Web document classification using topic modeling based document ranking

Youngseok Lee[1], Jungwon Cho[2]
[1]KNU College of Liberal Arts and Sciences, Kangnam University, Yongin-si, Republic of Korea
[2]Department of Computer Education, Jeju National University, Jeju-si, Republic of Korea

Article Info

Article history:

Received Jul 31, 2020
Revised Dec 10, 2020
Accepted Dec 21, 2020

Keywords:

Big data
Latent Dirichlet allocation
Machine learning
Text mining
Topic modeling

ABSTRACT

In this paper, we propose a web document ranking method using topic modeling for effective information collection and classification. The proposed method is applied to the document ranking technique to avoid duplicated crawling when crawling at high speed. Through the proposed document ranking technique, it is feasible to remove redundant documents, classify the documents efficiently, and confirm that the crawler service is running. The proposed method enables rapid collection of many web documents; the user can search the web pages with constant data update efficiently. In addition, the efficiency of data retrieval can be improved because new information can be automatically classified and transmitted. By expanding the scope of the method to big data based web pages and improving it for application to various websites, it is expected that more effective information retrieval will be possible.

Corresponding Author:

Jungwon Cho
Department of Computer Education
Jeju National University
102 Jejudaehakno, Jeju-si, Jeju-do, 63243, Republic of Korea
Email: jwcho@jejunu.ac.kr

1. INTRODUCTION

The Internet provides diverse information in the fields of education, politics, economy, sports, society, culture, science, and technology. Such information from the Internet is easy to access as web-based open data, but the data provided are complex in a nonhierarchical and heterogeneous form. Thus, data collection and analysis are becoming increasingly important.

Among big data, research to derive implications through text data has been conducted from various viewpoints [1, 2]. The development of information and communication technology has made possible the analysis of frequency of text data, clustering, classification, time series analysis, and network analysis [3]. A lot of attention has recently been focused on services that utilize artificial intelligence based on big data [4].

It is time-consuming and expensive to collect and analyze information in order to utilize different data from the web [5]. To extract valuable information from the web data and create value-added content, it is necessary to extract, transform, and rework data [6]. Recently, web crawlers have attracted attention as a way of extracting relevant information stored in a specific website. Web crawlers are also known as web spiders because their web crawling process works like a spider crawling over a spider web. The web crawler browses the web server, collects considerable information on each hyperlinked homepage, replaces the repetitive task of following each link and obtaining information, automatically analyzes the content of the web page, and connects to the web pages one by one to collect information [7].

There are two major problems related to the collected information. First is the problem of searching for a website address that is rich in good information. The conventional search tools do not take into consideration the possibility of renewing the information of a website because they search by title, introductory article, etc. The second problem to be addressed is the need for information classification. Existing crawlers only need to gather new information at a time, so the user must sort the new information to find the relevant information [8]. To reduce this effort, there is a need for a method by which new information can be automatically classified according to the fields set by the user through document ranking techniques [9].

Most of the search engines use Boolean and vector model variants for ranking. The proposed three ranking algorithms in addition to the classical tf-idf method [10, 11]. They are Boolean spread, vector spread, and most-cited. The first two are regular ranking algorithms of the extended Boolean and vector models to include the pages that a specific page in the response points to, or pages that have a page in the response. The third algorithm, most-cited, is based on terms contained in pages that have links to pages in the response. By comparing these techniques with 56 queries for a collection of 2400 web pages, the vector model produces a better recall-to-precision curve compared to the other algorithms, with an average accuracy of 75% [12, 13].

In hyperlink-induced topic search (HITS), hyperlinked information is used for ranking. The number of hyperlinks pointing to a page provides a level of popularity and quality. Moreover, many links common to pages or pages referenced by the same page often represent relationships between these pages [14]. HITS refers to the pages in the response or considers the set S of pages to which they refer. Pages that have many links pointing to themselves in S are called authorities, and pages that have several outgoing links are called hubs [15, 16].

PageRank was proposed to measure the relative importance of web documents [17]. Authority and hub values are calculated each time they are searched, whereas the page rank is calculated in advance when indexing, so that it can be searched quickly. Consequently, PageRank is used more for commercial information retrieval systems [18, 19]. PageRank is a graphical representation of Web documents linked through hyperlinks and the ranking of Web documents. Many documents assume that the document they are pointing to is more important and express its value in numerical terms.

To obtain the page rank of all web documents, a method of initially setting a start value and repeatedly calculating a page rank is applied. The page rank of the web document converges from the initial value to the approximate value of the original page rank through an iterative calculation process [20]. In this study, we analyze research trends in the field of big data; a topic modeling technique recently applied among text mining techniques is used. Document rankings can be used to collect various website addresses that can be utilized by users; the number of data and the number of updates can be set by the user. The problem of new information classification to be delivered to the user can be solved by suggesting an appropriate index selection method suitable for the field set by the user.

2. RESEARCH METHOD

2.1. Collection of web documents

The amount of web information available to web users has increased dramatically. Such web information is scattered across various portal sites. However, the web information is collected only by the user's own registration, and the information can be collected only through various efforts taken by the user. In addition, the search for current information is limited to the title and introduction part of the web document. In this paper, we propose a method to collect web document information by automating the crawler without interacting with the user and extracting the title, content, author, authoring time, and so on.

The considerations for theme-driven crawler implementation are as follows: First, web documents do not have a single formal structure. Second, important web address links usually reside on the main page of the website. Therefore, the range of the link search for searching the address of the web document should be limited to the number of links existing on the main page of the web site. To address the abovementioned issues and to increase the effect of collecting web documents, the crawler presented in this paper uses the random access module of the portal site. Through the call of the module, the crawler continuously obtains website addresses, restricts the size of the search queue by the number of links existing on the main page, and examines the links in a breadth-first manner. However, additional effort is required if the main page of the website is in frame form. As frames do not contain actual link information, it is necessary to access a page constituting a frame to actually find a web page in which a link exists.

2.2. Topic modeling

The learned website address dynamically constitutes a directory. The directory to which a web document belongs to can help users find the information they need, if they did not know the specific query keyword [6]. Topic modeling defines meaning as a relationship of words and seeks meaning through clusters to which words belong to [21]. At this time, the cluster is a sack of unordered words, so the frequency with which words appear simultaneously is a requirement for determining the subject. The topic modeling algorithm is a probabilistic model that extracts latent topics from a set of unstructured documents. It extracts important topics from unstructured data, text, and determines whether a relationship exists between a single topic and a single topic or between multiple topics and multiple topics.

Topic modeling can be regarded as similar to document clustering techniques from the viewpoint of clustering a large number of documents according to the topic [22]. However, it is to be noted that topic modeling fits the demands of the present scenario because one document can be mapped to multiple topics simultaneously. Topic modeling can be said to be of high value in the sense that it can analyze a large number of documents and provide insight into related fields, apart from using the analysis results to perform supplementary analysis.

Topic modeling is mainly used in industry and social computing, such as the proposed initial online user review analysis, and opinion mining in online blogs. Recently, it has also been used in various fields, including industry and academia. Topic modeling uses data that have a clear distinction between topics and relatively objective judgment on word relevance for model evaluation [23]. The main model of this study, latent Dirichlet allocation (LDA), is a Bayesian inference model that utilizes the Dirichlet probability distribution. The name is given because the Dirichlet distribution is used as a dictionary distribution for estimating the probability distribution of the topic and the word probability distribution [5]. Bayesian inference basically takes the form of a posterior probability distribution by multiplying the likelihood function obtained from the data with the prior probability distribution assumed by the researcher.

LDA2VEC was proposed by Moody in an attempt to improve the performance of topic modeling by integrating word embedding into LDA after Mikolov's Word2Vec was developed [10, 11]. Word embedding is suitable for expressing the relationship between words. Nevertheless, it is difficult to intuitively understand what each vector represents [11, 13]. LDA can easily interpret topics expressed through words; however, it does not reflect the meaning of individual words as well as word embedding [24, 25]. Therefore, a model was proposed to combine the semantic information inherent in word embedding with the interpretative advantage of LDA [26, 27].

2.3. Index word comparison and similarity calculation

The similarity of words extracted from the input document can be calculated based on a group of indexes by field as in (1) and the words are classified into the field with the highest similarity.

$$Sim(D,Q) = \sum_{i=1}^{n}(w_{di} \times w_{qi})$$

$$w_{di} = \frac{\log tf_{di}}{\sum_{i=1}^{n}[\log tf_{di}]^2}$$

$$w_{qi} = \frac{(\log tf_{qi})\cdot\log\frac{N}{n_i}}{\sum_{i=1}^{n}\left[(\log tf_{qi})\cdot\log\frac{N}{n_i}\right]^2} \tag{1}$$

D : document
Q : Query
w_{di} : Weight of the i-th index word t_i in document *D*
w_{qi} : Weight of the i-th index word t_i in query *Q*
tf_{di} : frequency of occurrence of the i-th index word in document *D*
tf_{qi} : Frequency of occurrence of the i-th index word in query *Q*

After examining the frequency of the index words and the frequency of the documents in which the index words appear, and selecting a group of index words by field, it is necessary to consider the overlap of index words with other fields. Therefore, a word that is selected as an index word in another field must be removed from the field to become a group of index words with efficiency when classifying in terms of field. The criteria for classifying duplicate documents based on index terms are as given in (2).

$$S_2 > S_1 \times 0.9 \tag{2}$$

In (2), S_1 is a first-order similarity value and S_2 is a second-rank similarity value. The weight for determining the similarity is set to 0.9 based on experiments on 20 documents randomly selected by field. That is, if the condition of the expression is satisfied, it is classified as another document and the subject area is normally classified.

2.4. Classification and ranking of web documents

Since documents such as SNS automatically deliver content, web document search tools search for titles and contents in Internet newspapers, SNS, and general websites, and they present a ranking regardless of the information update cycle. Therefore, it is necessary to prioritize and present documents to which information is updated rather than to present them as they are. To this end, a method of determining the ranking by calculating the similarity of documents is proposed, but it is difficult to prioritize the documents of the sites at a time. Furthermore, a study that calculates the similarity of word criteria based on a title using hyperlink information is not suitable because it does not consider whether information such as images and multimedia other than the characteristics of the document exists. The method of ranking a document based on a static link is not suitable for ranking SNS because hyperlink information with other documents changes dynamically and the number of links is limited in most cases.

In this study, priority is given to documents whose content is frequently updated though the data update cycle is short. To this end, we propose a ranking method for web documents that considers the update rate and update cycle of data, including user query words. The first criterion for setting the web document search and update period, and ranking the web documents is as shown in (3).

$$DocumentRank = TitleDR + TextDR$$

$$TitleDR = \sum_{i=1}^{n}\left(\frac{q}{m_i} \times 1\right) \tag{3}$$

$$TextDR = \sum_{i=1}^{n}\left(\frac{q}{m_i} \times 0.5\right)$$

DocumentRank : Priority value of web document
TitleDR : the location of the query in the case of the title and Priority value according to the number of posts
TextDR : the location of the query in the case of the body and Priority value according to the number of posts
n : Number of documents that matched the corresponding post among query words
q : User's query word count
mi : The number of posts in the i-th web document, including query words, during the web search period

The first criterion is ranking according to the number of the title and body text of the post, including the queried words that exist within the web document search period. If the results of the first criterion are the same, the ranking is determined by applying the second criterion, such as (4), to determine whether the web document update is consistent.

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - M)^2 \tag{4}$$

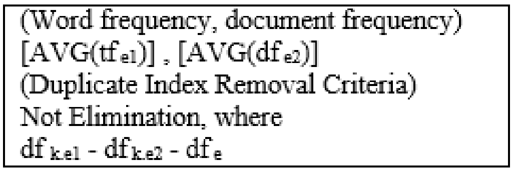
n : (Web search period)/(Web renewal cycle)
x_i : number of posts including query words in the i cycle
M : Average number of documents updated during the web search period

Using this priority method, documents in which posts are frequently updated within a web document search period are preferentially presented. If the web document search period and update period are the same, the document with a high match with the user's query word has a higher priority. For example, the search period for the web document is set to 45 days and that for the sample period to 5 days. The number of posts including query words in the search period of web document A and web document B is 1, 2, 2, 3 and 0, 5, 0, 5, 0, 3, 0, 2, 0 respectively. Assuming the number of posts including query words in the search period to be the same, web document A has a smaller variance value compared to web document B, so it has a higher ranking. It is significant to have such a high ranking because the amount of information updated within a specified search period is an important factor in providing suitable information to the user.

3. RESULTS AND DISCUSSION

We selected seven fields, i.e., education, politics, economy, culture, society, sports, and IT for this study, and prepared 100 news articles already classified for each field. There were three important considerations when extracting index words. The first was the word frequency (tf). The decision whether to treat words that appear more than once in a document is an important index. The second was the document frequency (df). In the same field, determining the number of times the document frequency of the word is a word that can represent the field. Finally, it was a criterion for eliminating overlapping index words in each field. If a word representing a field exists in another field, it should make a mistake in classification and set a criterion for removing it. If all are removed without proper criteria, it is likely that a word with a high document frequency in one field is deleted by a word in another field with a relatively low document frequency.

The ultimate goal in extracting index words for real-time content classification is to obtain the maximum classification effect with a minimum number of index words. The smaller the size of the index word group, the smaller the number of similarity comparison operations. Therefore, the extraction of the index word representing the field should select the optimal value of the aforementioned three factors and select the minimum set of index words with appropriate classification performance. Figure 1 shows the pseudo code of the proposed algorithm.



```
(Word frequency, document frequency)
$[AVG(tf_{e1})]$ , $[AVG(df_{e2})]$
(Duplicate Index Removal Criteria)
Not Elimination, where
$df_{k.e1} - df_{k.e2} - df_{e}$
```

Figure 1. Pseudocode of the proposed algorithm

In summary, the method of selecting the minimum indexing group for rapid classification is to maintain the correct accuracy and construct a group of indexing words by taking the average word frequency of each field and the average value of the average document frequency average. It is effective to adopt a scheme in which an index removal standard is applied. The proposed system automatically extracts a group of index keywords from the initial document with criteria of frequency of occurrence and elimination of redundant indexes in each field. In order to investigate the effectiveness of this method, we conducted a comparative experiment with the case of explicit index extraction for each field.

Five students were used to extract the index words for each field. When the actual users select the index words related to the field, they tend to select nouns representing each field in view of their personal experience. The number of selections ranged from 5 to 20. Based on the index words selected by the five participants of the experiment, 10-20 index words were assigned to each field. Figure 2 shows a comparison of the indexing method based on the frequency of the proposed system and the explicit indexing method of the user.

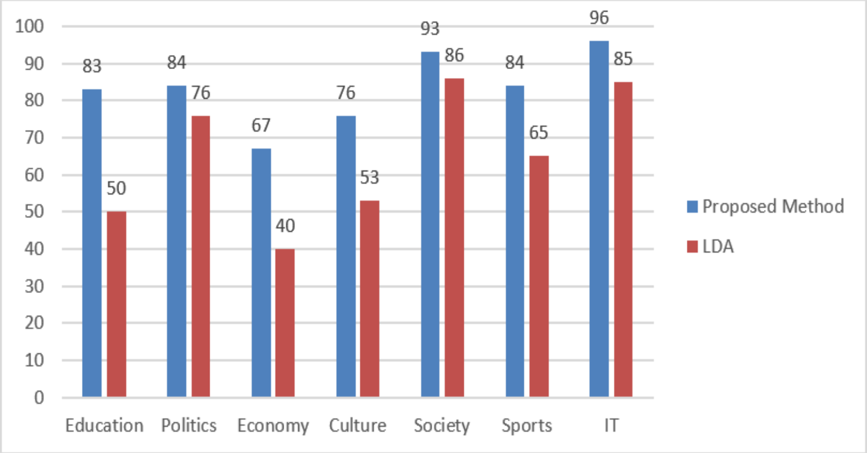


Figure 2. Performance comparison by index

In this case, the proposed method showed approximately 18% improvement in performance. This result shows that there are many cases where the number of index words set by the user is too small to be classified. In addition, there are many cases where the indexes are miscategorized because the index words were generated without elimination criteria of the duplicate indexes.

4. CONCLUSION

Recently, the amount of web information that users can use through search engines has become immense. The proposed document rankings based web crawler service uses a document ranking technique to avoid problems related to duplicate crawling. A classification method that can properly classify new information on websites was discussed through the proposed document ranking method. Users who use conventional web crawler services to search for web information can view relevant information by searching, filtering, bookmarking, and so on.

Users who use the web crawler service based on the proposed document ranking can select relevant web information from the database constructed by web crawling using a desired search keyword or a filtering function. It is also possible to provide a bookmark function such that when a user finds information they are interested in, the user can keep the information; the web crawler then notifies the user whenever new information is updated. In this study, we designed and implemented a web crawler service by crawling web pages for a certain site for a particular period of time. In future, if the web page range that can be applied to the web crawler service based on the proposed document rankings is increased and applied to various web sites at the same time, more efficient information retrieval can be achieved.

REFERENCES

[1] X. Wang, and H. C. Kim, “New Feature Selection Method for Text Categorization,” *Journal of Information and Communication Convergence Engineering*, vol. 15, no. 1, pp. 53-61, 2017.

[2] X. Wang and H. C. Kim, “Text Categorization with Improved Deep Learning Methods,” *Journal of information and communication convergence engineering*, vol. 16, no. 2, pp. 106-113, 2018.

[3] G. Wang and S. Y. Shin, “An Improved Text Classification Method for Sentiment Classification,” *Journal of information and communication convergence engineering*, vol. 17, no. 1, pp. 41-48, 2019.

[4] M. David, "Probabilistic topic models," *Communications of the ACM*, vol. 55, no. 4, pp.77-84, 2012.

[5] B. Piotr, et al., "Enriching word vectors with subword information," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 135-146, 2017.

[6] C. Jonathan, et al., "Reading tea leaves: How humans interpret topic models," *Advances in neural information processing systems*, pp. 288-296, 2009.

[7] D. Scott, et al., "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391-407, 1990.

[8] L. Siwei, et al., "How to generate a good word embedding," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 5-14, 2016.

[9] K. Dasom, et al., "A methodology for analyzing public opinion about science and technology issues using text analysis," *Journal of Information Technology Services*, vol. 14, no. 3, pp. 33-48, 2015.

[10] M. Tomas, et al., "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[11] M. Christopher E., "Mixing dirichlet topic models and word embeddings to make lda2vec," *arXiv preprint arXiv:1605.02019*, 2016.

[12] N. Masato, et al., "A bag of useful tricks for practical neural machine translation: Embedding layer initialization and large batch size," *Proceedings of the 4th Workshop on Asian Translation (WAT2017)*, 2017, pp. 99-109.

[13] P. Sungjoon, et al., "Subword-level word vector representations for Korean," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2429-2438.

[14] P. Jeffrey, et al., "Glove: Global vectors for word representation," *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[15] Q. Jipeng, et al., "Topic modeling over short texts by incorporating word embeddings," *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, Cham, 2017, pp. 363-374.

[16] S. Iulian V., et al., "Building end-to-end dialogue systems using generative hierarchical neural network models," *Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 3776-3783.

[17] B. David M., et al., "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993-1022, 2003.

[18] S. David, et al., "Mastering the game of go without human knowledge," *Nature*, vol. 550, pp. 354-359, 2017.

[19] Y. Jieming, et al., "A novel feature selection based gravitation for text categorization," *International Journal of Database Theory and Application*, vol. 9, no. 3, pp. 211-228, 2016.

[20] X. Guixian, et al., "Sentiment analysis of comment texts based on BiLSTM," *IEEE Access*, vol. 7, pp. 51522-51532, 2019.

[21] L. Ximing, et al., "Group topic model: organizing topics into groups," *Information Retrieval Journal*, vol. 18, no. 1, pp. 1-25, 2015.

[22] Hornik, Kurt, and Bettina Grün, "topicmodels: An R package for fitting topic models," *Journal of statistical software*, vol. 40, no. 13, pp. 1-30, 2011.

[23] Le, Quoc, and Tomas Mikolov., "Distributed representations of sentences and documents," *ICML'14: Proceedings of the 31st International Conference on International Conference on Machine Learning*, vol. 32, 2014, pp. 1188-1196.

[24] L. Fetty Fitriyanti, et al., "Topic Discovery of Online Course Reviews Using LDA with Leveraging Reviews Helpfulness," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 426-438, 2019.

[25] S. T. Park, and Y. K. Kim, "A study on deriving an optimal route for foreign tourists through the analysis of big data," *Journal of Convergence for Information Technology,* vol. 9, no. 10, pp. 56-63, 2019.

[26] S. J. Malebary and A. Shakeel, "Semi-supervised method for sensitivity based documents," *International Journal of Advanced and Applied sciences,* vol. 7, no. 5, pp. 20-26, 2020.

[27] S. W. Kim amd J. M. Gil, "Research paper classification systems based on TF-IDF and LDA schemes," *Human-centric Computing and Information Sciences,* vol. 9, pp. 30-50, 2019.

BIOGRAPHIES OF AUTHORS

Youngseok Lee is a Professor at the KNU College of Liberal Arts and Sciences, Kangnam University. He is an author of over 20 papers in refereed international journals and conference proceedings. His research interests include Computer Education, SMART learning, Intelligent System, Knowledge Informaton System, Liberal Education, and so on.

Jungwon Cho is a Professor at the Department of Computer Education, Jeju National University. He is chief editor and vice-president of the journal of the Korean association of computer education, South Korea. Also, He is the director of ‘Center of Intelligent Computing Education’ in Jeju National University. He is studying about Computing education, Intelligent information ethics, Intelligent information system and so on.