Exploring Cross-Language Differences in Wikidata-based Hyperlink Types for Enhanced Editorial Support on Wikipedia

Nhu Nguyen
nhunt@nii.ac.jp
Graduate University for Advanced Studies, SOKENDAI,
National Institute of Informatics
Tokyo, Japan

Hideaki Takeda
National Institute of Informatics
Tokyo, Japan
takeda@nii.ac.jp

ABSTRACT

Wikipedia has become one of the most widely used language resources in more than 330 languages, attracting contributions from editors around the world. However, a considerable gap still exists among language editions, encompassing variations in article count, subject coverage, and even the number of community editors. To bridge this gap, research efforts have sought to capitalize on diverse factors such as cross-language integration and hyperlink utilization. Despite of these efforts, the full potential of hyperlinks remains unexplored. Therefore, in this study, we introduced a novel approach to explore the hyperlinks by focusing on hyperlink types based on Wikidata. We aim to extract and analyze the patterns associated with these hyperlink types across different languages, then use them as the recommendation solution to enhance the editorial activities of editors. In our initial collaborative filtering experiment, we observed improved performance when combining multiple languages, rather than using a single language.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; • **Computing methodologies → Cross-validation**.

KEYWORDS

Wikipedia, Wikidata, Hyperlink, Recommendation, Collaborative Filtering

ACM Reference Format:
Nhu Nguyen and Hideaki Takeda. 2018. Exploring Cross-Language Differences in Wikidata-based Hyperlink Types for Enhanced Editorial Support on Wikipedia. In *Proceedings of The 12th International Joint Conference on Knowledge Graphs (IJCKG '23).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

In recent years, Wikipedia has become a big source of information in numerous languages thanks to the local communities of editors [15]. Wikipedia currently stores more than 20M topics in multi-domain. Notably, the English edition stands as a significant contributor, covering about 6M topics, constituting a noteworthy one-third proportion[21]. Furthermore, the top 15 languages (*https://meta.wikimedia.org/wiki/List_of_Wikipedias*) in terms of article count have collectively accounted for nearly half of Wikipedia's total articles [1]. It becomes evident that a significant disparity exists among language editions, with smaller language editions lacking a considerable number of topics and receiving comparatively less attention. In addition, this disparity also extends to articles with the same topics across different languages. The primary source of these differences largely stems from Wikipedia's open editing policy and the community of editors. Researchers have tried to bridge this gap by leveraging cross-language links and hyperlinks among languages, aiming to support weaker languages. Nevertheless, many aspects have not been thoroughly examined and explored [1].

Therefore, we propose the novel exploration of *Wikidata-based hyperlink type* in supporting languages and editors. A hyperlink type is a classification or category assigned to a hyperlink and determined through the "`P31: instance of`" relation in Wikidata. Each Wikipedia article is associated with a unique identifier in Wikidata. These identifiers reinforce Wikipedia's collaborative nature by connecting diverse language editions of the same article. This allows for easy exploration of topics and concepts on a global scale [21]. Our objective is to investigate the disparity in hyperlink types among cross-language articles. Additionally, we aim to leverage this information for the recommendation system and assess the effectiveness of our approach in various language aggregations. This enables us to offer relevant recommendations to support editors in creating and updating Wikipedia articles, particularly for small communities.

In the experimental phase, we employed collaborative filtering for our recommendation system on the acquired data from three languages: English, Japanese, and Vietnamese. Although both Japanese and Vietnamese are ranked within the top 15 Wikipedia editions by volume, they still lag significantly behind English. Furthermore, the communities of editors in these two languages are notably smaller when contrasted with the English community (refer to Fig. 1). Thus, using these languages can also be sufficiently representative to assess the effectiveness of our approach. The experimental results demonstrate that the collaborative filtering models perform better, and a greater number of recommendations are obtained on multilingual data.

The remainder of this paper is organized as follows. After reviewing the related works (Section 2) and introducing the problem statement (Section 3), we describe the statistical analysis of data in Section 4 and hyperlink type recommendation in section 5. The

IJCKG '23, December 08–09, 2023, Tokyo, Japan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/XXXXXXX.XXXXXXX

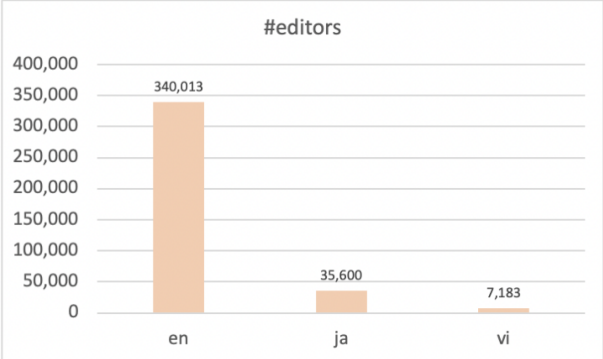

Figure 1: The number of editors in local communities. (*https://stats.wikimedia.org/#/all-projects)*

experiments are presented in Section 6, followed by concluding statements in Section 7.

2 RELATED WORKS

Previous research has identified the information gap among different Wikipedia editions and attempted to bridge it [10], [4],[5],[12],[9]. Hyperlinks and cross-language links are essential data elements leveraged in this task. [8] used hyperlinks for recommending related articles in Wikipedia. Jiawen et al used hyperlinks as bridge model the relationship between two documents [7] . [3] identifies the information gap between analyzing linked entities in a cross-lingual knowledge graph. While many approaches exist to finding missing cross-language links. Jong et al [17] proposes a language-independent method for discovering missing cross-language links between English and Japanese. .[19] also leverages cross-lingual analysis to study topical biases across various language editions of Wikipedia. However, hyperlinks and other relationships within Wikipedia represent an exceptional resource that has yet to be fully comprehended[20]. When it comes to language biases and disparities among different versions of Wikipedia, there are two main categories of studies: the editing behaviors of editors and readers' behaviors[9]. [2] explores navigation patterns by using hyperlinks to understand readers and address structural biases and knowledge gaps. Another research project delivers personalized recommendations to editors by acquiring expertise in effectively representing Wikipedia articles and offering relevant content suggestions[18].

Recommendation systems aimed at reducing the information gap within Wikipedia have been implemented using numerous methods, collaborative filtering is a common approach used [22], [6], [14].

In this work, we present the disparity of different languages by identifying the distribution of hyperlinks and types. Then, using these types for collaborative filtering to suggest useful topics to editors. As far as we know, our work is the first study that focuses on Wikidata-based types of hyperlinks in multiple language editions of Wikipedia. We aim to emphasize the impact of multilingualism on the recommendation outcomes.

Table 1: The number of pages in three languages

Language	#Wikipedia pages
en	6,438,267
ja	1,297,653
vi	1,268,004

3 PROBLEM STATEMENT

Each Wikipedia article in each language has a distinct title (e.g., "Tokyo"). Its content contains hyperlinks (or intra-language links) directing readers to other correlated articles within the same language (e.g., "Japan","Edo"). Although hyperlinks can connect resources across languages, websites, and various media, we only focus on links within the same language version of a resource. We also consistently refer to them using one unique term, "hyperlink" in this paper. Additionally, we introduce the concept of hyperlink types. These types are extracted based on the relation in Wikidata, named "P31: instance of". For example, the article "Tokyo" in the English version has a hyperlink to the article "Japan". Wikidata Qid of "Japan" item is $Q17$ and it has "instance of" values such as sovereign state and country. These values serve as types for the hyperlink "Japan". This "P31: instance of" relation is used in knowledge graphs related to Wikidata to assign one or several type(s) to an entity. In this research, we use it to retrieve Wikidata-based type of hyperlinks. We, as pioneers, consider this type as distinct concept, deploying it to evaluate language-based disparities to gain insight into the behaviors and preferences of local editors. Consequently, we explore its role in recommendation solution for supporting editors and bridging language gaps.

To do this, we leverage the cross-language link of Wikipedia articles in the three languages mentioned. These articles in different languages on the same topic are linked through cross-language links (or interlanguage links). For example, the article "Tokyo" in the English edition has a cross-language link to the article "東京都" in Japanese and "Tokyo" in Vietnamese. Thanks to Wikidata, these three articles share the same unique identifier in Wikidata ($Q1490$). As a result, this feature provides a straightforward way to retrieve information about an item in various language editions.

4 STATISTICAL ANALYSIS

4.1 Data Extraction

Data collection scope. As previously mentioned, our research is conducted on a case study involving the English (en), Japanese (ja), and Vietnamese (vi) languages. Table 1 presents the number of articles for these languages, sourced from data dumps of January 20, 2022, published by the Wikimedia Foundation. Based on this data, we only retrieved the articles that have cross-language links in all three languages.

Figure 2 provides an overview of creating the data in our study, including Processing and Data Extraction. The process begins by retrieving articles from a Wikipedia dump, followed by their processing using two essential tools: the Dump Parser[11] and Data Indexing[13]. To elaborate, this process is responsible for handling specific files from Wikipedia dumps. These dumps encompass

copies of all pages from three Wikipedia wikis and redirect files, all of which are accessible at *https://dumps.wikimedia.org*. For each article, critical details such as the title and ID information are meticulously extracted. After retrieving and parsing, we extract the relevant hyperlinks and hyperlink types based on Wikidata-lite [16]. As a result of this step, we obtain a collection of JSON files, each housing articles in three different languages, along with their corresponding hyperlinks and hyperlink types. Consequently, we found 130,000 shared Wikidata identifiers between the three languages. That is, we obtained 130,000 articles from each language to serve as a primary data source for our research.

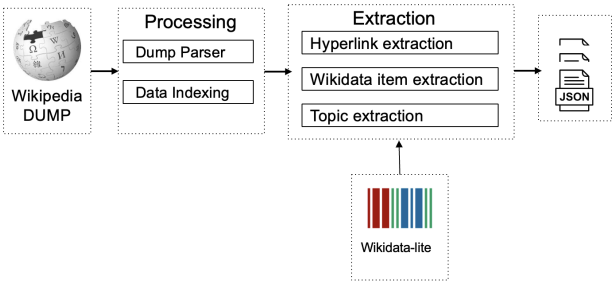

Figure 2: Pipeline for extracting hyperlinks and types in three languages

4.2 Statistical Analysis

In this section, we analyze the collected data to extract valuable insights. With the specified condition, there are 130,000 articles extracted in each language. We retrieved information from the article, such as Wikidata identifier, title, hyperlinks, hyperlink types, and their occurrences.

Table 2 presents a comparative analysis of the number of hyperlinks and hyperlink types across three languages on the same number of articles for each language. English stands out with a significant count of 1,827,775 hyperlinks, nearly 2.5 times that of Japanese and almost seven times that of Vietnamese. Correspondingly, English also demonstrates a notably higher count of hyperlink types at 35,004, whereas Japanese and Vietnamese exhibit comparatively lower counts of 21,690 and 12,376 hyperlink types, respectively. This distribution is consistent with the prominence of English, which holds the highest number of articles. Simultaneously, it also distinctly reflects the differences in content and the topics covered within the articles.

Additionally, it's important to note that various hyperlink types exhibit distinct frequencies in articles, and these trends can vary significantly across different languages. By analyzing the occurrences of these hyperlink types in our dataset, we identify unique patterns for each language. Specifically, we calculate the occurrence rate of each hyperlink type in the articles and filter out types with more than 50%.

In Figure 3, we provide a visual representation of these trends, highlighting the diverse array of subjects within each linguistic community. For instance, we observe a notable preference for hyperlinks related to topics: film, automobile model, album, single,

Table 2: Cross-Linguistic Analysis of Hyperlinks and Hyperlink Types

	en	ja	vi
#Hyperlinks	1,827,775	739,784	266,596
#Hyperlink Types	35,004	21,690	12,376

and city in English edition. Conversely, in Japanese articles, hyperlinks concerning various facets of Japan, including cities, towns, municipalities, and railway stations, along with subjects like film, earthquake, and manga series, capture considerable attention. Meanwhile, within the Vietnamese context, local interest predominantly centers around hyperlinks linked to essential medicine, sovereign state, country, and calendar year. These observations underscore the rich diversity of topics and interests within each language's community and provide valuable insights into content preferences and tendencies of users in different linguistic contexts.

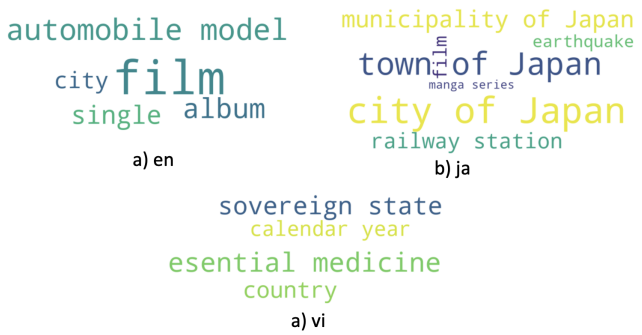

Figure 3: Local interest of hyperlink types in different languages

5 HYPERLINK TYPE RECOMMENDATION

In this study, we explore hyperlink types and their occurrences as a dataset for recommendation solutions to support editors with additional information. These types assist editors in taking preliminary topics to adjust content according to the recommendations.

We construct a user-item matrix, treating articles as users and hyperlink types as items. The occurrences of hyperlink types become the ratings within this matrix. This constructed matrix forms the basis for predicting missing items using collaborative filtering (CF). This approach constructs a model using the historical behavior of users. Although the occurrences of the hyperlink types in the article are based on real-world events, it differs from explicit ratings that are typically used to express preferences. However, these numerical values are definitely useful as they provide confidence in a specific observation. Section 4 illustrates the differences in the occurrences of hyperlink types across languages, leading us to believe that these values may indicate the preference in local editor communities. Our aim is to predict occurrence (rating r_{ui}) of hyperlink type *(item - i)* for an article *(user - u*. To do this, we employ CF methods such as matrix factorization and neighborhood-based methods.

Neighborhoods in CF can be categorized into two kinds: user-based approach and item-based approach. Over time, the latter become more popular. In item-based approach, a rating is predicted by leveraging known ratings provided by the same user for comparable items. Thus, a similarity measure between items is calculated. We employ the Mean Squared Difference (MSD) to identify the k items rated by user u that exhibit the highest similarity to item i. The predicted rating of r_{ui} is a weighted average of the ratings for neighboring items:

$$\hat{r}_{ui} = \frac{\sum_{j \in N(u)} \text{sim}(i,j) r_{ui}}{\sum_{j \in N(u)} \text{sim}(i,j)} \quad (1)$$

Whereas, $\hat{r}_{ui}$ is the predicted rating for user u on item i. $N(u)$ presents the set of items that user u has rated. $\text{sim}(i,j)$ is the similarity between items i and j based on the MSD similarity measure.

In the matrix factorization approach, Singular Value Decomposition (SVD) is one of the most common and successful techniques in collaborative filtering. It captures latent patterns within user-item interactions. In a recommendation system, the matrix representing user-item interactions is often sparse because not all users interact with all items. SVD is applied to decompose this matrix into three matrices: U (user matrix),Σ (diagonal singular value matrix), and $V^\intercal$(item matrix transpose). Singular value decomposition assumes a matrix M (for example, a matrix mxn) is decomposed as:

$$\text{M} = U\Sigma V^\intercal \quad (2)$$

SVD reduces the dimensions of the utility matrix A by extracting its latent factors, thereby mapping each user and item into an $r-$dimensional latent space. This mapping provides a clear representation of the relationships between users and items. That means the dot product of these two vectors gives you the predicted rating for the user-item pairs.

Data sparsity is a common challenge in recommendation systems. To overcome this, we selectively extracted data with cross-language links, which will be presented in the following Section 6. Through collaborative filtering models, we aim to uncover hidden connections between articles and hyperlink types so that missing hyperlink types can be recommended.

6 EXPERIMENT

We conducted experiments on collected data in three languages: English, Japanese, and Vietnamese.

6.1 Experimental Setting

Input Data.

From the data obtained in Section 4.1, we have 130,000 articles in each language that satisfied the criterion of having cross-lingual connections. This maybe improve the recognition rate of patterns when more languages are integrated. Then, we filter the articles with the number of hyperlink type exceeding 200, which served as a sample data for our models. This ensures that the chosen articles have high-quality content in all three languages, serving the dual purpose of maintaining experiment fairness and observing the results of language integration. After filtering, we retained approximately 1,000 articles in each language that satisfied the specified conditions. Consequently, our input data comprises 3,000 articles across three languages and encompasses more than 18,000 hyperlink types, resulting in a total of over 533,000 interactions between the acquired articles and hyperlink types.

Table 3: Language Interaction Statistics

Language	en	ja	vi
Number of Ratings (Occurrences)	210,384	155,004	111,807
Number of Users (Articles)	1,000	1,000	1,000
Number of Items (Hyperlink types)	16,019	11,116	7,197
Average number of Ratings per User	210.384	155.004	111.807
Average number of Ratings per Item	16.019	11.116	7.197

Because nearly 90% of the data falls within the rating range of 1 to 10. Furthermore, hyperlink types that have the high occurrences (ratings) are often types such as *taxon* (*Q*16521) and *Wikimedia list article* (*Q*13406463). They are abstract types or unrelated to topics. Therefore, for the convenience of evaluating CF techniques, we remove them and only assess the dataset with ratings ranging from 1 to 10. As mentioned above, the ratings in this context are essentially based on real events. In this experiment, our goal is to suggest whether types should appear, high occurrence is not necessarily the most important factor. Figure 4 shows the distributions of ratings by groups. Rating of 3 has the highest rating count among ratings 1-10. Table 3 shows the statistic of the dataset used for CF methods.

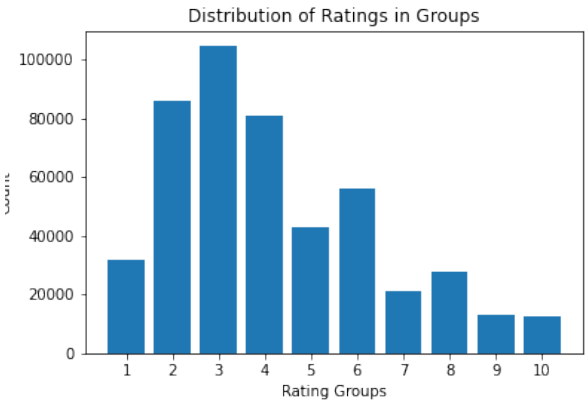

Figure 4: Distribution of Articles by Rating Group.

We use cross-validation technique to assess the performance of predictive models. The most common form of cross-validation is k-fold cross-validation, where the dataset is divided into k subsets or "`folds`". The model is trained on k-1 of these folds and tested on the remaining fold. The performance metrics used are the official metrics of collaborative filtering: MAE and RMSE.

Mean Absolute Error (MAE) is a metric for calculating the average of all absolute differences between the algorithm's predictions

and the actual ratings. A lower MAE indicates higher accuracy.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{3}$$

Where, y_i is the actual rating, $\hat{y}_i$ is the predicted rating and n is the amount of ratings.

Root Mean Squared Error (RMSE) computes the mean value of all squared differences between the true and predicted ratings. Subsequently, it calculates the square root of the result. RMSE is most useful when large errors are particularly unwanted.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

We performed 5-fold cross-validation by dividing the dataset into five equal parts. In each fold, 80% of the data were used for training, and the remaining 20% were used for testing. In addition, we also compare performance of CF approaches with baseline methods that predict a random rating based on the distribution of the training set.

The data used in this paper is publicly accessible, and our code has been made available to facilitate the replication of our experiments. [1]

6.2 Experimental Results

We first report the performance of collaborative filtering algorithms on all datasets, and then investigate how multilingual datasets improve the recommendation performance.

Specifically, we conducted experiments on three different language configurations as follows:

- Monolingual dataset
- Bilingual dataset
- Trilingual dataset

Table 4 shows MAE and RMSE scores of models on different datasets. The lower these values, the higher the performance of the models. SVD represents for matrix factorization approach. K-Nearest Neighbor (KNN) approach have KNNBaseline, KNNWithMeans, KNNWithZscore and KNNBasic models. NormalPredictor model represents Baseline method. Clearly, CF approaches have better performance than Baseline method. The reason is that NormalPredictor model only uses maximum likelihood estimation for prediction. CF approaches estimate another characters of items or relationship of users and items. Considering CF approaches, SVD and KNNBaseline model have competitive results in monolingual dataset. But, KNN models are better than SVD in multiple language. In detail, the lowest MAE and RMSE of KNNBasic **0.956** and **1.503**, respectively, are attained on the trilingual configuration.

Figure 5 displays the lowest MAE and RMSE values achieved by the KNN models. It is evident that these values are inversely proportional to the number of aggregated languages, indicating a positive impact of multiple languages on the experimental results.

After hyperparameter tuning, we generate recommendations on the datasets. Figure 6 illustrates the relative increase or decrease in the number of recommendations when multiple languages are used. For instance, the number of recommendations for Japanese

[1] https://github.com/nhunthp/Hlink_RS#readme

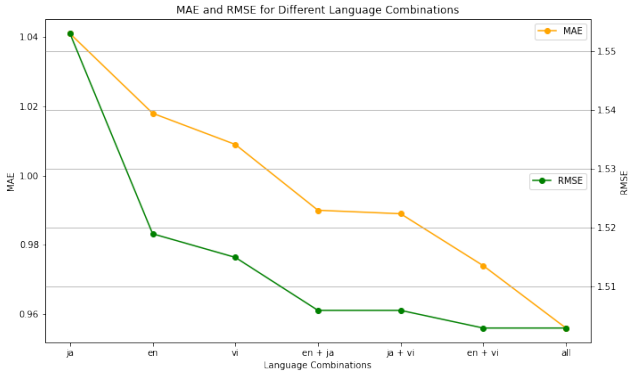

Figure 5: Comparison of the lowest MAE and RMSE values among different models on datasets.

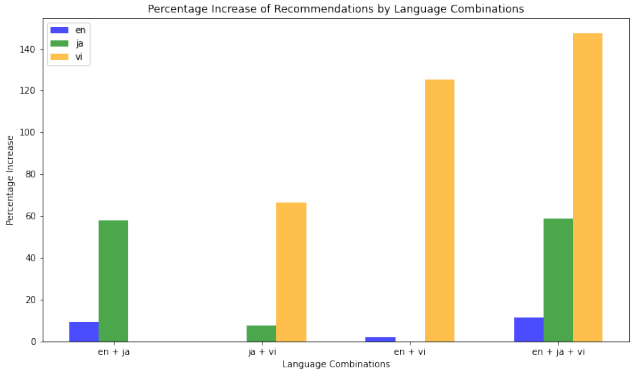

Figure 6: Multilingual recommendation impact chart

articles increases more 50% when combined with other languages (primarily due to English). Vietnamese articles experience a more significant increase, with recommendations reaching around 150% when combined with all three languages. These results are not surprising, as initial statistics indicated that Vietnamese articles had the lowest number of hyperlink types. However, they suggest that low-source languages can benefit in recommendation systems when combined with stronger languages. This enrichment of recommendation content offers more choices and supplements missing content, ultimately narrowing the gap between languages within Wikipedia.

7 CONCLUSION

We explore hyperlink types, revealing patterns across various languages. Furthermore, we introduced a pipeline for retrieving these types and performed a statistical analysis of their occurrences across three languages. The statistical findings reveal notable distinctions within the collected data, highlighting the potential for utilizing this information in recommendations to support editors. We evaluate across multiple experiments involving different language combinations. The experimental results show that multilingual aggregating

Table 4: MAE, RMSE values of models on the datasets

Model	ja		vi		en		en+vi		en+ja		ja+vi		all (en+ja+vi)	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
SVD	1.047	1.553	1.023	1.513	1.035	1.557	1.020	1.535	1.035	1.558	1.020	1.526	1.024	1.542
KNNBaseline	**1.041**	**1.539**	**1.009**	**1.506**	**1.018**	**1.519**	0.99	**1.503**	1.006	**1.515**	0.999	**1.506**	0.987	**1.503**
KNNWithMeans	1.198	1.683	1.161	1.636	1.183	1.672	1.123	1.620	1.132	1.630	1.121	1.611	1.100	1.602
KNNWithZScore	1.244	1.733	1.210	1.688	1.229	1.721	1.160	1.661	1.167	1.671	1.160	1.653	1.132	1.638
KNNBasic	1.065	1.704	1.026	1.653	1.035	1.671	**0.974**	1.638	**0.99**	1.66	**0.989**	1.656	**0.956**	1.641
NormalPredictor	2.47	3.076	2.440	2.450	2.451	3.060	2.465	3.074	2.466	3.075	3.058	3.049	2.459	3.068

allows us to overcome the limitations of data sparsity in individual languages.

By leveraging the characteristics of cross-language links from various editions of Wikipedia, we enrich the dataset. Consequently, the recommendation system becomes more robust in collaborative filtering models. These results hold promise for supporting low-resource languages and bridging the gap in Wikipedia.

In the future, we are considering exploring alternative approaches for the recommendation system, including examining graph-based data organization or incorporating additional text attributes.

REFERENCES

[1] Eneko Agirre, Ander Barrena, and Aitor Soroa. 2015. Studying the wikipedia hyperlink graph for relatedness and disambiguation. *CoRR*, abs/1503.01655. http://dblp.uni-trier.de/db/journals/corr/corr1503.html#AgirreBS15.

[2] Akhil Arora and et al. 2022. Wikipedia reader navigation: when synthetic data is enough. (Jan. 2022).

[3] Vahid Ashrafimoghari. 2023. Detecting cross-lingual information gaps in wikipedia. In *WWW '23 Companion: Companion Proceedings of the ACM Web Conference 2023*. ACM, 581–585.

[4] Armand Boschin and Thomas Bonald. 2021. Enriching wikidata with semantified wikipedia hyperlinks. In *CEUR Workshop Proceedings*. Wikidata@ISWC 2021.

[5] Iacer Calixto, Alessandro Raganato, and Tommaso Pasini. 2021. Wikipedia entities as rendezvous across languages: grounding multilingual language models by predicting Wikipedia hyperlinks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Online, (June 2021), 3651–3661. DOI: 10.18653/v1/2021.naacl-main.286.

[6] Fernández-Tobías et al. 2019. Addressing the user cold start with cross-domain collaborative filtering: exploiting item metadata in matrix factorization. *User Model User-Adap Inter*, 29, (Apr. 2019).

[7] Jiawen Wu et al. 2022. Pre-training for information retrieval: are hyperlinks fully explored? (Sept. 2022).

[8] Malte Schwarzer et al. 2016. Evaluating link-based recommendations for wikipedia. In *JCDL '16: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries*. ACM, 191–200.

[9] Volodymyr Miz et al. 2020. What is trending on wikipedia? capturing trends and language biases across wikipedia editions. In *WWW '20: Companion Proceedings of the Web Conference 2020*. ACM, 794–801.

[10] Laxmi Amulya Gundala and Francesca Spezzano. 2018. Readers' demanded hyperlink prediction in wikipedia. In *Proceedings of The Web Conference 2018*.

[11] Jeff Heaton. 2021. Dump parser. Retrieved 2022 from https://github.com/jeffheaton/present/tree/master/youtube/wikipedia.

[12] Seungho Kim Juram Kim and Changyong Lee. 2019. Anticipating technological convergence: link prediction using wikipedia hyperlinks. *Technovation*, 79, (Jan. 2019), 25–34.

[13] Jan-Christoph Klie. 2019. Dump parser. Retrieved 2023 from https://github.com/jcklie/wikimapper.

[14] Kanako Komiya and et al. 2014. Cross-lingual product recommendation system using collaborative filtering. In *CICLing 2014*, 141–152.

[15] Andrew Lih. 2009. *The Wikipedia revolution: How a bunch of nobodies created the world's greatest encyclopedia*. Hyperion (March 17, 2009).

[16] Phuc Nguyen and Hideaki Takeda. 2022. Wikidata-lite for knowledge extraction and exploration. In *IEEE International Conference on Big Data (Big Data)*. IEEE, 3684–3686.

[17] Jong-Hoon Oh and et al. 2008. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*. IEEE.

[18] Diego Saez-Trumper Oleksii Moskalenko Denis Parra. 2020. Scalable recommendation of wikipedia articles to editors using representation learning. (Sept. 2020).

[19] Tiziano Piccardi and Robert West. 2021. Crosslingual topic modeling with wikipda. In *WWW '21: Proceedings of the Web Conference 2021*. ACM, 3032–3041.

[20] Takashi Tsunakawa, Makoto Araya, and Hiroyuki Kaji. 2014. Enriching Wikipedia's intra-language links by their cross-language transfer. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. Dublin City University and Association for Computational Linguistics, Dublin, Ireland, (Aug. 2014), 1260–1268. https://aclanthology.org/C14-1119.

[21] Denny Vrandeic and Markus Krotzsch. 2014. Wikidata: a free collaborative knowledge base. *ACM Commun.*, 75, 2, (Apr. 2014), 78–85.

[22] Tao Zhou Yan-Li Lee. 2021. Collaborative filtering approach to link prediction. *Physica A: Statistical Mechanics and its Applications*, 578, (Sept. 2021).