Implementation of Fuzzy C-Means algorithm to classifying research topics in informatics department, UIN Sunan Gunung

M I N Saputra*, D Fauzy, R A Hakim, P Dauni, M D Firdaus and I Taufik
Department of Informatics, UIN Sunan Gunung Djati Bandung, Bandung, Indonesia

*indra@if.uinsgd.ac.id

Abstract. The more rapid and large number of electronic documents stored in University library repositories or Departments, such as scientific works from academicians including theses, research reports, etc., are available in the digital version. In the process of grouping the thesis title automatically it is hoped that it can help policy makers such as the head of the study program. In this case the researcher uses fuzzy c-means algorithm which can group data in clusters so that the data of a cluster has a high level of equality with each other. In the application there is a text mining method which is a development of data mining that can be applied to overcome thesis grouping problems or research data. The result is a web that can classify research data / thesis data with a total accuracy of 96%. This number is the accuracy given based on the performance of the fuzzy c-means algorithm on the system.

1. Introduction
There are many electronic documents saved in university or department library repository, like academic writing from academic community such as thesis, research report, report on practical work and many more available in digital version. However, this is not accompanied by knowledge that can extract the information needed from these electronic documents. Therefore a method is needed to make the classification of these documents ease and simplify [1]. One of the method that can be used is text mining [2,3].

Text mining method is a development of data mining that can be applied to overcome this problem [4]. The algorithm in text mining are made to be able to recognize semi-structured data such as synopsis, abstracts, and the content of documents [1,2,5].

Text mining applications have been applied in libraries, especially for searching text-based library materials [6]. However, not many applications have been developed for the purpose of analysis, so it is very difficult to be able to immediately find out popular research topics in a given year or the trends in student research interests in a department [7]. Documents are classified into several categories referring to the expertise group approved by the informatics department [8].

Clustering is the process of grouping data in classes or clusters so that data in a cluster has a high level of equality with each other, but it is very different from the data in other clusters. At present there are several clustering algorithms known as Fuzzy C-Means Clustering Algorithm. Fuzzy C-Means is one of the clustering methods that are part of hard K-Means. Fuzzy C-Means uses the fuzzy approach concept so that a data can become a member of all existing clusters. Membership matrix is formed with the level of membership of each data in each cluster that has a value between 0 and 1. The level of
existence of data in a cluster is determined by the degree of membership. With the grouping of similar documents in advance is expected that later it can improve the performance of text document categorization and save more computational time to choose the closest neighbor of the document. This is because the search for the nearest document is only sufficiently searched for within the scope of the group documents that are similar and do not need to be compared to the entire collection of documents [9,10].

In the process of grouping the title of thesis automatically it is hoped that it can help policy makers such as the head of department, dean, or the head of the library to be able to quickly analyze the topic of the thesis and the desired year.

2. Methodology
This study aims to develop and implement clustering methods using the Fuzzy C-Means Clustering algorithm to group data contained in the class [11].

The Fuzzy C-Means clustering algorithm is as follows:

- Input data to be clusters X, in the form of matrix size \( n \times m \) (\( n = \) data delicate, \( m = \) attribute of each data). \( X_{ij} \) = the first bit data (\( i = 1, 2, \ldots, n \)), the \( j \)-attribute (\( j = 1, 2, \ldots, m \)).
- Determine:
  - Number of cluster = \( c \);
  - Rank = \( w \);
  - Maximum iteration = \( \text{maxIter} \);
  - Expected smallest error = \( \varepsilon \);
  - Initial objective function = \( P_0 = 0 \);
  - Initial iteration = \( t = 1 \);
- Generate a random value of \( \mu_{ik} (i = 1, 2, \ldots, n), (j = 1, 2, \ldots, c) \). Position and matrix values are built randomly, where th value of membership lies in intervals of 0 and 1.

\[
Q_j = \sum_{k=1}^{c} \mu_{ik} = 1
\]

\( Q_j \) is the number of degrees of membership per column \( = 1 \) with \( j = 1, 2, \ldots, m \). Calculate:

\[
\mu_{ik} = \frac{\mu_{ik}}{Q_j}
\]

- Calculate the center of cluster \( V_{kj} (k = 1, 2, \ldots, c), (j = 1, 2, \ldots, m) \)

\[
V_{kj} = \frac{\sum_{i=1}^{n} (\mu_{ik})^w * X_{ij}}{\sum_{i=1}^{n} (\mu_{ik})^w}
\]

- Calculate the objective function in \( t \) iteration

\[
P_t = \sum_{i=1}^{n} \sum_{k=1}^{c} \left[ \sum_{j=1}^{m} (X_{t,j} - V_{kj})^2 (\mu_{ik}) \right]
\]

- Calculate changes to the matrix partition

\[
\mu_{ik} = -\frac{\left[ \sum_{j=1}^{m} (X_{ij} - X_{kj})^2 \right]^{w-1}}{\sum_{k=1}^{c} \left[ \sum_{j=1}^{m} (X_{ij} - X_{kj})^2 \right]^{w-1}}
\]

With \( i = 1, 2, \ldots, n \) and \( k = 1, 2, \ldots, c \)
• Check stop conditions:
  o If: \(|P_t - P_{t-1}| < \varepsilon\) or \(t > \text{maxIter}\) then stop
  o If not: \(t = t + 1\), and repeat fourth step

3. Result and discussion
The first step is input the document. The document needed for this is an abstract document. TF-IDF Algorithm is used first to determine the root word [10], and every word rearrange to make a keywords [1]. The second step, initialize the limit of cluster, maximum iteration, weight, and minimum error value. The limit of cluster is based on the number of expertise groups. The keyword formed according to the first step is grouped based on the expertise groups in the department shown in table 1.

Table 1. Clustering the keyword based on the expertise group.

| Keywords                                      | First expertise group | Second expertise group | Third expertise group | Forth expertise group |
|-----------------------------------------------|-----------------------|------------------------|-----------------------|-----------------------|
| naïve bayes classifier                        | 1                     | 0                      | 0                     | 0                     |
| data training                                 | 1                     | 0                      | 0                     | 0                     |
| Database                                      | 1                     | 0                      | 0                     | 0                     |
| Kecocokan                                     | 1                     | 0                      | 0                     | 0                     |
| Action Roleplaying Game Finite State Machine  | 0                     | 0                      | 0                     | 1                     |
| Kecerdasan Buatan                             | 0                     | 0                      | 0                     | 1                     |
| Random Number Generator                      | 0                     | 0                      | 0                     | 1                     |
| Logika Fuzzy                                  | 1                     | 0                      | 0                     | 0                     |
| K-Means Shift Clustering                     | 1                     | 0                      | 0                     | 0                     |
| Interpolasi Linier                           | 1                     | 0                      | 0                     | 0                     |
| Fisher Yates Shuffle                          | 1                     | 0                      | 0                     | 0                     |
| Linear Congruent Method                      | 1                     | 0                      | 0                     | 0                     |
| Fuzzy C-Means                                | 1                     | 0                      | 0                     | 0                     |
| Clustering                                    | 1                     | 0                      | 0                     | 0                     |
| Search Engine                                 | 1                     | 0                      | 0                     | 0                     |
| Pencarian                                     | 1                     | 0                      | 0                     | 0                     |
| Forward Chaining                              | 0                     | 0                      | 0                     | 1                     |
| Expert System                                 | 0                     | 0                      | 0                     | 1                     |
| string matching                               | 1                     | 0                      | 0                     | 0                     |
| Sistem Informasi                              | 1                     | 0                      | 0                     | 0                     |
| MySQL                                         | 1                     | 0                      | 0                     | 0                     |
| Sistem Informasi Geografis                    | 1                     | 0                      | 0                     | 0                     |
| Inventori                                     | 1                     | 0                      | 0                     | 0                     |
| Sistem Pakar                                  | 0                     | 0                      | 0                     | 1                     |
| Quality of Service                            | 0                     | 0                      | 1                     | 0                     |
| Jaringan                                      | 0                     | 0                      | 0                     | 1                     |
| Network                                       | 0                     | 1                      | 0                     | 0                     |
| Mikrotik                                      | 0                     | 1                      | 0                     | 0                     |
| Automate                                      | 0                     | 0                      | 1                     | 0                     |

Add the value for each keyword formed shown in table 2. For maximum iteration, rank and expected error value, used value of 100, 2, and 0.000001 respectively. Third step, generate random value for each cluster. The random value generated has a range from 0 to 1 with a total equal to 1 for the entire cluster. Fourth step, calculate the centre of cluster. Fifth, calculate the objective function that describes the distance from a given data point to the centre of the cluster. Sixth, calculate changes the the matrix partition. Seven, check whether the centre of the cluster is less than minimum error value or number of iteration is more than maximum iteration. If both of cases is true, then the steps is finish. If not, increase the number of iteration, and recalculate the centre of cluster until the centre get less than minumum error value or maximum iteration is reach.
Table 2. Weight value for every keyword formed.

| Keywords                                                                 | Value for expertise group |
|-------------------------------------------------------------------------|---------------------------|
| naive bayes classifier                                                  | 0.62                      |
| data training                                                           | 0.62                      |
| Database                                                                | 0.62                      |
| Kecocokan                                                               | 0.62                      |
| Action Roleplaying Game Finite State Machine                            | 0.21                      |
| Kecerdasan Buatan                                                       | 0.21                      |
| Random Number Generator                                                 | 0.21                      |
| Logika Fuzzy                                                            | 0.62                      |
| K-Means Shift Clustering                                               | 0.62                      |
| Interpolasi Linier                                                      | 0.62                      |
| Fisher Yates Shuffle                                                    | 0.62                      |
| Linear Congruent Method                                                 | 0.62                      |
| Fuzzy C-Means                                                          | 0.62                      |
| Clustering                                                              | 0.62                      |
| Search Engine                                                           | 0.62                      |
| Pencarian                                                               | 0.62                      |
| Forward Chaining                                                       | 0.21                      |
| Expert System                                                           | 0.21                      |
| string matching                                                        | 0.62                      |
| Sistem Informasi                                                        | 0.62                      |
| MySQL                                                                   | 0.62                      |
| Sistem Informasi Geografis                                             | 0.62                      |
| Inventori                                                               | 0.62                      |
| Sistem Pakar                                                            | 0.21                      |
| Quality of Service                                                      | 0.14                      |
| Jaringan                                                                | 0.14                      |
| Network                                                                 | 0.14                      |
| Mikrotik                                                                | 0.14                      |
| Automate                                                                | 0.03                      |

For testing purpose, 10 abstract documents are tested with different number of iterate for each document. The result of the test is shown at table 3.
Table 3. Testing of abstract document.

| No | Abstract titles                                                                 | First expertise group | Second expertise group | Third expertise group | Forth expertise group | Number of test | Result | Expected expertise group |
|----|--------------------------------------------------------------------------------|-----------------------|------------------------|-----------------------|-----------------------|----------------|--------|--------------------------|
| 1  | Analisis Perbandingan Algoritma Fisher Yates Shuffling dan Linear Congruent Method Untuk Pengacakan Soal Ilmu Nahwu | 2                     | 0                      | 0                     | 0                     | 2              | 100    | First expertise group    |
| 2  | Penerapan Kecerdasan Buatan Finite State Machine Dalam Permainan Action Roleplaying Game | 0                     | 0                      | 0                     | 3                     | 3              | 100    | Forth expertise group    |
| 3  | Aplikasi pembelajaran rumus kimia untuk kelas X sma berbasis android menggunakan Framework phonegap | 0                     | 2                      | 0                     | 0                     | 2              | 100    | Second expertise group   |
| 4  | Aplikasi Info Halal Menggunakan Barcode Scanner Pada Smartphone Android          | 0                     | 3                      | 0                     | 0                     | 3              | 100    | Second expertise group   |
| 5  | Menentukan Dress Up Matching Dengan Metode Naïve Bayes                           | 4                     | 0                      | 0                     | 0                     | 4              | 100    | First expertise group    |
| 6  | Sistem Pakar Diagnosa Penyakit Sayuran Tomat Dengan Menggunakan Metode Forward Chaining | 0                     | 0                      | 0                     | 2                     | 2              | 100    | Fourth expertise group   |
| 7  | Aplikasi Sistem Informasi Geografis (Sig) Kecamatan Pangalengan Kabupaten Bandung Untuk Data Kependudukan | 2                     | 0                      | 3                     | 0                     | 5              | 60     | Third expertise group    |
| 8  | Implementasi algoritma fuzzy C-Means pada perangkat lunak Bantu pengklaeteran hasil pencarian hadits shahih | 4                     | 0                      | 0                     | 0                     | 4              | 100    | First expertise group    |
| 9  | Rancang bangun aplikasi pencarian ayat alqur’an berbasis Mobile                  | 0                     | 3                      | 0                     | 0                     | 3              | 100    | Second expertise group   |
| 10 | Sistem Pakar Diagnosa Penyakit Kulit Pada Anak                                     | 0                     | 0                      | 0                     | 1                     | 1              | 100    | Fourth expertise group   |

Can be seen in table 3 in the result column there are still result that are still less than 100, because in the test there are two data that result in the accuracy of percentage is not 100. The test results produce an accuracy value in the form of a percentage then it can be concluded as the accuracy of the performance of the fuzzy c-means algorithm [12].

\[
Accuracy = \frac{\sum Result}{Number of documents}
\]

\[
Accuracy = \frac{960}{10}
\]

Accuracy is a formula for calculating the accuracy done after testing with 10 abstract documents with 4 clusters obtained total accuracy of 96%.
4. Conclusion
This Fuzzy C-Means Algorithm to classified the research topic have been tested with 10 documents. The content of the document is the abstract of research topic. The results of the test show that the accuracy obtained is 96%.

The result out after testing with 4 clusters and tested 29 times with a total accuracy of 96%. This number is the accuracy given based on the performance of the Fuzzy C-Means Algorithm.

Acknowledgments
Authors wishing to acknowledge Research and Publication Centre of UIN Sunan Gunung Djati Bandung that supports and funds this research publication.

References
[1] Zulfikar W B, Irfan M, Alam C N and Indra M 2017 The comparation of text mining with Naive Bayes classifier, nearest neighbor, and decision tree to detect Indonesian swear words on Twitter in 2017 5th International Conference on Cyber and IT Service Management (CITSM) 1–5
[2] Maylawati D S A 2015 Pembangunan library pre-processing untuk text mining dengan representasi himpunan frequent word itemset (hfwi) studi kasus: bahasa gaul Indonesia (Bandung Institute of Technology)
[3] Prilianti K R and Kunci K 2014 Aplikasi Text Mining untuk Automasi Penentuan Tren Topik Skripsi dengan Metode K-Means Clustering Daft. Pustaka 2(1) 1–6
[4] Maylawati D S, Aulawi H and Ramdhani M A 2018 The concept of sequential pattern mining for text IOP Conf. Ser. Mater. Sci. Eng. 434(1) 012042
[5] Weinberger M 2015 How to get a software developer to work for free Bus. Insid. 1(1) 60–76
[6] Sa’adillah D M and Putri Saptawati G A 2017 Set of Frequent Word Item sets as Feature Representation for Text with Indonesian Slang J. Phys. Conf. Ser. 801(1) 012066
[7] Maylawati D S, Rahman A, Saputra M I N, Darmalaksana W and Ramdhani M A 2018 Model of Cytation Network Analysis using Sequence of Words as Structured Text Representation IOP Conf. Ser. Mater. Sci. Eng. 288(1) 012048
[8] Neill G 2009 Sales on e-books ‘held back’ Bookseller 5367 6
[9] Ke- PS, Tiap, F. C-means, F. Teknik, and U. M. Kudus Fakultas Teknik – Universitas Muria Kudus 79 no. 2000, pp. 79–84, 2014.
[10] Irfan M, Jumadi W B, Zulfikar and Erik 2017 Implementation of Fuzzy C-Means algorithm and TF-IDF on English journal summary in 2017 Second International Conference on Informatics and Computing (ICIC) 1–5
[11] Kessel G, Pada C and Lq I 2015 1 , 2 , 3 I 4 543–551
[12] Widodo P, Putra J A, Afridhi S, Arifin A Z and Herumurti D 2016 Klasifikasi Kategori Dokumen Berita Berbahasa Indonesia Dengan Metode Kategorisasi Multi- Label Berbasis Domain Specific Ontology II(2)