New book classification based on *Dewey Decimal Classification* (DDC) law using tf-idf and *cosine similarity* method

To cite this article: Y Nurdiansyah *et al* 2019 *J. Phys.: Conf. Ser.* **1211** 012044

View the article online for updates and enhancements.
New book classification based on *Dewey Decimal Classification* (DDC) law using tf-idf and cosine similarity method

Y Nurdiansyah¹, A Andrianto¹, L Kamshal¹
¹,²,³ Faculty of Computer Science, University of Jember, Indonesia
E-Mail: yanuar_pssi@unej.ac.id, anang.uppti@unej.ac.id

**Abstract**

Classification new book is needed in facilitating students and lecturers to find books. The law used is Dewey Decimal Classification (DDC) classification. The application of the DDC classification requires a high level of accuracy and concentration in grouping books into appropriate classes. Errors that occur in the form of discrepancies in the provision of class books. Performance can be improved by the existence of an information system that can help classify classes in books according to DDC law. The process of giving classes to books by looking for the highest similarity between titles and synopsis of books with each DDC dictionary class. Adjusting to the process of giving classes to books at the University of Jember Library, the title, synopsis and DDC dictionary are processed using the text mining method. Text mining produces data in the form of basic words from the title, synopsis and DDC dictionary. The number of occurrences of each word is useful for measuring how important a word is in a document. The method that is suitable for calculating the importance of a word in a document is the method of weighting Term Frequency-Inverse Document Frequency (TF-IDF). The results of the TF-IDF weighting are used to find the highest similarity between the title and the synopsis with the class in the DDC dictionary. The appropriate method in calculating the similarity of two documents is Cosine Similarity. The biggest similarity value between the title and synopsis with the DDC dictionary using Cosine Similarity method is made a priority in determining the class of books. The results of the application of the method in the system there are 20 data books resulting in book classes in DDC 000 class there are 3 books, DDC 100 class is 1 book, DDC class 200 there is 1 book, DDC 300 class there are 6 books, DDC 400 class there are 4 books, DDC 500 class is 1 book, DDC 600 class there are 2 books and DDC 700 class there are 2 books. Testing book classification information system produces accuracy percentage of 35 %.

1. **Introduction**

In academic world, especially lecturers, is still in dire need of books. Books are beneficial for learning materials and references. The very important role of books makes one of the reasons for providing library facilities. Every educational institution usually has its library.

Jember University has library facility, namely UPT-Perpustakaan. UPT-Perpustakaan provides many books according to the needs of students and lecturers. Classification is needed in facilitating students and lecturers to find books. The UPT-Perpustakaan makes policies using the Dewey Decimal Classification (DDC)
classification, because the DDC classification is very suitable with the many types of books owned by the UPT-Perpustakaan and has been widely used on an international scale [1].

The development of science makes the availability of books should always be updated. The renewal of the availability of books at the UPT-Perpustakaan is always done by increasing the number of books. The application of the DDC classification requires a high level of accuracy and concentration in grouping books into appropriate classes [2].

The addition of a large number of books requires more time and concentration in grouping books. Errors that occur in the form of discrepancies in the provision of class books. The performance of UPT-Perpustakaan can be improved by the existence of a system that can help classify classes in books according to DDC law. The process of giving classes to books in the UPT-Perpustakaan is to look for the highest similarity between the title and the synopsis of the book with each class in the DDC dictionary. Adjusting to the process of giving classes to books in the UPT-Perpustakaan, the title, synopsis and DDC dictionary are processed using the text mining method [3]. Text mining produces data in the form of basic words from the title, synopsis and DDC dictionary [4] [5]. Each basic word can have an appearance on more than one document. The number of occurrences of each word is useful for measuring how important a word is in a document. The method that is suitable for calculating the importance of a word in a document is the method of weighting Term Frequency-Inverse Document Frequency (TF-IDF) [6].

The results of the TF-IDF weighting are used to find the highest similarity between the title and synopsis with the class in the DDC dictionary [7]. The appropriate method in calculating the similarity of two documents is Cosine Similarity. The biggest similarity value between the title and synopsis with the DDC dictionary using the Cosine Similarity method is made a priority in determining the class of books [8]. The above problems in the class classification process in books that require a lot of time and high concentration can be handled by building a system that applies the TF-IDF and Cosine Similarity methods. TF-IDF and Cosine Similarity can handle these problems by shortening the processing time and giving the appropriate class name [9] [10].

Term frequency (TF) is a measurement that is applied in the weighting method. Each term is assumed to have a proportion of interests by the number of occurrences in the document. Term frequency can correct recalled values in information retrieval, but does not always improve precision values. This is because the frequent terms tend to appear in a lot of texts, so these terms have a small distinguishing power (uniqueness). Inverse document frequency (IDF) is a weighting method that focuses more on the appearance of terms in the entire collection of text. In IDF, terms that rarely appear on all documents are considered more valuable. The value of importance for each term is assumed to be inversely proportional to the number of texts containing the term. Weighting can be obtained based on the number of occurrences of a term (word) in a document term frequency (TF) and the number of occurrences of the term in the inverse document frequency (IDF) in the document collection. The IDF value of a term can be calculated using the equation 1

\[ IDF = \log \frac{D}{df_t} \]  ...equation 1

Where D is the number of documents and df_t is the number of occurrences (frequency) terms of D.

The equation used to calculate the weight (W) of each document to the keyword (query), can be solved with equation 2:

\[ W_{d,t} = tf_{d,t} * IDF_t \]  ...equation 2

with the following information:
- d = d document
- t = t-term of keywords
- tf = term frequency / word frequency
Wd, t = the weight of the d-document against the t-term.

Cosine Similarity is a method used to calculate the level of similarity between two objects. For document clustering purposes, a good function is the Cosine Similarity function

\[ \text{Similarity}(X, Y) = \frac{|X \cap Y|}{|X|^2 |Y|^2} \quad \text{equation 3} \]

where |X ∩ Y| is the number of terms contained in document X and those in document Y, |X| is the number of terms contained in document X and |Y| is the number of terms contained in document Y. From the set notation above can be made a mathematical equation as in equation 4

\[ \text{Similarity}(X, Y) = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \quad \text{equation 4} \]

Where x and y are different documents, xi is the i term contained in document x which is the term contained in document y.

2. Result and Discussion

Research on the development of information systems to classify library books at the University of Jember Library used the System Development Life Cycle (SDLC) waterfall [11]. Waterfall model according to the stages as in Figure 1.

![Figure 1 The Phase-in Waterfall Method](image)

Stages of needs analysis in the research that had been carried out began with the data collection phase including literature study and observation. The next stage is the data processing stage and an overview of the system to be built.

a. Data Collecting Process

Data collection in this study was carried out in several stages of data collection as follows:

1. Conduct a literature study to be the part of data collection steps used in research. The sources used in literature studies are books, journals, scientific papers, previous research and scientific works relating to Text mining, TF-IDF, Cosine similarity and Dewey Decimal Classification.

2. Conduct interviews by giving questions to librarians at the University of Jember Library to obtain the data needed to find out the problems being faced by librarians at the University of Jember Library.
3. Observation is done by visiting the University of Jember Library. Observations were carried out with the aim to get the data sample title and the synopsis of books in the University of Jember Library.

b. Data Processing Stage

The data processing stage begins by reviewing the overall data that has been obtained from the data collection stage. The collected data is used to identify functional needs and non-functional needs of the system. Observation data is then processed into a database.

New Book Classification Information System By Dewey Decimal Classification (DDC) Law Using the TF-IDF Method and Cosine Similarity is a system that classifies new books in Jember University libraries. The purpose of system development is to assist library staffs in classifying new books. This system can classify new books based on DDC law using the TF-IDF and Cosine similarity methods. The data needed in determining the class is a DDC dictionary, book title and book synopsis. There are two system access rights, namely admin and operator. The system can manage user data, book data, DDC dictionary data and find out the results of new book classes. The system can record user activity history. The system is built based on the web to facilitate data communication between users.

The system design stage is a stage in system development planning. System design as follows:

a. business process

Business process is a diagram that describes the data needed by the system. Business process consists of five parts, namely:
1. Input: Data entered into the system
2. Output: Data generated by the system
3. Goal: The purpose of building a system
4. Uses: Platform that is the basis of the system
5. Process: System that works

b. Use case Diagram

Use case diagram is illustrated by interactions that can be carried out by actors with their access rights. Use case diagrams are shown in Figure 2.

![Use case Diagram](image-url)

Figure 2. Use case Diagram of Library Book Classification Information System
2.1. Calculation of TF-IDF Method and Cosine Similarity for library book classification

Calculation of TF-IDF and Cosine Similarity methods required e-DDC Edition 23 (Version 3.3). Attributes as important words from the dictionary of each class that will represent the process of calculating the Cosine Similarity. The determinant attributes used to process DDC dictionary data are as follows:

1. DDC Words Dictionary
   It is an attribute of the DDC dictionary word that is in the form of sentences in the condition of having many add-ins or not basic words.

2. DDC Words Class
   Is a class attribute for the DDC dictionary word that describes the DDC dictionary word included in the class with the range between classes 0 to 9.

DDC dictionary dataset is 103 data, sample dataset which includes class 0 is found in Table 1.

| DDC Dictionary Word                          | Class |
|----------------------------------------------|-------|
| General Science Work                         | 0     |
| Classify here discussions from various fields| 0     |
| Includes general science                     | 0     |
| Bibliography, bibliography with specific topics| 0    |
| Classify here analytical bibliography        | 0     |
| Library and information science              | 0     |
| Classify here archive science, archival science| 0    |

Information:
Word Dictionary DDC = Word obtained from the DDC dictionary
Class = DDC Data Dictionary Class

The initial stage of the workings of the TF-IDF calculation process is to retrieve DDC dictionary data that is processed into text mining to produce basic words. Classification can be determined by the following steps:

1. The basic words are calculated in frequency as in Table 2 below:

| Words          | D1 | D2 | D3 |
|----------------|----|----|----|
| Science        | 2  | 0  | 3  |
| Knowledge      | 2  | 0  | 0  |
| Class          | 1  | 1  | 1  |
| In             | 1  | 1  | 1  |
| Here           | 1  | 1  | 1  |
| Discussion     | 1  | 0  | 0  |
| From           | 1  | 0  | 0  |
| As             | 1  | 0  | 0  |
| Field          | 1  | 0  | 0  |
| Enter          | 1  | 0  | 0  |
| General        | 1  | 0  | 0  |
| Bibliography   | 0  | 4  | 0  |
| With           | 0  | 1  | 0  |
| Topic          | 0  | 1  | 0  |
| Specific       | 0  | 1  | 0  |
| Analytic       | 0  | 1  | 0  |
| Descriptive    | 0  | 1  | 0  |
| Library        | 0  | 0  | 1  |
| And            | 0  | 0  | 1  |
| Information    | 0  | 0  | 1  |
| Archive        | 0  | 0  | 2  |

Information:
Word = Word obtained from the DDC dictionary text mining process
TF = Term frequency

5
D1 = subclass 1 DDC dictionary 001 - 009
D2 = subclass 2 DDC dictionary 010 - 019
D3 = subclass 3 DDC dictionary 020 – 029

2. The results of the TF are calculated using the IDF value using the IDF equation which can be seen in equation 1 and 2.
3. After getting the IDF value for each word, it is calculated using the weight calculation formula.
4. The results of these weights are searched for overall averages.
5. The results of the average weights are compared to each weight of the word, words that have a weight greater than the average weight will be stored into important words that represent the class.
6. The next step is to process the book data which includes the title and synopsis that will be given a text mining process which will eventually produce a base word for the book's data. The following is a sample test book data that has been given the text mining process in
7. The basic words of the text mining result are then calculated TF (term frequency).
8. Look for the average of TF results.
9. Words with a frequency greater than the average value of TF will be considered as the basic words that represent the book.
10. Then applied the equation of Cosine Similarity by comparing the basic words in the book data compared to the results of the process from the fifth stage above.
11. The class that will be generated for the book data is from the biggest value of Cosine Similarity calculations.

2.2 Implementation of the TF-IDF Method and Cosine Similarity for library book classification

The research implementation on the TF-IDF and Cosine Similarity methods is used in the classification process to determine the new book class based on Dewey Decimal Classification (DDC) law. The attributes used for classification are DDC dictionaries, titles and synopsis of books. DDC dictionary, title and book synopsis are all in the form of text, so the text mining process is needed to change all words into basic words before applying to the TF-IDF and Cosine Similarity methods [12].

The first stage of classification is to process words into basic words using the text mining method. Each important word is saved in the database. After calculating the frequency of each word, the next step is to calculate the inverse document frequency (IDF) value for each word in each class. After calculating the IDF value of each word in each class, the next step is to calculate the weight value (W) of each word in each class. After each class has a word weight, the next process is to process the book data that the class wants to know. The data of the book is taken the title and synopsis to be processed using text mining to get the basic words. After gaining weight, the next process compares the weight of the words in the title and the book synopsis with the word weight of each class in the DDC law using the Cosine Similarity method.

The results of the Cosine Similarity calculation process will find out which class with the highest Cosine Similarity value. Where when the class has the highest Cosine Similarity value, then the class is appropriate for the book that has been previously processed by the title and synopsis.

2.3 Classification Validity Testing Using the TF-IDF Method and Cosine Similarity Validity testing is one of the stages in the data mining process, which is used to determine the percentage of validity of the classification results. The results of the tests that have been carried out on 20 books data can be seen in Table 3.
Table 3. Testing Using 20 books

| No | Book Titles                                                | Class System Results                                      | Manual Class Results                                      |
|----|------------------------------------------------------------|------------------------------------------------------------|------------------------------------------------------------|
| 1  | Philosophy and Ethics                                      | DDC 100 - Philosophy and Psychology                        | DDC 100 - Philosophy and Psychology                        |
| 2  | Practical Approach of the Statistical Analysis with Microsoft Excel | DDC 600 - Technology and Applied Sciences                   | DDC 000 - Generalities                                     |
| 3  | Easy Way to Calculate Corporate Income Tax with Tax Law    | DDC 000 - Generalities                                     | DDC 300 - Social Science                                   |
| 4  | Prime Dimensions of State Administration in the Age of Challenges | DDC 000 - Generalities                                     | DDC 300 - Social Science                                   |
| 5  | Dynamics of Ethics and Medical Law in the Age of Challenges | DDC 300 - Social Science                                   | DDC 100 - Philosophy and Psychology                        |
| 6  | English Competency Reading Comprehension                   | DDC 400 - Language                                         | DDC 400 - Language                                         |
| 7  | Filology                                                   | DDC 400 - Language                                         | DDC 400 - Language                                         |
| 8  | Teachers in Implementation of 2013 Curriculum               | DDC 300 - Social Science                                   | DDC 300 - Social Science                                   |
| 9  | Healthy Living with Disease Problems                       | DDC 600 - Technology and Applied Sciences                   | DDC 000 - Generalities                                     |
| 10 | Language Research Methods                                  | DDC 000 - Generalities                                     | DDC 400 - Language                                         |
| 11 | Modernity in the Village, The influence of group housing complexes on people's settlements in Semarang in the 20th century | DDC 300 - Social Science                                   | DDC 700 - The Art, Fine and Sport                           |
| 12 | Change Pioneers                                            | DDC 300 - Social Science                                   | DDC 900 - Geography and History                            |
| 13 | Integrated Education Paradigm, Preparing Altab Ulul Generation | DDC 200 - Religion                                         | DDC 200 - Religion                                         |
| 14 | C ++ Programming for beginners                             | DDC 400 - Language                                         | DDC 000 - Generalities                                     |
| 15 | Web Programming Based on HTML5, PHP, and JavaScript        | DDC 400 - Language                                         | DDC 000 - Generalities                                     |
| 16 | Economic Assessment of Mangrove Forest Ecosystems and their Application in Coastal Area Planning | DDC 300 - Social Science                                   | DDC 500 - Natural Science and Mathematics                  |
| 17 | Literature Performance: Reading Literary Work as a Art of Spectacle | DDC 700 - The Art, Fine and Sport                           | DDC 700 - The Art, Fine and Sport                           |
| 18 | Positive Sufism                                            | DDC 300 - Social Science                                   | DDC 200 - Religion                                         |
| 19 | Personality Theories                                       | DDC 500 - Natural Science and Mathematics                  | DDC 100 - Philosophy and Psychology                        |
| 20 | Understanding the Art of Psychiatry Interview              | DDC 700 - The Art, Fine and Sport                           | DDC 600 - Technology and Applied Sciences                  |

Based on the results of testing of 20 documents, the calculation of validity percentage can be seen as follows:

\[
\text{Validation Percentage} = \frac{\text{Correct Data}}{\text{Data Amount}} \times 100 = \frac{10}{20} = 50% 
\]

Validity test results show that the TF-IDF and Cosine Similarity method for library book classification is less than 50% so it is not suitable to be applied in case studies at the University of Jember Library.

3. Conclusion

The application of the TF-IDF Method and Cosine Similarity for classifying books at the Jember University Library is to determine the class of a book. The first step to classification is by processing the DDC dictionary data using the TF-IDF method on the DDC dictionary. Dictionary data used in this study were 103 data with ten DDC classes. The calculation is done based on the value of the dictionary's civil frequency so that it produces the word weight for each DDC class. Processing of the TF-IDF method is also applied to the title and synopsis of the book data to produce the word weight of the book data. The weight of the words from the book data is compared to the class weight of DDC using the Cosine Similarity method. The result of the Cosine Similarity value is the biggest value so that the class obtained from the largest value is the class for the book.
The results of the application of the method in the system there are 20 data books produce class books where DDC 000 class there are 3 books, DDC class 100 there is 1 book, DDC 200 class there is 1 book, DDC class 300 there are 6 books, DDC class 400 there are 4 books, class There is 1 DDC 500 book, DDC class 600 there are 2 books and DDC 700 class there are 2 books. Testing the book classification information system produces an accuracy percentage of 35% because the Cosine Similarity method is very influential on the difference in the number of words and the weight of the word in the book classification process, for example when there is a bibliographic word on book data with a frequency of 5 compared to the bibliographic words on dictionary data with frequency 1 then the Cosine Similarity value is smaller than if the bibliographic word on the book data and dictionary data is the same frequency 3. This research can be concluded that the Cosine Similarity method is less suitable to be applied to the object of this study.

The book classification information system at the University of Jember Library was built based on the website has been designed and built using the SDLC model of the waterfall development method, where the system requirements have been defined at the beginning of system development. Need analysis is done by defining the system in detail to get complete functional and non-functional needs. Complete needs analysis is very important, because the initial construction of the waterfall model. The process of working on the waterfall model is done sequentially; if there are changes or additional features, the analysis process will be carried out again. The next stage is development that includes analysis, design, writing program code and system testing. The design phase is done by designing the system by making business processes, use case diagrams, scenarios, activity diagrams, sequence diagrams, class diagrams and ERD which are used as references in writing program code. The writing stage of the program code uses the PHP programming language, Code Igniter framework and MySQL management database. The testing phase is carried out by means of the black box method, white box and validity test. In this study SDLC waterfall can be applied well, because the stages in this research are carried out in a sequential or systematic manner. Difficulties faced in this study when implementing SDLC waterfall if something goes wrong must see one step above, when there is an error it takes longer to fix it.

Acknowledgment

We express our gratitude to the Ministry of Research, Technology and Higher Education of the Republic of Indonesia and the University of Jember for grant and support so that this article published. We also acknowledge the anonymous reviewers for improving this article. We are solely responsible for the content of the paper.

References

[1] Landry, Patrice Dewey 2006 Decimal Classification (DDC) at the Swiss National Library Cataloging & Classification Quarterly 42 3–4 pp 135–45.

[2] Beall, Julianne, and Joan S Mitchell 2010 History of the Representation of the DDC in the MARC Classification Format Cataloging & Classification Quarterly 48 1 pp 48–63.

[3] Nurdiansyah Y, S Bukhori and R Hidayat 2018 Sentiment Analysis System for Movie Review in Bahasa Indonesia Using Naive Bayes Classifier Method In Journal of Physics: Conference Series 1008.

[4] Joorabchi, Arash, and Abdulhussain E Mahdi 2011 An Unsupervised Approach to Automatic Classification of Scientific Literature Utilizing Bibliographic Metadata Journal of Information Science 37 5 pp 499–514.

[5] Pong, Joanna Yi-Hang, Ron Chi-Wai Kwok, Raymond Yiu-Keung Lau, Jin-Xing Hao and Percy Ching-Chi Wong 2008 A Comparative Study of Two Automatic Document Classification Methods in a Library Setting Journal of Information Science 34 2 pp 213–30.
[6] Qaiser, Shahzad and Ramsha Ali 2018 Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents *International Journal of Computer Applications* 181 1 pp 25–29.

[7] Wang and Jun 2009 An Extensive Study on Automated Dewey Decimal Classification *Journal of the American Society for Information Science and Technology* 60 11 pp 2269–86.

[8] Nurdiansyah Y, F Nur Muharrom and F Firdaus 2018 Implementation of Winnowing Algorithm Based K-Gram to Identify Plagiarism on File Text-Based Document In *MATEC Web of Conferences* 164

[9] Kumari, Madhu, Akshat Jain and Ankit Bhatia 2016 Synonyms Based Term Weighting Scheme: An Extension to TF.IDF *Procedia Computer Science* 89 pp 555–61.

[10] Goswami, Puneet and Vidya Kamath 2014 The DF-ICF Algorithm-Modified TF-IDF *International Journal of Computer Applications* 93

[11] Nurdiansyah, Yanuar, Dwiretno S T Istiyadi and Ragil Illyandi I Erick Putra 2014 Konferensi Nasional Ilmu Komputer (KONIK) Implementasi Algoritma Aes-128 pada Mobile Learning Universitas Jember.

[12] Sejal, D T Ganeshsingh, K R Venugopal, S S Iyengar and L M Patnaik 2016 Image Recommendation Based on ANOVA Cosine Similarity *Procedia Computer Science* 89 pp 562–67.