AAAI workshop on Wikipedia and artificial intelligence (WIKIAI'08)*, pages 7–12. AAAI Press, 2008. 20

[52] Katherine Ehmann, Andrew Large, and Jamshid Beheshti. Collaboration in context: comparing article evolution among subject disciplines in wikipedia. *First Monday*, 13(10), 2008. URL `http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/viewArticle/2217/2034`. 37

[53] Tom Fawcett. Roc graphs: notes and practical considerations for researchers. Technical report, HP Laboratories, 2004. 75

[54] Edgardo Ferretti, Donato Hernández Fusilier, Rafael Guzmán Cabrera, Manuel Montes-y-Gómez, Marcelo Errecalde, and Paolo Rosso. On the use of PU Learning for quality flaw prediction in Wikipedia: notebook for PAN at CLEF 2012. In *Notebook Papers of CLEF'12 Labs and Workshops*, 2012. ISBN 978-88-904810-3-1. 79, 80, 88

[55] Oliver Ferschke, Iryna Gurevych, and Marc Rittberger. FlawFinder: a modular system for predicting quality flaws in Wikipedia: notebook for PAN at CLEF 2012. In *Notebook Papers of CLEF'12 Labs and Workshops*, 2012. ISBN 978-88-904810-3-1. 79, 80

[56] Flavio Figueiredo, Fabiano Belém, Henrique Pinto, Jussara Almeida, Marcos Gonçalves, David Fernandes, Edleno Moura, and Marco Cristo. Evidence of quality of textual features on the Web 2.0. In *Proceeding*

of the 18th ACM international conference on information and knowledge management (CIKM'09), pages 909–918. ACM, 2009. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646070. 17

[57] Rudolph Flesch. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233, 1948. doi: 10.1037/h0057532. 97

[58] Fabian Flöck, Denny Vrandečić, and Elena Simperl. Revisiting reverts: accurate revert detection in Wikipedia. In *Proceedings of the 23rd ACM conference on hypertext and social media (HT'12)*, pages 3–12. ACM, 2012. ISBN 978-1-4503-1335-3. doi: 10.1145/2309996.2310000. 103

[59] Andrea Forte and Amy Bruckman. Scaling consensus: increasing decentralization in Wikipedia governance. In *Proceedings of the 41st annual Hawaii international conference on system sciences (HICSS'08)*, pages 157–. IEEE Computer Society, 2008. ISBN 0-7695-3075-8. doi: 10.1109/HICSS.2008.383. 11

[60] Andrea Forte, Judd Antin, Shaowen Bardzell, Leigh Honeywell, John Riedl, and Sarah Stierch. Some of all human knowledge: gender and participation in peer production. In *Proceedings of the 2012 ACM conference on computer supported cooperative work (CSCW'12)*, pages 33–36. ACM, 2012. ISBN 978-1-4503-1051-2. doi: 10.1145/2141512.2141530. 12

[61] Wikimedia Foundation. Wikimedia Movement Strategic Plan: A collaborative vision for the movement through 2015. Published online. Retrieved June 5, 2012 from http://strategy.wikimedia.org/wiki/Wikimedia_Movement_Strategic_Plan_Summary, 2011. 1

[62] Evgeniy Gabrilovich and Shaul Markovitch. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *Proceedings of the 20th international joint conference on artifical intelligence (IJCAI'07)*, pages 1606–1611. Morgan Kaufmann Publishers Inc., 2007. 14, 83

[63] Loris Gaio, Matthijs den Besten, Alessandro Rossi, and Jean-Michel Dalle. Wikibugs: using template messages in open content collections. In *Proceedings of the 5th international symposium on wikis and open collaboration (WikiSym'09)*, pages 14:1–14:7. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641330. 2, 19, 21

[64] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070): 900–901, 2005. URL `http://www.nature.com/nature/journal/v438/n7070/full/438900a.html`. 18

[65] Ruediger Glott, Philipp Schmidt, and Rishab Ghosh. Wikipedia survey – overview of results. UNU-Merit and Collaborative Creativity Group, 2010. URL `http://www.wikipediastudy.org/docs/Wikipedia_Overview_15March2010-FINAL.pdf`. 1, 11, 12

[66] Thomas Gottron, Maik Anderka, and Benno Stein. Insights into explicit semantic analysis. In *Proceedings of the 20th ACM international conference on information and knowledge management (CIKM'11)*, pages 1961–1964. ACM, 2011. ISBN 978-1-4503-0717-8. doi: 10.1145/2063576.2063865. 7, 8, 14, 83

[67] Jerry Greenfield. *Classic readability formulas in an EFL context: are they valid for Japanese speakers?* Ed.d. thesis, University Microfilms International, 1999. 97

[68] Robert Gunning. *The Technique of Clear Writing*. McGraw-Hill, 1952. 97

[69] Mangesh Gupte, Pra Shankarvin, Jing Li, S. Muthukrishnan, and Liviu Iftode. Finding hierarchy in directed online social networks. In *Proceedings of the 20th international conference on World Wide Web (WWW'11)*, pages 557–566. ACM, 2011. ISBN 978-1-4503-0632-4. doi: 10.1145/1963405.1963484. 13

[70] Alexander Halavais and Derek Lackaff. An analysis of topical coverage of Wikipedia. *Journal of computer-mediated communication*, 13(2):429–440, 2008. ISSN 1083-6101. doi: 10.1111/j.1083-6101.2008.00403.x. 11

[71] Jingyu Han, Chuandong Wang, Xiong Fu, and Kejia Chen. Probabilistic quality assessment of articles based on learning editing patterns. In *Proceedings of the international conference on computer science and service system (CSSS'11)*, pages 564–570, 2011. ISBN 978-1-4244-9762-1. doi: 10.1109/CSSS.2011.5973947. 2, 19

[72] Xianpei Han and Jun Zhao. Named entity disambiguation by leveraging Wikipedia semantic knowledge. In *Proceedings of the 18th ACM international conference on information and knowledge management (CIKM'09)*, pages 215–224. ACM, 2009. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1645983. 14, 83

[73] Kathryn Hempstalk, Eibe Frank, and Ian H. Witten. One-class classification by combining density and class probability estimation. In *Proceedings of the european conference on machine learning and knowledge discovery in databases: part I (ECML/PKDD'08)*, pages 505–5192. Springer, 2008. ISBN 978-3-540-87478-2. doi: 10.1007/978-3-540-87479-9_51. 61, 69

[74] Tin Kam Ho. Random decision forests. In *Proceedings of the 3rd international conference on document analysis and recognition (ICDAR'95)*, pages 278–282. IEEE, 1995. doi: 10.1109/ICDAR.1995.598994. 70

[75] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artificial Intelligence Review*, 22(2):85–126, 2004. ISSN 0269-2821. doi: 10.1023/B:AIRE.0000045502.10941.a9. 61

[76] Jian Hu, Lujun Fang, Yang Cao, Hua-Jun Zeng, Hua Li, Qiang Yang, and Zheng Chen. Enhancing text clustering by leveraging Wikipedia semantics. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'08)*, pages 179–186. ACM, 2008. ISBN 978-1-60558-164-4. doi: 10.1145/1390334.1390367. 14, 83

[77] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in Wikipedia: models and evaluation. In *Proceedings of the 16th ACM international conference on information and knowledge management (CIKM'07)*, pages 243–252. ACM, 2007. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321476. 2, 19, 64, 66, 67, 103

[78] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, and Xiaohua Zhou. Exploiting Wikipedia as external knowledge for document clustering. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'09)*, pages 389–396. ACM, 2009. ISBN 978-1-60558-495-9. doi: 10.1145/1557019.1557066. 14, 83

[79] Adam Jatowt and Katsumi Tanaka. Is Wikipedia too difficult?: comparative analysis of readability of Wikipedia, Simple Wikipedia and Britannica. In *Proceedings of the 21st ACM international conference on information and knowledge management (CIKM'12)*, pages 2607–2610. ACM, 2012. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398703. 63

[80] Sara Javanmardi and Cristina Lopes. Statistical measure of quality in Wikipedia. In *Proceedings of the 1st workshop on social media analytics (SOMA'10)*, pages 132–138. ACM, 2010. ISBN 978-1-4503-0217-3. doi: 10.1145/1964858.1964876. 19

[81] Ian Jolliffe. *Principal component analysis.* John Wiley & Sons, Ltd., 2005. 62

[82] Joseph M. Juran and A. Blanton Godfrey. *Juran's quality handbook.* McGraw-Hill London, 1999. 16, 83

[83] J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of new readability formulas (automated readability index, fog Count and Flesch reading ease formula) for Navy enlisted personnel. Technical Report 8-75, Naval Technical Training Command, Naval Air Station Memphis - Millington, TN., 1975. 97

[84] Aniket Kittur and Robert E. Kraut. Beyond Wikipedia: coordination and conflict in online production groups. In *Proceedings of the 2010 ACM conference on computer supported cooperative work (CSCW'10)*, pages 215–224. ACM, 2010. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718959. 12

[85] Aniket Kittur, Ed H. Chi, Bryan A. Pendleton, Bongwon Suh, and Todd Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. In *Alt.CHI at the SIGCHI conference on human factors in computing systems*, 2007. 12

[86] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in Wikipedia. In *Proceedings of the 25th SIGCHI international conference on human factors in computing systems (CHI'07)*, pages 453–462. ACM, 2007. ISBN 978-1-59593-593-9. doi: 10.1145/1240624.1240698. 11, 21

[87] Aniket Kittur, Ed H. Chi, and Bongwon Suh. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. In *Proceedings of the 27th SIGCHI international conference on human factors in computing systems (CHI'09)*, pages 1509–1512. ACM, 2009. ISBN 978-1-60558-246-7. doi: 10.1145/1518701.1518930. 11

[88] Aniket Kittur, Bryan Pendleton, and Robert E. Kraut. Herding the cats: the influence of groups in coordinating peer production. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration (WikiSym'09)*, pages 7:1–7:9. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641321. 12

[89] Moshe Koppel and Jonathan Schler. Authorship verification as a one-class classification problem. In *Proceedings of the 21st international conference on machine learning (ICML'04)*. ACM, 2004. ISBN 1-58113-838-5. doi: 10.1145/1015330.1015448. 61

[90] Travis Kriplean, Ivan Beschastnikh, and David W. McDonald. Articulations of wikiwork: uncovering valued work in Wikipedia through barnstars. In *Proceedings of the 2008 ACM conference on computer supported cooperative work (CSCW'08)*, pages 47–56. ACM, 2008. ISBN 978-1-60558-007-4. doi: 10.1145/1460563.1460573. 11

[91] Stacey Kuznetsov. Motivations of contributors to Wikipedia. *ACM SIGCAS Computers and Society*, 36(2), 2006. ISSN 0095-2737. doi: 10.1145/1215942.1215943. 11

[92] David Laniado and Riccardo Tasso. Co-authorship 2.0: patterns of collaboration in Wikipedia. In *Proceedings of the 22nd conference on hypertext and hypermedia (HT'11)*, pages 201–210. ACM, 2011. ISBN 978-1-4503-0256-2. doi: 10.1145/1995966.1995994. 12, 13

[93] Andrew LaVallee. Jimmy Wales on Wikipedia quality and tips for contributors. The Wall Street Journal, Digits – technology news and insights, 2009. URL `http://blogs.wsj.com/digits/2009/11/06/jimmy-wales-on-wikipedia-quality-and-tips-for-contributors`. 1

[94] James Lee and Brent Ware. *Open source Web development with LAMP: using Linux, Apache, MySQL, Perl, and PHP.* Addison Wesley, 2002. ISBN 0-201-77061-X. 14

[95] Yang W. Lee, Diane M. Strong, Beverly K. Kahn, and Richard Y. Wang. AIMQ: a methodology for information quality assessment. *Information and Management*, 40(2):133–146, 2002. ISSN 0378-7206. doi: 10.1016/S0378-7206(02)00043-5. 16

[96] Bo Leuf and Ward Cunningham. *The Wiki way: quick collaboration on the Web.* Addison-Wesley, 2001. 9

[97] Mary Levis, Markus Helfert, and Malcolm Brady. Information quality management: review of an evolving research area. In *Proceedings of the 12th international conference on information quality (ICIQ'07).* MIT, 2007. 16

[98] Elisabeth Lex, Michael Völske, Marcelo Errecalde, Edgardo Ferretti, Leticia Cagnina, Christopher Horn, Benno Stein, and Michael Granitzer. Measuring the quality of Web content using factual information. In *Proceedings of the 2nd joint WICOW/AIRWeb workshop on Web quality (WebQuality'12)*, pages 7–10. ACM, 2012. ISBN 978-1-4503-1237-0. doi: 10.1145/2184305.2184308. 19

[99] Andrew Lih. Wikipedia as participatory journalism: reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the 5th international symposium on online journalism*, pages 16–17, 2004. URL `http://jmsc.hku.hk/faculty/alih/publications/utaustin-2004-wikipedia-rc2.pdf`. 2, 19, 66, 67

[100] Nedim Lipka and Benno Stein. Identifying featured articles in Wikipedia: writing style matters. In *Proceedings of the 19th international conference on World Wide Web (WWW'10)*, pages 1147–1148. ACM, 2010. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772847. 2, 19, 64, 94

[101] Nedim Lipka, Benno Stein, and Maik Anderka. Cluster-based one-class ensemble for classification problems in information retrieval. In *Proceedings of the 35th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'12)*, pages 1041–1042. ACM, 2012. ISBN 978-1-4503-1472-5. doi: 10.1145/2348283.2348459. 7, 8

[102] Bing Liu, Yang Dai, Xiaoli Li, Wee Sun Lee, and Philip S. Yu. Building text classifiers using positive and unlabeled examples. In *Proceedings of the 3rd IEEE international conference on data mining (ICDM'03)*, pages 179–186. IEEE Computer Society, 2003. ISBN 0-7695-1978-4. doi: 10.1109/ICDM.2003.1250918. 79, 88

[103] Yue Lu, Panayiotis Tsaparas, Alexandros Ntoulas, and Livia Polanyi. Exploiting social context for review quality prediction. In *Proceedings of the 19th international conference on World Wide Web (WWW'10)*, pages 691–700. ACM, 2010. ISBN 978-1-60558-799-8. doi: 10.1145/1772690.1772761. 17

[104] Stuart E. Madnick and Hongwei Zhu. Improving data quality through effective use of data semantics. *Data & Knowledge Engineering*, 59(2): 460–475, 2006. ISSN 0169-023X. doi: 10.1016/j.datak.2005.10.001. 16

[105] Stuart E. Madnick, Richard Y. Wang, Yang W. Lee, and Hongwei Zhu. Overview and framework for data and information quality research. *Journal of data and information quality*, 1(1):1–22, 2009. ISSN 1936-1955. doi: 10.1145/1515693.1516680. 16

[106] Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497, 2003. ISSN 0165-1684. doi: 10.1016/j.sigpro.2003.07.018. 61

[107] G. Harry McLaughlin. SMOG grading: a new readability formula. *Journal of Reading*, 12(8):639–646, 1969. 98

[108] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM international conference on information and knowledge management (CIKM'07)*, pages 233–242. ACM, 2007. ISBN 978-1-59593-803-9. doi: 10.1145/1321440.1321475. 13, 83

[109] David Milne and Ian H. Witten. Learning to link with Wikipedia. In *Proceeding of the 17th ACM international conference on information and knowledge management (CIKM'08)*, pages 509–518. ACM, 2008. ISBN 978-1-59593-991-3. doi: 10.1145/1458082.1458150. 13, 83

[110] Thomas M. Mitchell. *Machine learning.* McGraw-Hill, 1997. 70

[111] Domas Mituzas. Wikipedia: site internals, configuration and code examples, and management issues. MySQL conference & expo, 2007. URL `http://www.scribd.com/doc/43868/Wikipedia-site-internals-workbook-2007`. 14

[112] Santiago M. Mola-Velasco. Wikipedia vandalism detection. In *Proceedings of the 20th international conference on World Wide Web (WWW'11)*, pages 391–396. ACM, 2011. ISBN 978-1-4503-0637-9. doi: 10.1145/1963192.1963349. 2, 20

[113] Jonathan T. Morgan, Robert M. Mason, and Karine Nahon. Negotiating cultural values in social media: a case study from Wikipedia. In *Proceedings of the 45th annual Hawaii international conference on system sciences (HICSS'12)*, pages 3490–3499. IEEE Computer Society, 2012. ISBN 978-0-7695-4525-7. doi: 10.1109/HICSS.2012.443. 13

[114] Andreas Neus. Managing information quality in virtual communities of practice. In *Proceedings of the 6th international conference on information quality (ICIQ'01)*, pages 119–131. MIT, 2001. 9

[115] Oded Nov. What motivates Wikipedians? *Communications of the ACM*, 50(11):60–64, 2007. ISSN 0001-0782. doi: 10.1145/1297797.1297798. 11

[116] Felipe Ortega and Jesus M. Gonzalez-Barahona. Quantitative analysis of the Wikipedia community of users. In *Proceedings of the 3rd international symposium on wikis and open collaboration (WikiSym'07)*, pages 75–86. ACM, 2007. ISBN 978-1-59593-861-9. doi: 10.1145/1296951.1296960. 12

[117] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. On the inequality of contributions to Wikipedia. In *Proceedings of the 41st annual Hawaii international conference on system sciences (HICSS'08)*, pages 304–. IEEE Computer Society, 2008. ISBN 0-7695-3075-8. doi: 10.1109/HICSS.2008.333. 12

[118] Felipe Ortega, Jesus M. Gonzalez-Barahona, and Gregorio Robles. Quantitative analysis of the top ten Wikipedias. In *Software and Data Technologies*, volume 22 of *Communications in Computer and Information Science*, pages 257–268. Springer, 2009. ISBN 978-3-540-88654-9. doi: 10.1007/978-3-540-88655-6_19. 12

[119] Katherine Panciera, Aaron Halfaker, and Loren Terveen. Wikipedians are born, not made: a study of power editors on Wikipedia. In *Proceedings of the ACM 2009 international conference on supporting group work (GROUP'09)*, pages 51–60. ACM, 2009. ISBN 978-1-60558-500-0. doi: 10.1145/1531674.1531682. 12

[120] Katherine Panciera, Mikhil Masli, and Loren Terveen. "How should I go from ___ to ___ without getting killed?": motivation and benefits in open collaboration. In *Proceedings of the 7th international symposium on wikis and open collaboration (WikiSym'11)*, pages 183–192. ACM, 2011. ISBN 978-1-4503-0909-7. doi: 10.1145/2038558.2038587. 11

[121] Ari Pirkola and Tuomas Talvensaari. A topic-specific web search system focusing on quality pages. In *Proceedings of the 14th european conference on research and advanced technology for digital libraries (ECDL'10)*, pages 490–493. Springer, 2010. ISBN 3-642-15463-8. doi: 10.1007/978-3-642-15464-5_64. 16

[122] Martin Potthast, Benno Stein, and Maik Anderka. A Wikipedia-based multilingual retrieval model. In *Advances in information retrieval. 30th european conference on information retrieval research (ECIR'08)*, pages 522–530. Springer, 2008. ISBN 978-3-540-78645-0. doi: 10.1007/978-3-540-78646-7_51. 7, 8, 13, 83

[123] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in Wikipedia. In *Advances in information retrieval. 30th european conference on information retrieval research (ECIR'08)*, pages 663–668. Springer, 2008. ISBN 978-3-540-78645-0. doi: 10.1007/978-3-540-78646-7_75. 2, 20, 46, 88

[124] Martin Potthast, Benno Stein, and Teresa Holfeld. Overview of the 1st international competition on Wikipedia vandalism detection. In *Notebook Papers of CLEF'10 Labs and Workshops*, 2010. ISBN 978-88-904810-2-4. 2, 20, 46, 88

[125] Reid Priedhorsky, Jilin Chen, Shyong (Tony) K. Lam, Katherine Panciera, Loren Terveen, and John Riedl. Creating, destroying, and restoring value in Wikipedia. In *Proceedings of the 2007 international ACM conference on supporting group work (GROUP'07)*, pages 259–268. ACM, 2007. ISBN 978-1-59593-845-9. doi: 10.1145/1316624.1316663. 9, 12

[126] Laura Rassbach, Trevor Pincock, and Brian Mingus. Exploring the feasibility of automatically rating online article quality. In *Wikimania 2007*, 2007. URL `http://upload.wikimedia.org/wikipedia/wikimania2007/d/d3/RassbachPincockMingus07.pdf`. 2, 19

[127] Lucy Holman Rector. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review*, 36(1):7–22, 2008. ISSN 0090-7324. 18

[128] Roy Rosenzweig. Can history be open source? Wikipedia and the future of the past. *The journal of American history*, 93(1):117–146, 2006. doi: 10.2307/4486062. 18

[129] Alessandro Rossi, Loris Gaio, Matthijs den Besten, and Jean-Michel Dalle. Coordination and division of labor in open content communities: the role of template messages in Wikipedia. In *Proceedings of the 43rd annual Hawaii international conference on system sciences (HICSS'10)*, pages 1–10. IEEE Computer Society, 2010. ISBN 978-0-7695-3869-3. doi: 10.1109/HICSS.2010.122. 2, 19, 21

[130] Cindy Royal and Deepina Kapila. What's on Wikipedia, and what's not ...? *Social Science Computer Review*, 27(1):138–148, 2009. ISSN 0894-4393. doi: 10.1177/0894439308321890. 11

[131] Bernhard Schölkopf, Robert C. Williamson, Alex J. Smola, John Shawe-Taylor, and John C. Platt. Support vector method for novelty detection. In *Advances in neural information processing systems 12. 13th annual neural information processing systems conference (NIPS'99)*, pages 582–588. MIT Press, 1999. ISBN 0-262-19450-3. 62

[132] Joachim Schroer and Guido Hertel. Voluntary engagement in an open Web-based encyclopedia: Wikipedians and why they do it. *Media Psychology*, 12(1):96–120, 2009. ISSN 1521-3269. doi: 10.1080/15213260802669466. 11

[133] Aaron Schwartz. Who writes Wikipedia? Online blog, 2006. URL `http://www.aaronsw.com/weblog/whowriteswikipedia`. 12

[134] R. J. Senter and E. A. Smith. Automated readability index. Technical Report AMRL-TR-66-220, Wright-Patterson Air Force Base, Ohio, 1967. 96

[135] Shyong (Tony) K. Lam, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R. Musicant, Loren Terveen, and John Riedl. WP:clubhouse?: an exploration of Wikipedia's gender imbalance. In *Proceedings of the 7th international symposium on wikis and open collaboration (WikiSym'11)*, pages 1–10. ACM, 2011. ISBN 978-1-4503-0909-7. doi: 10.1145/2038558.2038560. 12

[136] Koen Smets, Bart Goethals, and Brigitte Verdonk. Automatic vandalism detection in Wikipedia: towards a machine learning approach. In *Proceedings of the AAAI workshop on Wikipedia and artificial intelligence (WIKIAI'08)*, pages 43–48. AAAI Press, 2008. 45

[137] Diomidis Spinellis and Panagiotis Louridas. The collaborative organization of knowledge. *Communications of the ACM*, 51(8):68–73, 2008. ISSN 0001-0782. doi: 10.1145/1378704.1378720. 10

[138] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the american society for information science and technology*, 60(3):538–556, 2009. ISSN 1532-2882. doi: 10.1002/asi.v60:3. 94

[139] Benno Stein and Maik Anderka. Collection-relative representations: a unifying view to retrieval models. In *Proceedings of the 6th international workshop on text-based information retrieval (TIR'09)*, pages 383–387. IEEE, 2009. ISBN 978-0-7695-3763-4. doi: 10.1109/DEXA.2009.50. 7, 8

[140] Benno Stein, Nedim Lipka, and Peter Prettenhofer. Intrinsic plagiarism analysis. *Language Resources and Evaluation*, 45(1):63–82, 2011. ISSN 1574-020X. doi: 10.1007/s10579-010-9115-y. 61

[141] Jorge Stolfi. A two-phase model for Wikipedia growth. Technical Report IC-09-45, Institute of Computing, State University of Campinas, 2009. 10

[142] Jeff Stuckman and James Purtilo. Measuring the wikisphere. In *Proceedings of the 5th international symposium on wikis and open collaboration (WikiSym'09)*, pages 11:1–11:8. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641326. 12

[143] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the 10th international conference on information quality (ICIQ'05)*, pages 442–454. MIT, 2005. 2, 19, 64, 65, 66, 67

[144] Besiki Stvilia, Michael B. Twidale, Linda C. Smith, and Les Gasser. Information quality work organization in Wikipedia. *Journal of the american society for information science and technology*, 59(6):983–1001, 2008. ISSN 1532-2882. doi: 10.1002/asi.v59:6. 2, 19, 28

[145] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: slowing growth of Wikipedia. In *Proceedings of the 5th international symposium on wikis and open collaboration (WikiSym'09)*, pages 1–10. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641322. 11, 46

[146] James Surowiecki. *The wisdom of crowds*. Doubleday, 2005. ISBN 978-0-385-50386-0. 12

[147] David M. J. Tax. *One-class classification*. Ph.d. thesis, Delft University of Technology, 2001. 61, 69, 71, 87

[148] Hang Ung and Jean-Michel Dalle. Project management in the Wikipedia community. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration (WikiSym'10)*, pages 13:1–13:4. ACM, 2007. ISBN 978-1-4503-0056-8. doi: 10.1145/1832772.1832790. 12

[149] C. J. van Rijsbergen. *Information retrieval*. Butterworth, 1979. ISBN 0-408-70929-4. 73

[150] Iraklis Varlamis. Quality of content in Web 2.0 applications. In *Proceedings of the 14th international conference on knowledge-based and intelligent information and engineering systems: part III (KES'10)*, pages 33–42. ACM, 2010. ISBN 3-642-15392-5. 17

[151] Fernanda B. Viegas. The visual side of Wikipedia. In *Proceedings of the 40th annual Hawaii international conference on system sciences (HICSS'07)*. IEEE Computer Society, 2007. ISBN 0-7695-2755-8. doi: 10.1109/HICSS.2007.559. 11

[152] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on human factors in computing systems (CHI'04)*, pages 575–582. ACM, 2004. ISBN 1-58113-702-8. doi: 10.1145/985692.985765. 4, 9, 12, 20, 45

[153] Fernanda B. Viegas, Martin Wattenberg, Jesse Kriss, and Frank van Ham. Talk Before You Type: Coordination in Wikipedia. In *Proceedings of the 40th annual Hawaii international conference on system sciences (HICSS'07)*. IEEE Computer Society, 2007. ISBN 0-7695-2755-8. doi: 10.1109/HICSS.2007.511. 9, 12

[154] Jakob Voss. Measuring Wikipedia. In *Proceedings of the 10th international conference of the international society for scientometrics and informetrics (ISSI'05)*, pages 221–231, 2005. 10

[155] Pu Wang and Carlotta Domeniconi. Building semantic kernels for text classification using Wikipedia. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'08)*, pages 713–721. ACM, 2008. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401976. 13, 83

[156] Richard Y. Wang and Diane M. Strong. Beyond accuracy: what data quality means to data consumers. *Journal of management information systems*, 12(4):5–33, 1996. ISSN 0742-1222. 16, 17, 83

[157] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in Wikipedia. In *Proceedings of the 3th international symposium on wikis and open collaboration (WikiSym'07)*, pages 157–164. ACM, 2007. ISBN 978-1-59593-861-9. doi: 10.1145/1296951.1296968. 2, 10, 19, 66, 67

[158] D. Randall Wilson and Tony R. Martinez. Bias and the probability of generalization. In *Proceedings of the international conference on intelligent information systems (IIS'97),*, pages 108–114. IEEE, 1997. 70

[159] Thomas Wöhner and Ralf Peters. Assessing the quality of Wikipedia articles with lifecycle based metrics. In *Proceedings of the 5th international symposium on wikis and open collaboration (WikiSym'09)*, pages 1–10. ACM, 2009. ISBN 978-1-60558-730-1. doi: 10.1145/1641309.1641333. 19

[160] Heng-Li Yang and Cheng-Yu Lai. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377–1383, 2010. ISSN 0747-5632. doi: 10.1016/j.chb.2010.04.011. 11

[161] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. Computing trust from revision history. In *Proceedings of the international conference on privacy, security and trust (PST'06)*, pages 1–1. ACM, 2006. ISBN 1-59593-604-1. doi: 10.1145/1501434.1501445. 20

[162] Yun Zhou and W. Bruce Croft. Document quality models for Web ad hoc retrieval. In *Proceedings of the 14th ACM international conference on information and knowledge management (CIKM'05)*, pages 331–332. ACM, 2005. ISBN 1-59593-140-6. doi: 10.1145/1099554.1099652. 16

[163] Xiaolan Zhu and Susan Gauch. Incorporating quality metrics in centralized/distributed information retrieval on the World Wide Web. In *Proceedings of the 23rd annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'00)*, pages 288–295. ACM, 2000. ISBN 1-58113-226-3. doi: 10.1145/345508.345602. 16, 94

[164] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):016115, 2006. doi: 10.1103/PhysRevE.74.016115. 10, 11

About the Author

Maik Anderka was born in Arolsen (now Bad Arolsen), Germany, on the 7th of February 1981. He finished his secondary education in 2000. After completing his military service in 2001, he studied computer science at the University of Paderborn, Germany. His areas of specialization were software engineering and formal verification, and he studied psychology as a secondary subject. He received a Bachelor of Computer Science in August 2006 and a Master of Computer Science in October 2007. Since November 2007, he worked as a research associate at Bauhaus-Universität Weimar, Germany, in the Web Technology and Information Systems Group. His research interests include various topics in the fields of machine learning, data mining, and information retrieval.

He published at international conferences, including WWW, CIKM, SIGIR, ECIR, and CLEF. He has regularly served as program committee member of several conferences and workshops (e.g., SIGIR, I-KNOW, DETECT, and Wikipedia Academy), and he was reviewer for different journals (e.g., ACM TOIS). From 2009 to 2012 he was an organizing committee member of the workshop series on Text-based Information Retrieval (TIR), and he is co-chair of TIR'13. He initiated and co-organized the "1st Competition on Quality Flaw Prediction in Wikipedia", held in conjunction with the PAN'12 Lab at the CLEF'12 conference.