Identify the Relevant Pages of Book to be Indexed Using Naive Bayes Classification Method

S Christina¹, and D Ronaldo²

¹²Informatic Departement, Faculty of Engineering, University of Palangka Raya, Palangka Raya, Indonesia

E-mail: sherly.christina.upr@gmail.com

Abstract. The back-of-book index is a component that is often found in non-fiction books. The back-of-book index contains important terms and page numbers where the terms appear in the book, it is useful for helping readers find certain term directly, without having to search on every page. Usually, the back-of-book index is compiled by the author or a professional indexer, although currently there are several automatic indexing applications available. To determine a term on a page worth to be indexed requires knowledge and expertise of the author whom better understand the book’s context. Hence, generating a good back-of-book index is a task that requires great effort, knowledge and cost. Therefore, the back-of-book index is very possible containing indexed terms that refer to irrelevant page numbers due to human error or the weakness of the indexing application. This study aims to identify the relevance of the pages to be indexed in the back-of-book index, using Naive Bayes Classification. The testing result shows that the approach with Naive Bayes Classification produces an average precision value of 74.02 percent and 100 percent for the recall value. The average precision value that more than 50 percent indicates that the naive bayes classification approach is capable to identify the relevant and irrelevant pages.

1. Introduction
The Non-fiction books are usually equipped with a list of indexed terms to help readers find important terms. The following is an example of the indexed terms in a book entitled "Big Data, Data Mining and Machine Learning".

binary state, 225
binning, 199
Bluetooth, 6
boosting, 124–125

The list of indexed terms that usually called the back-of-book index, contains terms along with page numbers that indicate where the terms appear in the book. The list of indexed terms facilitates the reader to look for important words directly without having to open every page in the book [1]. Therefore, the back-of-book index is an important component of a book, thus much research has been done to find the optimal indexing method [2] [3] [4].

The list of indexed terms is usually created by the author or a professional book indexer that provided by the publisher. To compile an excellent list of indexed terms, the author must identify all concepts that are important to be indexed, then the author must scan each page of the book, and decide which pages are relevant to be indexed in concern to the context of the book.
Therefore, there are many software applications available to help the authors and indexing experts to generate a list of indexed terms from a book [5], but although a software application can produce a set of index lists, there is still a problem regarding the relevance of an indexed page. The page numbers which referenced in the index list must lead the reader to the text which containing indexed terms and containing information which is relevant to the indexed terms. Donald and Anna Cleveland defined that a good index will guide the reader to find the information needed with no hurdles and no irrelevant material [6].

In this study, we aim to identify the pages which containing the indexed term that relevant to be indexed using Naive Bayes classification (NBC) method. The NBC method is implemented to identify the pages that are relevant and not relevant to be indexed. In this study, the data sets are the texts containing the indexed terms and indexed terms from the book's page, along with the values of semantic relatedness and cosine similarity between the text and the indexed term. To our knowledge, no research has been conducted to identify the relevant pages from a back-of-book index, by processing the semantic relatedness and cosine similarity in the data set and identifying the relevance of book pages using NBC method.

2. The Relevant Page and Irrelevant Page

The following is example of relevant and irrelevant page to be indexed, in the book titled C Programming for Arduino there is a term "double" indexed on page 57.

Page 57: double - It generally stores float values with a precision two times greater than the float value

The term "double" found on page 57 has a definition that related to data type in C programming. The term "double" on page 57 is closely related to the context of the book, therefore the term "double" on page 57 is important and relevant to be indexed. Then on page 49, also found the term "double" which is not referred by in the index list.

Page 49: Imagine that you have followed carefully the Blink250ms project, everything is wired correctly, you double-checked that, and the code seems okay too, but it doesn't work.

The term "double" that appears in the text on page 49 does not have an important meaning related to the context of the book in C Programming for Arduino, thus page 49 is not relevant to be referred by in the index list.

A good back-of-book index must contain page number that refers to the page that indeed contain the indexed term that is relevant to the book's context, because the terms defined in the index list represent the subject or concepts conveyed by the book [7]. Nevertheless, to produce a good back-of-book index, it requires accuracy, creativity and knowledge from the authors or book indexing experts [7] [8]. Therefore, indexing the terms is a task that requires expertise and knowledge, human error or the weakness of the indexing application can produce the irrelevant list of indexed terms.

3. Methodology

In the methodology section we briefly explain the methods implemented in this study.

3.1. Semantic Relatedness

The semantic relatedness approach [9] is a method to measure the semantic relatedness between text (S) containing the indexed term and the indexed terms (T) itself. In our previous studies [10][11], the semantic relatedness approach show a reliable performance that has been evaluated using Kappa values [12]. The following are the stages in the semantic relatedness measurement approach:

- The text mining stage is the step of text preparation for data sets, such as doing tokenization process, removing the stop words, and stemming the text (S) and indexed term (T).
The speech tagging stage is the next stage after the text mining stage. The speech tagging stage defines the part of speech of S and T, such as noun, verb or adverb of each word in S and T. The Word Sense Disambiguation stage, is the stage to find senses of each words in S and T. The word sense disambiguation stage in this study use the wordnet lexical database. Create semantic similarity matrix M in mxn dimension, where M \[i,j\] is the semantic relatedness value between the most suitable sense of the word at the position i of T with the most suitable sense of the word at the position j of S. Calculate the semantic relatedness with the heuristic method, as in equation (1).

$$Similarity^{(T,S)} = \frac{\sum_{maxR} + \sum_{maxC}}{\text{number of rows and columns}}$$

Where maxR is the maximum value of each row, and maxC is the maximum value of each column.

### 3.2 Cosine Similarity

The cosine Similarity method [13][14] is a syntactic approach to measure the relatedness of S and T. The value of cosine similarity (C) is the term-frequency vector of T and S, as in equation (2).

$$C^{(T,S)} = \frac{T \cdot S}{|T||S|}$$

The distance between vectors T and vector S is an indication of the relatedness between the indexed term and the text on the page which number has been indexed.

We use the cosine similarity approach in this study because the number of the occurrences of the indexed term and the number of the words in the text (S), also determines the relevance of a page to be indexed, although in previous studies [15], the testing result on the cosine similarity approach shows the Kappa value that is lower than the semantic relatedness approach.

### 3.3 Data Set

In this study, we obtained the data sets from several electronic books with different knowledge domains, such as social-politics, economy-business, religion, computer science, education, general and biology. From each e-book, we took texts containing indexed terms in the back-of-book index (S\(_1\)). Then we took other texts containing the same indexed terms from other pages that are not indexed in the back-of-the-book index (S\(_2\)). In this study, S\(_1\) is considered as a relevant page to be indexed, because it is a page that is referenced in the back-of-book index. Whereas S\(_2\) is considered an irrelevant page to be indexed because it is not referenced in the back-of-book index.

S\(_1\) and S\(_2\) are measured in relation to the indexed term (T) using the semantic relatedness and cosine similarity methods. The following is example of indexed term “double” from the "C Programming for Arduino" book.

T: double

S\(_1\) (page 57): double - It generally stores float values with a precision two times greater than the float value. 

Semantic Relatedness Value \((T,S_1) = 0,52\) 

Cosine Similarity Value \((T,S_1) = 0,25\)

S\(_2\) (page 49): Imagine that you have followed carefully the Blink250ms project, everything is wired correctly, you double-checked that, and the code seems okay too, but it doesn't work. 

Semantic Relatedness Value \((T,S_2) = 0,45\) 

Cosine Similarity Value \((T,S_2) = 0,3\)

The number of the data sets in this study are 1664 data, consisting of S\(_1\) and S\(_2\) along with its features, that are semantic relatedness values and cosine similarity values.
3.4. Naive Bayes Classification

The Naive Bayes Classification (NBC) method [16] [17] is used as an approach to identify the relevant and irrelevant pages to be indexed. The NBC method has been proven effective and reliable to produce high accuracy in several studies [18] [19] [20].

On NBC, the hypothesis is a class label which is the target of classification, while evidence is a feature which is input into the classification model. If X is an input vector that contains features and C is the class label, then NBC is written as P(C|X). P(C|X) means the probability of class C is obtained after X features are observed. P(C|X) is also called the posterior probability for C, while P(C) is called the prior probability.

In the training phase, we build a model during the learning process which is done on P(C|X) for each combination of X and C based on the information obtained from the training data. Then the model is used to classify the testing data. Hence, the testing data that has the X features can be classified to find the value of C by maximizing the value of P(X|C) from the testing data. The NBC method in this study is shown by equation (3).

\[
P(C|X) = \frac{P(C)\pi_{i=1}^{k} P(X_i|C)}{P(X)}
\]

(3)

Where C is a class that affirms the relevance of S, C consists of Relevant Class (C_1) and Not Relevant Class (C_2). X is the attribute vector set of S_1 and S_2 consist of semantic relatedness (X_1) and cosine similarity (X_2) values.

\[P(C|X)\] is the probability of data with vector X in class C. P(C) is the initial probability of class C. π_{i=1}^{k} P(X_i|C) is the independent probability class C of all X attributes. The value of P(X) is always fixed, therefore, to identify the relevant class, only calculate P(C)π_{i=1}^{k} P(X_i|C) and only choose the largest value to be the chosen class as a result of predicting the relevant page. Then the independent probability π_{i=1}^{k} P(X_i|C) is the effect of all the features of the data on each class C.

3.5. Method to Identify the Relevant and Irrelevant Page

Figure 1 shows the method for identifying the relevant and irrelevant page. In this study a number of the training data (X,C) have been provided as the input data for constructing the identifying model. The model is a black box that accepts input, then does reasoning on the input data and provides answer as the output. The training process for building models uses the NBC algorithm. The model is used to identify the relevance class from the test data (X,?) thus the true relevant class C(X|C) is known.
4. Experiment
In the experiment section, we explain the testing scenario in our study. In the testing scenario, the dataset is divided into training data and testing data. The amount of training data and testing data is divided in the equal proportions for each book domain, as shown in table 1.

| Book Domain      | Data Set | Data Training | Data Testing |
|------------------|----------|---------------|--------------|
| Social Politics  | 434      | 217           | 217          |
| Economy Business | 286      | 143           | 143          |
| Religion         | 122      | 61            | 61           |
| Computer Science | 346      | 173           | 173          |
| Education        | 84       | 42            | 42           |
| General          | 144      | 72            | 72           |
| Biology          | 248      | 124           | 124          |

We created an identifier model that tested on each book domain and the input data for the identifier model are the training data from the related domain. The model is used to identify the relevant class of the testing data for each book domain. Furthermore, we evaluated the performance on NBC using the identification results, by calculating the precision and recall value of the model outputs.

5. Result and Discussion
In the result and discussion section we explain the results of our experiment. In every testing scenario, the NBC approach works very satisfactorily in finding all data training and data testing, it is indicated by the recall value of 100% from every test result.

Table 2 shows the precision values of the NBC approach in identifying the relevant and irrelevant pages to be indexed. The precision average value from each book domain is more than 50 percent, namely 74.02%. The average precision that exceeds the 50% accuracy, indicates that the NBC approach result in reliable accuracy and can contribute to identifying the relevance of a page for constructing the excellent back-of-book index.

| Domain           | Precision Value |
|------------------|-----------------|
| Social Politics  | 69.59           |
| Economy Business | 59.44           |
| Religion         | 88.52           |
| Computer Science | 78.61           |
| Education        | 66.67           |
| General          | 81.94           |
| Biology          | 73.39           |
| Average Precision Value | 74.02 |

Subsequently the graph in figure 2 shows the precision values of each book domain. It can be seen the accuracy of the NBC approach in each book domain. The NBC approach yields high precision values in the religion, general and computer science domains. In the books from the biology, social politics, education, and economy-business domains the accuracy of the NBC approach is also quite high, more than 50%.
However figure 2 also shows that the NBC approach produces the lowest precision value in the economy-business domain. It is because of the semantic relatedness value feature in the economic-business dataset. In the economy-business books, many $S_1$ data have low semantic relatedness values that are even lower than $S_2$. Whereas $S_1$ is the text from the pages which number is indexed in the back-of-book index, so it should have a higher semantic relationship with the indexed term. This anomaly occurs because of many terms such as product names, companies or trademarks from economy-business books that are not found in the WordNet database. Although in general the NCB can identify the relevance of pages in economy-business books, the problem of semantic relatedness in the economy-business books dataset needs to be solved by providing a more complete lexical database.

6. Conclusion

The results of this study indicate that Naive Bayes Classification in our model can be used to identify the relevant and irrelevant pages to be indexed, hence our model could contribute in creating the excellent back-of-book index. The accuracy of the Naive Bayes Classification results, is influenced by the reliability of the identification model, the more training data being trained, the higher the accuracy level of the Naive Bayes Classification approach [18].

Therefore, for future studies, it is necessary to analyze more data sets and apply the other machine learning approach beside Naive Bayes Classification method, in order to find the more reliable methods for identifying the relevant and irrelevant pages for the back-of-book index. The availability of a lexical database that has a broader scope of knowledge also needed for improving the accuracy result.

References

[1] M. Coe, “What do readers expect from book indexes and how do they use them? An exploratory user study,” *Indexer*, vol. 33, no. 3, pp. 90–101, 2015.

[2] C. Gottschalg-duque and H. Lobin, “Ontology Extraction for Index Generation,” *Electron. Publ.*, no. June, pp. 111–120, 2004.

[3] O. Medelyan, “Thesaurus-based index term extraction,” in *In: The 6th Agricultural Ontology Service (AOS) Workshop*, 2005, pp. 1122–1129.
[4] T. Lahtinen, “Automatic indexing: an approach using an index term corpus and combining linguistic and statistical methods Timo Lahtinen," Helsinki, 2000.
[5] M. Coe, “Where is the evidence? A review of the literature on the usability of book indexes,” Indexer, vol. 32, no. 4, pp. 161–168, 2014.
[6] D. Cleveland and A. Cleveland, Introduction to Indexing and Abstracting 4th Edition, 4th ed. Libraries Unlimited, 2013.
[7] I. Fourier, “Book Indexing: A Reflection on the Contrasting Complexities and Ease of Conceptualization and How We Can Deepen Our Understanding.,” Mousaion, vol. 26, no. 1, pp. 111–125, 2008.
[8] Z. Wu and C. L. Giles, “Measuring Term Informativeness in Context,” in Proceedings of NAACL-HLT 2013, 2013, no. June, pp. 259–269.
[9] T. Dao, N. and T. Simpson, “Measuring Similarity between sentences,” 2005.
[10] S. Christina, “Kinerja Cosine Similarity dan Semantic Similarity dalam Pengidentifikasian Relevansi Nomor Halaman pada Daftar Indeks Istilah,” in Seminar Nasional Teknologi Informasi dan Komunikasi 2014 (SENTIKA 2014), 2014.
[11] D. Siahaan and S. Christina, Using semantic similarity for identifying relevant page numbers for indexed term of textual book, vol. 516. 2015.
[12] M. Banerjee, M. Capozzoli, L. Mcsweeney, and D. Sinha, “Beyond kappa: A review of interrater agreement measures *,” Can. J. Stat., vol. 27, no. 1, pp. 3–23, 1999.
[13] A. Singhal, “Modern information retrieval: A brief overview,” IEEE Data Eng. Bull., pp. 1–9, 2001.
[14] A. C. A. B, A. Deraman, M. Masila, and A. Jalil, “An Ensemble Similarity Model for Short Text Retrieval An Ensemble Similarity Model for Short,” in 17th International Conference on Computational Science and Its Applications, 2017, no. March 2018.
[15] S. Christina and E. D. Oktaviyani, “Identifying the relevant page numbers that referred by the back-of-book index using syntactic similarity and semantic similarity,” in Proceedings of the 2nd International Conference on Informatics and Computing, ICIC 2017, 2018, vol. 2018–January.
[16] N. Friedman, “Bayesian Network Classifiers,” Mach. Learn., vol. 29, no. 2–3, pp. 131–163, 1997.
[17] J. Dean, Big Data, Data Mining,and Machine Learning. Hoboken, New Jersey: John Wiley & Sons, Inc., 2014.
[18] B. Liu, E. Blasch, Y. Chen, D. Shen, and G. Chen, “Scalable Sentiment Classification for Big Data Analysis Using Naïve Bayes Classifier.pdf,” in Big Data, 2013 IEEE International Conference, 2013.
[19] I. Rish, “An empirical study of the naive Bayes classifier,” in Proceedings of IJCAI-2001 workshop on Empirical Methods in AI, 2001, pp. 41–46.
[20] V. Rizqiyan, A. Mulwinda, and D. Mahadji, “Klasifikasi Judul Buku dengan Algoritma Naive Bayes dan Pencarian Buku pada Perpustakaan Jurusan Teknik Elektro,” J. Tek. Elektro, vol. 9, no. 2, pp. 60–65, 2017.