INFORMATION RETRIEVAL OF TEXT DOCUMENT WITH WEIGHTING TF-IDF AND LCS

Munjiah Nur Saadah, Rigga Widar Atmagi, Dyah S. Rahayu, and Agus Zainal Arifin

Department of Informatics Engineering, Faculty of Information Technology, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

E-mail: munjiah.nur11@mhs.its.ac.id

Abstract

Information retrieval of text document requires a method that is able to restore a number of documents that have high relevance according to the user's request. One important step in the process is a text representation of the weighting process. The use of LCS in Tf-Idf weighting adjustments considers the appearance of the same order of words between the query and the text in the document. There is a very long document but irrelevant cause weight produced is not able to represent the value relevance of documents. This research proposes the use of LCS which gives weight to the word order by considering long documents related to the average length of documents in the corpus. This method is able to return a text document effectively. Additional features of word order by normalizing the ratio of the overall length of the document to the documents in the corpus generate values of precision and recall as well as the method of Tasi et al.

Keywords: *LCS, Information Retrieval, Tf-Idf*

Abstrak

Sistem temu kembali dokumen teks membutuhkan metode yang mampu mengembalikan sejumlah dokumen yang memiliki relevansi tinggi sesuai dengan permintaan pengguna. Salah satu tahapan penting dalam proses representasi teks adalah proses pembobotan. Penggunaan LCS dalam penyesuaian bobot Tf-Idf mempertimbangkan kemunculan urutan kata yang sama antara *query* dan teks di dalam dokumen. Adanya dokumen yang sangat panjang namun tidak relevan menyebabkan bobot yang dihasilkan tidak mampu merepresentasikan nilai relevansi dokumen. Penelitian ini mengusulkan penggunaan metode LCS yang memberikan bobot urutan kata dengan mempertimbangkan panjang dokumen terkait dengan rata-rata panjang dokumen dalam korpus. Metode ini mampu melakukan pengembalian dokumen teks secara efektif. Penambahan fitur urutan kata dengan normalisasi rasio panjang dokumen terhadap keseluruhan dokumen dalam korpus menghasilkan nilai presisi dan *recall* yang sama baiknya dengan metode Tasi dkk.

Kata Kunci: *LCS, sistem temu kembali informasi, Tf-Idf*

1. Introduction

Currently, most information is stored in digital form on an electronic media so that information systems should be able to recover a large amount of text data that users need. Technology text document retrieval system provides a way to retrieve the required information. This technology is capable of finding text documents stored in the system according to the specific query entered by the user, either by word or phrase query. Text documents are displayed in descending order starting from the documents that have the highest relevance value to the query in question [1]. In the development of document classification is also used so that will help facilitate the process information retrieval of text documents.

Two stages of technology text document retrieval system is a pre-text processing and text representation. Pre-processing of the text consists of many stages, such as tokenizing, stemming, and stop listing. While the text representation stage commonly known as text weighting stage. There have been many previous studies that propose new methods for text weighting. Weighting method which is still commonly used, namely Term frequency - inverse document frequency (Tf-Idf) considering the frequent appearance of the term in the document and the ratio of the length of the documents in the corpus [2]. In addition, there are also considering the BM25 weighting long documents than the average length of documents in the corpus along with some parameter adjustment [3]. Erenel & Altincay [4] using a weighting that utilizes linear transformations on term frequency as a feature to categorize text documents. There is also a weighting based on the return value from the previous discrimination [5]. As with the Luo and Xiong [6] which uses weighting based on semantic concepts in the research.

Tasi et. al. [7] proposed a weighting method with the additional features of word order using the Longest Common subsequence (LCS) in the text document retrieval system that is integrated. Features the word order is added to the Tf-Idf weighting which has been established previously. However, the use of feature weighting word order still has shortcomings. That is because there are some very long documents but irrelevant. These documents usually do a lot of repetition of words. Instead, there is a short document, but in it contains important information so that the sequence of query words in the document is less. Therefore, the necessary process of adjustment to the features of the word sequence involving length feature produced the document so that the document actually had relevant information.

In this study proposed a method of text representation features LCS word order involving normalization of the ratio of the length of the entire document in the document corpus. With the better text representation of expected documents retrieve by the system has a high relevance to the user's wishes.

2. Vector Based Method

Similarity level measurement is an important element in the mechanism of a text document retrieval system. In this study, the focus there is exposure to vector-based method and sequence. In vector-based method, each word of the query and the document represented as elements in the vector. Documents representation will be used to calculate the degree of similarity of documents. The procedure vector-based method is divided into two phases, text representation and similarity computation.

Tf-Idf is a popular method used in determining the weight of each word. The weight may reflect its importance in a document. Weights each word will be mapped in a vector, so it will form the n-dimensional vector. Weighting the LCS is also one of the vector-based methods that have been proposed in previous studies. The order of words between document and query vectors used values.

The next stage after the text representation is the calculation of similarity. Computing the similarity aimed to calculate the degree of similarity between the query and the document. The three most common method used is Cosine, Dice, and Jaccard.

2.1 Weighting Tf-I df

Term frequency inverse document frequency (Tf-Idf) is a calculation that illustrates the importance of the word (term) in a document and a corpus. This process is used to assess the weight of term relevance of a document to all documents in the corpus. Term-frequency is a measure of the frequent appearance of a term in a document and also in the whole document in the corpus. Term frequency is calculated using equation (1) with is the i-th term frequency and is the frequency of occurrence of the i-th term in the j-th document. While the inverse document frequency is the logarithm of the ratio of the total number of documents in the corpus by the number of documents that have the term is mathematically written as in equation (2) [8]. The value obtained by multiplying both formulated in equation (3).

$$tf(i) = \frac{freq_i(d_j)}{\sum_{i=1}^{k} freq_i(d_j)}, \qquad (1)$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|} \qquad (2)$$

$$(tf - idf)_{ij} = tf_i(d_j).idf_i \qquad (3)$$

2.2 Weighting LCS

LCS is used to calculate the length of sequential relationships between queries and documents. In the study [7], LCS adopted in the text document retrieval systems as a feature weighting. LCS value between query *q* with the *j-th* document that has been obtained is then normalized by equation (4) where *m* is the number of terms in the query and *n* is the number of terms in the document.

$$LCS(q, j)_{normalisasi} = \frac{LCS(q, j)}{m + n} \qquad (4)$$

LCS normalized value is then used to adjust the weighting adjusted previously existing, with weights obtained from the *Tf-Idf.* Final weights for documents that have the word order is higher than the corresponding query documents that do not have a word sequence matching the query. Thing is to impact the value of similarity between the query and documents. Documents that have the word order weights have a higher similarity value.

2.3 Similarity Calculation

Similarity calculations are commonly used in vector-based methods such as Dice, Jaccard, and Cosine. By proving mathematically by [7] note that this method is not appropriate Cosine similarity weights used in calculating involving word order features. In this study, [7] concluded that a suitable similarity measure for text document retrieval system with Tf-Idf weighting and LCS is the Dice and Jaccard.

3. Modified LCS

The use of the word order features into consideration in the determination of the weight of the text document retrieval system is a contribution. However, it still has shortcomings that could cause incorrect results return documents. Inaccuracies due to a non-standard size document information content. There is a very long document, but does not have sufficient information content is important. In this document there is a possibility of a lot of repetition words that have the same order query. This leads to an increase in the weight of LCS which will impact the value of document similarity to the query. On the other hand, there is a document that has important information but lengthy document weights LCS is small so the document will also be of low value. This causes the document has a small similarity value. It eventually occur inaccuracies documents returned by the system.

Therefore, the weight of LCS used for text document retrieval system should be adjusted to the length of the document. In this study proposed a modification of the weighting sequence of words considering the length of the document. Modifications LCS is calculated using equation (5) with LCS (q, j) is the value of a document to a query LCS-j, j is the ratio of the length NDL jth document with an average length of all the documents in the corpus.

$$LCS_{\mathrm{mod}ifikasi}(j) = \log \frac{(1+LCS(q,j))}{(1+ndl_j)} \qquad (5)$$

4. Gathering System Return Document text

Proposed new weighting is applied to the text document retrieval system intact. In short, the methodology and process flow that exist in the text document retrieval systems is illustrated in Figure 1.

First, the text pre-processing phase of the query term extraction and text documents to be stored in the repository. The first process is to separate the query and the document text into tokens and ignore punctuation and numbers. The process is called tokenizes. Then stopper will delete the words in the text are included in the stop list. The words included in the stop list are words that do not contain unnecessary information. The next process is to get the word stemming the basis of the remaining term of the previous two processes. Stemmer used is Indonesian [9].

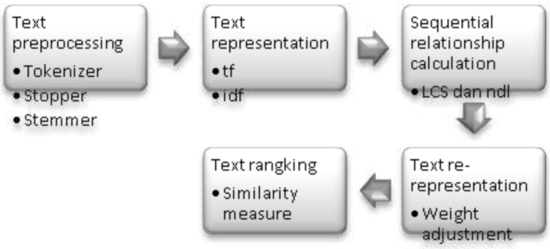

Fig 1. Gathering System Methodology Return Document Text

Second, it provides Tf-Idf weighting term in the query and the document text in the text representation stage. Tf-Idf weighting calculations made on the basis of the results of the phase word pre-processing text. Third, the assessment of long-CSF modification documents used to calculate sequential relationships in the query and text.

Fourth, the relationship is used to adjust the weights sequential Tf-Idf at the stage of re-representations of text [7]. This stage is used to add Tf-Idf weighting of terms based on the sequential relationship between the query and the text resulting from the modification LCS value. There are two concepts to adjust the weight of Tf-Idf. The first concept is the process of adjusting the term sequential because it deals with the sequential information. Sequential Term is a term contained in the sequential relationship between the query and text. The second concept is to reduce the difference in sequential terms. In this way, sequential term weight increases, so the difference in sequential term will be reduced to increase the similarity between the query and text.

The final step, similarity measure is used to calculate the similarity between the query and the text in the text ranking stage. The measurement used is the Dice similarity according to the results of the study [7].

5. Test Result and Analysis

Tests conducted on 140 pieces of documents belonging to Indonesian language news in 11 fruit categories. To determine the ability of the proposed system is calculating the value of precision and recall of 57 trials with different queries. Values of precision and recall of the proposed system is 30% and 96%. This value indicates that the system can work effectively in restoring a number of text documents that are required according to the user's query.

The results of testing the proposed system is also compared with the results of the test system [7]. Table 1 shows the average values of precision and recall of the proposed system compared to the system. Precision and recall values between the two systems together indicate that weight normalization method of word order and Tf-Idf not differ significantly, although the weight given by the proposed method is greater than the method [7]. Values of precision and recall value method similar to the method [7], due to the condition of the search text documents using existing search query sequence. Tf-Idf weights summed with weight LCS only generate linear comparison though larger value. The addition of weights occurred in almost all documents that have a value of Tf-Idf weighting proportionately high.

TABLE I
THE AVERAGE VALUES OF PRECISION AND RECALL OF THE PROPOSED SYSTEM WITH THE SYSTEM [7].

The average value	The average value	The average value
Precision (%)	30,32	30,32
Recall (%)	96,84	96,84

6. Conclusion

Text document retrieval system in this study utilizes the enhanced features word order. Experiments were conducted to prove the use of the word order features as effective as an integrated retrieval system used [7]. The proposed method is able to restore a number of text documents that are relevant to the user's query. Values of precision and recall of the proposed method and the method of [7] same value so it can be represented as robustness and effectiveness of the similarity retrieval of text documents.

Reference

[1] van Rijsbergen, CJ, Information Retrieval, 2 nd ed. London: Butterworths; 1979.

[2] Salton, G., Mc.Gill, MJ Introduction to Modern Information Retrieval. New York: Mc Graw Hill Book. Co; 1983.

[3] Robertson, SE, Walker, S., Jones, S., Hancock-Beaulieu, MM, & Gatford, M. 1995. Okapi at Trec-3, In D. Harman (Ed.) Proceedings of the third Text Retrieval Conference (Trec-3). Pp. 109-126.

[4] Erenel, Z. & Altincay, H. 2012.Nonlinear transformation of term frequencies for term weighting in text categorization (2012) Engineering Application of Artificial Intelligence 25. Pp. 1505-1514.

[5] Song, S. & Myaeng SH 2012. A novel term weighting scheme based on discrimination power obtained from past retrieval results. Information Processing and Management. 48. Pp. 919-920.

[6] Luo, Q., Chen, E. And Xiong, H. 2011.A semantic term weighting scheme for text categorization. 38. Pp. 12708-12716.

[7] Tasi, Cheng-Shiun, Huang, Yong-Ming, Liu, Chien-Hung Huang, Yueh-Min. 2012.Applying VSM and LCS to develop an integrated text retrieval mechanism. Expert Systems with Applications, Pp. .3974-3982.

[8] Jones, KS, 1973. Indexing term weighting, Inf. Storage Retr, 9, Pp. 619-633.

[9] Asian, J., Wiliams, H.E., Tahaghoghi, S.M.M., 2005. Stemming Indonesian. Proc. 28th Australasian Conference on Computer Science, vol. 38, pp. 307-314.