Tf*Idf and Random Walk For Term Candidate Selection On Automatic Subject Indexing

Nurseno Bayu Aji[1], Musta'inul Abdi[2], and Ardon Rakhmadi[3]

[1] Department of Electrical Engineering, Politeknik Negeri Semarang, Indonesia

[2] Department of Computer and Information Technology, Politeknik Negeri Lhokseumawe, Indonesia

Abstract—Subject indexing is the act of describing or classifying a document by index terms or other symbols to indicate what the document is about, to summarize its content, or to increase its findability. The selection of term candidates on automatic subject indexing is very important because it can influence the result of topic extraction on the document. Recently, automatic subject indexing especially in the term candidate selection only considers terms in the document collection. In contrast, the indexer prefers to choose a general term on manual subject indexing for the selection of term candidates. In this paper, we proposed a new strategy for selecting term candidates on automatic subject indexing for extraction of the main topic of the document. The proposed method uses a combination of Term Frequency Inverse Document Frequency (TF*IDF) and Random Walk on the structure of the thesaurus. Experimental results show that the proposed method can select the term candidate that relevant to the topic of the document with an F-Measure of 0.24.

Index Terms—TF*IDF, Random Walk, Thesaurus, Automatic Subject Indexing, Term Candidate.

1. Introduction

Study about manual subject indexing using controlled vocabulary began in the 1950s and 1960s, since then the researchers develop a method for automatic subject indexing documents [1]. In the manual document indexing, the indexer (persons/experts who do the indexing of documents) using an external resource to determine the term to accord with the terms that they read before.

Based on that concept, many researchers are trying to develop a document indexing automatically. The approach to automatic indexing is divided into two categories: rule-based and statistic (machine learning) [2]. As a rule-based approach that has been used by much previous research, where the researchers use the formal language for indexing term-based document collections [1, 3 - 4]. That formal language usually refers to the "semantic vocabulary" or "lexical dictionary" which contains the rules to help with natural language indexing term or from a document collection into indexing with the controlled term (thesaurus, dictionary, etc.).

The statistical approach focuses on the co-occurrence of the terms in the document collection. As the use of co-occurrence of the terms that have been indexed with keywords related to the citation [5]. Another development uses statistical techniques to gather the terms controlled natural language in the title and abstract [6 - 7].

The research about the calculation of the value of the co-occurrence between words in the titles and abstracts of documents is as well as the terms that have been in the index to determine the level of the journal [8].

Both approaches describe the relationship between natural language in the document and the term that input manually and controlled without ignoring the structure in the vocabulary. Several studies are using certain terms about vocabulary as a feature in the process of learning a statistical approach. Such features include the TF*IDF [9 - 10].

The concept is applied by Loose and Willis, who suggested the random walk method in thesaurus structure to successfully perform automatic subject indexing using controlled terms (not referring only to the relevant documents) [11]. So getting the index results from a term that near results of manual indexing. Use of Random Walk in the search for a term that is more likely to be used by the indexer manually based on the structure graph of the thesaurus [11].

The use of random walk on the structure of a thesaurus is for the term candidate selection proved to increase the extraction of relevant topics in the document. Thesaurus structures as the relationship between the broader and narrower terms are explored in depth so that the topic of a document can be seen even if the term in question is not contained in the document.

Automatic system subject indexing commonly used technique frequency term. TF*IDF techniques refer to the terms contained in the documents in question and do not yet include the terms of the relationship (broader and narrower) in the thesaurus structure. So the potential term contained within the structure of the thesaurus is not explored in depth.

In this study, we proposed a new strategy for selecting term candidates on automatic subject indexing. The proposed method uses a combination of TF*IDF and Random Walk on the structure of the thesaurus. This method may recover the problem of ranked term and common word relationships (broader) or specific words (narrower) issues.

2. Term Selection Using TF*IDF and Random Walk

There are several stages in the subject automatic indexing using TF*IDF and Random Walk. TF*IDF is a combination of two methods: Term Frequency (TF) and Inverse Document Frequency (IDF) [12 - 13].

*2.1. TF*IDF*

TF*IDF is an information retrieval technique that weighs a term's frequency (TF) and its inverse document frequency (IDF). Each word or term that occurs in the text has its respective TF and IDF score. The product of the TF and IDF scores of a term is called the TF*IDF weight of that term.

TF*IDF evolved from IDF proposed by Sparck Jones namely the reduction of the dominance term that frequently appears on many documents and gives a weight of less than one[14]. TF*IDF method is a method for calculating the weight of each word most commonly used in information retrieval. This method is also well known efficient, easy, and has accurate results [9]. Weighting TF*IDF is a combination of methods Term Frequency (TF) and Inverse Document Frequency (IDF) for each token (word) in every document in the corpus. Eq.1 is the classic formula for weighting TF*IDF

$$W_{i,j} = tf_{i,j} * log(\frac{N}{df_i}),$$

where is a term for weight i document i, N is the number of collection documents, is a term of i term within j document and is document frequency of term i in the collection. In manual subject indexing, the indexer uses items varied and the term may not exist in the document being indexed. The variation of the term that gets by the indexer based on other sources such as dictionaries or other documents had been read.

TF*IDF is an efficient method and has had great results in automatic subject indexing, however, TF*IDF is still affected on one document to be indexed. Thus the random walk method is needed to address the problem. The Random Walk will be discussed in the next chapter.

2.2. Random Walk

Thesaurus which has a broader and narrower relationship can be represented as a graph, where each term has a relationship with another term [11]. The term can be represented by vertex and relations between them can be represented by the edge [15]. Random walk is a method used to perform the movement of a graph, such as moving from a term to another term in the thesaurus nearby. In this case, the random walk will be used to search for terms that are more common in the thesaurus. The starting point of the random walk is a term derived from the thesaurus centric matching the TF*IDF. Of the term that is used as a starting point, the agent will do a random walk in the random walk expedition to get a more general term, which is expected to be the same as the manual indexing in general.

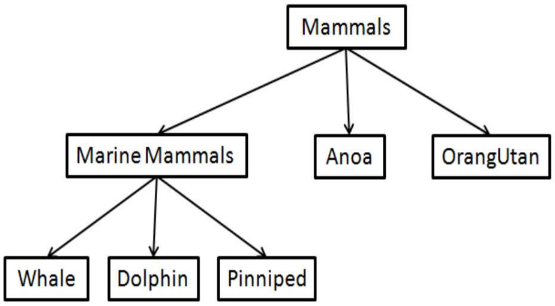


Fig. 1. Sample terms and relationships from a thesaurus.

Figure 1. illustrates a simple set of terms and the relationship between terms in the thesaurus. If we have starting points e.g. "Whale," "Dolphin," and "Pinnipeds," from the starting point that be expected to end in "Whale Mammals." If the starting point "Anoa" and "orangutan" it will be expected that one of the agents will run random and ends at "Mammals." Because the process is done randomly, some agents will end on terms that are not related. The idea is to identify the terms that are most often bypassed by the agent, with the term that is most often bypassed by the agent, we assume that term is a very common term that can be found by the agent.

The effect from thesaurus structure on the process of indexing the subject, we begin with S, the set of one or more terms produced by the process of matching thesaurus centric and TF * IDF are used as a starting point for browsing and selection.

```
Input  : list term S from TF*IDF
Output : list term from vertex ranking
DoRandomWalk (N, K)
for each starting term s
  for n ← 1 to N
     for k ← 1 to K
          Select next term t using
          weighted random selection
          endpoints[t]++
     end for K
  end for N
Vertex ranking bypassed by agent
```

Fig. 2. Random Walk Algorithm

There is a Graph (G) that builds a thesaurus with the vertex (V) for each term in the thesaurus. Edge (E) is the relationship between terms that are broader and narrower. For each starting term S! represents a match between the document and thesaurus. Figure 2 describes the process where N is the number of terms that became the starting point of the agent is running, and K represents the maximum length of expeditions agent. Figure. 2. Random Walk algorithm In this implementation, if there is no relationship from the initial term, the agent will remain in the initial vertex.

2.3. Proposed Method

The first data that has been in the form of clusters of documents is done in the form of a text Figure. 1. Sample terms and relationships from a thesaurus Figure. 3. Automatic subject indexing phase preprocessing tokenization, stemming, and stopwords to facilitate the processing of data while also extracting words and phrases that are in each document.

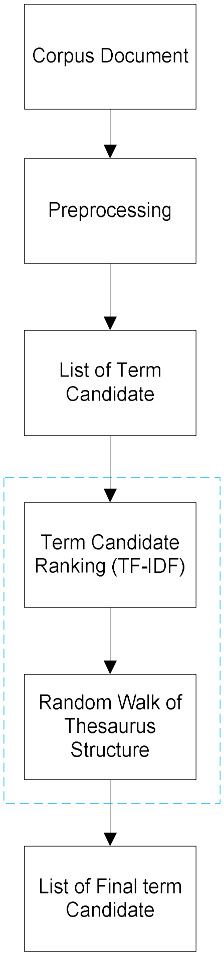
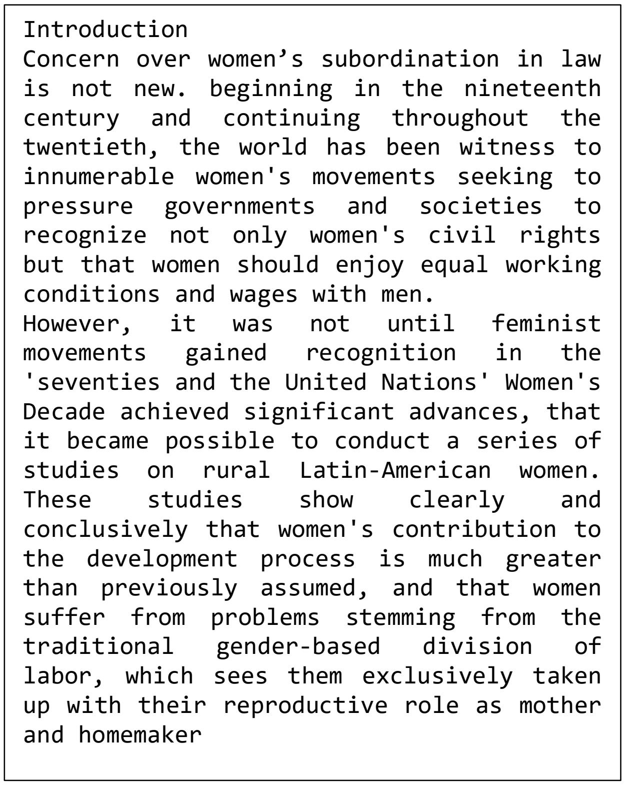

Corpus Document
Preprocessing
List of Term Candidate
Term Candidate Ranking (TF-IDF)
Random Walk of Thesaurus Structure
List of Final term Candidate

Fig. 3. Automatic Subject Indexing Phase.

The next stage is the data from the preprocessing matched with words that were in the thesaurus resulting in some candidate list terms that have been potential to describe the topic of a document. The term candidate list then its weight will be calculated using the TF*IDF. Some term candidates that have top ranking or rating then explored the structure of the thesaurus to know the terms that relate broader and narrower by using a random walk. Term candidate selected through the exploration process of thesaurus structure by using the Random Walk is then added as a new feature for Subject Automatic Indexing System.



Introduction
Concern over women’s subordination in law is not new. beginning in the nineteenth century and continuing throughout the twentieth, the world has been witness to innumerable women's movements seeking to pressure governments and societies to recognize not only women's civil rights but that women should enjoy equal working conditions and wages with men.
However, it was not until feminist movements gained recognition in the 'seventies and the United Nations' Women's Decade achieved significant advances, that it became possible to conduct a series of studies on rural Latin-American women. These studies show clearly and conclusively that women's contribution to the development process is much greater than previously assumed, and that women suffer from problems stemming from the traditional gender-based division of labor, which sees them exclusively taken up with their reproductive role as mother and homemaker

Fig. 4. Example of Dataset.

The quality of the extracted terms in this study was evaluated with precision, recall, and F-Measure. Method of precision, recall, and F-Measure is used in the dataset FAO-780 to evaluate summaries of automatic text. Method of precision, recall, and F-Measure is effectively used to evaluate the quality of the extraction of terms relevant to the topic of a document [11]. In this study, good extracted terms are terms or a phrase of automatic subject indexing that matches the term or phrase that is indexed manually by humans. While the extracted term is a term or phrase that is generated by the automatic system of subject indexing.

Table I
Comparison Result

Method	Precision	Recall	F-Measure	Extracted Term	Number of Walk
TF*IDF	23.08	20.91	21.94	7	6
RW(Broader)	2.18	2.17	2.18	7	6
RW(Narrower)	1.97	1.97	1.97	7	6
TF*IDF and RW(Broader)	25.63	23.23	24.37	7	6
TF*IDF and RW(Narrower)	23.99	21.78	22.78	7	6

Testing of methods performed to test or run the system with some of the existing parameters on the method. In this study, the test uses two parameters: length of walk K and many topics were extracted. The purposes of the test parameters are to get a value of parameters that are most optimal to provide the best testing results. The parameters contained in the proposed method are shown in Table 1. The purpose of a testing parameter is the flow of the testing system from the parameter to the testing system. In this study, the performance of the proposed method is evaluated based on the value of precision, recall, and F-Measure.

3. Result and Analysis

The dataset used in this research is from FAO and AGROVOC thesaurus with total datasets are 780 documents [11]. The example of the FAO dataset is shown in Figure 4. The AGROVOC (UN Food and Agriculture Organization) thesaurus and 780 pre-indexed documents are from the UN Food and Agriculture Organization (FAO) digital library. AGROVOC thesaurus is encoded using the simple knowledge organization system (SKOS) format.

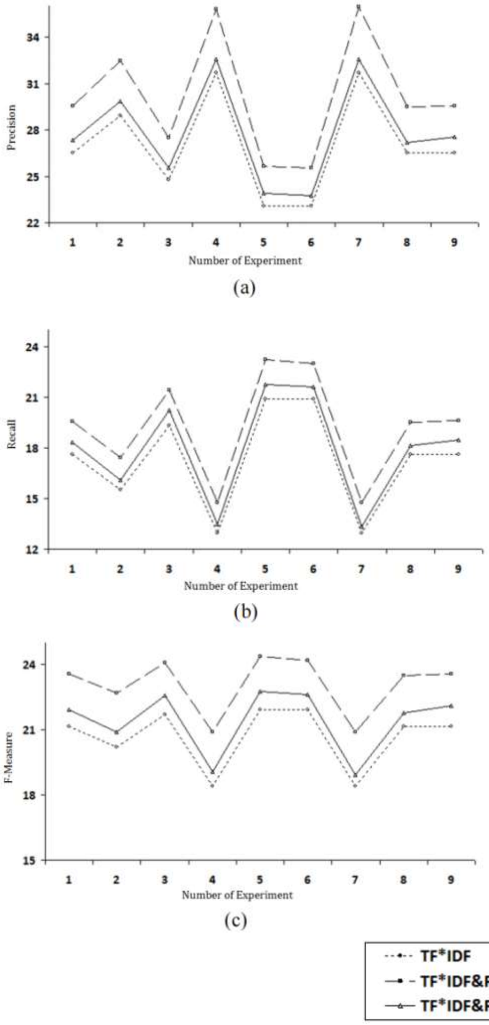
Evaluation methods used are precision, recall, and F-Measure to measure the relevance of the term to the topic of the document. Experimental results show that characteristics and relationships among terms in the thesaurus influence the performance of automatic subject indexing.

$$precision = \frac{\#good\ extracterd\ terms}{\#extraction\ terms},$$

$$precision = \frac{\#good\ extracterd\ terms}{\#manually\ assigned\ terms},$$

$$F - Measure = 2 * \frac{precision * recall}{precision + recall},$$

On the term extraction process, the extracted terms are determined based on the extracted term parameters. The highest term ranking measured by TF*IDF is supposed to present the topic of a document. At the testing phase, the dataset is divided with the composition of 66% to 34% for training and testing. The composition is taken from previous studies [10]. If the composition of the divisions of the dataset too much data training it will be dataset will experience overfitting. If the composition of the dataset more testing data than the training data, then there is a possibility of low accuracy test results because automatic subject indexing cannot extract all the potential long-term candidates.



(a)

(b)

(c)

Fig. 5. Experimental result : (a) precision, (b) recall, (c) F-Measure.

The next phase is the exploration of the structure of the thesaurus of terms extracted by the automatic subject indexing system. Random Walk is used in the exploration of the relationships within the structure of thesaurus terms (broader, narrower). The depth of exploration of the Random Walk on the structure of the thesaurus will be determined by the parameter number of walks. According to the experiment, the most optimal value in the parameter number of walks is 7. Optimal parameter values were obtained by way of trial error and the results of the highest F-Measure.

Experimental results show that the structure of the thesaurus can enhance the performance of candidates selection on Automatic Subject term Indexing. Experiment results also show the difference in performance on each of the relationships in the structure of the thesaurus as shown in Figure 5. The factor the affected the precision and recall selection was the number of extracted terms.

The precision that increasing shows that the comparison between the terms generated by automatic subject indexing relevant to the topic that generated a total of the term which is extracted. The smaller the number of parameters extracted term used then the Precision will be increased and vice versa.

The next phase is the exploration of the structure of the thesaurus of terms extracted by an automatic subject indexing system. Random Walk is used for the exploration of the relationships in the structure of thesaurus terms (broader, narrower). The depth of exploration of the Random Walk on the structure of the thesaurus will be determined by the parameter number of the Walk. According to the experiment, the most optimal value of the parameter number of the walks is 7. Optimal parameter values were obtained by way of trial error and the results of the highest F-Measure.

Experimental results show that the structure of the thesaurus can enhance the performance of candidate selection on automatic subject indexing. Experiment results also show the difference in performance on each of the relationships in the structure of the thesaurus as shown in Figure 5. The factor that affected the precision and recall selection was the number of extracted terms.

The precision that increasing shows the comparison between the terms generated by automatic subject indexing relevant to the topic that generated a total of the term which is extracted. The smaller the number of parameters extracted term used then the Precision will be increased and vice versa.

Recall that increasing shows the term generated by automatic subject indexing relevant to the topic of the document that generated the more. Experimental results show that the smaller the number of parameters extracted term used and the Recall will progressively decrease and vice versa as shown in Figure 5.

Experiment results also show that the broader thesaurus structure (relationships of general words) gives an average precision, recall, and F-Measure better than any structure of the thesaurus narrower (specific word relationships). This is because the topics of documents typically contain common words (broader) from words that are available to the document.

The proposed method can obtain a higher average value of precision, recall, and F-Measure than the TF*IDF method as shown in Figure 5 because it manages to explore the structure of a thesaurus of terms that are extracted by the TF*IDF method. Term extracted by the TF*IDF method combined with the term extracted by a Random Walk method exploration of the structure of the thesaurus. A combination of the terms gives higher accuracy than the TF*IDF method so that the value of precision, recall, and F-Measure also increases in average F-Measure of 0.20 to 0.23.

4. Conclusion

Candidate term generated by the method that proposed gives better results at a level of relevance to the topic of the document that proves based on the value of precision-recall and F-Measure. The level of relevance was measured by the comparison between the results of term extraction from the system and the topic of the manually indexed documents contained in the dataset.

The proposed method gives the average value of precision, recall, and F-Measure bigger than the value of the TF*IDF method, increased F-Measure from 0.20 to 0.23 in average. This indicates that the proposed method can choose a better candidate term based on the correlation between the frequency term with thesaurus in terms of suitability or relevance of the candidate term to the topic document. Based on the result of the experiment, it can be concluded that the structure of the thesaurus broader giving an average result of precision, recall, and F-Measure is better than the thesaurus narrower structure. It is because the topics in the document usually contain the general term (broader) of the term contained within the document.

Future work may develop the strategy for candidate term selection on automatic subject indexing for supporting multiple vocabularies.

References

[1] J. P. Silvester, M. T. Genuardi, and P. H. Klingbiel, “Machine-Aided Indexing at NASA,” *Inf. Process. Manag.*, vol. 30, no. 5, pp. 631–645, 1994.

[2] F. Sebastiani, “Machine Learning in Automated Text Categorization,” vol. 34, no. 1, pp. 1–47, 2002.

[3] Cleveland, Donald B. And Cleveland, Ana B., Introduction to Indexing and Abstracting, Libraries Unlimited Inc, 2014.

[4] Gil-Leiva, Isidoro. SISA: Automatic Indexing System for Scientific Articles: Experiments with Location Heuristics Rules versus TF-IDF Rules. Knowledge Organization 44(1): xx-xx. 2017.

[5] Desale, Sanjay K. and Kumbhar, Rajendra. Re-search on automatic classification of documents in li-brary environment: a literature review. Knowledge organi-zation, 40: 295-304. 2013.

[6] Li, X. Keyphrase Extraction and Grouping Based on Association Rules.Master thesis. University of Guelph, canada. 2013.

[7] Mutschke, P., & Mayr, P. Science models for search. A study on combining scholarly information retrieval and scientometrics. Scientometrics. 2015.

[8] Stevenson M, Agirre E, Soroa A. . Exploiting domain information for word sense disambiguation of medical documents. J Am Med Inform Assoc ; 19:235–40. 2012.

[9] Adrien Bougouin, Florian Boudin, and Beatrice Daille. Topicrank: Graph-based topic ranking for keyphrase extraction. In Proceedings of the 6th International Joint Conference on Natural Language Processing, pages 543–551. 2013.

[10] M. Sun, Y.-N. Chen, and A. I. Rudnicky, "An intelligent assistant for high-level task understanding," in Proceedings of The 21st Annual Meeting of the Intelligent Interfaces Community (IUI),pp. 169–174, 2016.

[11] C. Willis and R. M. Losee, "A Random Walk on an Ontology : Using Thesaurus," *J. Am. Soc. Inf. Sci.*, vol. 64, no. 7, pp. 1330–1344, 2013.

[12] K. F. H. Holle, A. Z. Arifin, and D. Purwitasari, "Preference Based Term Weighting For Arabic Fiqh Document Ranking," *J. Ilmu Komput. dan Inf.*, vol. 8, no. 1, pp. 45–52, 2015.

[13] M. N. Saadah, R. W. Atmagi, D. S. Rahayu, and A. Z. Arifin, "Information Retrieval of Text Document with Weighting TF-IDF and LCS," *J. Ilmu Komput. dan Inf.*, vol. 6, no. 1, pp. 34–37, 2013.

[14] W. Zhang, T. Yoshida, and X. Tang, "Expert Systems with Applications A comparative study of TF*IDF , LSI and multi-words for text classification," *Expert Syst. Appl.*, vol. 38, no. 3, pp. 2758–2765, 2011.

[15] S. Hassan, R. Mihalcea, and C. Banea, "Random walk term weighting for improved text classification," *Int. J. Semant. Comput.*, vol. 1, no. 4, pp. 421– 439, 2007.