A document recommendation system of stemming and stopword removal impact: A web-based application

W G S Parwita
Teknik Informatika, STMIK STIKOM Indonesia, 80225 Denpasar-Bali, Indonesia

*gede.suka@gmail.com

Abstract. Stemming and stopword removal is process that requires a lot of resources in the text pre-processing. The resources used in stemming and stopword removal are directly proportional to the amount of stopword, text, and document. Elimination of stemming and stopwords is one of many options which can reduce process in document-based recommendation system. However, the elimination of stemming and stopword removal have an impact on the recommendation system accuracy. This study determines the impact of stemming and stopword removal in document-based recommendation system. The system used in this study is recommendation system that recommend final project supervisor based on similarity student preliminary research proposal (UPP) document and lecturers’ scientific publications. Student preliminary proposal and lecturers’ scientific publications are document in Bahasa Indonesia. This study begins with analysis to map out the components to be used in each recommendation systems testing. Then proceed with the rearrangement of recommended system components based on testing focus. The study result is precision, recall, and f-measure values comparison between each recommendation system component elimination. Elimination of stemming and stopwords show that the elimination generates precision and f-measure value worse than the system with stemming and stopword removal. However, system with elimination give a better result at recall value. In the future study, recommendation system needs some development to improve the precision, recall, and f-measure value with modification in stemming method using Sarawi project and increase the amount of lecturer’s publication.

1. Introduction
Recommendation system are software tools and techniques providing suggestions for items to be of use to a user [1]. Items are term used for suggestion results of the recommendation system for the user. The suggestion is based from various decision making such as what items to buy, what songs you want to listen to, and what books to read.

To generate recommendations, recommendation system requires user interest that explicitly expressed in the item ranking or concluded by guessing the user's actions such as click view or time that user spend on the page. Besides these two methods, recommendation in the document recommendations can be generated by finding the similarity between document content. The similarity of document is calculated based on the keyword of the compared documents. Keyword extraction is a process to extract documents keywords which include process such as tokenization, stemming, and stopword removal process.

Student final project supervisor recommendations in Sintesys (Synchronized Student’s Final Project Management System) comparing student preliminary research proposal (UPP) with supervisor reasearch
and publication documents as a basis for determining recommendations [2]. The documents keywords and keyword frequency will be extracted from these documents and then the similarity will be calculated with cosine similarity. The increasing of UPP documents, lecturers, and lecturer’s research documents annually will impact to the increase of document comparison process that is conducted by the system. With this increase, the system will also work harder because the document increase that is compared is directly proportional with the resource that is needed.

Elimination of stemming and stopwords removal process in keyword extraction system is one of many options which can reduce process in document-based recommendation system. However, the elimination of stemming and stopword removal have an impact on the recommendation system result. This study determines the impact of stemming and stopword removal in document-based recommendation system in precision, recall, and f-measure value.

2. Method

UPP document and lecturer’s research was extracted to find keyword which was in each document. The extraction process would through some stages which were tokenization, stopword removal, stemming, and weighting. Stopword removal process used stopword from Tala research [3] and stemming process was going to used algorithm that was developed by Tala [3]. The lecturer’s research document keyword would be saved on database so that the extraction repetition would not happen. Weighting on keyword would be based on term frequency assumption. Keyword and frequency were going to be used for similarity calculation based on cosine similarity approach. Overall, system in the previous study [2] was shown in Figure 1.

![Figure 1. Document recommendation schema.](image-url)

2.1. Tokenisms

Electronic text was a linear symbol order (characters, words or phrases). Before the tabulation, text needed to be segmented into linguistic units such as words, punctuation, number, alpha-numeric, and etc. This process called tokenization. Simple tokenization (white space tokenization) was separating the word by the character space, tab and new line [4]. But, not every language did this (for example Chinese, Japanese, Thailand language). In Bahasa Indonesia, beside simple tokenization it needed also tokenization that separated words with character such as “/” and “-“.
2.2. Stopword removal
Stopword removal was a basic approach in pre-processing that omitted words that often appear (stopword). The main function was to prevent the effect by the stopword in the next process. A lot of the stopword was useless in the Information Retrieval (IR) and text mining because those words did not deliver information (such as to, from, and, or). The usual way to decide whether it stopword or not was by using stop list. Stop list was a group of word or dictionary that contained by stopword list. The omitting stopword removal process was the important and useful step [5].

2.3. Stemming
Stemming algorithm was process that was done by morphology variant mapping which was different from words into general word (stem). Stemming was useful in many linguistic computation field and information retrieval [6]. In Bahasa Indonesia, so far there were only two algorithms to do the process of stemming which were algorithm that was developed by Nazief and Adriani also algorithm that was developed by Tala. Nazief and Adriani Algorithm were developed by using confix stripping approach with scanning in dictionary. While stemming that was developed by Tala used rule-based approach.

Tala algorithm processed prefix, suffix, and both of them that were combined in derivative word. Even though in Indonesian had infix, the word that was derive used very few infixes. Because of this and for the sake of simplification, the infix would be ignored. Figure 2 showed the process steps in Tala algorithm.

![Figure 2. Tala stemming.](image)

In Bahasa Indonesia, the smallest unit of the word was syllable. The fewest syllable consisted of one vowel. Implementation design of Tala algorithm could not identify all syllables. This was because two vowels that considered as one syllable/diphthong which are ai, au, and oi. The combination of two vowels (especially ai, oi) could be a problem, moreover when it was in the final syllable. This was because the difficulty on differentiate the word with suffix –i. this made the vowel combination of ai/oi would be treated like derivative word. The last character (-i) would be omitted in the stemming process. Many basic words consisted of minimum two syllables. This was the reason why the word that was going to be processed had minimum two syllables.
2.4. Term weighting
This stage purpose was giving weight on every term that can be found in the document. Term was a word or more that is directly chosen from original document corpus using term-extraction method. Term level feature only consist of some words and expressions that were found in the original document [7]. In text categorization and other application in information retrieval or machine learning, term qualification usually handled through searching text method, which was meant not involving learning stage [8]. There were three monotone assumption that appeared in the most of the qualification method, in a form or another, they were [9]:

- Rare term was also important rather than the usual-appeared term (IDF assumption).
- Appearing repetition of the term in the document was also important rather than single appearance (TF assumption).
- For the term matching with the same amount of matching, long document was not really important rather than short document (normalization assumption).

Quality was needed to decide that the term was important or not. The quality that was given on a term depend to the method that was used for the qualification. This study used Term Frequency (TF) assumption.

2.5. Cosine similarity
Cosine similarity approach often used to find out proximity between text documents. Cosine similarity calculation was started by calculating dot product. Dot product was a simple calculation for every component in each two vectors. Vector was representation of every document with term amount on each document as the dimension of vector [10]. Vector was showed by notation (1) and (2). The result of dot product was not in the form of vector but in scalar. Equation (3) was the dot product calculation where \( n \) was the dimension of vector [11].

\[
\mathbf{a} = (a_1, a_2, a_3, ..., a_n) \\
\mathbf{b} = (b_1, b_2, b_3, ..., b_n) \\
\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i = a_1 b_1 + a_2 b_2 + \cdots + a_n b_n
\]

\( a_n \) and \( b_n \) was the component of vector (term quality of each document) and \( n \) was dimension of vector. Cosine similarity was the calculation that measure cosine value from the angle of between two vectors (or two documents in vector space). Cosine similarity could be seen as comparison between documents because not only considering the size of each amount of words (quality) of each document, but also the angle between documents. Equation (4) and (5) were the notation from the cosine similarity method where \( \|\mathbf{a}\| \) was Euclidean norm of vector \( \mathbf{a} \) and \( \|\mathbf{b}\| \) was Euclidean norm vector \( \mathbf{b} \) [12].

\[
\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \|\mathbf{b}\| \cos \theta
\]

\[
\cos \theta = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|}
\]

From notation (4) and (5) can be form into mathematic equation that was shown by equation (6).

\[
\text{Similarity}(x,y) = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{i=1}^{n} a_i^2} \sqrt{\sum_{i=1}^{n} b_i^2}}
\]

Where:
ai: the \( i \) term that was in the document a.
bj: the \( i \) term that was in the document b.
3. Testing
This study is started by the system analysis that is built in the previous study [2]. This analysis maps the component that is going to use in the system recommendation test. After system analysis, then is done the data collection such as stopword that is received by Tala research [3], lecturer’s recommendation, and lecturer’s research document. Testing implementation is done by building test algorithm using the system recommendation component that is already built.

The test is implemented by using components in the recommendation system that is already built. Implementation is going to use 3 test schemes they are system scheme completed with stemming and stopword removal process (scheme 1), system scheme without stemming process (scheme 2), and system scheme without stopword removal process (scheme 3). The recommendation result is going to be compared to the real recommendation that is given in the form of questionnaire.

The test is done by calculating Precision, Recall, and F-Measure between recommendation result and the real recommendation. Recommendation test that is resulted by the system is started by looking for similarity value in every research document to each UPP documents. When the lecturer has more than 1 (one) research, then the highest similarity value is used as lecturer with student similarity value. Figure 3 is preview of the test algorithm.

![Testing schema](image)

**Figure 3.** Testing schema.

3.1. Precision
Precision along with recall is one of the basic tests and often used for determining the effectivity of information retrieval system or recommendation system. True positive (tp) in information retrieval is relevant item that is resulted by the system. While false positive (fp) is all item that is resulted by the system. So, in the information retrieval, precision is calculated with equation (7) [10].

\[
Precision = \frac{tp}{tp+fp} = \frac{relevant\ \text{item\ retrieved}}{retrieved\ \text{item}} \quad (7)
\]

The term positive and negative refers to prediction that is done by the system. While the term true and false refers to prediction that is done by outsiders or the party that does the observation. The condition division is shown in the Table 1 [10].
Table 1. Condition division.

|               | Relevant          | Nonrelevant       |
|---------------|------------------|-------------------|
| Retrieved     | True positive (tp) | False positive (fp) |
| Not retrieved | False negative (fn) | True negative (tn) |

3.2. Recall

Recall is used as relevant document measurement that is resulted by the system. False negative (fn) is all relevant item that is not resulted by the system. In information retrieval system evaluation, recall is calculated by equation (8) [10].

\[
Recall = \frac{tp}{tp + fn} = \frac{\text{relevant item retrieved}}{\text{relevant item}} \quad (8)
\]

3.3. F-measure

F-measure is single value that is resulted by combining the precision value and recall value. F-measure can be used to measure the performance of recommendation system or information retrieval system. Because of the harmonic average of precision and recall, F-measure can give performance assessment that is more equal. Equation (9) is the way to calculate F-measure [13].

\[
F_{measure} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (9)
\]

4. Result

Every lecturer’s research document is going to be compared to the student UPP. The highest value in the one of the lecturer’s research document is going to be used as similarity value between the lecturer’s research and the student UPP. Similarity value is used to determine the recommendation that is resulted by the system. The system recommendation system result is going to be compared to the real recommendation result that is collected through questionnaires. This comparison is going to result precision, recall, and f-measure value of each lecturer. The value of each lecturer is going to be sum up and then divided to the total number of the lecturer to look for precision, recall, and f-measure value averages. These value averages are going to be used to measure the quality of the recommendation.

Similarity level that is used is in between 5% to 70% with the increase of 5%. The recommendation that is compared resulted by 3 keyword extraction schemes they are keyword extraction with stemming and stopword removal process (schema 1), keyword extraction without stemming (schema 2), and keyword extraction without stopword removal (schema 3). Table 2 shows the recommendation result with scheme 1. In this scheme, the highest precision value is in the similarity of 30%. For the highest recall value is on similarity 5% and for f-measure is on similarity 20%.

Table 2. Schema 1 result.

| Similarity (%) | Precision | Recall | F-Measure |
|----------------|-----------|--------|-----------|
| 5              | 0.0797    | 0.9870 | 0.1390    |
| 10             | 0.1142    | 0.8394 | 0.1855    |
| 15             | 0.1588    | 0.6987 | 0.2325    |
| 20             | 0.2882    | 0.6245 | 0.3202    |
| 25             | 0.3455    | 0.4816 | 0.3131    |
| 30             | 0.3903    | 0.3495 | 0.2782    |
| 35             | 0.2291    | 0.1908 | 0.1577    |
Table 2. Cont.

|   | Precision | Recall | F-Measure |
|---|-----------|--------|-----------|
| 40 | 0.2402    | 0.1369 | 0.1245    |
| 45 | 0.2068    | 0.0991 | 0.1128    |
| 50 | 0.1951    | 0.0468 | 0.0613    |
| 55 | 0.1742    | 0.0297 | 0.0437    |
| 60 | 0.0682    | 0.0070 | 0.0126    |
| 65 | 0.0000    | 0.0000 | 0.0000    |
| 70 | 0.0000    | 0.0000 | 0.0000    |

Table 3 is the result of the calculation with the recommendation scheme without stemming process. In this scheme, the highest precision value is on similarity 55%. For the highest recall value is on similarity 5-25% and for f-measure is on similarity 55%.

Table 3. Schema 2 result.

| Similarity (%) | Precision | Recall | F-Measure |
|----------------|-----------|--------|-----------|
| 5              | 0.0757    | 1.0000 | 0.1320    |
| 10             | 0.0757    | 1.0000 | 0.1320    |
| 15             | 0.0757    | 1.0000 | 0.1320    |
| 20             | 0.0757    | 1.0000 | 0.1321    |
| 25             | 0.0760    | 1.0000 | 0.1326    |
| 30             | 0.0763    | 0.9805 | 0.1331    |
| 35             | 0.0824    | 0.9675 | 0.1413    |
| 40             | 0.0815    | 0.9153 | 0.1413    |
| 45             | 0.0999    | 0.7877 | 0.1665    |
| 50             | 0.1318    | 0.6278 | 0.2004    |
| 55             | 0.2636    | 0.4367 | 0.2395    |
| 60             | 0.2150    | 0.2633 | 0.1993    |
| 65             | 0.2471    | 0.1099 | 0.1215    |
| 70             | 0.1999    | 0.0501 | 0.0643    |

Table 4 is the result of calculation of recommendation scheme without stopword removal process. In this scheme, the highest precision value is on similarity 25%. For the highest recall value is on similarity 5% and for f-measure is on similarity 15%.

Table 4. Schema 3 result.

| Similarity (%) | Precision | Recall | F-Measure |
|----------------|-----------|--------|-----------|
| 5              | 0.0815    | 0.9361 | 0.1418    |
| 10             | 0.1319    | 0.7235 | 0.2015    |
| 15             | 0.2519    | 0.5689 | 0.2965    |
| 20             | 0.3091    | 0.3781 | 0.2745    |
| 25             | 0.3506    | 0.2275 | 0.2181    |
| 30             | 0.1909    | 0.1002 | 0.0984    |
| 35             | 0.1995    | 0.0608 | 0.0799    |
| 40             | 0.1136    | 0.0122 | 0.0205    |
| 45             | 0.0606    | 0.0103 | 0.0174    |
| 50             | 0.0000    | 0.0000 | 0.0000    |
| 55             | 0.0000    | 0.0000 | 0.0000    |
| 60             | 0.0000    | 0.0000 | 0.0000    |
| 65             | 0.0000    | 0.0000 | 0.0000    |
| 70             | 0.0000    | 0.0000 | 0.0000    |
By the figure 4, the highest precision value is in the scheme 1 that is the complete process scheme in similarity of 30%. Testing scheme with stemming and stopword removal process is tend to be higher than the other.

![Precision graph](image)

**Figure 4.** Precission graph.

In the recall value, the highest value is on scheme 2 that is the scheme without stemming process with the value of 1 (figure 5). This is because total recommendation that is resulted by the system is too many and almost all document is recommended by the system.

![Recall graph](image)

**Figure 5.** Recall graph.

F-measure is harmonic value between precision and recall shows the optimum value is in scheme 1 at 20% similarity. Figure 6 shows the f-measure value.
5. Conclusion

This study, aimed to find the impact of stemming and stopword removal process on document recommendation system. The testing in this study found:

- Optimum precision has been achieved at schema with stemming and stopword removal process that is 0.39 at 30% similarity.
- Optimum recall is 1 at without stemming schema, which mean all the human recommendation is generated by system. Optimum recall has been archived at 5%-25% similarity.
- Optimum F-measure generated at system that used stemming and stopword removal process which is 0.32 at 20% similarity.
- F-measure value shows that the stopword removal and stemming process is important process that when it is omitted, the impact is on the decrease of f-measure optimum value.

In the future study, recommendation system needs some development to improve the precision, recall, and f-measure value with modification in stemming method using Sarawi project and increase the amount of lecturer’s publication.

References

[1] F Ricci, L Rokach, B Shapira and P B Kantor 2011 Recommender System Handbook.
[2] W G S Parwita, M H P Swari and W Welda 2018 Perancangan Sistem Rekomendasi Dokumen Dengan Pendekatan Content-Based Filtering CESS Journal Comput. Eng. Syst. Sci. 3(1) 65–74.
[3] F Z Tala 2003 A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia M.Sc. Thesis, Append. D.
[4] S M Weiss, N Indurkhya, T Zhang and F J Damerau 2005 Text mining: Predictive methods for analyzing unstructured information.
[5] V Srividhya and R Anitha 2010 Evaluating preprocessing techniques in text categorization Int. J. Comput. Sci. Appl. 2010 49–51.
[6] J B Lovins 1968 Development of a stemming algorithm Mech. Transl. Comput. Linguist. 11(June) 22–31.
[7] R Feldman and J Sanger 2006 The Text Mining Handbook.
[8] F Debole and F Sebastiani 2003 Supervised term weighting for automated text categorization Proceedings of the 2003 ACM symposium on Applied computing - SAC 03 784.
[9] J Zobel and A Moffat 1998 Exploring the similarity space ACM SIGIR Forum.
[10] C D Manning, P Ragahvan and H Schutze 2009 An Introduction to Information Retrieval.
[11] S Axler, F Gehring and K Ribet 1997 Linear Algebra Done Right, Second Edition *Comput. Eng.* \textbf{15}(251).

[12] H Jiawei, M Kamber, J Han, M Kamber and J Pei 2012 *Data Mining: Concepts and Techniques*.

[13] D Jannach, M Zanker, A Felfernig and G Friedrich 2011 *Recommender systems: an introduction*. 