Journal of Information Processing Vol.23 No.1 22–30 (Jan. 2015)

[DOI: 10.2197/ipsjjip.23.22]

Regular Paper

Quality Assessment of Wikipedia Articles Using *h*-index

Yu Suzuki[1,a)]

Received: April 13, 2014, Accepted: October 8, 2014

Abstract: In this paper, we propose a method for assessing quality values of Wikipedia articles from edit history using *h*-index. One of the major methods for assessing Wikipedia article quality is a peer-review based method. In this method, we assume that if an editor's texts are left by the other editors, the texts are approved by the editors, then the editor is decided as a good editor. However, if an editor edits multiple articles, and the editor is approved at a small number of articles, the quality value of the editor deeply depends on the quality of the texts. In this paper, we apply *h*-index, which is a simple but resistant to excessive values, to the peer-review based Wikipedia article assessment method. Although *h*-index can identify whether an editor is a good quality editor or not, *h*-index cannot identify whether the editor is a vandal or an inactive editor. To solve this problem, we propose *p*-ratio for identifying which editors are vandals or inactive editors. From our experiments, we confirmed that by integrating *h*-index with *p*-ratio, the accuracy of article quality assessment in our method outperforms the existing peer-review based method.

Keywords: Wikipedia, *h*-index, peer-review, quality

1. Introduction

Collaborative projects, such as Wikipedia [*1] and OpenStreetMap [*2], have become important for constructing wide-ranging, high quality knowledge bases. In Wikipedia's policy, anyone can edit any articles. A merit of this policy is that Wikipedia has high comprehension and fresh information, but a demerit is that Wikipedia has many low quality texts, because articles in Wikipedia are not reviewed. There are not only high quality editors but also low quality editors who are referred to as "vandals." Vandals are editors who post false or insufficient descriptions, or editors who make an article blank. When users browse Wikipedia articles, they do not always have enough knowledge about the articles to find which texts are high quality or not. For these reasons, there is a need for assessing the quality of Wikipedia articles.

In this paper, a quality of article is defined as an approval rate for the article from Wikipedia readers. When many editors approve an article, we judge that the article is high quality. Therefore, for assessing the quality of an article, we should acquire the users' approval for the articles.

One approach for gathering users' approval is an explicit voting method. The Wikimedia Foundation, which operates Wikipedia, implements Article Feedback Tools [*3] for the purpose of gathering votes about the approval or the disapproval of articles from users. However, this tool only collects a small number of votes, because many users do not decide which articles are high quality or not. The number of votes are not enough for assessing Wikipedia articles accurately. From the results, gathering trustworthy users' approval is difficult using these explicit voting method.

To solve this problem, many researchers such as Adler et al. [1], Hu et al. [2] and the author [3] proposed an implicit voting method, which we call a survival ratio based method. In this method, we extract the approval of editors from edit history. We assume that if an editor leaves a text, the editor approves the text, but if the editor deletes the text, the editor disapproves the text. Therefore, if a text survives beyond multiple edits, we consider that the text is approved by many editors, and the quality of the text should be high.

In this method, we proceed in the following four steps:

(1) Extract positive ratings by editors from Wikipedia edit history.
(2) Calculate text quality using the positive ratings.
(3) Calculate editor quality using the text quality.
(4) Calculate article quality using the editor quality.

At step (2), we calculate the text quality using the editor reputation based method as we mentioned above. In this step, we confirmed the accuracy of the text quality described at Ref. [4]. However, at step (3), we used a simple arithmetic mean value for calculating editor quality. As a result, the accuracy of the editor quality decreases. This is because, if an editor posts texts to multiple articles, and if a small number of texts earn excessively high positive ratings in a small number of articles, the quality of the editor is dependent on these excessive text quality values. To solve this problem, we should develop a method for calculating editor quality values which are resistant to excessive text quality values.

When we measure a researcher's quality, we face a similar problem. Citation count is generally used for measuring the quality of papers and researchers. Generally, researcher quality is de-

[1] Graduate School of Science and Technology, Nara Institute of Science and Technology, Ikoma, Nara 630–0192, Japan
[a)] ysuzuki@is.naist.jp
[*1] http://www.wikipedia.org/
[*2] http://www.openstreetmap.org/
[*3] http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool

fined by the arithmetic mean value of citation counts. However, if a small number of papers written by the researchers are referred to an excessive number of papers, the researcher's quality depends on these excessively high citation counts. To solve this problem, Hersch proposed h-index [5] for calculating a researcher's quality index which is resistant to excessive citation counts. In this paper, Hersch writes as follows:

Definition 1 (h-index for scientists) A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have no more than h citations each.

h-index has some novel features, such as being a simple method, intuitively clear [6], easy to calculate [7], and resistant to excessive citation counts. For these reasons, h-index has become one of the most popular indexes for evaluating scientists.

A remarkable feature of h-index is that this index is resist to excessive values. This means that if we integrate a set of values, and the set contains excessive values, the integrated value does not depend on the excessive values. We consider that when we apply this idea of h-index to calculate Wikipedia editor quality values, the accuracy of calculating editor quality values should improve.

When we calculate h-index for Wikipedia editors, we should extract positive ratings which correspond to the citation counts. Survival ratio based approach is a popular approach for extracting positive ratings. If many readers feel excellent for a text, the quality of this text is good, but if many readers feel that a text should be removed, the quality of this text is poor. Adler et al. [1] found that 79% of poor quality texts are short-lived. We can estimate from this result that if editors find poor quality texts, the editors should remove them. Therefore, when an editor leaves a text, we consider that the editor gives a positive rating to the text.

However, the problem of h-index is that if there is a low h-index editor, we cannot distinguish whether the editor is a vandal or not. In **Fig. 1**, we show three types of editors, such as (A) high quality editors, (B) vandals, and (C) inactive editors. The editors in the cluster (A) have a high h-index, but the editors in both (B) and (C) have a low h-index. The editors in (B) do not change the quality of articles, because these editors do not post texts to Wikipedia. However the editors in (C) decrease the quality of articles, because these editors post many texts, and many part of these texts are not good. Therefore, the clustering of (B) and (C) is important.

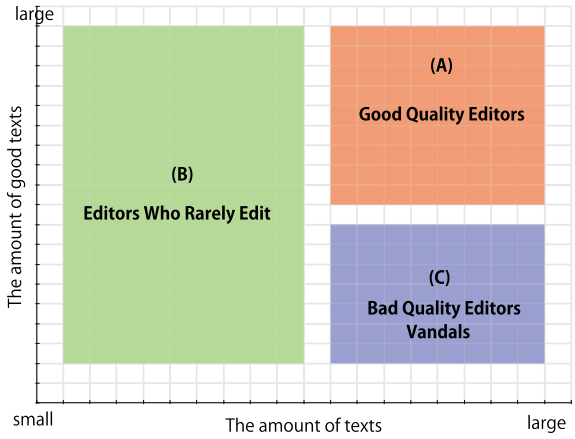

Fig. 1 Clustering of editors using editor quality.

To solve this problem, we propose p-ratio for clustering editors who have a low h-index. If an editor posts many texts, the editor faces many chances to increase his/her h-index. Therefore, if an editor's h-index is low, the editor is not approved by the other editors. However, if an editor posts a small number of texts, the editor does not have enough chances to increase his/her h-index, so we cannot decide whether the editor is a vandal or not. We define p-ratio as the value of h-index divided by the number of all edited articles. When we use h-index with p-ratio, the accuracy of the article quality calculated by our proposed method should improve.

Our two contributions are as follows:

(1) We propose h-index based editor assessment method which can identify high quality editors.
(2) We propose p-ratio for identifying whether the editors are vandals or rarely editing editors.

2. Related Work

In this section, we introduce several methods for assessing Wikipedia articles. Then, we introduce assessment method for research papers, because our proposed method is inspired by the assessment methods of research papers.

2.1 Assessment of Wikipedia Articles

There has been many researches for assessing quality of products, people, and objects using reputation based method [8]. A key concept for evaluating Wikipedia articles is *peer review based process*. Wikipedia is not thought to have a peer review system because most texts are instantly made and saved, though no one reviews these texts. However, Stivila et al. [9] mentioned that the open edit system is a kind of peer review system where editors of the system vote on implicit features of the texts.

In these investigations, many features are extracted from Wikipedia data, and they can be divided into two types: explicit and implicit features. Explicit feature is the user's decision which are directly input to the system by users, and implicit features are the user's decision which the system presumes from their behaviors. In this section, we describe the studies that have used explicit and implicit features and also describe why we choose to use implicit features.

One of the most important purposes of these researches is for improving the quality of articles in Wikipedia, which is related to the governance of Wikipedia. However, there is no standard for measuring the effectiveness of these methods, we cannot find how these methods are effective for improving articles. Geiger et al. [10] shows a case study of the tool called Huggle. In this research, they discover how Huggle is effective to improve the quality of articles.

For improving the quality of articles in Wikipedia, vandalism detection methods are proposed by many researchers. For example, West et al. [11] and Smets et al. [12] proposed a machine learning based method. These methods can identify vandals and low quality articles, but cannot identify good quality editors or articles. In the research introduced the following section, they try to identify good articles and editors.

2.1.1 Explicit Features

Explicit feature is commonly used to evaluate quality of information, products, and objects. For example, many on-line shopping sites like Amazon.com [*4] have voting systems for users to evaluate products. When users want to evaluate how satisfied or not they are with a product they have bought, they give the product 1–5 stars and submit review texts. Then, the system presents the average number of stars along with the reviews. If the other users want to know the quality or the satisfaction of the products, they refer to the number of stars and reviews, and decide whether to buy it or not. This system has been implemented as a part of many on-line Web service, such as YouTube [*5] and Google+ [*6], because it is easy to implement and the process of calculating the number of stars is easy and clear.

Kramer [13] implemented the voting system on MediaWiki [*7] for educational use, and also implemented in the English version of Wikipedia as Article Feedback Tool [*8]. Using these systems, users easily understand which are good quality articles by referencing these votes. However, one problem with this system is that not every user always appropriately evaluates or reviews. In fact, according to statistics about YouTube, almost all users who vote gave the highest score to almost all videos they votes [*9]. Moreover, the survey of the Article Feedback Tool by Wikipedia [*10] shows that 90.9% rates are the highest score. From this statistic, we find that users rate only good targets, but they do not rate poor targets.

The advantage of this system is that users can directly evaluate the quality of targets. However, the disadvantage is that only a small number of users input negative ratings. Therefore, if there is a target with a small number of voters, we cannot identify the target as either the poor quality target or the non-reviewed target. One reason of this problem is a lack of negative ratings, which is hard to recover by analysis of ratings. Therefore, we do not use explicit features.

2.1.2 Implicit Features

Implicit features are the user's decisions which the system presumes from their behaviors. When the system uses these features, users do not need to input the evaluation of items. Our proposed method uses this method. However, how can users' evaluations be presumed from their behavior?

Life-cycles of texts are used for calculating qualities of texts or articles. Wöhner et al. [14] calculate Wikipedia article qualities using the life-cycle of texts. In this method, they discovered that the quality and life-cycle of a text have a relationship. Halfaker et al. [17] presume implicit features from contribution degrees for editors. In this method, they proposed six heuristic assumptions about why editors contribute to Wikipedia. These ideas are appropriate when the articles are frequently edited. However, the frequency of edits is different, then the life-cycle of a text is different for every article. In addition, when edit warring occurs, this method cannot calculate appropriate qualities. Our method can calculate appropriate qualities if articles are not only infrequently edited but also suffering an edit war because we consider editor qualities.

Adler et al. [1], [18], [19] and Wilkinson et al. [20] proposed a method for calculating quality values from edit histories. This method is based on survival ratios of texts. Hu et al. [2] also proposed a method for calculating article quality using editor quality, which is similar to our proposed method. This method focuses on unchanged content, and they assumed that if an editor unchanged texts, the editor treated the texts as good texts. However, this method does not consider editors. Therefore, if there is an article which has only one version, the text in the article is not edited by the other editors, we cannot calculate the quality values of the text using the existing methods. In our method, we consider editors. Therefore, if the editor of the text edits another texts, and these texts are leaved and deleted by the other editors, we can calculate the quality of the new text.

Stvilia et al. [21] and Warncke-Wang et al. [22] proposed a method for classifying articles by article qualities using machine learning techniques. In these methods, they extract multiple features, such as reputation, completeness and complexity. However, these methods are not effective if each feature is related to the quality of articles. In our research, we confirm that reputation of editors are effective for assessing quality of articles. Therefore, if their system use our proposed article quality measure with the other features, the accuracy of article quality values should improve.

2.2 Assessment of Researchers

Impact Factor [23] is one of the most famous index for measuring quality of academic journal. In this index, if the papers in an academic journal are referred by many papers, the quality of the journal is high. This index does not express a quality of researchers, but this index is sometimes applied for measuring a quality of researchers.

Hirsch [5] proposed h-index for evaluating quality of researchers. This measure is simple, intuitively clear, and easy to calculate. For these reasons, h-index is one of the most popular index for evaluating researchers and scientists. However, this index has several weak points which is discussed at Ref. [24]. For example, h-index cannot calculate appropriate quality values for young researches. There are a lot of young researchers who write great papers. However, young researchers spend shorter time in research than senior researchers, young researchers submit smaller number of papers than senior researchers, generally. Therefore, if a young researcher submit a small number of great papers, and these papers are cited by many papers, the h-index of the young researcher is less than the number of the papers. As a result, h-index of young researchers are relatively smaller than that of senior researchers.

To solve this problem, many indices, such as g-index, A-index, and R-index, are proposed. Antonakis et al. [24] proposed IQp, which can solve this problem using Impact Factor. However, the procedure of calculating IQp is very complex, and the value of

[*4] http://www.amazon.com/
[*5] http://www.youtube.com/
[*6] https://plus.google.com/
[*7] http://www.mediawiki.org/
[*8] http://en.wikipedia.org/wiki/Wikipedia:Article%20Feedback%20Tool
[*9] http://www.techcrunch.com/2009/09/22/youtube-comes-to-a-5-star-realization-its-ratings-are-useless/
[*10] http://www.mediawiki.org/wiki/Article_feedback/Survey

this index is not intuitive, hard to understand. Consequently, *h*-index is still a major index for evaluating quality of researchers.

In our research, we introduce *p*-ratio with *h*-index to solve this problem. We assume an editor who posts a small amount of texts to Wikipedia articles, and who is equivalent to a young researcher in academic research field. If the texts are high quality, many editors give positive ratings to the texts. Similarly, if the papers are high quality, many papers cite the papers. In this case, the value of *h*-index is very small, at most the number of articles in the case of Wikipedia editors, the number of papers in the case of researchers. However, *p*-ratio is close to 1, the maximum value of *p*-ratio. On the other hand, if there is a vandal or a low quality researcher, the value of *p*-ratio is very small. As a result, when we use *p*-ratio and *h*-index, the value of these indices are intuitive, and we can solve the problem of *h*-index for editors who edit a small amount of articles, and young researchers. This *p*-ratio is similar to several indexes proposed by Iglesias [25] and Jensen [26]. However these indexes are used to identify high quality researchers, but they do not try to identify low quality researchers. We propose *p*-ratio for identifying low quality editors.

3. Proposed Method

In this section, we describe a method for assessing Wikipedia articles using edit history. Our proposed method consists of the following four steps:

(1) Peer reviewing of editors.
(2) Calculates editor quality using two aspects.
- *h*-index
- *p*-ratio

(3) Integrate two editor quality values.
(4) Calculate quality values of Wikipedia editors.

3.1 Peer Reviews of Editors

In this section, we introduce how to extract editors' positive ratings to texts from the edit history.

In the original definition of *h*-index at Ref. [5], Hirsch assume that if paper p_a refers another paper p_b, they consider that p_a gives positive ratings to p_b. When we calculate *h*-index, we need positive ratings of editors. Therefore, when we calculate *h*-index for Wikipedia editors, we should extract positive ratings for texts from edit history. In our method, we treat the editors' actions of leaving as positive ratings, which is similar to Adler's method [1]. In this idea, when editors leave texts, we treat that the editors who leave the texts give positive ratings to the texts. This is because, we assume that editors should delete the texts if the texts are inappropriate, and editors should leave the texts if the texts are appropriate to leave.

However, editors do not always browse the whole articles, the editors cannot delete texts that the editors do not browse, even if these texts are inappropriate for leaving. Therefore, we cannot treat all leaving actions of texts by editors as positive ratings. To solve this problem, we use a section based extraction method which is proposed at Ref. [27]. In this method, if an editor adds or removes several texts, the system considers that the editors gives positive ratings to the texts in the section that the editor adds or removes.

Using this method, we identify the peer review results of two editors. Let us consider two editors e_a, e_b in all editors E, $e_a \neq e_b$ who are the editors of article d in all articles D, and e_a leaves e_b's texts in d. E is a set of Wikipedia editors, and D is a set of Wikipedia articles. In this case, positive rating $p(d, e_a \rightarrow e_b)$ of e_b from e_a in d is defined as follows:

$$p(d, e_a \rightarrow e_b) = \begin{cases} 1 & \left(\text{if } \dfrac{|t_{e_b}| - |r_{e_a \rightarrow e_b}|}{|t_{e_b}|} \geq \alpha\right) \\ 0 & (\text{else}) \end{cases} \quad (1)$$

where t_{e_b} is the texts written by e_b in d, and $|t_{e_b}|$ is the number of characters in t_{e_b}. $r_{e_a \rightarrow e_b}$ is the texts written by e_b and deleted by e_a, and $|r_{e_a \rightarrow e_b}|$ is the length of $r_{e_a \rightarrow e_b}$. α $(0 < \alpha \leq 1)$ is a parameter for deciding whether the action of e_b is the leaving or not. When e_a leaves more than $\alpha \cdot |t_{e_b}|$ characters, $p(d, e_a \rightarrow e_b)$ is set to 1, then we judge that e_a gives positive ratings to e_b.

One important policy is that editors do not evaluate themselves. Therefore, even if e_a leaves texts written by e_a himself/herself, e_a does not give positive ratings to e_a. This is because, if editors can evaluate themselves, editors can arbitrary increase their *h*-index.

In our method, when e_a or e_b edits the same articles more than once, and if e_a gives positive ratings to e_b more than once, we judge that e_a gives positive ratings to e_b. Therefore, if e_a edits d twice, and e_a gives positive ratings once, we judge that e_a gives positive ratings to e_b. Moreover, if e_b edits more than twice, and e_b earns positive ratings more than once, we judge that e_b earns positive ratings from e_a. However, this ratings is calculated for each article respectively. Therefore, if e_b earns positive ratings from e_a at two articles d_1 and d_2, e_b earns positive ratings from e_a twice.

3.2 Two Editor Quality Values

In this section, we introduce two kinds of editor quality values: *h*-index and *p*-ratio. *h*-index indicates a degree of editor quality. *p*-ratio is a complementary index of *h*-index, and *p*-ratio is used to identify whether an editor is a vandal or an inactive editor.

3.2.1 *h*-index

Here, we define *h*-index of editor $e \in E$. We already show the original definition of *h*-index at Section 1. We modify this definition for assessing Wikipedia editors.

The definition of *h*-index $h(e)$ of editor e is as follows:

Definition 2 (*h*-index of editor *e*) $h(e)$ is the index of editor e who edits more than $h(e)$ articles, and in $h(e)$ article, e get positive ratings from at least $h(e)$ editors, and in the other $N - h(e)$ articles e does not get positive ratings from less than $h(e)$ editors. N is the number of articles which are edited by e.

In this definition, if there are two editors e_a and $e_b \in E$, and $p(d, e_a \rightarrow e_b)$ is 1 by Eq. (1), e_b get positive ratings from e_a.

Example 1 (*h*-index of editor e_a) In **Fig. 2**, we show the example of the behavior of editors e_a, e_b, e_c, e_d. In this figure, circle means editor, rectangle means article, white arrow means edits, and black arrow means positive rating. This figure shows that there are four editors e_a, e_b, e_c and e_d. e_a edits three texts to three articles d_1, d_2, and d_3. In d_1, editors e_b and e_c leave e_a's text. In d_2, editor e_d leaves e_a's text. In d_3, no editor leaves e_a's text. The number of editors who leave e_a's texts are shown in **Table 1**.

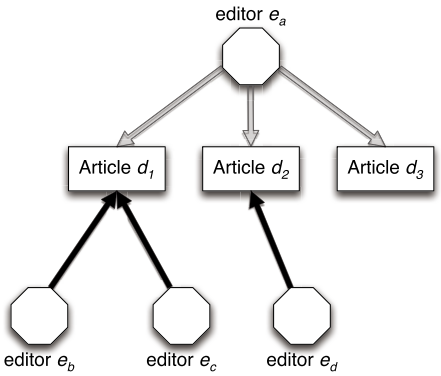

Fig. 2 Example of e_a's behavior.

Table 1 The number of editors who give positive ratings to editor e_a.

	d_1	d_2	d_3
# of editors who approve e_a's text	2	1	0

We check whether h-index $h(e_a)$ of e_a is at least one or not. From Table 1, e_a receives more than 1 positive ratings in two articles d_1 and d_2. Therefore, h-index of e_a is at least 1.
Next, we check whether $h(e_a)$ is at least two or not. From Table 1, e_a receives more than 2 positive ratings in one article d_1. Therefore, e_a does not get positive ratings more than two articles, h-index of e_a is less than 2. From these results, h-index of e_a is 1.

This definition of h-index shows that if an editor edits many articles, and the editors get positive ratings from many editors, the editor's h-index increases. However, if the editors only edits many articles and do not get positive ratings, or if the editors edit small number of articles and get many positive ratings, the h-index does not increase. Therefore, h-index is resistant to excessive number of positive ratings.

h-index is resistant to edit warring. For example, we consider the case that good editors edits vandals' texts, and the vandals revert the good editors' edits. In this case, the vandals do not give positive ratings to the good editors, and the good editors do not give positive ratings to the vandals. Therefore, we cannot identify which editors are good or not from this edit history. However, in Wikipedia, there are many editors, and the grater part of the editors are good editors. Many good editors should give positive ratings to the good editors, and should not give positive ratings to vandals. As a result, we can divide good editors and vandals using many editors' decisions.

3.2.2 *p*-ratio

h-index can identify high quality editors, but cannot identify rarely edited editors and low quality editors. To explain this, we consider two editors e_1 and $e_2 \in E$. h-index of both editors are 1. e_1 can be considered vandal, because e_1 edits many articles, but e_1 earn positive ratings from one editor in one article. e_2 can be considered good editor, because e_2 edits only one article, but e_2 earn positive ratings from many editors in one article. In this case, the quality value of e_2 should be better than that of e_1. However, when the system use only h-index, the quality values of e_1 and e_2 is the same.

To solve this problem, we use p-ratio with h-index. p-ratio is a ratio of the number of articles which increase h-index to the number of articles which is edited by the editor. We assume that if an editor is a vandal like e_1, even if the editor edits texts in many articles, the texts are rarely approved by the other editors. Then, p-ratio of the editor is low. On the other hand, if an editor is a good editor like e_2, even if the editor edits texts in a small number of articles, the texts are approved by many editors. Then, p-ratio of the editor should be high. Using p-ratio, we can identify whether an editor is a vandal or an editor who rarely edits Wikipedia articles.

We define p-ratio $p(e)$ of e as follows:

$$p(e) = \begin{cases} \dfrac{h(e)}{|w(e)|} & (\text{if } |w(e)| > 0) \\ 0 & (\text{if } |w(e)| = 0) \end{cases} \tag{2}$$

where $w(e)$ is the articles which e edits, and $|w(e)|$ is the number of articles in $w(e)$. The range of $w(e)$ is between 0 and 1. We define that if an editor edits no article,

Example 2 (p-ratio of editor e_a) We use the same example in Section 3.2.1. As we described, h-index of e_a is 1, and the number of articles which e_a edits is 3. Therefore, p-ratio $p(e_a)$ of e_a is $\frac{1}{3} \approx 0.33$.

3.2.3 Editor Quality

We calculate editor quality $u(e)$ of editor e using h-index and p-ratio as follows:

$$u^h(e) = h(e) \tag{3}$$

$$u^{p+}(e) = h(e) + p(e) \tag{4}$$

$$u^{p*}(e) = h(e) \cdot p(e) \tag{5}$$

When we calculate the editor quality, we select one of three equations. $u^h(e)$ at Eq. (3) is defined using h-index, ignore p-ratio. u^{p+} at Eq. (4) is defined as $h(e)$ plus $p(e)$, which is inspired by the idea of extended boolean model. Using this equation, if either $h(e)$ or $p(e)$ is high value, u^{p+} is high value. u^{p*} at Eq. (5) is $h(e)$ multiplied by $p(e)$, which is also inspired by the idea of extended boolean model. In this case, u^{p*} is high value if both $h(e)$ and $p(e)$ are high value. We compared these three equations whether the operators "OR" and "AND" are effective or not for integrating $h(e)$ and $p(e)$.

Example 3 (Editor quality of e_a) We also use the same example as Section 3.2.1. From this example, $h(e_a)$ is 1 and $p(e_a)$ is 0.33. Therefore, $u^h(e_a)$ is 1, $u^{p+}(e_a)$ is $1 + 0.33 = 1.33$, and $u^{p*}(e_a)$ is $1 \cdot 0.33 = 0.33$. The quality values of e_b and e_c are described at **Table 2**.

Table 2 The quality values of e_a, e_b, and e_c.

	$h(e)$	$p(e)$	$u^h(e)$	$u^{p+}(e)$	$u^{p*}(e)$
e_a	1	0.33	1	1.33	0.33
e_b	1	1	1	2	1
e_c	2	0.33	2	2.33	0.66

3.2.4 Article Quality

Finally, we calculate the quality value $q(d)$ of article d using a weighted sum as follows:

$$q(d) = \frac{\sum_{e \in E(d)} t(e) \cdot u(e)}{T(d)} \tag{6}$$

where $E(d)$ is the editors who edit d, $t(e)$ is the number of characters inserted by e, and $T(d)$ is the total number of characters in

d. $q(d)$ means one of three kinds of article qualities $q^h(d)$, $q^{p+}(d)$, and $q^{p*}(d)$.

Using this formula and three kinds of editor qualities defined at Eqs. (3)–(5), we calculate three kinds of article qualities. For example, there are three editors e_a and e_c who edit article d, and the numbers of characters of e_a and e_c are 200 and 100 respectively. In this case, $q^h(d)$ is $\frac{200\cdot1+100\cdot2}{100+200} \approx 1.33$. In the same way, we can calculate that $q^{p+}(d) \approx 1.66$ is, and $q^{p*}(d) \approx 0.44$.

4. Experimental Evaluation

To confirm that our proposed method can calculate accurate quality values for Wikipedia articles, we did two experiments. In these experiments, we used two kinds of test collections. One test collection is a set of rated articles from all Wikipedia articles, and another test collection is multiply rated articles in the specific category.

In our experiment, we compared four methods: *hindex*: the system using h-index only ($q^h(d)$), *hindex+pr*: the system using h-index plus p-ratio ($q^{p+}(d)$), *hindex* $\cdot$ *pr*: the system using h-index multiplied by p-ratio ($q^{p*}(d)$), and *remain*: the baseline system based on the proposed method proposed by Adler et al. [1].

We used edit history dump data of Japanese Wikipedia at March 28, 2013 [*11]. In this data, there are 1,362,653 articles, 2,654.683 editors, and 38,371,993 versions. The latest version of all articles have at least 1 characters. The editor set contains bots [*12], an automated or semi-automated tools.

In our experiments, we use two kinds of article sets as correct answers. In experiment 1, we use the answer set as "featured" and "good" articles selected from all articles. In experiment 2, we use the answer set as the seven graded articles, such as "featured," "good," "A," "B," "C," "Start," and "Stub," from the articles in the two categories.

We use two different answer sets, because we cannot find an ideal answer set. The answer set in experiment 1 is a large article set, then we can observe the effectiveness of our method for a whole article. However, the coverage of featured and good articles are small, many good quality articles are not selected as featured or good articles. This is because, the reviewers of these articles does not always browse a whole article. To complement the defect of this article set, we use two small article sets in experiment 2. In these sets, almost all high quality articles are selected as high grade articles. This is because, the reviewers read all articles in the category. We did two experiments in a macro-viewpoint at experiment 1 and in a micro-viewpoint at experiment 2 to complement the defects of each answer set.

The experimental procedure of experiment 1 and 2 is as follows:

(1) Our proposed method extracts statistical data about positive ratings from edit history of all articles in Wikipedia.
(2) We set a correct answer set, and set scores to target articles.
- In experiment 1, target articles are all articles in the Wikipedia. We set the score "2" to the featured articles, "1" to the good articles, and "0" to the other articles.
- In experiment 2, target articles are the articles in the category "Islam" and "Sports." We set the feature articles, good articles, A-class, B-class, C-class, Start, Stub class, to score $7, 6, \cdots, 0$, respectively.

(3) The system computes four kinds of editor quality values, such as *hindex*, *hindex+pr*, *hindex* $\cdot$ *pr*, *remain*. We set α from 0.1 to 1.0 in 0.1 increments and calculate ten kinds of editor quality values for *hindex*, *hindex+pr*, *hindex* $\cdot$ *pr*.
(4) The system computes the article quality values using the editor quality values.
(5) The system makes article list in decedent order of the article quality values.
(6) The system makes article list in decedent order of the article score described at step (2).
(7) We compare two article lists at step (5) and (6), and calculate nDCG.

We use evaluation measure as nDCG [28], [29] which is widely used for evaluating multi-grade test collections. nDCG is defined as follows:

$$nDCG = \frac{DCG}{IDCG} \tag{7}$$

where DCG is defined as follows,

$$DCG = r_1 + \sum_{i=2}^{p} \frac{r_i}{\log_2(i)} \tag{8}$$

where r_i is the graded relevance of the result list at position i. $IDCG$, which stands for Ideal DCG, is the value of DCG about the result list which is sorted by the score of the correct answer set. In this measure, if nDCG of a method is high, the article list ordered by the method is similar to the golden standard.

4.1 Experiment 1: Featured and Good Article Based Evaluation

4.1.1 Experimental Setup

In this experiment, we set "featured articles" and "good articles" as a correct answer set. Featured and good articles are selected by the votes of Wikipedia editors, and are evaluated by "Featured article criteria." The editors select 68 featured articles and 765 good articles from 1,362,653 articles. When we evaluate, we should convert the ratings into the numbers. Therefore, we convert the featured article into 2, the good article into 1 and the other article into 0.

4.1.2 Experimental Result

Figure 3 shows the experimental results. The horizontal axis shows the threshold value α which is used at Section 3.1, and the vertical axis shows nDCG. From this figure, we discover that three methods *hindex*, *hindex+pr*, *hindex* $\cdot$ *pr* can calculate more accurate article quality values than the baseline method *remain*. Moreover, the accuracies of *hindex* and *hindex+pr* are almost the same, and are higher than *hindex* $\cdot$ *pr*.

We also discover that if α is 0.5, the accuracy of *hindex* and *hindex+pr* is effective. However, we cannot find the appropriate value of α for *hindex* $\cdot$ *pr*.

We note that this correct answer set does not cover all high quality articles. However, we manually browse all featured and

[*11] http://dumps.wikimedia.org/jawiki/20130328/
[*12] http://en.wikipedia.org/wiki/Wikipedia:Bots

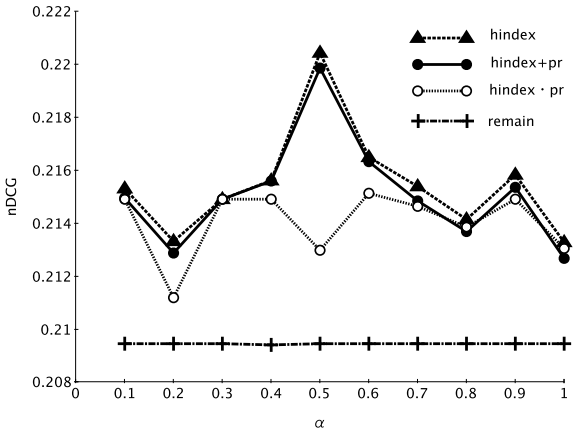

Fig. 3 Experiment result using "Featured articles" (FA) and "Good articles" (GA).

good articles, and we confirm that all featured and good articles are high quality. Therefore, from this result, we confirm that when we use *hindex* or *hindex+pr*, we can identify high quality articles. However, we cannot conclude that these methods can identify low quality articles, because there are many high quality articles which are not tagged as featured and good articles. Therefore, in the next experiment, we confirm which methods are adequate for identifying low quality articles.

In Fig. 3, the shapes of the curves of *hindex*, *hindex+pr*, *hindex · pr* are not smooth. This is because, the ratio of leaving mentioned at Eq. (1) is not always related to the quality of text. For example, there are two texts, and an editor leaves 80% and 20% of the texts respectively. In this case, we assume that the former text should have better quality than the latter text. However, in some cases, the latter text is better quality than the former text. As a result, the relationship of the leaving ratio of text and the quality of text is not always fixed, then the shapes of the curves in Fig. 3 are not smooth.

4.2 Experiment 2: Evaluation Using Articles in the Specific Categories

In experiment 1, we use the featured and good articles as correct answer articles, and we use target articles as a whole article in Wikipedia. However, from this experiment, we cannot confirm whether the methods can identify low quality articles or not. To solve this problem, we use target articles as a set of articles in specific categories. The number of articles in the target of experiment 2 is very small in comparison to the number of all articles in Wikipedia. Therefore, the reviewers can browse all articles in the target articles, the low quality articles identified by the reviewers are truly low quality articles. Using this correct answer set, we can confirm which methods can identify low quality articles.

4.2.1 Experimental Setup

In Japanese Wikipedia, there are several Wiki Projects q[*13] which are the working groups for specific topics. We picked up two projects such as "Islam"[*14] and "Sports"[*15], because in only these two projects, the editors assess the quality of articles, and categorize the articles in the projects into multiple grades.

[*13] http://ja.wikipedia.org/wiki/Wikipedia:ウィキプロジェクト
[*14] http://ja.wikipedia.org/wiki/Portal:イスラーム
[*15] http://ja.wikipedia.org/wiki/Wikipedia:ウィキプロジェクト_スポーツ

Table 3 Test collection - Evaluations of articles and the number of articles.

	F	G	A	B	C	Sta	Stu	Total
Islam	2	12	19	77	0	105	113	328
Sports	1	25	0	178	403	969	946	2,522

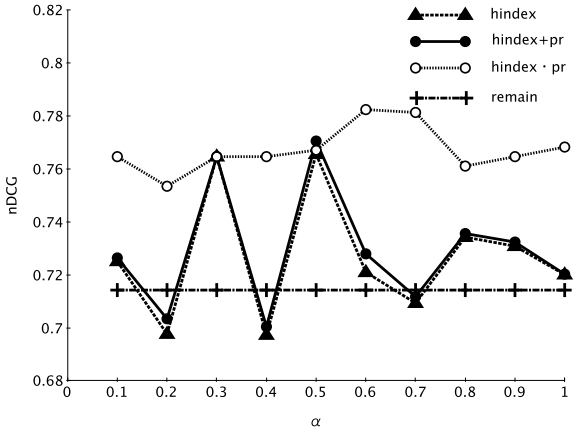

Fig. 4 Evaluation results of the articles in category "Islam."

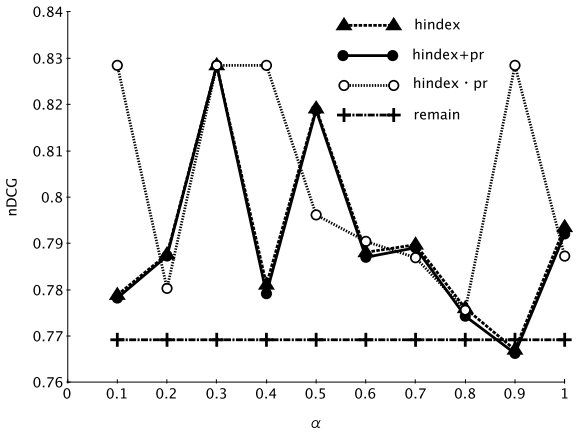

Fig. 5 Evaluation results about articles in category "Sports."

In these projects, the users evaluate the articles about the projects, and categorize into seven grades such as *Featured*, *Good*, *A*-class, *B*-class, *C*-class, *Start*, and *Stub*[*16]. When we evaluate our proposed method, we convert these grades into the number $6, 5, \cdots, 0$. **Table 3** shows the number of articles. In this table, F means featured articles, G means good articles, A, B, and C means A-class, B-class, and C-class articles respectively, Sta means Start-class articles, and Stu means Stub-class articles.

4.2.2 Experimental Result

Figures 4 and **5** show our experimental results. The meanings of horizontal and vertical axes are the same as Fig. 3 of experiment 1. From these figures, we can confirm that our proposed methods *hindex*, *hindex+pr*, and *hindex · pr* are more accurate than the baseline method *remain*. Moreover, in many cases, *hindex · pr* is the most accurate methods than the other methods.

From Fig. 4, we confirmed that when we use the category "Islam," *hindex · pr* is the most accurate method. Especially, when we use *hindex · pr*, the system can identify whether an article is low quality, such as Start or Stub, or high quality, such as Featured article, Good Article, A-class, and B-class. Therefore, we

[*16] http://en.wikipedia.org/wiki/Wikipedia:Version_1.0_Editorial_Team/Assessment

confirm that *hindex* · *pr* is effective for identifying low quality articles, and *p*-ratio is effective for identifying low quality editors.

However, we cannot clearly confirm best methods from our proposed three methods when we use the category "Sports" which is described at Fig. 5. Therefore, we discover the detail of the result list. As you see in Table 3, about 91% of all articles in this category are tagged as C-class, Start or Stub. C-class articles are written by a small number of editors in many cases. Therefore, our proposed methods misjudge C-class articles to be Start or Stub articles, the accuracy of our proposed methods decreases.

We also discover that when we use the category "Islam," if the threshold α is set to between 0.6 and 0.7, *hindex* · *pr* is the most accurate. However, when we use the category "Sports," we cannot find the most appropriate value of α. From the results of experiment 1 and 2, appropriate value of α should depends on target articles. Therefore, we should develop a method for setting the most appropriate value of α.

From these results, we conclude that if the users try to identify high quality articles, *hindex* or *hindex+pr* are the best solution. On the other hand, if the users try to identify low quality articles, *hindex* · *pr* is the best solution.

4.3 Discussions

To analyze the detail of results, we pick up several editors, and analyze these editors.

First, we discuss whether *h*-index is effective or not. "At by at"[*17] is an editor, and we observe that he is a low quality editor. His interests are mainly related to soccer, therefore he have posted many texts to a small number of articles about soccer. His texts were occasionally approved by many editors, then the quality value of *remain* is very high due to excessive text quality values. However, using our proposed method, these excessive text quality values do not affect his editor quality value. Therefore, his *hindex* is very low. We can observe many cases like this editor when we did our experiments.

Next, we discuss the effectiveness of *p*-ratio. The editor "ホイップ" (Whip)[*18] is a vandal, and a blocked editor, who prevent from editing Wikipedia. This editor have posted 23,575 times, but he posted a small number of texts to a large number of articles. Almost all his texts are meaningless. However, his texts are not seen by the other editors in many cases, his texts are occasionally approved by the editors. Therefore, his *h*-index is 53, which is very high value when we compare to the other editors. Therefore, when we use *hindex* or *hindex+pr*, he is considered as a high quality editor. However, his *p*-ratio is 0.0022, which is very low value when we compare to the other editors. Then *hindex* · *pr* is a very low value, he is considered as a low quality editor. As a result, the accuracy of editor quality improves, then the accuracy of article quality also improves.

However, our proposed method is not effective for sock puppetry, who is a user who use multiple Wikipedia accounts. If an editor posts a text, and the other editor approve the text, and these two editors are different name but same person, the quality value of the text increases against our intention. In future work, we should construct a method against these kinds of group attacks.

5. Conclusion

In this paper, we proposed a method for assessing Wikipedia articles from edit history using *h*-index and *p*-ratio. The articles in Wikipedia are written by unspecified large number of editors, and these articles are not reviewed by the other editors. Therefore, there are a lot of low quality articles. Remain ratio based methods are proposed in the past, but the article quality values calculated by these methods are affected by several excessively high and low values of remain ratios. To solve this problem, we proposed *h*-index based method, which is resistant to the excessive values.

From two experimental evaluations, we confirmed that our proposed article assessing technique is more useful for assessing Wikipedia articles than the existing method. Especially, when we identify good quality articles, *hindex* and *hindex+pr* are the adequate methods, and when we identify low quality articles, *hindex* · *pr* is the adequate method.

However, from experimental results, we cannot develop a perfect measure which can identify both high and low quality articles. This is a limitation of our proposed measures. Therefore, we should combine these two or three measures, and develop one perfect measure which can identify both high and low quality articles, in the future work.

The techniques of Wikipedia article assessment can be applied to many research areas, such as knowledge base construction and data cleaning. Recently, large scale knowledge databases such as DBPedia[*19] and YAGO2s[*20] are constructed using Wikipedia. When the researchers of DBPedia and YAGO2s construct these databases, they use texts in the Wikipedia. However, there are many incorrect information in these databases. This is because, the texts in Wikipedia are not always correct. Therefore, using our proposed system for data cleaning, the quality of these knowledge databases will improve.

Finally, we describe our future works. *h*-index is considered as a simple and effective index by many researchers, and we confirmed that *h*-index is also effective for Wikipedia article assessment. However, *h*-index is not perfect. For example, good quality editors who posts a small number of articles are evaluated as low quality editors. If *h*-index of an editor is 100, the editor should posts at least 100 articles. Even if an editor posts only 10 articles to many good quality texts, *h*-index of the editor is at most 10, which is a very low value. To solve this problem, many indices like *g*-index [30], *A*-index, *R*-index, IQp [24] are proposed. However, several indices have the same problem as *h*-index, and several indices, especially IQp, is very complex, and is not intuitive. Therefore, we should develop a simple but effective method for Wikipedia article assessment.

In these *h*-index improvement methods, for evaluating young researchers or journals fairly, citations of recently edited papers are considered as more important citations than citations of papers written in the past. However, we do not use this idea, because the history of Wikipedia is shorter than the history of aca-

[*17] http://ja.wikipedia.org/wiki/利用者:At_by_At
[*18] http://ja.wikipedia.org/wiki/利用者:ホイップ
[*19] http://dbpedia.org/
[*20] http://www.mpi-inf.mpg.de/yago-naga/yago/

demic research papers, and there are few editors who continuously edit Wikipedia for a long time. Therefore, almost all editors are treated as young editors, we do not care a problem that senior editors are tend to be evaluated as high quality editors and young editors are evaluated as low quality editors. However, if the history of Wikipedia becomes longer, we cannot ignore this problem. Therefore, in the near future, we should develop a method for fairly evaluating young editors, who edit Wikipedia a small number of times.

References

[1] Adler, B. and de Alfaro, L.: A content-driven reputation system for the Wikipedia, *Proc. 16th International Conference on World Wide Web* (*WWW '07*), pp.261–270 (online), DOI: http://doi.acm.org/10.1145/1242572.1242608 (2007).

[2] Hu, M., Lim, E., Sun, A., Lauw, H.W. and Vuong, B.: Measuring Article Quality in Wikipedia: Models and Evaluation, *Proc. ACM International Conference on Information and Knowledge Management* (*CIKM 2007*), pp.243–252 (2007).

[3] Suzuki, Y. and Yoshikawa, M.: Assessing quality score of Wikipedia article using mutual evaluation of editors and texts, *Proc. 22nd ACM International Conference on Information and Knowledge Management*, *CIKM '13*, pp.1727–1732, ACM (online), DOI: 10.1145/2505515.2505610 (2013).

[4] Suzuki, Y.: Assessing Quality Values of Wikipedia Articles Using Implicit Positive and Negative Ratings, *Proc. 13th International Conference on Web-Age Information Management* (*WAIM 2012*), pp.127–138 (2012).

[5] Hirsch, J.E.: An index to quantify an individual's scientific research output, *Proc. National Academy of Sciences*, Vol.102, pp.16569–16572 (online), DOI: 10.1073/pnas.0507655102 (2005).

[6] Egghe, L. and Rousseau, R.: An informetric model for the Hirsch-index, *Scientometrics*, Vol.69, pp.121–129 (online), DOI: 10.1007/s11192-006-0143-8 (2006).

[7] Batista, P.D., Campiteli, M.G., Kinouchi, O. and Martinez, A.S.: Is it possible to compare researchers with different scientific interests?, *Scientometrics*, Vol.68, pp.179–189 (online), DOI: 10.1007/s11192-006-0090-4 (2006).

[8] Stvilia, B., Gasser, L., Twidale, M.B. and Smith, L.C.: A Framework for Information Quality Assessment, *Journal of the American Society for Information Science and Technology*, Vol.58, No.12, pp.1720–1733 (2007).

[9] Stvilia, B., Twidale, M., Smith, L. and Gasser, L.: Information quality work organization in wikipedia, *J. Am. Soc. Inf. Sci. Technol.*, Vol.59, No.6, pp.983–1001 (online), DOI: 10.1002/asi.v59:6 (2008).

[10] Geiger, R.S. and Ribes, D.: The Work of Sustaining Order in Wikipedia: The Banning of a Vandal, *Proc. 2010 ACM Conference on Computer Supported Cooperative Work*, *CSCW '10*, pp.117–126, ACM (online), DOI: 10.1145/1718918.1718941 (2010).

[11] West, A.G., Kannan, S. and Lee, I.: Detecting Wikipedia Vandalism via Spatio-temporal Analysis of Revision Metadata?, *Proc. 3rd European Workshop on System Security*, *EUROSEC '10*, pp.22–28, ACM (online), DOI: 10.1145/1752046.1752050 (2010).

[12] Smets, K., Goethals, B. and Verdonk, B.: Automatic vandalism detection in Wikipedia: Towards a machine learning approach, *WikiAI '08: Proc. AAAI Workshop on Wikipedia and Artificial Intelligence* (2008).

[13] Kramer, M., Gregorowicz, A. and Iyer, B.: Wiki Trust Metrics based on Phrasal Analysis, *Proc. International Symposium on Wikis* (*WikiSym '08*) (2008).

[14] Wöhner, T. and Peters, R.: Assessing the quality of Wikipedia articles with lifecycle based metrics, *Proc. International Symposium on Wikis and Open Collaboration* (*WikiSym '09*), pp.1–10 (online), DOI: http://doi.acm.org/10.1145/1641309.1641333 (2009).

[15] Priedhorsky, R., Chen, J., Lam, S.T.K., Panciera, K., Terveen, L. and Riedl, J.: Creating, destroying, and restoring value in wikipedia, *Proc. 2007 International ACM Conference on Supporting Group Work*, *GROUP '07*, pp.259–268, ACM (online), DOI: 10.1145/1316624.1316663 (2007).

[16] Panciera, K., Halfaker, A. and Terveen, L.: Wikipedians are born, not made: A study of power editors on Wikipedia, *Proc. ACM 2009 International Conference on Supporting Group work*, *GROUP '09*, pp.51–60 (online), DOI: 10.1145/1531674.1531682 (2009).

[17] Halfaker, A., Kittur, A., Kraut, R. and Riedl, J.: A Jury of Your peers: Quality, Experience and Ownership in Wikipedia, *Proc. International Symposium on Wikis and Open Collaboration* (*WikiSym '09*), pp.1–10 (online), DOI: http://doi.acm.org/10.1145/1641309.1641332 (2009).

[18] Adler, B.T., Chatterjee, K., de Alfaro, L., Faella, M., Pye, I. and Raman, V.: Assigning trust to Wikipedia content, *Proc. 4th International Symposium on Wikis*, *WikiSym '08*, pp.26:1–26:12, ACM (online), DOI: 10.1145/1822258.1822293 (2008).

[19] Adler, B.T., de Alfaro, L., Pye, I. and Raman, V.: Measuring author contributions to the Wikipedia, *Proc. 4th International Symposium on Wikis*, *WikiSym '08*, pp.15:1–15:10, ACM (online), DOI: 10.1145/1822258.1822279 (2008).

[20] Wilkinson, D.M. and Huberman, B.A.: Cooperation and quality in Wikipedia, *Proc. 2007 International Symposium on Wikis* (*WikiSym '07*), pp.157–164, ACM (online), DOI: 10.1145/1296951.1296968 (2007).

[21] Stvilia, B., Twidale, M.B., Smith, L.C. and Gasser, L.: Assessing information quality of a community-based encyclopedia, *Proc. 2005 International Conference on Information Quality* (2005).

[22] Warncke-Wang, M., Cosley, D. and Riedl, J.: Tell Me More: An Actionable Quality Model for Wikipedia, *Proc. 9th International Symposium on Open Collaboration*, *WikiSym '13*, pp.8:1–8:10, ACM (online), DOI: 10.1145/2491055.2491063 (2013).

[23] Garfield, E.: Citation analysis as a tool in journal evaluation, *Science*, Vol.178, No.60, pp.471–479 (1972).

[24] Antonakis, J. and Lalive, R.: Quantifying Scholarly Impact: IQp Versus the Hirsch H, *J. Am. Soc. Inf. Sci. Technol.*, Vol.59, No.6, pp.956–969 (online), DOI: 10.1002/asi.v59:6 (2008).

[25] Iglesias, J.E. and Pecharromán, C.: Scaling the h-index for different scientific ISI fields, Vol.73, No.3, pp.303–320, Springer (2007).

[26] Jensen, P., Rouquier, J.-B. and Croissant, Y.: Testing bibliometric indicators by their prediction of scientists promotions, *Scientometrics*, Vol.78, No.3, pp.467–479, Springer (2009).

[27] Suzuki, Y.: Effects of Implicit Positive Ratings for Quality Assessment of Wikipedia Articles, *Journal of Information Processing*, Vol.21, No.2, pp.342–348 (2013).

[28] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *ACM Trans. Inf. Syst.*, Vol.20, No.4, pp.422–446 (online), DOI: 10.1145/582415.582418 (2002).

[29] Järvelin, K. and Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents, *Proc. 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, *SIGIR '00*, pp.41–48, ACM (online), DOI: 10.1145/345508.345545 (2000).

[30] Egghe, L.: An Econometric Property of the G-index, *Inf. Process. Manage.*, Vol.45, No.4, pp.484–489 (online), DOI: 10.1016/j.ipm.2009.04.001 (2009).

Yu Suzuki was born in 1977. He received his M.E. and Ph.D. degrees from Nara Institute of Science and Technology in 2001 and 2004 respectively. He became an assistant professor at Ritsumeikan University in 2004, a researcher at Kyoto University in 2009, and an assistant professor at Nagoya University in 2010. He is currently an associate professor at Nara Institute of Science and Technology. His current research interests include Social Web analysis and data mining. He is a member of IPSJ, IEICE, DBSJ, IEEE-CS, and ACM.