Indonesian Journal of Electrical Engineering and Computer Science
Vol. 23, No. 2, August 2021, pp. 1049~1058
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v23.i2.pp1049-1058

Determining subject headings of documents using information retrieval models

Evi Yulianti, Laksmita Rahadianti
Faculty of Computer Science, Universitas Indonesia, Indonesia

Article Info

Article history:

Received Apr 30, 2021
Revised Jul 4, 2021
Accepted Jul 7, 2021

Keywords:

Document retrieval
Information retrieval
Language model
Subject heading
Vector space model

ABSTRACT

Subject heading is a controlled vocabulary that describes the topic of a document, which is important to find and organize library resources. Assigning appropriate subject headings to a document, however, is a time-consuming process. We therefore conduct a novel study on the effectiveness of information retrieval models, i.e., language model (LM) and vector space model (VSM), to automatically generate a ranked list of relevant subject headings, with the aim to give a recommendation for librarians to determine the subject headings effectively and efficiently. Our results show that there are a high number of our queries (up to 61%) that have relevant subject headings in the ten top-ranked recommendations; and on average, the first relevant subject heading is found at the early position (3rd rank). This indicates that document retrieval methods can help the subject heading assignment process. LM and VSM are shown to have comparable performance, except when the search unit is title, VSM is superior to LM by 8-22%. Our further analysis exhibits three faculty pairs that are potential to have research collaboration as their students' thesis often have overlap subject headings: i) economy and business-social and political sciences, ii) nursing-public health, and iii) medicine-public health.

Corresponding Author:

Evi Yulianti
Faculty of Computer Science
Universitas Indonesia
Kampus UI Depok, Jawa Barat, Indonesia
Email: evi.y@cs.ui.ac.id

1. INTRODUCTION

Universitas Indonesia (UI) Library archives some document collections, such as: thesis, and book. It is crucial to organize these collections in order to help users easily find the resources that they need. One way to organize library collection is called cataloguing, conducted by creating bibliographic records to represent documents [1]. A part of cataloguing process that is usually performed by librarians is subject analysis (or subject cataloguing), which require them to find the subject headings of a particular library resource [1]-[4].

Subject heading is a controlled topical vocabulary that is assigned to bibliographic records to find and organize library materials based on topics [5]-[7]. It is important especially when searchers have limited information about the title or authors of documents, so that they can still find documents by their topics. A few studies have investigated the benefit of subject headings for library users [8], [9]. Lee-Smeltzer and Hackleman [8] found that library users preferred keyword searching in subject fields over other catalog searching methods. later, sapon-white and hansbrough [9] showed that the dissertations with subject headings are found to be more likely to circulate than those without subject headings. Some researchers have also studied subject heading for various purposes, such as: analysing subjects listed in different subject heading

list [10]-[12], generating map of science [13], [14], developing simplified subject heading list [15], [16], and assigning subject heading automatically [17]-[19].

Determining an accurate subject heading for a document is a complex process. It is because a librarian needs to determine the topic/subject of the document by conducting a thorough subject analysis; then find controlled vocabularies from the subject heading list to assign appropriate subject headings [1]. Therefore, not all librarians are capable to do this task accurately [20]. In addition, different persons may use different interpretations of the subject as they use different vocabularies to represent it. This issue is confirmed in the results of previous study that shown that humans often assigning classes or keywords to bibliographic records in an inconsistent manner, both compared to themselves and compared to other humans [21].

Based on the aforementioned problem, this research aims to help librarians to conduct the subject heading assignment process by providing a recommendation list of subject headings for a given document. The librarians then can look at the recommendation list and decide which subject headings are relevant to be assigned to a document. It is expected that librarians can find relevant subject headings from the recommendation list, or at the very least they can obtain some insight about the topics of similar documents in the collection. Finally, the goal of this study is to enable the subject heading assignment process to run more effectively and efficiently by librarians.

Our work is different to text categorization [22], [23] since we do not categorize documents into certain topics. Some previous work that are mostly related to our work are summarized in Table 1. They investigated the ways to assign subject headings automatically [17]-[19]. Our work is different to them in which we formulate our task as a subject heading recommendation/retrieval instead of subject heading assignment or classification. Therefore, we do not assign a subject heading to a document, but we rank the subject headings of similar documents in the collection in order to generate subject headings recommendation list. In a typical document retrieval system, the search results are the ranked list of documents, however in this work, the results are the ranked list of subject headings from similar documents. The evaluation is therefore performed on the unit of subject headings instead of documents, which further differentiates this work with a standard document retrieval task.

Table 1. Comparison with some mostly related work

	Task	Method	Type
Kavuluru and He [17]	Subject heading assignment	Using named entity recognition, relationship extraction, and output label co-occurrence frequencies of medical subject heading term pairs.	Unsupervised
Kavuluru and Lu [18]	Subject heading assignment & classification	Using term co-occurrence frequencies and latent term association	Unsupervised Supervised
Golub *et al.* [19]	Subject heading classification	Using machine learning algorithm (e.g., SVM)	Supervised
Our work	Subject heading recommendation / retrieval	Using information retrieval methods, i.e., LM and VSM	Unsupervised

Our idea is supported by Golub *et al.* [19] who concluded that automatic subject heading assignment should never be implemented on its own; instead, a system should combine the efficiency of automatic suggestions with quality of human decisions at the final stage. They found that applying purely automatic subject heading classification does not work, because there are a large number of subject headings classes. For this reason, they limit to use only top three levels in dewey decimal classification (DDC) subject heading list (by using 803 instead of 14,413 classes), which therefore does not reflect the operational system. Based on this reason, we propose to produce a system that can recommend a list of relevant subject headings, and then let the librarians to decide which of those are appropriate to be assigned to a given document. Here, we use information retrieval methods, i.e., query likelihood language model (LM) [24], [25] and vector space model (VSM) [26], [27], to generate the recommendation list.

In summary, our contributions are as follows: i). We conduct a novel study on the effectiveness of information retrieval models, i.e., LM and VSM, to automatically generate a ranked list of relevant subject headings, with the goal to give a recommendation for librarians in determining the subject headings effectively and efficiently. To the best of our knowledge, none of the previous published work have reported about this study. ii). We perform an empirical evaluation to test the effectiveness of LM and VSM retrieval methods in generating subject headings recommendation using our collection of bachelor thesis from Universitas Indonesia and three different search units: title, abnstract, and title+abstract. iii). We analyze different faculties that often share similar subject headings. The result of this analysis can suggest the potential research collaboration between these faculties.

2. PROBLEM STATEMENT

The task that is studied in this work is subject heading recommendation/retrieval problem, aimed to help librarians in assigning subject headings more effectively and efficiently. Our general framework is illustrated in Figure 1. Given a document metadata (that will be assigned the subject headings) as a query, we will retrieve a ranked list of subject headings from the most similar document metadata in the collection using LM and VSM retrieval methods as the output. It is expected that the output can display relevant subject headings in the early position of the recommendation list, so that librarians can find them quickly. The basic idea of utilizing document retrieval methods to generate subject heading recommendation is that "two documents that are assigned similar subject headings usually describe about the same topics and therefore they may have high document similarity". The similarity between documents is assessed using title and abstract text representation, since our dataset does not have the full text of documents (see Section 4.1).

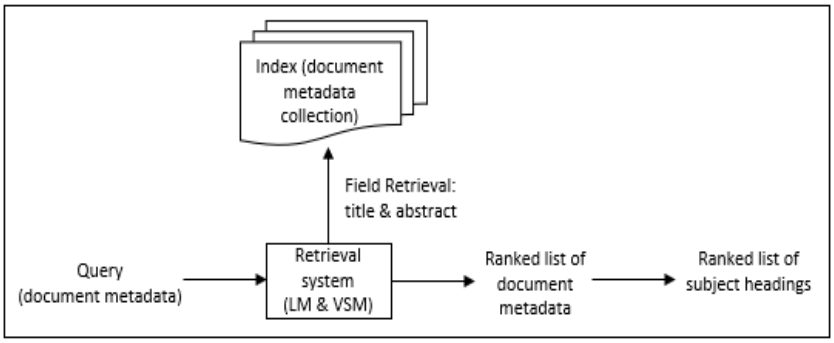

Figure 1. Our general framework

Initially our indexing data is indexed using field indexing [28], [29], so we have title and abstract fields in the index. This enables us to search on title and abstract parts individually. More specifically, the title of queries is compared against the title field of the index, and the abstract of queries is compared against the abstract field of the index. This way, in general we calculated similarity on the titles and abstracts of documents. The LM and VSM retrieval methods will rank *k* document metadata in the collection that are most similar to the query according to the titles and abstracts. Then, the subject headings of the *k* top-ranked documents are extracted as the subject heading recommendation list. Note that a metadata may have several subject headings. As a result, there could be more than one subject heading at every rank in the top *k* recommendation list. The number of metadata retrieved in the top rank (*k*) can be set according to people needs. In our experiments, we retrieve 1, 5, and 10 documents in the top rank, and take their subject headings in the recommendation list.

3. RETRIEVAL METHODS

3.1. Query likelihood language model (LM)

Language model (LM) uses probabilistic concept to estimate the probability of relevance of documents to user queries [24], [25]. A language model M_d for each document *d* is inferred and then used to estimate the probability of generating the query according to the model $p(q|M_d)$. This probability is also referred to as query generation probability. Documents in the collection are then ranked according to these probabilities. The higher the probability of a document, the more probable the query can be generated from the document, which implies that the more relevant the document to the query. The probability of producing the query q given the language model of document *d*, i.e., M_d, using Maximum Likelihood Estimate and Dirichlet smoothing can be computed as follows (1):

$$\log p(q|M_d) = \sum_{t \in q} \log \frac{tf_{t,d} + \mu \frac{cf_t}{|C|}}{|d| + \mu} \quad (1)$$

where *t* is a term in query q, $tf_{t,d}$ denotes the frequency of term *t* in query q, $p(t|M_d)$ is probability of generating term *t* from document language model M_d, and $|d|$ denotes total number of terms in document *d*, $|C|$ denotes the total terms in the collection, cf_t denotes the frequency of term *t* in the collection. Dirichlet smoothing is a default smoothing technique provided in the search engine library Indri (default μ value is 2500).

3.2. Vector space model (VSM)

Vector space model (VSM) represents documents as vectors of terms in the vector space [26], [27]. Cosine similarity is then used to calculate the similarity between two document vectors in the vector space. When two documents share many similar words, then their vectors in the vector space is close (the angle between them is small). As a result, the cosine of the angle between them is high. Two identical documents will have cosine similarity score of 1. The greater the similarity between two documents, the cosine similarity score will be getting closer to 1. The range of cosine similarity score is between 0 and 1. The formula to compute cosine similarity is as follows (2):

$$sim(d_a, d_b) = \frac{\sum_{k=1}^{max\,(na,nb)} w_{ka} . w_{kb}}{\sqrt{\sum_{k=1}^{na} {w_{ka}}^2 . \sum_{k=1}^{nb} {w_{kb}}^2}} \quad (2)$$

where w_{ka} is the weight of term k in document d_a; w_{kb} is the weight of term k in document d_b; na is the number of terms in document d_a; and nb is the number of terms in document d_b. The weight of terms is calculated using TF-IDF formula.

4. EXPERIMENT

4.1. Dataset

This work uses a metadata collection of bachelor thesis of UI students as our dataset. This metadata collection was crawled from UI library website (http://lib.ui.ac.id/) in the section "UI - Skripsi (Open)" using a program written in Perl and LWP module. This collection contains document metadata of bachelor thesis from 14 faculties in UI during 2008-2011, with a total of 14,722 metadata. It is important to note that we used document metadata, instead of full text, because most of documents in the website do not have full text (and some full text of documents cannot be accessed since they are securely encrypted).

We only crawl four information for each metadata that is needed by our method: title, abstract, faculty, and subject heading. Abstract contains the summary of the thesis. Title and abstract information are important as the metadata representation for indexing and querying purpose. Related metadata in the collection are determined by our methods based on the similarity on these text. Faculty information is important for our analysis on cross-faculty similar documents. Subject heading information is surely important as this will be extracted from related metadata to produce recommendation. In addition, it is also used as a gold standard in the evaluation process. It is important to note that one document may be assigned several subject headings.

Figure 2 describes an example of subject headings from two metadata in our dataset. To help readers understand the context of the metadata where this subject heading come from, we also display the title of metadata. These two metadata have more than one subject heading that are delimited with ";" character. For the second metadata, it has the subject heading "Maternal-Child Nursing" that is composed of main subject heading "Maternal" and additional subject heading "Child Nursing".

Title: *Analisis selectivity dan market timing pada reksa dana saham di Indonesia: periode 2005-2007* (Analysis of the selectivity and market timing on equity funds in Indonesia: 2005-2007 period) **Subject headings:** Mutual funds; Shares
Title: *Hubungan tingkat pengetahuan suami dengan praktik pemberian ASI eksklusif di RW 25 Baktijaya Sukmajaya Depok* (The correlation between husband's knowlegde levels of exclusive breastfeeding practice in RW 25 Baktijaya, Sukmajaya, Depok) **Subjet headings:** Breastfeeding; Maternal--Child Nursing

Figure 2. Examples of subject headings of documents

As we require each metadata in our dataset to have title, abstract, subject heading, and faculty information, so we exclude all metadata that do not have any of this information. It results in 9,261 metadata in our dataset. The statistics of total documents for each faculty is described in Table 2. We can see from the table that most metadata in our dataset comes from Faculty of Engineering. This may be due to the high number of study program in this faculty which results in the high number of students. According to the data of UI's bachelor student admission in 2020, the Faculty of Engineering has the highest number of students among all faculties in UI. To ensure that each faculty has enough metadata, we decide to exclude the ones that have a small number of metadata. We make a justification of 100 documents as our lower bound of the

number of metadata for each faculty. This causes the metadata from four faculties (i.e., Faculty of Pharmacy, Faculty of Computer Science, Faculty of Dentistry, and Faculty of Administrative Science) to be filtered out from our dataset. They respectively have 95, 94, 84, and 35 documents. Finally, there are 8,953 metadata that comes from 10 faculties at UI in our final dataset.

Table 3 describes the term statistics in in our dataset. On average, title contains around 23 terms, which is quite long for a title. We found this happened because some metadata has a mixed of Indonesian and English titles in the website, which makes the overall title long. The average term in abstract is around 205 words. The average number of subject headings in a metadata is 1.5. This denotes that most of metadata are assigned more than one subject headings that consists of 3 terms on average.

Table 2. The statistics of total metadata for each faculty in our dataset

Faculty	Abbreviation	#study program	#document metadata
Faculty of Engineering	Eng	13	2,400
Faculty of Humanities	Hum	15	1,487
Faculty of Mathematics and Natural Science	Math	9	1,188
Faculty of Social and Political Sciences	Soc	7	1,081
Faculty of Public Health	PubH	4	971
Faculty of Law	Law	1	712
Faculty of Economy and Business	Eco	5	521
Faculty of Psychology	Psy	1	253
Faculty of Medicine	Med	1	219
Faculty of Nursing	Nurs	1	121
Total			9,261

Table 3. The statistics of terms and subject headings in our dataset

Average terms for each title	23.12
Average terms for each abstract	205.16
Average terms for each subject heading	3.17
Average number of subject headings for each metadata	1.51

Our dataset is split into 2 parts: indexing data and queries. This is important because we have to ensure that the metadata to be queried (evaluated) do not appear in the index. Otherwise, the retrieval system will always obtain perfect match as the most-relevant metadata found is basically the metadata input itself. For this reason, the dataset has to be split. We take a random sample of 30 metadata for each faculty, resulting 300 metadata as the queries. Then the remaining data with the total of 8,653 metadata is used for indexing.

4.2. Implementation

We use a search engine library Indri version 5.12 for indexing and retrieval using LM retrieval method. Then, we use a search engine library Lucene version 8.7.0 to implement indexing and retrieval using VSM retrieval method. For each method, index is built for title and abstract fields. When the search unit is title, we compute title similarities between query's title and documents' title in the index. When the search unit is abstract, we compute abstract similarities between query's abstract and documents' abstract in the index. When the search unit is title+abstract, then we compute both title similarities and abstract similarities between query and documents in the index. To examine which one is more important between title relevance and abstract relevance, we experimented with different γ values in the range of 0-1 with the step of 1. Indri query operators *#combine* and *#weight* is used for computing title and abstract relevance using LM. #combine operator is used to combine all terms in the title or abstract, while #weight operator is used to give the importance of title relevance and abstract relevance. The following is Indri query formula for field retrieval:

$$\#weight(\; (1-\gamma)\ \#combine[title](t_{1,title}\ t_{2,title}\ \dots\ t_{m,title}) \; \gamma\ \#combine[abstract](t_{1,abstract}\ t_{2,abstract}\ \dots\ t_{n,abstract}))$$

The $t_{i,j}$ indicates the *i*-th term in the *j* field. The γ value denotes the importance of title relevance and abstract relevance in the relevance. The lower the γ value, the more importance the title relevance to the final relevance score. When γ equals to 0, then the querying process only consider the title relevance. Otherwise, when γ equals to 1, then we only take into account the abstract relevance.

To compute title and abstract relevance using VSM method, we use *BoostQuery* class in Lucene library. It needs two parameters: instance of *TermQuery* which contains the query terms and the index field in which the search will be performed, and γ value which denotes the weight of the relevance score for the query terms.

4.3. Evaluation

Evaluation is performed by comparing the subject headings in the recommendation list with the ground truth subject headings, which are obtained from the actual subject headings of the queries. Recall that a metadata may have several subject headings. Therefore, if subject headings retrieved in the specified position of the recommendation list contains any ground truth subject headings of the queries, then we consider such retrieved subject headings as relevant. For example: given the actual subject headings for a query is "*Breastfeeding; Maternal--Child Nursing*", then we consider there are three ground truth subject headings for this query: "*breastfeeding*", "*maternal*", and "*child nursing*". If subject headings retrieved in the first rank of the recommendation is "*Breastfeeding--Social aspects; Breast milk*", then it means that it consists of three subject headings: "*breast feeding*", "*social aspect*s", and "*breast milk*". Since it contains one ground truth subject heading for the given query, i.e., "*breast feeding*", then it is considered as relevant.

Some retrieval measures are used to evaluate the effectiveness of LM and VSM retrieval methods to retrieve the k top-ranked subject headings in the recommendation list R_k. In our experiment, we set k value to 1, 5, and 10 as we focus on the subject headings extracted from highly ranked documents.

Found@k measure is proposed by authors to demonstrate the number of queries that the relevant subject headings can be found in the recommendation list R_k. The higher the Found@k score, the more probable the system to retrieve relevant subject headings for queries. This measure may suggest the applicability of the system to help in the subject heading assignment process.

$$Found@k = \#query\ that\ has\ relevant\ subject\ headings\ in\ R_k$$

Precision@k measure highlights how accurate the method to put many relevant subject headings in the top rank of the recommendation list R_k. This is a standard metric that has been used in many information retrieval researches in previous work.

$$P@k = \frac{\#relevant\ subject\ headings\ in\ R_k}{k}$$

MRR measure observes the reciprocal rank of the position where the first relevant subject heading is found. The higher the MRR score indicates that the relevant subject headings can be found in the early position in the recommendation list. The lower the score indicates that the relevant subject headings are retrieved in the lower rank in the recommendation list, which causes users need to scroll down the list further in order to find the relevant results. In this work, we limit the MRR calculation to the 10 top-ranked subject headings as this is the highest number of retrieved documents in our experiment. This is a standard metric that has been used in many information retrieval researches in previous work.

$$MRR = \frac{1}{rank\ of\ the\ first\ relevant\ subject\ heading\ found}$$

5. RESULTS AND DISCUSSION

5.1. Effectiveness of LM and VSM methods in generating subject heading recommendation list

The *Found@k* scores for LM and VSM methods are shown in Figure 3. The results for Title+Abstract are the best results among all γ value between 0-1 with the step of 1. The results using each of these γ values are presented later in Section 5.2. For both methods, we can see that using abstract as a search unit produces more effective recommendation rather than using title in most cases. This is because taking the similarities between abstracts are more accurate to determine the similarity between two metadata documents, as opposed to taking the similarities between titles. Since abstracts are more detail, therefore two metadata with high similarity in abstracts tend to be more similar than those with high similarity in title. Using both title and abstract as a search unit can slightly improve the results of using abstract only. This indicates that combining similarities between titles and similarities between abstracts is more accurate to find relevant metadata. When using title only as the search unit, VSM is shown to be little superior to LM for all Found@k metrics by 8-22%. This can be understood because LM uses probabilistic concept in which it is less accurate when the size of data used to produce the probability score is small. Note that when the search unit is title only, then the language model used to compute query generation probability is built only using

metadata titles in the collection. When the search unit is abstract only or title+abstract, both LM and VSM methods seem to be comparable. This result demonstrates that LM and VSM are comparable when using longer text as the search unit, but VSM is preferred when using smaller text.

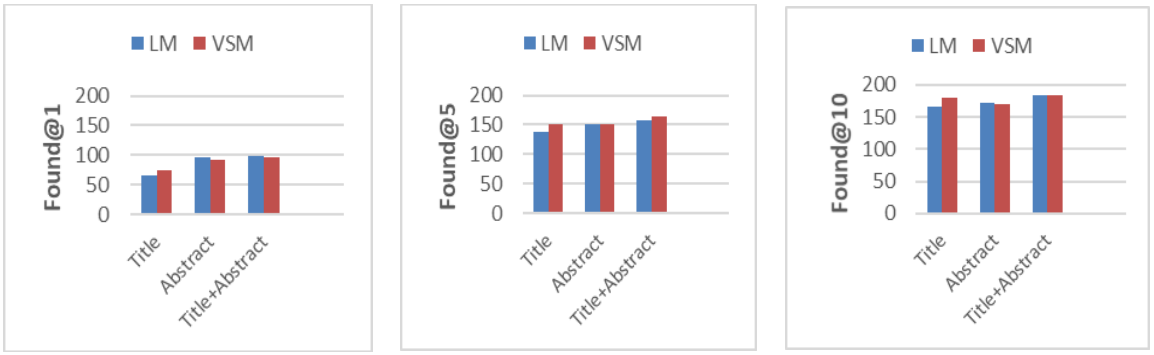

Figure 3. The number of queries that the relevant subject heading can be found in the recommendation list

Table 4 describe the effectiveness of LM and VSM methods based on standard retrieval measures. We can see that in general, this result is consistent to that presented in Figure 3. For both methods, using abstract as a search unit gains higher retrieval scores than using title. Then, combining similarities between titles and similarities between abstracts can further slightly enhance the results. An important finding from the table is that based on MRR scores, on average the first relevant subject heading is found at the 3rd rank in the recommendation list. This can benefit librarians as they can find relevant subject in the top rank of the recommendation list, so it could save their time.

An additional experiment was performed by performing stopping and stemming (using Sastrawi tools) in the pre-processing step to understand the effect of these steps to the effectiveness of subject headings recommendation results. However, we found that the effect of this steps to the results is not clear. In some cases, it can slightly increase the scores (but it is not statistically significant), and in some cases, it can decrease the score. Because of page length constraint, the results of this experiment are not shown in this paper.

Table 4. The retrieval effectiveness of LM and VSM according to standard retrieval measures

Search Unit	LM				VSM			
	MRR	P@1	P@5	P@10	MRR	P@1	P@5	P@10
Title	0.324	0.220	0.164	0.154	0.359	0.247	0.200	0.167
Abstract	0.407	0.323	0.215	0.184	0.391	0.310	0.219	0.184
Title+Abstract	0.410	0.330	0.229	0.199	0.409	0.320	0.232	0.197

5.2. Effectiveness of LM and VSM methods on different γ values

To examine the most optimal γ value for LM and VSM methods, we conduct an experiment using varying γ values between 0 and 1 with a step of 0.1. The results are reported in Table 5. We found different γ values that results in the best performance for LM and VSM methods. The most optimal γ values for LM are between 0.5 and 0.9, while for VSM it is between 0.1 and 0.3. So, the optimal γ value for LM tend to be middle or high, and it tends to be low for VSM. Note that the higher the γ values, the more importance the abstract similarities. Otherwise, the lower the γ values, the more importance the title similarities.

The optimal γ value for LM tends to be middle or high because as has been shown earlier, using title similarities in LM produces much worse results than using abstract similarities. This is because LM uses probabilistic concept in which it is less accurate when the size of data used to produce the probability score is small. Therefore, using abstract similarities gives much improvement over using title similarities, which is by 30.5% on average, since the query generation probability is built using bigger data. Then for VSM, the optimal γ value for LM tend to be low because as has been shown in the earlier section, using title similarities in VSM already produced fairly good results. This is because VSM can capture well the document similarities even by using small text unit, since it is based on term overlap between two documents that is computed using dot product of the documents' vectors, i.e., cosine similarity. Therefore, using abstract similarities only give slight improvement over using title similarities, which is by 14.6% on average.

Table 5. The effectiveness of LM and VSM on different γ values

	γ	Found @1	Found @5	Found @10	MRR	P@1	P@5	P@10
LM	0.1	73	144	172	0.343	0.243	0.182	0.163
	0.2	77	148	173	0.354	0.257	0.193	0.172
	0.3	83	152	179	0.374	0.277	0.205	0.182
	0.4	88	152	179	0.387	0.293	0.210	0.190
	0.5	91	151	184	0.396	0.303	0.219	0.198
	0.6	94	158	183	0.404	0.313	0.229	0.199
	0.7	94	157	177	0.401	0.313	0.227	0.198
	0.8	97	150	175	0.407	0.323	0.222	0.191
	0.9	99	154	171	0.410	0.330	0.223	0.188
VSM	γ	Found @1	Found @5	Found @10	MRR	P@1	P@5	P@10
	0.1	88	164	184	0.397	0.293	0.227	0.196
	0.2	92	161	180	0.404	0.307	0.231	0.195
	0.3	96	159	180	0.409	0.320	0.232	0.197
	0.4	93	155	177	0.398	0.310	0.225	0.196
	0.5	94	153	175	0.397	0.313	0.223	0.194
	0.6	93	155	171	0.395	0.310	0.227	0.192
	0.7	93	152	173	0.396	0.310	0.223	0.188
	0.8	93	153	169	0.393	0.310	0.221	0.188
	0.9	93	150	170	0.392	0.310	0.218	0.185

5.3. Analysis of subject headings overlaps of documents from different faculties

It is interesting to examine whether it is possible that a query has relevant subject headings that come from documents of different faculties. The findings from this analysis may give insights on different faculties that may share similar topics in their students' thesis. For this analysis, we examine the subject headings recommendation generated from top 10 retrieved metadata for all queries. We analyze the faculty of queries and the faculty of those documents metadata that share similar subject headings. As there are 10 retrieved metadata and 300 queries, therefore in total, there are 3000 faculty pairs to be examined. We found there are 4% cases where the faculty of queries is different to the faculty of top 10 retrieved metadata, but they share similar subject headings. Here, there are 133 cases (using LM) and 128 cases (using VSM).

Both LM and VSM agree on the top five pairs of faculties that often have overlap subject headings as shown in Table 6. This finding can therefore encourage the possibility of research collaboration between these faculties. The pair of faculties that is most often to have similar subject headings are Faculty of Economy and Business and the Faculty of Social and Political Science. We analyze that although two metadata from these two faculties contain similar topics, but they discuss them on different perspectives. Figure 4 shows an example of two metadata from Faculty of Economy and Faculty of social and political science that share similar subject heading, i.e., mutual funds. We can see from the titles of these metadata that although both metadata contains information about market timing, stock selection, and equity funds, but the first metadata discuss them on economy perspectives, while the second metadata discuss them on social perspectives.

Table 6. Five pairs of different faculties that are most often to have similar subject headings in our results

LM		VSM	
Faculty Pairs	Total	Faculty Pairs	Total
Eco - Soc	29	Eco - Soc	31
Nurs-PubH	28	Nurs-PubH	27
Med-PubH	13	Med-PubH	10
Med-Nurs	11	Med-Nurs	10
Nurs-Psy	7	Nurs-Psy	6

Query (Document metadata input):
Doc ID: 20320502
Title: *Analisis selectivity dan market timing pada reksa dana saham di Indonesia: periode 2005-2007* (Analysis of the selectivity and market timing on equity funds in Indonesia: 2005-2007 period)
Faculty: Faculty of Economy and Business
Ground truth subject heading: Mutual funds; Shares

Subjet headings recommendation at rank #1: Mutual funds—Indonesia
This is extracted from the following document metadata:
Doc ID: 20320288
Title: *Pengaruh Market Timing dan Stock Selection terhadap kinerja Reksa Dana Saham di Indonesia periode 2007-2011* (The effect of Market Timing and Stock Selection on the performance of Equity Funds in Indonesia 2007-2011)
Faculty: Faculty of Social and Political Sciences

Figure 4. Example of two metadata of different faculties that share similar subject headings

In future, user studies could be performed to test whether librarians can find relevant subject headings from the recommendation list or not. We can also ask their preference on the subject headings recommendation generated using LM and VSM methods. Different retrieval methods, such as: probabilistic BM25 [30], [31] and sequential dependence model (SDM) [32], [33] can be considered for further study. The possibility of combining semantic information [25], [34], [35] or social media [36], [37] into the retrieval model to address lexical mismatch problem is a great challenge to be explored.

6. CONCLUSION

We conduct a novel study on the use of document retrieval methods, i.e., query likelihood language models (LM) and vector space models (VSM), to generate subject headings recommendation list for a given document. The aim of this study is to help librarians to assign appropriate subject headings to a document more effectively and efficiently. Our results shows that information retrieval methods can help the subject heading assignment process. There are a high number of queries (up to 61%) that have relevant subject headings in the ten top-ranked recommendation list, which suggest the applicability of the system. Next, on average, the first relevant subject heading is found in the 3rd rank in the recommendation list. This is an interesting finding that indicates that librarians can find relevant subject headings quickly as they do not need to scroll down too far in the recommendation list. Both LM and VSM has comparable effectiveness when using longer text as the search unit. However, VSM is superior to LM by 8-21% when using smaller text as the search unit. Our analysis then reveals three pairs of faculties with the highest number of subject headings overlap are: i) Faculty of Economy and Business-Faculty of Social and Political Sciences, ii) Faculty of Nursing-Faculty of Public Health, and iii) Faculty of Medicine-Faculty of Public Health. This finding may suggest the potential research collaboration between these faculties.

ACKNOWLEDGEMENTS

This research is supported by the PUTI Q3 grant number NKB-4379/UN2.RST/HKP.05.00/2020 from Universitas Indonesia.

REFERENCES

[1] G. L. Hoffman, *Organizing Library Collections: Theory and Practice*, Rowman & Littlefield, 2019.
[2] L. Hoover, "A beginners' guide for subject analysis of theses and dissertations in the hard sciences," *Cataloging & classification quarterly*, vol. 41, no. 1, pp. 133–161, 2005, doi: 10.1300/J104v41n01_07.
[3] A. J. Spencer and J. D. Eldredge, "Roles for librarians in systematic reviews: a scoping review," *Journal of the Medical Library Association: JMLA*, vol. 106, no. 1, pp. 46-56, 2018, doi: 10.5195/jmla.2018.82.
[4] R. Sapon-White, "Subject analysis training for cataloging paraprofessionals: A model for ongoing learning and support," *Technical Services Quarterly*, vol. 26, no. 3, pp. 183–193, 2009, doi: 10.1080/07317130802520013.
[5] C.-A. Julien, B. Asadi, J. D. Dinneen, and F. Shu, "Library of congress subject heading (LCSH) browsing and natural language searching," *Proceedings of the Association for Information Science and Technology*, vol. 53, no. 1, pp. 1–4, 2016, doi: 10.1002/pra2.2016.14505301116.
[6] C. D. Batty, *An introduction to the Dewey decimal classification*, Asia Publishing House (1966), 2017.
[7] A. M. Ferris, "Birth of a Subject Heading," *Library Resources & Technical Services*, vol. 62, no. 1, p. 16, 2018.
[8] J. Lee-Smeltzer and D. Hackleman, "Access to OSU Theses and Dissertations in Kerr Library: How They Are Used or Are They?" *Technical Services Quarterly*, vol. 12, no. 4, pp. 25–43, 1995, doi: 10.1300/J124v12n04_03.
[9] R. E. Sapon-White and M. Hansbrough, "The impact of subject heading assignment on circulation of dissertations at Virginia Tech," *Library resources & technical services*, vol. 42, no. 4, pp. 282–291, 1998, doi: 10.5860/lrts.42n4.282.
[10] M. Adler, J. T. Huber, and A. T. Nix, "Stigmatizing disability: Library classifications and the marking and marginalization of books about people with disabilities," *The Library Quarterly*, vol. 87, no. 2, pp. 117–135, 2017, doi: 10.1086/690734.
[11] Y.-L. Lu and D. W. Bianchi, "Trends in prenatal diagnosis: An analysis of 40 years of Medical Subject Heading (MeSH) terms in publications," *Prenatal Diagnosis*, vol. 40, no. 13, pp. 1636–1640, 2020, doi: 10.1002/pd.5871.
[12] Y.-J. Kim, "Comparison of author key words and Medical Subject Heading terms in the Journal of Korean Society of Dental Hygiene from 2001 to 2015," *Journal of Korean society of Dental Hygiene*, vol. 18, no. 6, pp. 1047-1055, 2018, doi: 10.13065/jksdh.20180090.
[13] F. Shu, J. D. Dinneen, B. Asadi, and C.-A. Julien, "Mapping science using library of congress subject headings," *Journal of Informetrics*, vol. 11, no. 4, pp. 1080–1094, 2017, doi: 0.1016/j.joi.2017.08.008.
[14] F. Shu, J. Qiu, and V. Larivière, "Mapping the Life Science using Medical Subject Headings (MeSH)," in *17th International Conference on Scientometrics & Informetrics*, 2019, pp. 1927–1932.
[15] J. D. Dinneen, B. Asadi, I. Frissen, F. Shu, and C.-A. Julien, "Improving exploration of topic hierarchies: Comparative testing of simplified library of congress subject heading structures," in *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*, 2018, pp. 102–109, doi: 10.1145/3176349.3176385.

[16] C.-A. Julien, P. Tirilly, J. D. Dinneen, and C. Guastavino, "Reducing subject tree browsing complexity," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 11, pp. 2201–2223, 2013, doi: 10.1002/asi.22915.
[17] R. Kavuluru and Z. He, "Unsupervised medical subject heading assignment using output label co-occurrence statistics and semantic predications," in *International Conference on Application of Natural Language to Information Systems*, 2013, pp. 176–188, doi: 10.1007/978-3-642-38824-8_15.
[18] R. Kavuluru and Y. Lu, "Leveraging output term co-occurrence frequencies and latent associations in predicting medical subject headings," *Data & Knowledge Engineering*, vol. 94, pp. 189-201, 2014, doi: 10.1016/j.datak.2014.09.002.
[19] K. Golub, J. Hagelbäck, and A. Ardö, "Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches," *Journal of Data and Information Science*, vol. 5, no. 1, pp. 18–38, 2020, doi: 10.2478/jdis-2020-0003.
[20] E. Svenonius and D. McGarry, "Objectivity in evaluating subject heading assignment," *Cataloging & classification quarterly*, vol. 16, no. 2, pp. 5–40, 1993, doi: 10.1300/J104v16n02_02.
[21] K. Leininger, "Interindexer consistency in PsycINFO," *Journal of Librarianship and Information Science*, vol. 32, no. 1, pp. 4–8, 2000, doi: 10.1177/096100060003200102.
[22] Y. Su, Y. Huang, and C. J. Kuo, "Efficient Text Classification Using Tree-structured Multi-linear Principal Component Analysis," *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 585-590, doi: 10.1109/ICPR.2018.8545832.
[23] D. A. Meedeniya and A. S. Perera, "Evaluation of Partition-Based Text Clustering Techniques to Categorize Indic Language Documents," *2009 IEEE International Advance Computing Conference*, 2009, pp. 1497-1500, doi: 10.1109/IADCC.2009.4809239.
[24] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *SIGIR '98: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, 1998, pp. 275–281, doi: 10.1145/290941.291008.
[25] E. Yulianti, R.-C. Chen, F. Scholer, and M. Sanderson, "Using semantic and context features for answer summary extraction," in *Proceedings of the 21st Australasian Document Computing Symposium*, 2016, pp. 81–84, doi: 10.1145/3015022.3015031.
[26] G. Salton, A. Wong, and C.-S. Yang, "A vector space model for automatic indexing," *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975, doi: 10.1145/361219.361220.
[27] R. C. Belwal, S. Rai, and A. Gupta, "Text summarization using topic-based vector space model and semantic measure," *Information Processing & Management*, vol. 58, no. 3, 2021, doi: 10.1016/j.ipm.2021.102536.
[28] D. Metzler, T. Strohman, H. Turtle, and W. B. Croft, "Indri at TREC 2004: Terabyte track," Massachusetts Univ Amherst Center For Intelligent Information Retrieval, 2004.
[29] Y. Wang and H. Fang, "Combining Term-based and Concept-based Representation for Clinical Retrieval," *TREC*, 2017.
[30] S. Robertson, H. Zaragoza, and M. Taylor, "Simple BM25 extension to multiple weighted fields," in *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, 2004, pp. 42–49, doi: 10.1145/1031171.1031181.
[31] C. Kamphuis, A. P. de Vries, L. Boytsov, and J. Lin, "Which BM25 do you mean? A large-scale reproducibility study of scoring variants," in *European Conference on Information Retrieval*, 2020, pp. 28–34, doi: 10.1007/978-3-030-45442-5_4.
[32] D. Metzler and W. B. Croft, "A Markov random field model for term dependencies," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 2005, pp. 472–479, doi: 10.1145/1076034.1076115.
[33] E. Yulianti, R.-C. Chen, F. Scholer, W. B. Croft, and M. Sanderson, "Ranking documents by answer-passage quality," in *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018, pp. 335–344, doi: 10.1145/3209978.3210028.
[34] T. V. Rampisela and E. Yulianti, "Semantic-Based Query Expansion for Academic Expert Finding," *2020 International Conference on Asian Language Processing (IALP)*, 2020, pp. 34-39, doi: 10.1109/IALP51396.2020.9310492.
[35] T. V. Rampisela and E. Yulianti, "Academic Expert Finding in Indonesia using Word Embedding and Document Embedding: A Case Study of Fasilkom UI," *2020 8th International Conference on Information and Communication Technology (ICoICT)*, 2020, pp. 1-6, doi: 10.1109/ICoICT49345.2020.9166249.
[36] E. Yulianti, R. Chen, F. Scholer, W. B. Croft and M. Sanderson, "Document Summarization for Answering Non-Factoid Queries," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 1, pp. 15-28, 2018, doi: 10.1109/TKDE.2017.2754373.
[37] E. Yulianti, S. Huspi, and M. Sanderson, "Tweet-biased summarization," *Journal of the Association for Information Science and Technology*, vol. 67, no. 6, pp. 1289–1300, 2016, doi: 10.1002/asi.23496.