# WPUA-search: A Method to Discover Wikipedia Unusual Articles

**Tomoki Tada***
Hokkaido University

**Katsuhiko Hayashi***
Hokkaido University

**Hidetaka Kamigaito**
NAIST

**Yuya Taguchi**
The Asahi Shimbun Company

## Abstract

Wikipedia has high-quality articles on a variety of topics and has been used in diverse research areas. In this paper, we present a novel method using Wikipedia editor information to effectively discover unusual articles.

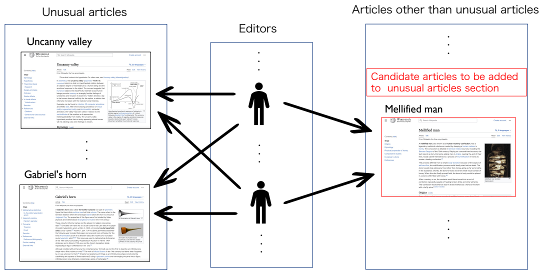
**Keywords:** Wikipedia, unusual articles, collaborative filtering, editor information, user preferences

## Introduction

Wikipedia contains many articles on a wide range of things, and some sections of Wikipedia contain articles on things not covered in ordinary encyclopedias. One such section is the unusual articles section. This paper focuses on a specific type of article, i.e., "unusual articles," which are characterized by being high-quality and slightly unconventional encyclopedia articles[1]. Since they cannot be found in ordinary encyclopedias, many of them are interesting and characteristic. The unusual articles section has a unique strength that can only be found on Wikipedia.

The problem we addressed is that the number of unusual articles is small: the Wikipedia dump data used in our experiments contains only 2,469 unusual articles compared with 22,329,081 total articles. Our aim is to develop methods for discovering articles that can be added to the unusual articles section and thereby expand it. Achieving this aim will lead to an improvement in Wikipedia's originality.

Wikipedia is an encyclopedia, and one of its basic policies is that content must be written from a "neutral point of view"[2]. Therefore, Wikipedia content information is less likely to reflect the personal opinions and preferences of the editors, and is mostly superficial attribute information. It is difficult to determine whether an article in Wikipedia is unusual by using only Wikipedia content information. We thus considered a different approach: using Wikipedia editor information instead of using content information. Using the editor information has been shown to be effective for the recommendation task (Takeuchi and Hayashi, 2022). Editor information reflects the preferences and knowledge of the editor whereas content information does not. The use of editor information should thus be useful in determining whether articles can be added to unusual articles section.


Unusual articles
Uncanny valley
Gabriel's horn
Editors
Articles other than unusual articles
Candidate articles to be added to unusual articles section
Mellified man

souce：https://en.wikipedia.org/wiki/Uncanny_valley、https://en.wikipedia.org/wiki/Gabriel%27s_horn、https://en.wikipedia.org/w/index.php?curid=6136647

Figure 1: Hypothesis based on collaborative filtering

We have developed a method for using editor information to discover articles that can be added to unusual articles section. Considering that editor information reflects editors' preferences and knowledge, we can assume that editors who edit a certain article are interested in articles similar to that article. We can interpret this assumption, to mean that "editors of unusual articles prefer funny articles similar to the unusual articles." In other words, if there are multiple editors who have edited the unusual articles, we can hypothesize that they prefer and edit articles that can be added to the unusual articles section.

This hypothesis, illustrated in Figure 1, is an application of collaborative filtering, and is the central idea of the proposed method. Collaborative filtering (Goldberg et al., 1992) is a method of recommending what other users with similar preferences have purchased or done and is based on the user's purchase history and/or action history. We conducted experiments to evaluate whether the use of editor information based on collaborative filtering is effective.

## Proposed Methods

To discover new unusual articles, we require a measure of how unusual articles are and define an article impor-

*equal contribution

[1] https://en.wikipedia.org/wiki/Wikipedia:Unusual_articles

[2] https://en.wikipedia.org/wiki/Wikipedia:Neutral_point_of_view

tance score. For this purpose, we first introduce some definitions:

$$\begin{aligned} Ar_{editor} &\equiv \text{(set of articles edited by an editor)} \\ Ar_{unusual} &\equiv \text{(set of all unusual articles)} \\ Ed_{article} &\equiv \text{(set of all editors of an article)} \\ Ed_{unusual} &\equiv \text{(set of all editors of unusual articles)}. \end{aligned}$$

We evaluate an editor ($editor$) on the basis of the editor's importance score ($SCORE_{editor}$), which reflects how many unusual articles an editor has edited. We evaluate an article ($article$) on the basis of the article's importance score ($SCORE_{article}$), which is the sum of the importance scores of editors who have edited unusual articles. $SCORE_{article}$ reflects how unusual an article is. The article's importance score is formally defined as

$$SCORE_{article} = \sum_{editor \in Ed_{article}} \begin{cases} SCORE_{editor} & editor \in Ed_{unusual} \\ 0 & editor \notin Ed_{unusual} \end{cases}.$$

We devised three methods for discovering new unusual articles. In the non-weighted method, the editor's importance score is set to 1 (no weighting).

- Non-Weighted Method

$$SCORE_{editor} = 1.$$

In the count method, the editor's importance score is based on the number of unusual articles the editor has edited.

- Count Method (weighted by number of unusual articles)

$$SCORE_{editor} = |Ar_{editor} \cap Ar_{unusual}|.$$

In the ratio method, the editor's importance score is based on the ratio of unusual articles edited by the editor to the total number of articles edited by the editor.

- Ratio Method (weighted by ratio of unusual articles)

$$SCORE_{editor} = \frac{|Ar_{editor} \cap Ar_{unusual}|}{|Ar_{editor}|}.$$

## Experiments

In our experiments, we used Wikipedia dump data for September 20, 2022, and Wikipedia editor data for October 2022. The two datasets shared 22,329,081 articles (including 2,469 unusual articles).

### Automatic Evaluation

We conducted two experiments to automatically evaluate whether the article's importance was an appropriate indicator for identifying it as unusual. A group of unusual articles was separated into training data (1,975 articles) and test data (494 articles) at a ratio of 8:2. The dataset used for automatic evaluation consisted of 100,000 articles including test data. For the training data, we calculated the article's importance score and generated a ranking of articles that could be added to the unusual articles section. On the basis of precision@k and recall@k, we evaluated the degree to which there were test data in the top k cases (k=50,100,200,500). This is experiment 1. In addition, we also conducted the other experiment with a restriction of a small number of editors (less than 165 editors). This is experiment 2. The value of 165 is the mean of the number of editors for unusual articles, and much higher than the median of 96. It is appropriate to exclude articles with a large number of editors since they are generally well-known articles. Tables 1 and 2 show the results of experiments 1 and 2, respectively. They clearly show that ratio method achieved the best performance.

### Human Evaluation

Given the improvement in top precision in Experiment 2, we extracted the top 30 most important articles from all articles with 165 or fewer editors (22,167,184 articles). The author then subjectively judged whether they could be added to the unusual articles section. Three example articles that the authors subjectively judged are listed in Table 3. As shown in Figure 2, ratio method had the best performance.

## Contributions and Findings

We summarize our contributions as follows:

- We presented a method that uses Wikipedia editor information to discover unusual articles.
- We demonstrated that the use of editor information based on collaborative filtering is effective for discovering unusual articles.

## References

[Goldberg et al.1992] David Goldberg, David Nichols, Brian M Oki, and Douglas Terry. 1992. Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, 35(12):61–70.

[Takeuchi and Hayashi2022] Koki Takeuchi and Katsuhiko Hayashi. 2022. Entity similarity estimation by collaborative filtering with wikipedia editor information. *Proceedings of Web Intelligence and Interaction*, 18:43–50.

| Evaluation index | Non-weighted | Count | Ratio |
|---|---|---|---|
| Precision@50 | 0.120 | 0.160 | **0.300** |
| Precision@100 | 0.170 | 0.150 | **0.360** |
| Precision@200 | 0.175 | 0.165 | **0.390** |
| Precision@500 | 0.216 | 0.194 | **0.332** |
| Recall@50 | 0.0121 | 0.0162 | **0.0300** |
| Recall@100 | 0.0344 | 0.0304 | **0.0729** |
| Recall@200 | 0.0709 | 0.0668 | **0.158** |
| Recall@500 | 0.219 | 0.196 | **0.336** |

Table 1: Results of experiment 1

| Evaluation index | Non-weighted | Count | Ratio |
|---|---|---|---|
| Precision@50 | 0.160 | 0.0800 | **0.560** |
| Precision@100 | 0.160 | 0.0700 | **0.500** |
| Precision@200 | 0.125 | 0.0750 | **0.445** |
| Precision@500 | 0.108 | 0.0800 | **0.296** |
| Recall@50 | 0.0162 | 0.00810 | **0.0567** |
| Recall@100 | 0.0324 | 0.0142 | **0.101** |
| Recall@200 | 0.0506 | 0.0304 | **0.180** |
| Recall@500 | 0.109 | 0.0810 | **0.300** |

Table 2: Results of experiment 2

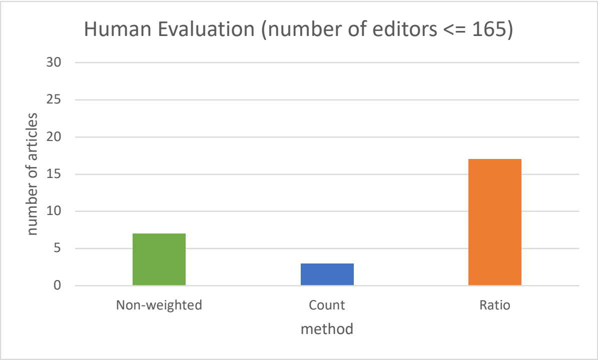


Figure 2: Number of articles with less than 165 editors that can be added to unusual articles section (out of top 30 most important articles)

| article | description |
|---|---|
| INTERCAL[3] | The Compiler Language With No Pronounceable Acronym (INTERCAL) is an esoteric programming language. It satirizes aspects of the various programming languages at the time. |
| Golden Arches[4] | The Golden Arches are the symbol of McDonald's, the global fast food restaurant chain. |
| Major League Quadball[5] | Major League Quadball (MLQ) is an amateur quidditch league based in the United States and Canada. Quidditch[6] is a team sport inspired by the fictional game Quidditch in the Harry Potter books. |

[3] `https://en.wikipedia.org/w/index.php?curid=15075`

[4] `https://en.wikipedia.org/w/index.php?curid=855186`

[5] `https://en.wikipedia.org/w/index.php?curid=48774528`

[6] `https://en.wikipedia.org/wiki/Quidditch_(real-life_sport)`

Table 3: Examples of articles that the author subjectively judged whether they could be added to the unusual articles section