4M1-IOS-3c-1

Thai Wikipedia Quality Measurement using Fuzzy Logic

Kanchana Saengthongpattana[1] Nuanwan Soonthornphisaj[*1]

[1] Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand.

Wikipedia is widely known as an online encyclopedia. The open access model is a key success for Wikipedia, however the quality of articles is a problem. Since the articles are collaboratively written and maintained by online volunteers. The flaws are normally detected and removed by Wikipedia users who encounter the problem. They use the cleanup tags to tag the article. Therefore, the quality detection tool can automatically help readers to identify articles that are of good quality. Many current techniques rely on exactly quality feature and identify quality article by classification or clustering method. The aim of this research is to classify the articles of Thai Wikipedia into two classes namely featured article and normal article by using Fuzzy Logic. We believe that the degree of the article's quality is ambiguous. Our dataset consists of 88 Thai featured articles and 100 normal articles. Our evaluation is based on a corpus comprising of human labeled the degree of each quality this articles. We found that the degree of quality articles obtained from Fuzzy Logic provide the accuracy close to the expert inspection.

1. Introduction

The Thai Wikipedia assigns quality level that includes featured articles and good articles. As of April 2012, only 88 articles out of a total of 72,826 articles on the Thai Wikipedia are labeled as the featured articles. From this number, it's interesting to consider why the good articles have so small number. Can we assume that the Thai Wikipedia is unreliable?

Wikipedia relies mostly on human editors and administrators to provide the quality content. But the magnitude of Wikipedia content makes locating all instances of article very time consuming. In addition, the article's quality is ambiguous due to some conditions cannot be clearly determined in examples such as advertisement's style writing, missing neutrality, and less content. These characteristics are difficult to judge the quality although it is done by professionals. Many researches try to construct automatic method base on classification or clustering techniques. To accomplish the task the algorithm needs informative features to determine a quality article.

The objective of this paper is to use fuzzy logic as an approach to evaluate and predict the degree of the quality of the Thai Wikipedia articles.

2. Feature Categories

2.1 User Features

The user features get the mining history revisions up to time and based solely on history data of user who edit each article [Javanmardi 2011] [Wikkinson 2007].

Several studies have tried to assess the quality of Wikipedia's content and the reputation of its contributors by analyzing historical contribution patterns [Javanmardi 2007] [Hu 2007]. They found that the user features represented by the history of user contributions are the most important.

Contact: N. Soonthornphisaj, Department of Computer Science, Faculty of Science, Kasetsart University. Bangkok, Thailand, Phone: +662 562 5444, Fax +662 9428488, Email: fscinws@ku.ac.th

2.2 Textual Features

The value of the textual features is focus on the content. Lower quality content has high frequency of vulgar words than the high quality content. On the contrary, more insertion and deletion of words are found in high quality content. These are a signal for the quality of article and it is the indicator of legitimate editing [Mola-Velasco 2010].

2.3 Meta Data Features

The Meta data category is extracted from the comments associated with the edits. The extraction comments are base on unigrams, bigrams, and trigrams. For example, Javanmardi proposed that the short time interval between the old and the new revisions might be an indicator of an editing quality [Javanmardi 2011]. In addition, some of these features included in textual feature.

In addition, these features include many features in the texture feature. Although some of the Meta data features are similar to textual features but it is extracted based on the comment.

2.4 Language Model Features

The language model features build from characteristic writing which is based on statistic method and natural language processing. Chin, et al., constructed statistical language model from new versions article. Their model was compared with language model built from previous versions and found that their model can detect vandalism instance better than the previous one [Chin 2010].

Javanmardi proposed three language models using features based on Kullback Leibler Distance (KLD) between the old revision and the new revision, the KLD between the inserted words and the new revision, and the KLD between the deleted words and the new revision. This method was applied to detect Wikipedia vandalism [Javanmardi 2011].

3. Wikipedia Quality Detection Technique

3.1 Features Based Techniques

Much efforts have been spent to tackle the quality problem of Wiki pages. Many tools are available for Wiki users such as, Vandal Fighter[1], Huggle[2] and Twinkle[3]. These tools are used to monitor the recently changes of articles and they revert changes if it deems vandalistic. Checklinks[4] are used to check external links. However, these tools help user to determine the quality in some criteria. Wikipedia relies mostly on human editors and administrators who check the right content. Many current techniques rely on features selection and identify quality article based on classification or clustering technique.

The selection of appropriate features can also enhance the quality detection of the article. The previous researches considered quality flaw and vandalism characteristic features to predict the quality of Wikipedia article. New metric for quality measurement was proposed by [Wöhner 2009]. This metrics is based on lifecycles of low and high quality articles, which refer to the changes of a persistent and transient contribution throughout a life span. The assumption is that the high quality articles should be more persistently edited than the low quality articles. Furthermore, the analysis showed that the length of an article is highly relevant to the quality measurement.

User reputation and user's action are useful features to determine the article's quality. The reputation assessment relies on the survival of contributed text and contributed edits. [Adler 2011] found that using the user reputation together with other features can improve the prediction performance and they showed that it is the strong predictors to locate low quality content.

However, User's reputation feature seems to be bias since the length of articles is vary. Therefore long articles might have been edited more often compared to short articles.

3.2 Classification Based Techniques

Several classification approaches include decision tree, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques.

We provide a review of different classification techniques applied for Wikipedia quality detection. For example, a PeerReview model applies Naive Baye method to automatically derive Wikipedia article quality rankings. In addition, this method is based on the interaction data and contributors.

Anderka, *et al.*, did research about the detection of text quality flaws as a one-class classification problem. They studied the feature based technique and found that the number of articles which does not cite to any references or sources is high. The class probability estimators in this research apply bagged random forest classifiers with 1,000 decisions trees and ten bagging iterations [Anderka 2011].

The classification techniques usually get high accuracy, however preparing the label data from both normal and featured article class is a time consuming task.

3.3 Clustering Based Techniques

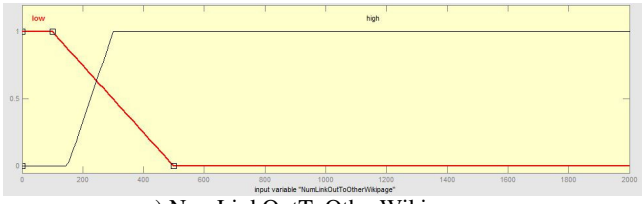
The clustering technique divides data into groups of similar objects such as quality and flaw articles. This technique needs no label data but if normal points are not created to represent data, this the clustering techniques may fail.

Roles of the contributors and collaboration patterns of each Wikipedia article were studied by [Liu 2011]. They proposed a mechanism to identify the roles of contributors and collaboration patterns of each Wikipedia article. They examined the quality of the articles to determine the impact of collaboration patterns on quality of the Wikipedia articles. This research applied the k-means method repeatedly by using k values ranging from 2 to 10. They found that the collaboration of contributor's pattern was a critical factor driving the quality of Wikipedia articles

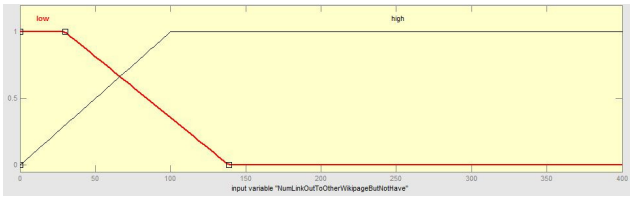
We found that many current techniques rely exactly on the quality feature and they identify quality article by classification or clustering method. But the fact is that the degree of the article's quality is ambiguous due to some conditions cannot be clearly determined such as advertisement's style writing, missing neutrality, and less content. These characteristics are difficult to judge for the quality although it is done by professional. Therefore, the quality of Wikipedia articles should be graded more than two values (good or not good). We believe that using fuzzy logic should be an advantage.

4. Our Approach

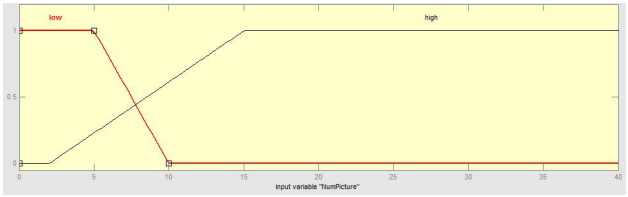
We propose to use sixteen features as shown in Table 1 and create 26 rules for fuzzy inference (see Table 2)



a) NumLinkOutToOtherWikipage



b) NumLinkOutToNullWikipage



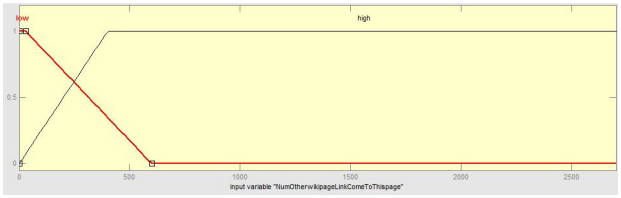
c) NumPicture

[1] http://en.wikipedia.org/wiki/User:Henna/VF

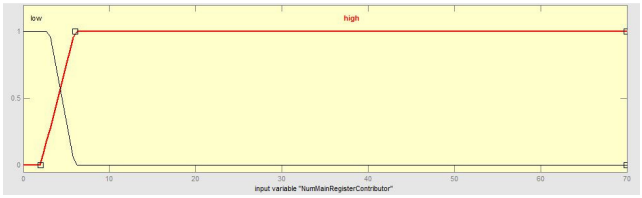
[2] http://en.wikipedia.org/wiki/Wikipedia:Huggle

[3] http://en.wikipedia.org/wiki/Wikipedia:Twinkle

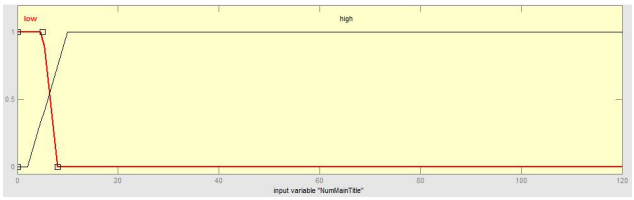
[4] http://en.wikipedia.org/wiki/Wikipedia:Checklinks



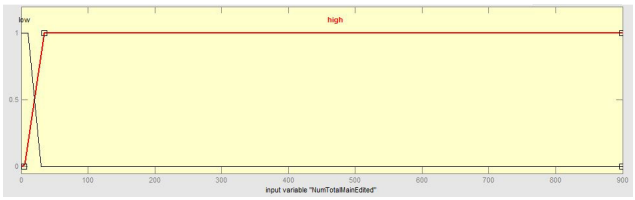
d) NumOtherwikipageLinkComeToThispage



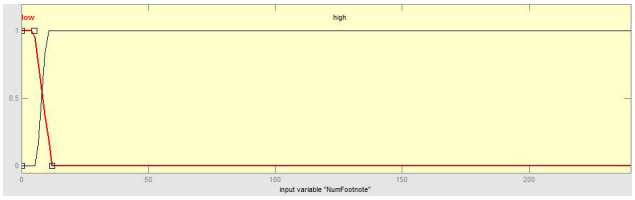
k) NumMainRegisterContributor



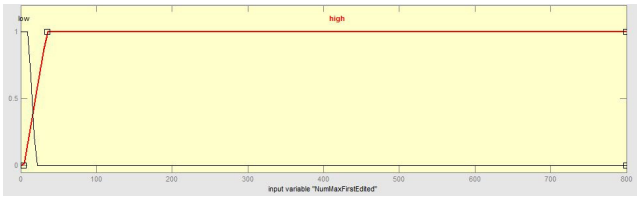
e) NumMainTitle



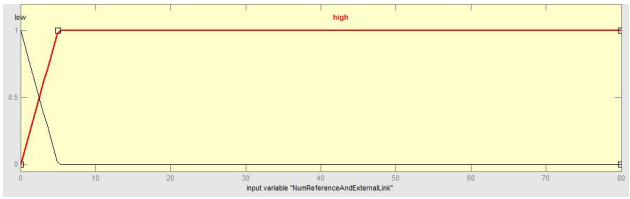
l) NumTotalMainEdited



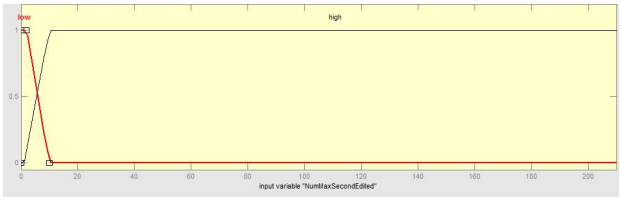
f) NumFootnote



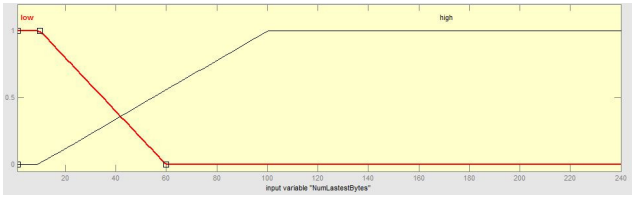
m) NumMaxFirstEdited



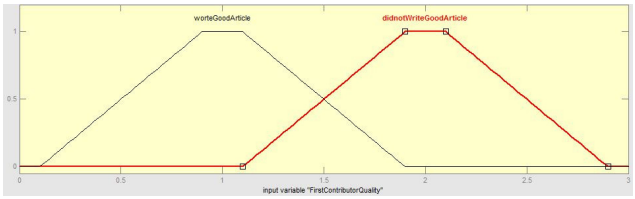
g) NumReferenceAndExternalLink



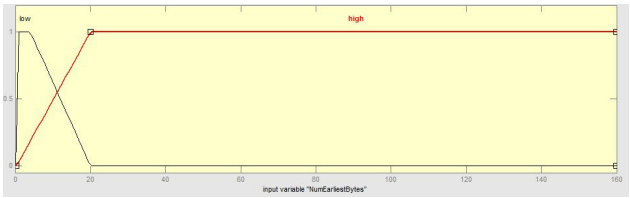
n) NumMaxSecondEdited



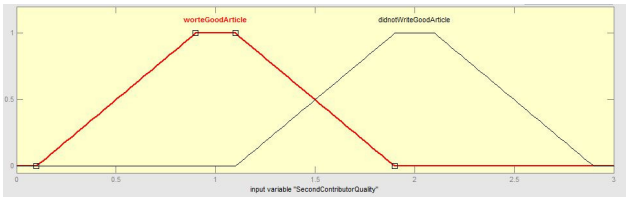
h) NumLastestBytes



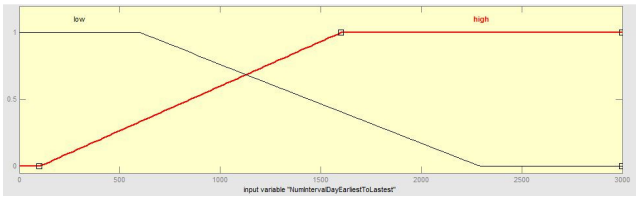
o) FirstContributorQuality



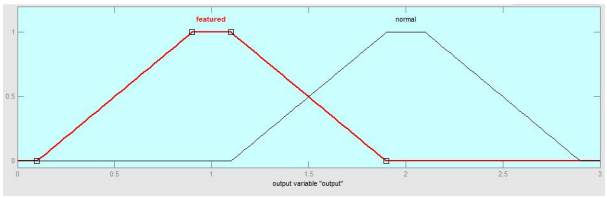
i) NumEarliestBytes



p) SecondContributorQuality



j) NumIntervalDayEarliestToLastest



q) Output

Table 1: The meaning of attribute used in our work.

No	**Attribute Name**	**Description**
1	NumLinkOutToOtherWikipage	Number of Wiki links from the current page that point to other Wiki pages
2	NumLinkOutToNullWikipage	Number of Wiki links from the current page that point to null page
3	NumPicture	Number of images found in the current page
4	NumOtherwikipageLinkComeToThispage	Number of other Thai wiki pages that point back to the current page
5	NumMainTitle	Number of main topics found in the current page
6	NumFootnote	Number of footnotes found in the current page
7	NumReferenceAndExternalLink	Number of references and external links found in the current page
8	NumLastestBytes	Number of the lastest Bytes
9	NumEarliestBytes	Number of the original Bytes when the article was first created
10	NumIntervalDayEarliestToLastest	Time interval measure from the creation date to the last edited date (no. of days)
11	NumMainRegisterContributor	Number of registered contributors who edited the page
12	NumTotalMainEdited	Total number of editing frequency found in the page
13	NumMaxFirstEdited	Editing frequency of the most editing contributor.
14	NumMaxSecondEdited	Editing frequency of the second most editing contributor.
15	FirstContributorQuality	Does the most editing contributor ever create the quality wiki page? (YES/NO)
16	SecondContributorQuality	Does the second most editing contributor ever create the quality wiki page? (YES/NO)

Table 2: Set of Rules used in Fuzzy Inference.

Rule no.	**Condition**	**Class**
1	(NumMaxFirstEdited is high)	Featured
2	(NumFootnote is high)	Featured
3	(NumFootnote is high) and (NumMaxFirstEdited is low)	Featured
4	(NumPicture is high) and (NumMaxFirstEdited is low)	Featured
5	(NumLinkOutToOtherWikipageButNotHave is high)	Featured
6	(NumLinkOutToOtherWikipageButNotHave is high) and (NumMainTitle is high)	Featured
7	(NumFootnote is high) and (NumReferenceAndExternalLink is high)	Featured
8	(NumFootnote is low) and (NumReferenceAndExternalLink is low)	Featured
9	(NumMainRegisterContributor is high) and (NumMaxFirstEdited is low)	Featured
10	(NumLinkOutToOtherWikipageButNotHave is low) and (NumMainTitle is high) and (NumTotalMainEdited is high)	Featured
11	(NumOtherwikipageLinkComeToThispage is high) and (NumIntervalDayEarliestToLastest is high) and (NumTotalMainEdited is high)	Featured
12	(NumTotalMainEdited is high) and (FirstContributorQuality is wroteGoodArticle) and (SecondContributorQuality is wroteGoodArticle)	Featured
13	(NumTotalMainEdited is low) and (FirstContributorQuality is didnotWriteGoodArticle) and (SecondContributorQuality is didnotWriteGoodArticle)	Featured
14	(NumLastestBytes is high) and (NumEarliestBytes is low) and (NumIntervalDayEarliestToLastest is high)	Featured
15	If (NumIntervalDayEarliestToLastest is high) and (NumMainRegisterContributor is high) and (NumTotalMainEdited is high) and (NumMaxFirstEdited is high) and (FirstContributorQuality is didnotWriteGoodArticle)	Featured
16	(NumOtherwikipageLinkComeToThispage is high) and (NumIntervalDayEarliestToLastest is low) and (NumTotalMainEdited is low)	Normal
17	(NumFootnote is low) and (NumMaxFirstEdited is low)	Normal
18	(NumTotalMainEdited is low)	Normal
19	(NumFootnote is low)	Normal
20	(NumLinkOutToOtherWikipageButNotHave is low) and (NumMainTitle is low)	Normal
21	(NumLinkOutToOtherWikipageButNotHave is low) and (NumMainTitle is low) and (NumTotalMainEdited is high)	Normal
22	(NumPicture is low) and (NumMaxFirstEdited is low)	Normal

23	(NumMainRegisterContributor is low) and (NumMaxFirstEdited is low)	Normal
24	(NumLastestBytes is low) and (NumEarliestBytes is low) and (NumIntervalDayEarliestToLastest is high)	Normal
25	(NumLastestBytes is low) and (NumEarliestBytes is low) and (NumIntervalDayEarliestToLastest is low)	Normal
26	(NumIntervalDayEarliestToLastest is high) and (NumTotalMainEdited is low) and (NumMaxFirstEdited is low)	Normal

5. Clustering Approach

In this study, we use a partitioning algorithm namely, *k*-means clustering . The *k* value is set to be 2. The *k*-means clustering algorithm [Jiawei, 2001] is implemented as follows:

Step 1. The number of clusters *c* is chosen a priori. The controids are chosen by randomly picking Wiki pages from the data set.

Step 2. For each pages in the dataset, the similarity measure between the Wiki page and the centroids are computed and the Wiki page is assigned to the cluster with which it exhibits the similarity measure.

Step 3. New cluster centers are computed as the centroids of the clusters.

Step 4. Repeat step 2 and 3 until there is no further change in the centroids.

6. Experimental Result

Our data set contains 188 Thai Wiki pages which consists of two classes. The target class is the featured article (88 pages) and the normal Wiki pages (100 pages). We did feature extraction process in order to get the attributes as shown in Table 1.

To evaluate the performance of fuzzy logic and clustering approach, we use Precision, Recall and F-Measure (see equation 1-3)

$$Precision = \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Positive} \quad (1)$$

$$Recall = \frac{No.of\ True\ Positive}{No.of\ True\ Positive + No.of\ False\ Negative} \quad (2)$$

$$F\ Measure = \frac{2x\ Precisionx\ Recall}{Precision + Recall} \quad (3)$$

Table 3: The performance obtained from Fuzzy logic and Clustering approach.

Approach	Class	precision	recall	F1
Fuzzy	featured	0.86	1.00	0.92
	normal	1.00	0.86	0.92

Table 4: The class distribution in each cluster obtained from *k*-mean.

Approach	Class	actual class	Cluster1	Cluster2
			99	89
Clustering	Featured	88	73	15
	Normal	100	26	74

The experiments are set up using the same attributes in both algorithms, fuzzy and *k*-mean.

As shown in Table 3, Fuzzy logic gets good performance at precision value 86% for featured class. No featured Wiki page is classified as normal page (recall = 1). Consider the class 'normal', we found that Fuzzy logic can predict all instances correctly.

We obtain 2 clusters, each of which contains mixed instances from 2 classes. There are 99 Wiki pages in the first cluster (cluster1). Most instances are in the featured class (73% of instances in the featured class). Consider the second cluster, most instances are from the normal class. 83.15% of cluster members are in the normal class.

We believe that the set of rules proposed in this work plays an important role since fuzzy logic approach outperforms the clustering algorithm.

6. Conclusion

In this paper, preliminary experimental test have shown a promising result in the Quality classification of Wiki pages. Our experiences indicate that the classification problem based on the labeled instance obtained from user's opinion should be deal with a kind of inference algorithm. Since the idea for determining the target class can be delivered by the domain experts via a set of rules. Moreover the continuous value of attributes can be transformed in terms of membership functions that encode the linguistic term from domain expert. This research is still ongoing and several future challenges are needed to deal with.

Acknowledgements

This research is partially supported by Department of Computer Science, Faculty of Science, Kasetsart University, Bangkok, Thailand.

References

[Adler 2011] Adler, B., de Alfaro, L., Mola-Velasco, S., Rosso, P., West, A., Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features Computational Linguistics and Intelligent Text Processing. vol. 6609, A. Gelbukh, ed: Springer Berlin / Heidelberg, pp. 277-288. 2011.

[Anderka 2011] Anderka M., B. Stein, Lipka N., Detection of text quality flaws as a one-class classification problem, The Proceedings of the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK, 2011.

[Chin 2010] Chin S.C., Street W.N., Srinivasan P., Eichmann D., Detecting Wikipedia vandalism with active learning and statistical language models, The Proceedings of the 4th

workshop on Information credibility, Raleigh, North Carolina, USA, 2010.

[Hu 2007] Hu M., Lim E.P. , Sun A., H. W. Lauw and Vuong B.Q., Measuring article quality in wikipedia: models and evaluation, The Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal, 2007.

[Javanmardi 2011] Javanmardi S., Measuring Content Quality in User Generated Content Systems: a Machine Learning Approach, Doctor of Philosophy, Information and Computer Science University of California, Irvine, 2011.

[Javanmardi 2007] Javanmardi S. and Lopes C., Modeling trust in collaborative information systems, International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 299-302, 2007,

[Jiawei, 2001] Jiawei, Han, Data Mining: concepts and techniques, Morgan Kaufmann Publishers, 2001.

[Liu 2011] Liu J. and Ram S., Who does what: Collaboration patterns in the wikipedia and their impact on article quality, ACM Trans. Manage. Inf. Syst., vol. 2, pp. 1-23, 2011.

[Mola-Velasco 2010] Mola-Velasco S. M., Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals, Lab Report for PAN at CLEF, 2010.

[Wikkinson 2007] Wikkinson D.M., Huberman B.A.F.M., Assessing the value of cooperation in wikipedia, 2007.

[Wöhner 2009] Wöhner T. and Peters R., Assessing the quality of Wikipedia articles with lifecycle based metrics, The Proceedings of the 5th International Symposium on Wikis and Open Collaboration, Orlando, Florida, 2009.