DEIM2021 F33-2

Product Recommendation based on Visual Search and Customer Satisfaction Analysis of Images and Reviews

Houji ZHONG† and Yuanyuan WANG†

† Graduate School of Sciences and Technology for Innovation, Yamaguchi University
2–16–1 Tokiwadai, Ube, Yamaguchi, 755–8611 Japan
E-mail: †{a087vgu, y.wang}@yamaguchi-u.ac.jp

Abstract In today's era, with the popularization of the Internet, e-commerce has attracted extensive attention. However, users' dissatisfaction with traditional product recommendation systems has gradually increased. With increased user needs, product recommendations based on image search are becoming more and more popular. In this paper, we propose a method that combines two similarities: one is the similarity of the product image and the user input image with the product description, and the other one is the similarity of the surrounding text of the user input image based on the image recognition. Furthermore, we perform a sentiment analysis on product reviews and combine users' repeated purchasing behaviors to recommend products of high user satisfaction from multiple aspects. Finally, we evaluated and discussed our proposed product recommendation method using real e-commerce product data.

Key words recommendation system, visual search, image analysis, review analysis, e-commerce

1 Introduction

Due to the popularization of the Internet, the explosive growth of online information available to people, and the consumer purchase behaviors on e-commerce websites have also increased considerably. As shown in Figure 1, according to a survey conducted by the global statistics website Statista [1], retail e-commerce sales in 2020 have reached 4,206 billion U.S. dollars and will continue to increase in the future.

Consumers often refer to the product information provided by e-commerce websites when purchasing products. Due to the excessive amount of information, people will spend a great deal of time choosing the appropriate products. As shown in Figure 2, searching and navigation (easy to find products) has become the preferred source of inspiration for online shoppers, with a proportion of 61% [2]. A recommended system is an effective method to resist this kind of consumer over-selection.

On the e-commerce website, users can input keywords about products they want, and the recommendation system can show products related to their input keywords. It means that searching for products requires users to grasp the product information in advance, which is not easy for indescribable products or consumers who are unfamiliar with recommendation systems. It also requires a great effort for consumers, which are not that proficient in online shopping. For example, given you see a piece of clothing or backpack

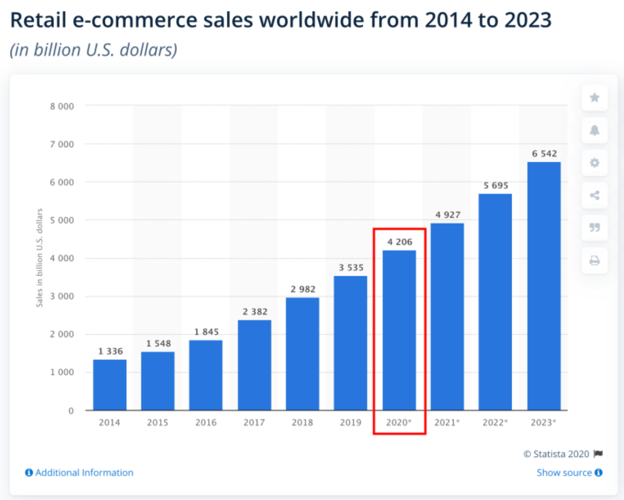

Figure 1 Retail e-commerce sales worldwide from 2014 to 2023.

you like, you may be able to specify its category, color, and other keywords. It is a challenge to describe the style of it that the results recommended by the system will be different from what is expected by the user as well. In the case of a specified product, the user wants to buy it, but he/she does not know the product name, which will also make the product search more difficult for the user.

The image-based product search has become a new method to address the needs of users in recent years. As part of previous works, there have been many relatively mature algorithms for image recognition, and they have also been

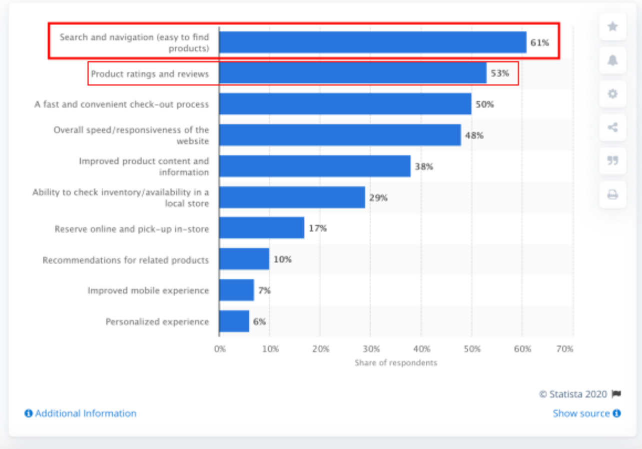

Figure 2 Preferred sources of inspiration for online shoppers worldwide as of March 2020.

used in real-world recommended systems. At present, the visual search recommendation system has been practically used based on the similarity between the user input image and the shape, color, and other elements of the product, and part of it combined with the sales and rating of the product for comprehensive analysis; however, the effect of recommendation is not yet very satisfying. For example, there are multiple colors for the same product sold by the seller, but the product image for display mainly uses only one of the colors. However, if the user input image does not match the color of the product's display image, then the user may have missed the product. Furthermore, the quality of recommended products cannot be guaranteed when searching for products based on the similarity of images or scores that do not necessarily match the real ones. As shown in Figure 2, the second source of inspiration for online shoppers is Product ratings and reviews, which accounted for 53% [2].

Therefore, in this work, to measure user satisfaction with each product, we use not only the similarity of the user input image and the image of each product but also combine with the sentiment analysis of the reviews and the repeat purchase rate of each product. Our goal is to recommend high-quality products with high satisfaction to users based on satisfying high similarity of products to consumers through image-based product search.

The remainder of this paper is structured as follows. The next section reviews some prior work on image recognition and the field of sentiment analysis. Section 3 explains the problem definition and discusses our proposed product recommendation method. Afterward, Section 4 introduces the data-set and discusses the experimental results of our proposed method.Finally, we conclude the paper in Section 5.

2 Related Work

2.1 Image Analysis

Within the field of image recognition, there are already many mature algorithms. Zauner et al. [3] proposed a benchmark framework called Rihamark, which is used for the perception of the image-based hash functions. The perceptual image hash function generates a hash value based on the visual appearance of the image. And it then uses a sufficient distance or similarity function to compare the two perceptual hash values to determine whether the two images are perceptually different. Zagoruyko et al. [4] showed how to learn a general similarity function for comparing image patches directly from the image data through a convolutional neural network.

In addition to hashing-based and convolutional neural network-based similarity matching algorithms, local feature detection algorithms have a great significance in similarity calculations, image retrieval, and object recognition. Compared with the pixel-level global features, the local features are more flexible in describing the image features. Among them, SIFT (Scale-invariant Feature Transform) is more commonly used, and it has the advantage of scale invariance. Even if the rotation angle or the shooting angle is changed, it may still be possible to get better detection. André Araujo et al. [5] introduced a new retrieval architecture, which can directly compare the queries to images of database data. To compare different image recognition algorithms, they proposed an asymmetric comparison technique for Fisher vectors and systematically explored queries or database items with different amounts and types of clutters. Lastly, it found that similarity computation based on local invariant features has an adversarial capability, its accuracy is better than the traditional perceptual hash algorithm, and support rotation invariance is better than a convolutional neural network.

2.2 Sentiment Analysis

In sentiment analysis, many tasks required for semantic labeling of phrases and texts rely on word lists with some semantic features, Andreevskaia et al. [6] proposed a method of extracting emotional adjectives from Word-Net using the Sentiment Tag Extraction Program (STEP). Hassan et al. [7] proposed a method that uses features such as lexical items, part-of-speech tags, and dependency relations to consider overall sentence structure at different levels or generalizations to predict whether a sentence exhibits attitude towards the recipient of a text. Yohan et al. [8] proposed Sentence-LDA (SLDA), a probabilistic generative model that assumes that all the words in a single sentence are generated from a single aspect. Additionally, they extend SLDA to a model called the Aspect and Sentiment Unification Model (ASUM), which incorporates both aspects and sentiment together to model the sentiments towards different aspects.

The Google AI Language in 2018 released a new language representation model, called BERT, which represents Bidirectional Encoder Representations from Transformers (BERT). The BERT model was designed to pre-train deep bidirectional representations from the unlabeled text by jointly conditioning on both left and right context in all layers and set a new entry in the performance of 11 NLP tasks [9]. As an infrastructure of these tasks, the BERT can be used for sentiment analysis, question answering system, spam filtering, named entity recognition, document clustering, and other tasks.

3 Product Recommendation

In this section, we explain our proposed product recommendation method by analyzing image recognition and reviews. It is possible to recommend products with high satisfaction related to the user input image through the image-based product search on e-commerce websites.

Therefore, from the perspective of image-based product search, we use a similarity computation based on local invariant features to analyze the similarity between the user input image and product image. Furthermore, we perform the label detection on the user input image, and we compute the similarity between the image labels and the text of the product description. It can comprehensively evaluate the results of the image-based product search. On the aspect of review analysis, we perform sentiment analysis on product user reviews, and we combine repeated purchase behaviors of users of the same product. It can comprehensively evaluate product satisfaction. Moreover, we further elaborate on our method of these two kinds of features.

3.1 Image Analysis of Products

In this section, we explain the image analysis method to compute the image scores of the products. Conventionally, the user input image does not contain only the product itself, which is the one that the user wants. For example, users want to purchase cups of the same style through the image-based product search, but often the background contains other irrelevant items, such as pens and notebooks on the table. Therefore, for analyzing an image, we need to extract the main object of the image and then remove the irrelevant background, to avoid affecting the image similarity calculation to the target product.

Then, we extract SIFT features from the user input image and each product image and match them with the SIFT features in advance to evaluate the similarity between these images as IS. The SIFT (Scale-invariant Feature Transform) is a machine vision algorithm published by David in 1999 [10]. In this way, it is a local feature used to describe the image and has a better scale and invariance to the rotation.

For the image-based search of the product recommendation system, when the product image is a yellow backpack, the same product in blue is sold; the product image is a round-neck shirt, but the actual sale also has the same shirt with V-neck. However, the main product images released by the seller cannot completely cover all the products. Only the image similarity cannot meet the user's needs of the user. Having obtained the similarity between images, we use the Google Cloud Vision API(注1) to extract the user input image labels, and we use the Word2Vec model to add synonyms for all image labels. Tomas et al. [11] proposed the Word2Vec model to detect the similarity between words, which can compute continuous vector representations of words from large-scale datasets. Hence, to obtain the similarity between the image and its surrounding text, we calculate the cosine similarity between two vectors of all image labels and the product description.

Hence, we define the word vector of the image labels as IL, the word vector of the product description as PDK, and compute the similarity between the image labels of the user input image and each product description as LDS by using the cosine similarity between IL and PDK as follows:

$$LDS = \frac{\sum_{i=1}^{n} IL_i \times PDK_i}{\sqrt{\sum_{i=1}^{n} IL_i^2} \times \sqrt{\sum_{i=1}^{n} PDK_i^2}} \tag{1}$$

Finally, we compute the image score of each product (S) by combining the similarity of the user input image and the image of each product (IS), and the similarity of the labels of the user input image and the description of each product (LDS), using the following formula.

$$S = \alpha \times IS + (1 - \alpha) \times LDS \tag{2}$$

Here, the value of S is between 0 and 1. In this paper, we set α to be 0.7 based on our preliminary experiment.

3.2 Review Analysis of Products

In this section, we explain the review analysis method to compute the review scores of the products. The existing product recommendation system generally provides the product ranking based on the user rating with a 5-point rating scale. However, in this case, it is easy to find that, although the user satisfaction with the product is not high, since there is no obvious shortcoming or because of user habits, high scores of 4 or 5, causing the product to score too high and lose referential relevance. In this work, we applied the Google Natural Language API(注2) to perform the sentiment analysis on the user's reviews based on the BERT

(注1)：https://cloud.google.com/vision
(注2)：https://cloud.google.com/natural-language

model. Moreover, we increase the proportion of user rating scores and the sentiment analysis on the user reviews based on the repeated purchase behavior.

Hence, we define the positive score of sentiment analysis for each review as PS, the weight of repeated purchase behavior as RPW, and the total number of product reviews as NoR. Then, the review score of each product (RS) is calculated as follows:

$$RS = \frac{\sum_{i=1}^{NoR}(PS_i \times RPW_i)}{NoR} \quad (3)$$

Here, we normalize the review score RS between 0 and 1 applying Min-Max.

$$\frac{RS - RS_{min}}{RS_{max} - RS_{min}} \quad (4)$$

3.3 Ranking Method based Image Score and Review Score

According to the above, we obtain the image score of the user input image and each product image, and we get the normalized review score of each product. In case the sum of two scores is directly used for the product recommendation when the product review value is too high, it may lead to a situation where the recommended product is different from the user input image. Therefore, we divide our product recommendation method into two stages: (1) confirm the number of product recommendations, and (2) select the similar products with the highest similarity score based on twice that number. When the number of recommended products is 10, we choose the top 20 most similar products based on the given similarity scores. From these top 20 products, we recommend the products with the high image scores to the user by combining them with the products' review scores.

Then, we compute the overall score of each product (OS) by combining the image score (S) with the review score (RS) of each product as follows:

$$OS = \beta \times S + (1 - \beta) \times RS \quad (5)$$

The value of OS is between 0 and 2. In this paper, we set β to be 0.5 based on the results of our preliminary experiment.

4 Evaluation

In this section, we verify the feasibility of our proposed product recommendation method with three steps: the data collection and processing, the experimental design and results, and discussion.

4.1 Data collection and Processing

In our evaluation, we used the product data and the review data of Rakuten Ichiba data from "Rakuten Dataset" provided by Rakuten, Inc. via IDR Dataset Service of National Institute of Informatics (NII) [12]. It includes about 156 million items of total product data and about 64 million review data from August 4, 2010, to April 1, 2014. For this work, we chose products that contain three types of metadata: the product image, the product description, and the user reviews. In this paper, we extracted the product data for 53 backpacks and 51 hats to conduct the experiment, respectively, if the product data has three types of data and the number of user reviews is more than 10.

Figure 3 User-inputted image for bag.

Figure 4 User-inputted image for hat.

4.2 Experimental Design

For product data, we use backpack and hat data, which satisfy the conditions extracted from Rakuten Ichiba data, to compute image scores and review scores of products by our proposed image and review analysis methods, and then perform four types of rankings based on image scores, review scores, and their overall scores, respectively. The bag and hat images inputted by the user in the experiment are shown in Figure 3 and Figure 4, respectively.

4.2.1 Ranking by Image Scores

Firstly, we extract the SIFT features of the product images obtained in the Rakuten Ichiba data and store them in a database. Secondly, we remove the background of the product images, match the SIFT features of the product images in the database, and calculate the similarity between the user input image and the product image. Thirdly, we use the Google Cloud Vision API to extract the labels for the user input image and use the Word2Vec model to add synonyms for all image labels. Then, we compute the similarity between the user input image and each product image by using the cosine similarity between two vectors of all image labels and the description text of each product. Taking the weights of these two similarity scores and adding them, we obtain the image scores between the user input image and each product image.

Figure 5 Four types of the rankings: 1) Ranking by Image Scores, 2) Ranking by Review Scores, 3) Ranking by Unrestricted Overall Scores, and 4) Ranking by Overall Scores based on Top 20 Image Scores for bag (good results in red frames, and bad results in blue frames).

4.2.2 Ranking by Review Scores

The users' reviews of products obtained in Rakuten Ichiba data are classified according to the products and use the Google Natural Language API to perform sentiment analysis on each product's review and detect whether each review contains a "repeat" tag. If the review with a "repeat" tag, we set the weight of it as 3, and if the review does not contains a "repeat" tag, we then set the weight as 1. After that, we normalize the review scores of all products. Then, we rank the products based on the normalized review scores (*NRS*).

4.2.3 Ranking by Unrestricted Overall Scores (Proposed Method)

We calculate overall scores by summing the image scores and normalized review scores of all products, and we provide the ranking by the overall scores without stage splitting.

4.2.4 Ranking by Overall Scores based on Top 20 Image Scores (Proposed Method)

We calculate overall scores by summing the image scores and the normalized review scores of all products based on the top 20 most similar products according to the image scores, and we provide the ranking by the overall scores based on the top 20 image scores.

4.3 Experimental Results

The four types of ranking results are shown in Figure 3 (bag) and Figure 4 (hat). To evaluate the usefulness of four types of rankings, we conducted a questionnaire survey on ten college students who frequently used e-commerce sites to purchase products. Based on their responses, we produced a ground-truth product review rating system. In this experiment, we used *MRR* (Mean Reciprocal Rank), *MAP* (Mean Average Precision) and *nDCG* (Normalized Discounted Cumulative Gain) to compare these four types of rankings.

Figure 6 Four types of the rankings: 1) Ranking by Image Scores, 2) Ranking by Review Scores, 3) Ranking by Unrestricted Overall Scores, and 4) Ranking by Overall Scores based on Top 20 Image Scores for hat (good results in red frames, and bad results in blue frames).

The indicators MRR and MAP are used to evaluate the relevance of the products. Here, MRR evaluates the ranking for the most relevant products, and MAP for the query set is the mean of the average precision scores for each query [13] as follows:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

$$MAP = \frac{1}{|Q|}\sum_{q=1}^{|Q|}AveP(q)$$

where Q is the number of queries.

The DCG (Discounted Cumulative Gain) is an index that measures the ranking quality of a recommendation system and is used to evaluate the effectiveness of a recommendation system, as follows:

$$DCG_p = \sum_{i=1}^{p}\frac{rel_i}{\log_2(i+1)}$$

Since the value of DCG depends on the recommendation system, DCG alone cannot compare different recommendation systems, thus it is necessary to normalize them. To sort the products that satisfy the criteria most according to their relevance, the largest DCG is generated by the location, which is called $IDCG$ [14]. Formally, the normalized DCG ($nDCG$) is defined as follows:

$$nDCG_p = \frac{DCG_p}{IDCG_p}$$

Hence, the evaluation results of the four rankings are listed in Table 1.

Additionally, we asked ten college students to compare the rankings of 3) Ranking by Unrestricted Overall Scores and 4) Ranking by Overall Scores based on Top 20 Image Scores, and select the better one. The ranking selection results are shown in Figure 7.

Table 1 Metrics evaluation results

Ranking	*MRR*	*MAP*	*nDCG@10*
1)	0.529	0.516	0.625
2)	0.556	0.668	0.679
3)	0.561	0.626	0.745
4)	**0.594**	**0.757**	**0.936**

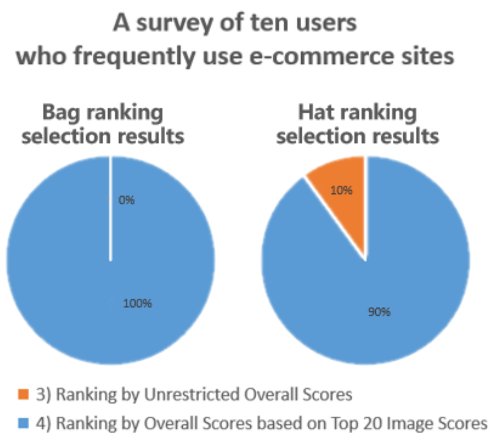

Figure 7 Ranking selection results.

4.4 Discussion

In this section, we compared and discussed the recommendation results by the four types of ranking models that are described in the previous subsection to select the best ranking model for recommending products. After that, we conducted a questionnaire survey with ten college students who frequently use e-commerce sites, comparing our proposed recommendation methods, 3) Ranking by Unrestricted Overall Scores and 4) Ranking by Overall Scores based on Top 20 Image Scores, with the conventional recommendation methods focused on image scores or review scores only, from the users' subjective judgment. And we made the ground-truth based on their answers to compare the four types of ranking models for the objective evaluation of the metrics.

In the case of 3) Ranking by Unrestricted Overall Scores in Figure 3 and Figure 4, we found that the products in the blue frames are not similar to the user input images, as shown in Figure 1 and Figure 2 because they have high review scores. Therefore, we considered that our proposed product recommendation method based on ranking model 3) is not suitable for the image-based product search. On the other hand, in the case of 4) Ranking by Overall Scores based on Top 20 Image Scores in Figure 3 and Figure 4, we found that the products in the red frames have high image scores for the user input images and also have high review scores. After the experimental comparison, we confirmed that our proposed method for product recommendation based on 4) Ranking by Overall Score based on Top 20 Image Scores is the most useful among the four types of ranking models.

From the results in Table 1, we found that 4) Ranking by Overall Scores Based on Top 20 Image Scores achieved the best results among the four types of ranking models, which are evaluated by *MRR*, *MAP*, and *nDCG@10*.

From the survey results of ten college students who frequently use e-commerce sites (Figure 5), we found that the majority of users learn towards the choice of 4) Ranking by Overall Scores based on Top 20 Image Scores in the two rankings by overall scores.

Overall, we confirmed that 4) Ranking by Overall Score based on Top 20 Image Scores is the best among the four types of ranker proposed in this work and meets the goal of this work, which achieves a product recommendation system for recommending high satisfaction products related to the user input image.

5 Conclusion

In this work, we proposed a product recommendation method for recommending products, which have high satisfaction related to the user input image. For this, we combined the similarity of each product's representative image and user input image with the similarity between each product's description and the surrounding text of the user input image based on image recognition to address the problem of not being able to search some product without images on e-commerce websites using only image recognition. Furthermore, we extracted positive information and repeated purchase behavior in relevant product reviews not only use a 5-point rating scale but consider more diversified ratings of product quality to compute the review score for each product as the degree of satisfaction. Finally, we evaluated the product recommendation method proposed in this paper by utilizing product image, product description text, and user review data from Rakuten Ichiba data, which proved superior to the product recommendation method by using the image recognition only.

Future issues are the improvements of the similarity computation of the images and the surrounding text and the repeated purchase behavior detection. Currently, we computed the similarity of the user input image with other product information, such as the product description and the textual labels for the product image, but there is still much room for improvement in this regard. Moreover, the current repeated purchase behavior is detected based on the label "repeat" of the Rakuten Ichiba data. Furthermore, it is possible to match specific words in user reviews or the same user IDs within the same product review to improve the accuracy of detecting the user repeated purchase behavior, which becomes another research direction in the future.

Acknowledgment

In this paper, we used “Rakuten Dataset” provided by Rakuten, Inc. via IDR Dataset Service of National Institute of Informatics.

References

[1] Tugba Sabanoglu (2020): Global retail e-commerce sales 2014-2023. Statista, accessed January 31, 2021. https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales.

[2] Tugba Sabanoglu (2020): Sources of shopping inspiration for digital shoppers worldwide 2020. Statista, accessed January 31, 2021. https://www.statista.com/statistics/1034149/online-shoppers-sources-of-inspiration.

[3] Christoph Zauner (2010): Implementation and benchmarking of perceptual image hash functions. http://www.phash.org/docs/pubs/thesis_zauner.pdf.

[4] Sergey Zagoruyko and Nikos Komodakis (2015): Learning to compare image patches via convolutional neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4353–4361.

[5] Andre Araujo and Bernd Girod (2017): Large-scale video retrieval using image queries. In IEEE transactions on circuits and systems for video technology, vol. 28, no. 6, pp. 1406–1420.

[6] Alina Andreevskaia and Sabine Bergler (2006): Mining wordnet for a fuzzy sentiment: Sentiment tag extraction from wordnet glosses. In Proceedings of the 11th conference of the European chapter of the Association for Computational Linguistics, pp. 209–216.

[7] Ahmed Hassan, Vahed Qazvinian, and Dragomir Radev (2010): What ’s with the attitude? identifying sentences with attitude in online discussions. In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, pp. 1245–1255.

[8] Yohan Jo and Alice H. Oh (2011): Aspect and sentiment unification model for online review analysis. In Proceedings of the fourth ACM international conference on Web search and data mining, pp. 815–824.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018): Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[10] David G. Lowe (1999): Object recognition from local scale-invariant features. In Proceedings of the seventh IEEE international conference on computer vision, vol. 2, pp. 1150–1157.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean (2013): Efficient estimation of word representations in vector space. In arXiv preprint, arXiv:1301.3781.

[12] Rakuten, Inc. (2020): Rakuten Ichiba data. Informatics Research Data Repository, National Institute of informatics. (dataset). https://doi.org/10.32130/idr.2.1

[13] Ricardo Baeza-Yates and Berthier Ribeiro-Neto (1999): Modern Information Retrieval. New York: ACM press, Vol. 463.

[14] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu (2013): A theoretical analysis of NDCG ranking measures. In Proceedings of the 26th annual conference on learning theory (COLT 2013), Vol. 8, 6 pages.