DOI: 10.14421/ijid.2020.09104 IJID (International Journal on Informatics for Development), *e-ISSN: 2549-7448*
Vol. 9, No. 1, 2020, Pp. 21-28

An Efficient Journal Articles Searching using Vector Space Model Algorithm

Azis Alvriyanto[1], M Taufiq Nuruzzaman[2]*, Maria Ulfah Siregar[3], Rahmat Hidayat[4]
Teknik Informatika
Universitas Islam Negeri Sunan Kalijaga
Yogyakarta, Indonesia
[1]aalvriyanto@gmail.com, {[2]m.taufiq, [3]maria.siregar, [4]rahmat.hidayat}@uin-suka.ac.id
**Corresponding author*

Article History
Received July 28th, 2020
Revised Aug 31st, 2020
Accepted Sep 3rd, 2020
Published Sep, 2020

***Abstract*— One of the main feature of digital library is a search engine which depends on keywords submitted by a user. However, in the traditional algorithm, the computation performance, searching speed, significantly relies on the number of journal articles stored in the databases. Some irrelevant search results also increase the speed of article searching process. To solve the problem, in this paper we propose vector space model (VSM) algorithm to search for relevant journal articles. The VSM algorithm considers a term frequency - inversed document frequency (TF-IDF). The VSM algorithm will be compared to the baseline algorithm namely traditional algorithm. Both algorithms will be evaluated using combination of keywords which can be a synonym, phrase, error typography, or suffix and prefix. By using the data consist of 635 journal articles, both algorithms are compared in terms of 11 evaluation criteria. The results show that VSM algorithm is able to obtain the intended journal at 5th rank on average as compared to the traditional algorithm which can obtain the intended journal at rank of 171st on average. Therefore, our proposed algorithm can improve the performance to accurately sort the journal articles based on the submitted keywords as compared to traditional algorithm.**

Keywords- digital library; scraping; tf-idf; vector space model

1 INTRODUCTION

The Indonesian constitution number 43/2007 state that library is an institution who manage journal articles, book, and theses for the purpose of education, research, information, and recreation. Therefore, library should be able to adapt with the needs of users. Library users want to obtain information they needed as soon as possible and as relevant as possible. Recently, the digital library becomes more popular where all journal articles, book, and theses will be available online.

The aims of digital library is to give an easy way for readers to obtain the documents they need by submitting some keywords. In the traditional algorithm, the computation performance, searching speed, significantly relies on the number of journal articles stored in the databases. Some irrelevant search results also increase the speed of article searching process. Obviously, the high numbers of documents may lead to longer searching process. Moreover, the obtained irrelevant document also increase the searching time. List of irrelevant documents attract the readers' attention resulting in overall longer searching process.

Therefore, we propose to employ vector space model (VSM) algorithm to speed up the searching process in the database of journal articles. VSM algorithm calculates the relevancy of all keywords submitted by the readers. The relevancy will be compared and associated with the documents in the database [1]. As compared to the traditional algorithm which considers a string matching only, VSM algorithm is able to achieve a better performance in terms of the rank of relevant results.

VSM algorithm has been employed for some specifics problem areas such as e-book [1]. In this work, 100% recall and 80% precision was obtained. However, how the performance of VSM algorithm to search for journal articles has not been evaluated. Another work employed the VSM algorithm for the purpose of finding the impact of online game based on textual symptoms where some words have different text but similar meaning [2]. In [3], the algorithm also has been employed for searching the relevant music or song in the karaoke machine. It gives a weighted value for each song relevant to the submitted keywords.

Document searching using VSM has been initially proposed in [4]. In this work, VSM has been employed to search for the abstract of students final projects written in Indonesian language. It does not only show the recall and precision but also computation time and the location of the relevant documents. Note that this study only consider a simple abstract as the searched documents. VSM also can be used as plagiarism checker [5]. In this work, the term frequency-inversed document frequency (TF-IDF) is used to check the similarity of many submitted documents based on the keywords in the documents. Based on the aforementioned works above, the VSM has not been employed as journal articles searching algorithm.

The rest of this paper is organized as follows. VSM algorithm is then detailed in Section 2. Section 3 shows the performance comparison results. Finally, Section 4 concludes this paper.

2 VECTOR SPACE MODEL ALGORITHM

2.1 Data

In this section, the detail process of VSM algorithm will be described. The VSM algorithm will be compared to traditional string matching to make a rank of searching results. The data used in the experiments are journal articles taken from institutional repository of UIN Sunan Kalijaga, which can be accessed here http://digilib.uin-suka.ac.id/view/divisions/ejour and are already textually preprocessed. The journal articles are published from 2003 to 2020. These both algorithms are evaluated using combination of synonym keywords, keywords with error typography, phrase, and keywords with suffix or prefix.

The journal article data in the PDF format are obtained through scrapping method from the website. The data then are converted into text format. The information collected from the data are title, authors, publication, date, number of pages, volume, subject, keywords, abstract, URL, the address of pdf file, date of submission, date of modification, and the content of the journal articles. The data then is stored in the MySQL database.

2.2 Scraping

Web scrapping is a process to crawl semi-structured documents from the Internet. The documents are normally in the markup language such as HTML or XHTML. The web scrapping process has four steps [6][7]:

a) *Create Scraping Template*: The structure of HTML document is examined. Based on the tag in document, a specific information can be obtained,
b) *Explore Site Navigation*: The website must be explored to find the right page. Once the page is obtained, it will put into the web scrapping software,
c) *Automate Navigation and Extraction*: Based on the steps 1 and 2 above, the software to extract the information automatically can be developed, and
d) *Extracted Data and Package History*: The obtained information will be stored in the database for future process of VSM algorithm.

Information obtained via web scrapping are more focus than that of manual process [8] [9].

2.3 Text Preprocessing

Because the textual data are unstructured, the data must be processed first before they can be processed by the text mining algorithm. In the text preprocessing process, the data will be selected for each document. The process of text preprocessing are:

a) *Case Folding*: all the characters will be converted into lowercase. All symbol will be removed including comma or semicolon [10].

b) *Tokenizing*: all the text will be tokenized resulting in a list of words only [11].

c) *Stop Word Removal*: *Stop words* are vocabularies which are not unique or have a little meaning for the sentences or documents [12]. The stop words can be removed because they can increase the computation time and sometime they have a contra productive effect for the result of text mining. The example of stop words in Indonesian language are "sebuah", "oleh", "pada", etc [13].

d) *Stemming*: In this step, all words with suffix or prefix are converted into their basic form. For example, the words 'membuka', 'terbuka', 'pembuka' will be converted into a single word 'buka'. This step increases the sensitivity and ability to find the most relevant documents. However, this also decrease the selectivity where grouping some words into a single word may remove some meanings in the documents. This stemming process is expected to increase the recall but may decrease precision as a tradeoff [14][15].

2.4 Term frequency - inverse document frequency (TF-IDF)

TF-IDF will be employed for weighting process in order to obtain the value for each extracted word obtained from previous steps. In this step, each document is considered as vectors with words/terms as elements extracted from previous steps. The vector consists of weight value from each term calculated based on TF-IDF. Term frequency is a method to calculate the weight for each term in the documents. In this method, how important the term is calculated based on its frequency in the documents. The more the term in the document is, the higher the term value is [16]. TF-IDF is also used in some applications, such as sentiment analysis [17][18]. The formulae of the term frequency can be seen in Formulae 1 as follows:

$$W(d,t) = TF(d,t) \quad (1)$$

Unlike term frequency which focus on the frequency of terms/words in a document, inverse document frequency (IDF) focus on the frequency of terms/words in all collected documents. Therefore, in IDF the terms/words which are rarely shown will have a higher value or more important as shown in Formulae 2 [16]:

$$IDF(t) = log\left(\frac{N}{df(t)}\right) \quad (2)$$

The combination between TF and IDF is used to give weighted value for each term/word. It is expected that the combination can improve the performance of any text mining algorithm. The factor of TF and IDF is supposed to be able to contribute the recall and precision value [19]. The combination of TF and IDF is formulated in Formulae 3 as follows:

$$TFIDF(d,t) = TF(d,t).IDF(t) \quad (3)$$

2.5 Vector Space Model Algorithm

The similarity among sentences plays important role for many research works related to text and application. VSM algorithm is employed as representative of a collected dataset textual document. The weight of each documents shows the importance and contribution of the words for the document and a collected document. How important the term/words is represented by the frequency of the term/words in the documents. One of the most popular text similarity measurement is cosine similarity [20][21]. The concept of cosine similarity is to calculate the cosine value of angle from both two vectors. In this case, given a document represented by vector d_j, query q, and term t, the cosine can be calculated. An angle between vector d_j and query q can be calculated. The smaller the angle is, the more similar vector d_j and query q is. The cosine of the angle is a coefficient which can represent the similarity between vector d_j and query q. The cosine formulae can be seen in Formulae 4 as follows [22]:

$$cos(q,d_j) = \frac{\sum_k [TFIDF(t_k,q)] \cdot [TFIDF(t_k,d_j)]}{\sqrt{\sum |TFIDFq|^2} \cdot \sqrt{\sum |TFIDFd_j|^2}} \quad (4)$$

The formulae $cos(q, d_j)$ is a notation of cosine between query q and document j. As for TF-IDF(tk, q) dan TF-IDF(tk,dj), each of them shows the weighted value for term/word *tk* in the query q and the weighted value for term/word *tk* in the document j. As for $|TF\text{-}IDF_q|$ and $|TF\text{-}IDF_{dj}|$, each of them is the length of vector query j and length of vector document j. The structured data is stored in Firebase database provided by Google to avoid calling the text preprocessing and weighting procedures many times.

3 RESULT AND DISCUSSION

3.1 Benchmark

In this section, VSM algorithm will be compared with traditional algorithm. In the traditional algorithm, normally standard searching form based on key words is available. Once the keywords are entered the string matching algorithm is performed. The results obtained from the string matching process will be sorted based on the subject such as: name, date, or title.

3.2 Evaluation Tools

For the purpose of performance comparison, a program written in JavaScript and ReactJs was developed. Two programs for traditional algorithm and VSM algorithm are prepared for comparison. Both program will show the list in rank mode of any searching results. Fig. 1 shows the keyword submission menu for both traditional and VSM algorithms. In the traditional algorithm the users must input the type of data such as name or title. As for VSM algorithm, this procedure is not needed.

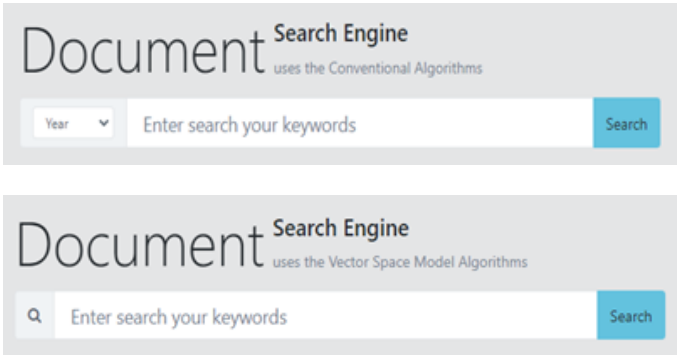

Figure 1. The keyword submission menu for both traditional and VSM algorithms

Traditional algorithm employs SQL command to select data based on the submitted keywords. The SQL query results are sorted based on author's name, journal title or year of publication. The SQL command uses LIKE function to select the data in the column of authors, title, abstract, and content. As for VSM algorithm, the results depend on the cosine similarity as mentioned before which is run on client side using JavaScript.

3.3 Evaluation Results

Both algorithms are performed on 635 journal articles using 11 document sample and its possible keywords using combination of synonym, typography error, pgrase, and suffix and prefix. Table 1 shows that the samples used in this evaluation. Both traditional and VSM algorithms are compared based on the rank of the intended journal article on the results list. Phrase is a union of words which is non predictive where one of the words is not a verbs [23]. In Table 1, the sample of phrase is shown by sample document number 3 and 4. The result for both document samples are depicted by Fig. 8 and Fig. 9.

Table 1. Samples documents (in Indonesian)

No.	Sample document		Keywords	Number of searching results
	Authors	Title		
1	Faisal Ismail	Perkembangan Islam di Amerika Serikat	islam di amerika	422
			islam di amerika serikat	422
			islam di usa	418
2	Early Maghfiroh Innayati	Dzikir Sebagai Kendali Emosi Bagi Remaja	fungsi dzikir	168
			fungsi zikir	166
			fungsi dikir	164
3	Sri Rohyanti Zulaikha	Kontribusi Islam atas Perkembangan Peradaban: Sikap dan Kaitan Islam dengan Perpustakaan dalam Pendistribusian Informasi	islam dan perkembangan peradaban	417
			perkembangan peradaban islam	417
4	Rahmat	Implementasi Nilai-Nilai Islam dalam Pendidikan	implementasi islam dalam kehidupan sehari-hari	467
			penerapan islam dalam kehidupan sehari-hari	462
			islam di kehidupan sehari-hari	461
5	Moch. Fatkhan	Kearifan Lingkungan Masyarakat Lereng Gunung Merapi	kebiasaan masyarakat lereng gunung merapi	352
			keseharian masyarakat lereng merapi	350
			kehidupan masyarakat lereng gunung merapi	394
6	Mohammad Zamroni	Perkembangan Teknologi Komunikasi dan Dampaknya Terhadap Kehidupan	kemajuan teknologi komunikasi	253
			kemajuan tekhnologi komunikasi	181
			perkembangan teknologi komunikasi	222
7	Lathifatul Izzah	Melihat Potret Harmonisasi Hubungan Antar Umat Beragama di Indonesia	hubungan antar umat beragama	379
			hubungan umat beragama	379
			hubungan antar kaum beragama	378
8	Rahmat Fajri	Sejarah Keuangan Islam	sejarah kelembagaan keuangan islam	472
			sejarah lembaga keuangan islam	472
9	Shofiyullah MZ	*E-commerce* dalam Hukum Islam (Studi Atas Pandangan Muhammadiyah dan NU)	*e-commerce* dalam islam	483
			commerce dalam islam	407
10	H. Mardjoko Idris MA	Teologis atau Bahasa?	majaz teologis bahasa	267
			majas teologis bahasa	267
11	Ade Ratnasari	Teknologi Informasi untuk Masyarakat Pedesaan	teknologi informasi di masyarakat	394
			technologi inpormasi di masyarakat	347

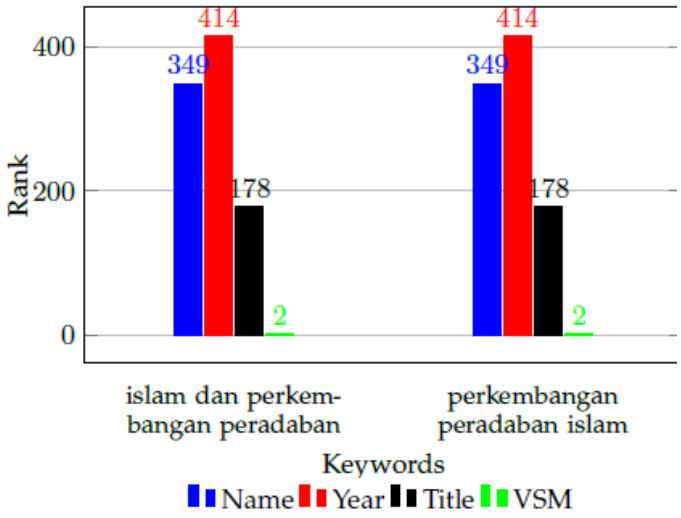

Figure 2. Rank for searching the journal title "Kontribusi Islam atas Perkembangan Peradaban"

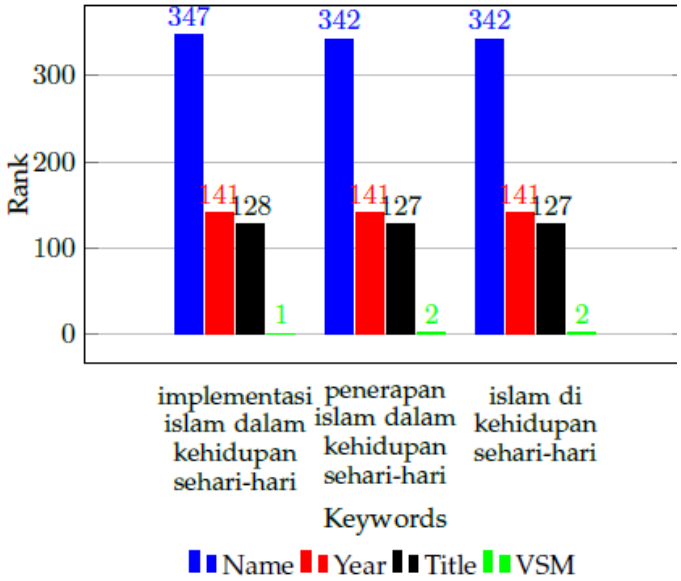

Figure 3. Rank for searching the journal title "Implementasi Nilai-Nilai Islam dalam Pendidikan"

Figs. 8 and 9 show that the use of phrase does not significantly influence the rank of the searched document because the results are not significantly change when the order of keyword is changed. However, Fig. 8 shows that the rank of the document searched by using VSM algorithm is different but only 1 rank different. Synonym is a group of words that has many different text but has similar or the same meaning. In Table 1, keywords that have a combination of synonym is shown by document sample number 5, 6, and 7.

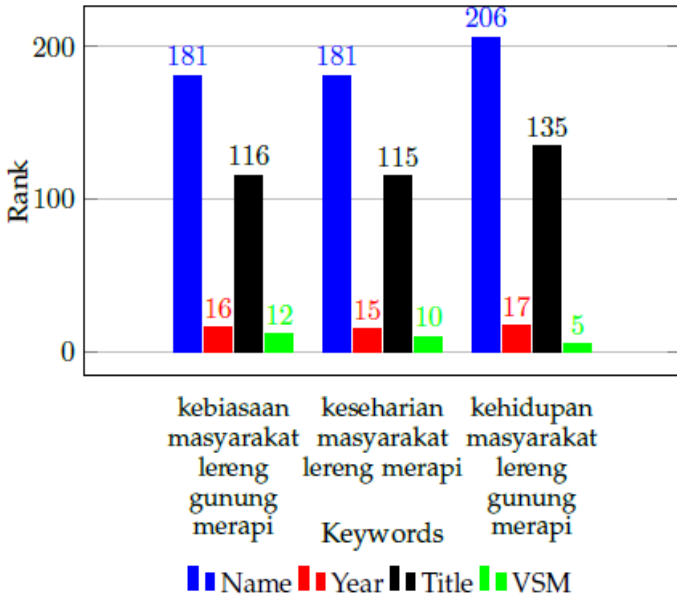

Figure 4. Rank for searching the journal title "Kearifan Lingkungan Masyarakat Lereng Gunung Merapi"

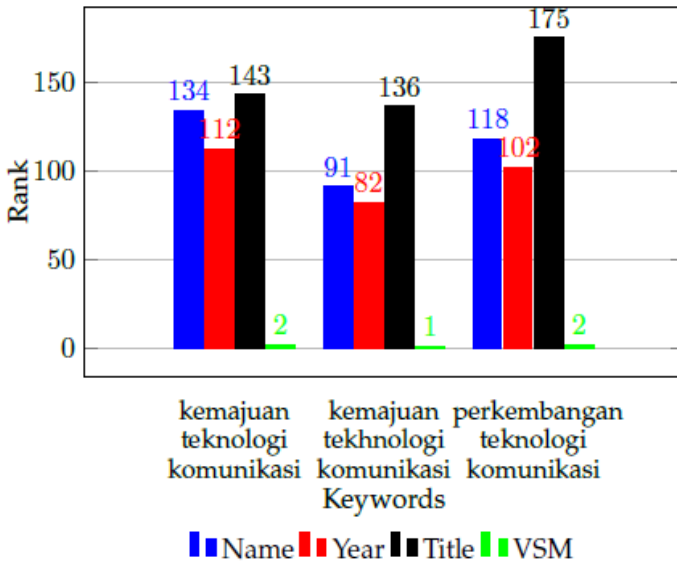

Figure 5. Rank for searching the journal title "Perkembangan Teknologi Komunikasi dan Dampak Terhadap Kehidupan"

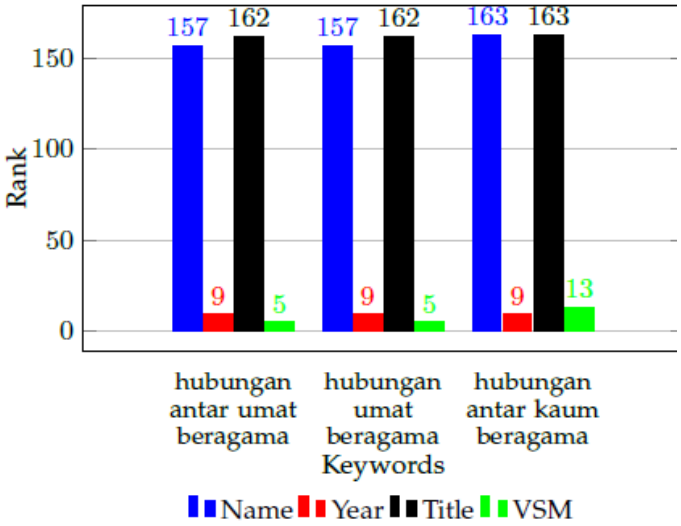

Figure 6. Rank for searching the journal title "Melihat Potret Harmonisasi Hubungan Antar Umat Beragama di Indonesia"

Figs. 10-12 show that the used of synonym significantly influence the rank of intended journal article. The figs show that the rank of the searched document is different along with the different synonym. However, for VSM algorithm the changing in rank of document results is relatively stable because the different is only 1 to 8 rank different. As for typography error, the documents for sample is number 1, 2, 10, and 11 in Table 1. Error typography significantly influences the rank as shown in Figs. 13-16. In the worst case, typography error may result in zero result for both traditional and VSM algorithms as shown in Fig. 14. The influence of suffix and prefix of the keywords is shown in Figs. 17-18. By using the document samples number 8 and 9, it is shown that additional prefix or suffix does not significantly influence the rank of the documents results.

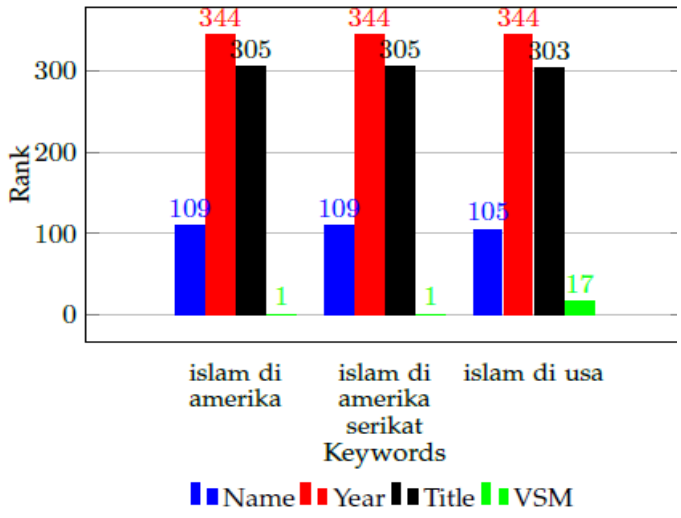

Figure 7. Rank for searching the journal title "Perkembangan Islam di Amerika Serikat"

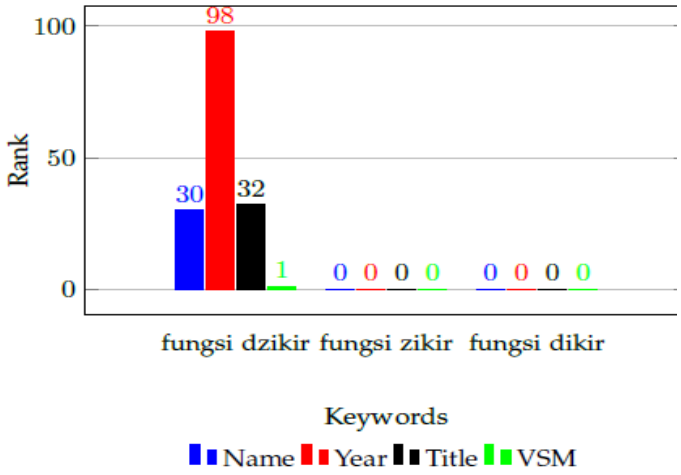

Figure 8. Rank for searching the journal title "Dzikir sebagai Kendali Emosi bagi Remaja"

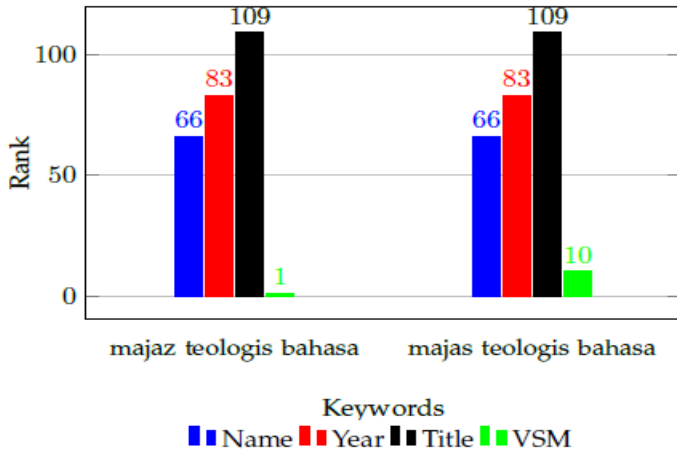

Figure 9. Rank for searching the journal title "Teologis atau Bahasa?"

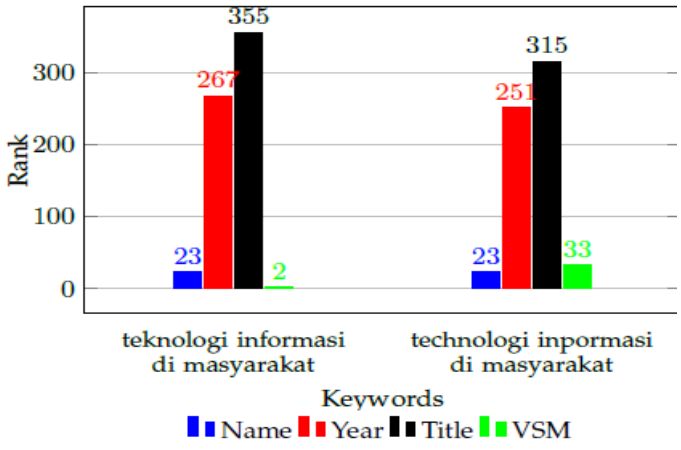

Figure 10. Rank for searching the journal title "Teknologi Informasi untuk Masyarakat Pedesaan"

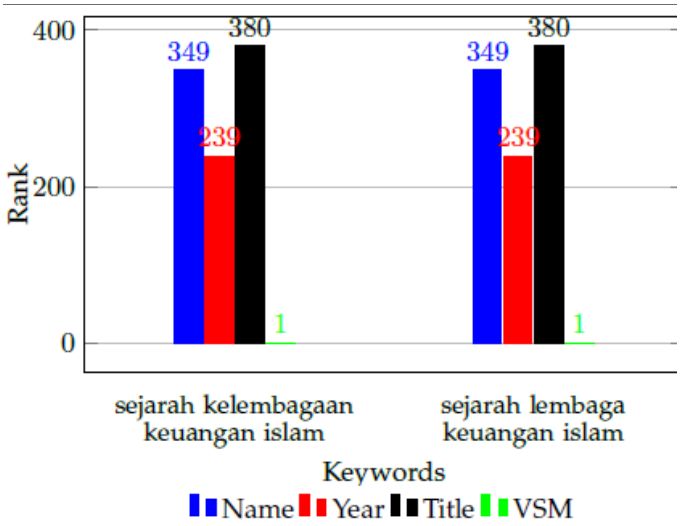

Figure 11. Rank for searching the journal title "Sejarah Keuangan Islam"

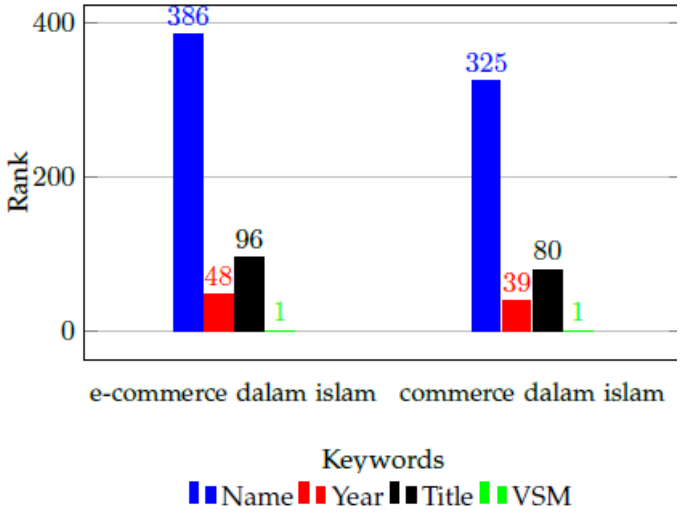

Figure 12. Rank for searching the journal title "*E-Commerce* dalam Hukum Islam"

Our current work focus on comparison analysis between traditional algorithm and VSM algorithm. Overall, VSM algorithm outperforms traditional algorithm in term of rank of the searched journal article. The results show that VSM algorithm is able to obtain the intended journal at 5th rank on average as compared to the traditional algorithm which can obtain the intended journal at rank of 171st on average. It is mainly because VSM algorithm sort the document based on the similarity between the keywords and the documents. In contrast, the traditional algorithm sorts the result based on similarity of the keyword and author's name, year of publication, and title. An accurate rank can ease the readers to find the document because they do not have to scroll down the bar to search for the document.

For traditional algorithm, the submitted keywords are only useful of the information of the document contains the keywords. As for the ranking process, it depend on the chosen field which can be author names, year of publication, or title putting the searched document into the lower rank. Moreover, VSM algorithm is not influenced by the keywords with phrase or prefix or suffix. For VSM algorithm, only the typography error and synonym which can decrease its performance. However, this problem does not significantly influence the rank. As for typography error, it can significantly decrease the performance because the error can change the meaning of the submitted keywords.

4 CONCLUSIONS

Based on the experiment previously mentioned, the VSM algorithm can significantly outperform the traditional algorithm to find the journal articles in the digital library. In any cases (synonym combination, typography error, phrase, and prefix and suffix), the results obtained through VSM algorithm are better than the results obtained through traditional algorithm. Therefore, we recommend to implement the VSM algorithm for searching the journal article in the digital library.

For future works, the VSM algorithm can be improved using query expansion (QE). In this solution, the queries from the readers are predicted before to avoid typography error. The typography error is still the main problem because it can change the meaning of keywords.

REFERENCES

[1] N. Annisa, W. Nengsih, and Ananda, "Implementasi Algoritma Vector Space Model dalam Pencarian E-Book," *J. Aksara Komput. Terap.*, vol. 3, no. 2, pp. 1–7, 2014.

[2] Bania Amburika, Y. H. Chrisnanto, and W. Uriawan, "Teknik Vector Space Model (VSM) dalam Penentuan Penanganan Dampak Game Online Pada Anak," *Pros. SNST ke-7 Tahun 2016*, vol. 1, no. 1, pp. 10–27, 2016, doi: 10.1103/PhysRevC.6.1023.

[3] Anna and A. Hendini, "Implementasi Vector Space Model Pada Sistem Pencarian Mesin Karaoke," *Evolusi J. Sains dan Manaj.*, vol. 6, no. 1, pp. 1–6, 2018, doi: 10.31294/evolusi.v6i1.3535.

[4] F. Amin, "Sistem Temu Kembali Informasi dengan Pemeringkatan Metode Vector Space Model," *J. Teknol. Inf. Din.*, vol. 18, no. 2, pp. 122–129, 2017, doi: 10.22441/fifo.v9i1.1444.

[5] C. M. Pasma, U. D. Rosiani, and R. Ariyanto, "Pengembangan Sistem Pendeteksi Kemiripan Karya Pada Inaicta 2013," *J. Inform. Polinema*, vol. 1, no. 4, p. 14, 2015, doi: 10.33795/jip.v1i4.117.

[6] M. Turland, *php|architect's Guide to Web Scraping with PHP*. Toronto: Marco Tabini & Associates, Inc., 2010.

[7] A. Josi, L. A. Abdillah, and Suryayusra, "Penerapan teknik web scraping pada mesin pencari artikel ilmiah," *ArXiv14105777 Cs*, pp. 159–164, 2014.

[8] "Web Site Scraper - The Most Effective Tool for Web Data Extraction," *The Computer Advisor*. .

[9] N. Juliasari and J. C. Sitompul, "Aplikasi Search Engine Dengan Metode Depth First Search (DFS)," *J. Tek. Inform. Univ. Budi Luhur. ISSN 1693 -9166*, vol. 9, no. 1, pp. 9–12, 2012, doi: 10.1109/20.312267.

[10] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. 2004.

[11] C. D. Manning, P. Raghavan, and H. Schütze, "Introduction to Modern Information Retrieval (2nd edition)," *Cambridge Univ. Press*, vol. 53, no. 9, pp. 462–463, 2009, doi: 10.1108/00242530410565256.

[12] E. Dragut, F. Fang, A. Sistla, C. Yu, and W. Meng, "Stop Word and Related Problems in Web Interface Integration.," *PVLDB*, vol. 2, pp. 349–360, Aug. 2009, doi: 10.14778/1687627.1687667.

[13] L. Bradji and M. Boufaida, "A Rule Management System for Knowledge Based Data Cleaning," *Intell. Inf. Manag.*, vol. 3, pp. 230–239, Jan. 2011, doi: 10.4236/iim.2011.36028.

[14] M. Adriani, J. Asian, B. Nazief, S. M. M. Tahaghoghi, and H. E. Williams, "Stemming Indonesian: A Confix-Stripping Approach," *ACM Trans. Asian Lang. Inf. Process.*, vol. 6, no. 4, pp. 1–33, Dec. 2008, doi: 10.1145/1316457.1316459.

[15] K. Akromunnisa and R. Hidayat, "Klasifikasi Dokumen Tugas Akhir (Skripsi) Menggunakan K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 4, no. 1, pp. 69–75, 2019.

[16] M. A. Hall and L. A. Smith, "Feature Selection for Machine Learning: Comparing a Correlation-Based Filter Approach to the Wrapper," in *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, 1999, pp. 235–239.

[17] R. Hidayat and S. Minati, "Comparative Analysis of Text Mining Classification Algorithms for English and Indonesian Qur'an Translation," *IJID (International J. Informatics Dev.*, vol. 8, no. 1, pp. 47–51, 2019.

[18] A. Deviyanto and M. D. R. Wahyudi, "Penerapan Analisis Sentimen Pada Pengguna Twitter Menggunakan Metode K-Nearest Neighbor," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 3, no. 1, p. 1, 2018, doi: 10.14421/jiska.2018.31-01.

[19] G. Salton, *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley, 1989.

[20] S. Tata and J. Patel, "Estimating the Selectivity of tf-idf based Cosine Similarity Predicates," *Sigmod Rec.*, vol. 36, Jun. 2007, doi: 10.1145/1361348.1361351.

[21] M. Habibi and P. W. Cahyo, "Journal Classification Based on Abstract Using Cosine Similarity and Support Vector Machine," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 4, no. 3, pp. 48–55, 2020.

[22] E. Garcia, "The Classic TF-IDF Vector Space Model," 2006.

[23] A. Chaer, *Linguistik umum*. Jakarta: Rineka Cipta, 1994.