Computer Engineering and Applications Vol. 5, No. 1, February 2016

Comparison Jaccard similarity, Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method

Lisna Zahrotun

Department of Informatics Engineering, Faculty of Industrial Technology, Universitas Ahmad Dahlan
lisna.zahrotun@tif.uad.ac.id

ABSTRACT

Text Mining is the excavations carried out by the computer to get something new that comes from information extracted automatically from data sources of different text. Clustering technique itself is a grouping technique that is widely used in data mining. The aim of this study was to find the most optimum value similarity. Jaccard similarity method used similarity, cosine similarity and a combination of Jaccard similarity and cosine similarity. By combining the two similarity is expected to increase the value of the similarity of the two titles. While the document is used only in the form of a title document of practical work in the Department of Informatics Engineering University of Ahmad Dahlan. All these articles have been through the process of preprocessing beforehand. And the method used is the method of document clustering with Shared Nearest Neighbor (SNN). Results from this study is the cosine similarity method gives the best value of proximity or similarity compared to Jaccard similarity and a combination of both

Keywords: ***shared nearest neighbour, text mining***, ***jaccard similarity, cosine similarity***

1. I NTRODUCTION

Data mining is often referred to as knowledge discovery in databases (KDD) is an activity that includes the collection, use historical data to find regularities, patterns of relationships in large data sets [1]. Text mining together with data mining, if data mining consists of data stored in the data base of the text mining the data in the form of documents such as emails, documents and other news.

According to Feldman and Dagan one thing in common in text mining-related research is to represent the text as a collection of words, or better known as Bag-of-Words approach or the Bow [2]. Text mining has penetrated in various fields including the field of security, biomedical, software and applications, online media applications, marketing applications, sentiment analysis and academic applications [3]. One technique in data mining is clustering or grouping. Clustering is useful for finding a group of data in order to obtain data more easily analyzed [4]

Research on the problem of grouping the document has been done through various methods. For example the use of the method of K-Nearest Neigbour (KNN) for categorization of text [5], grouping text data with the fuzzy c-means [6], the grouping with AHC single linked and K-Means [7], and text mining grouping titles practical work with using the K-Means that dikolaborisakan with AHC method for determining the center point initially [8]

A good similarity matrix is greatly responsible for the performance of spectral clustering algorithms [9]. So the purpose of this study was to find the most optimum value similarity. The method used combined similarity of Jaccard and cosine similarity. By combining the two similarity is expected to increase the value of the similarity of the two titles. While the document is used only in the form of a title document of practical work in the Department of Informatics Engineering University of Ahmad Dahlan. All these articles have been through the process of preprocessing beforehand. And the method used is the method of document clustering with Shared Nearest Neighbor (SNN).

2. LITERATURE REVIEW

2.1 TEXT CLUSTERING

Text Clustering Is Unsupervised Learning Process (Learning Process Itself) That Classify Documents Based On Similarity Relationship And Split Them Into Several Groups [10]

2.1.1 PREPROCESSING

Preprocessing the initial processing of documents in order to obtain a value that can be learned by the system clustering [11]

2.1.2 TOKENIZATION

Tokenization is the process of cutting the entire sequence of characters into a single piece of word [10].

2.1.3 STOPWORD REMOVAL

Stopword removal is the process for deleting all the words that have no meaning [10]

2.1.4 Stemming

Stemming is the process of forming a word being said essentially.

2.2 SHARED NEAREST NEIGHBOR

Nearest Neighbor (NN) becoming one of the methods in the top 10 methods of data mining the most popular used [12]. In some cases, clustering techniques that rely on standardized approach towards similarity and clustering density does not produce the desired results. SNN approach developed by Jarvis and Pat rick an indirect approach there are similarities [4]. Similarity between two points is shown in Figure 1 [13].

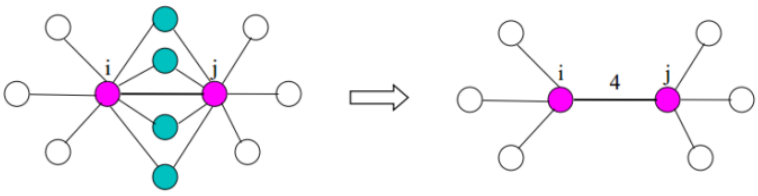

Figure 1. Similarity between two points

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

The key idea of this algorithm is to take the number of data points to determine the similarity measurement. The similarity in the SNN algorithm based on the number of neighbors held jointly for both objects contained in the list of nearest neighbors respectively as shown in Figure 1. Process SNN similarity is very useful because it can address some of the problems posed by the similarity calculation directly. Because include the contents of an object by using the number of nearest neighbors were held in common, SNN can resolve the situation, which is an object close to other objects of different classes. This algorithm works well for large dimension of data and in particular work well in finding dense clusters. Step algorithm SNN

1. Calculate the value of the similarity of the data set
2. Form a list of k nearest neighbors of each point of the data for the data k
3. Forms adjacency graph of the results list k nearest neighbors
4. Discover the density for each data
5. Form a cluster of dots representative [4]

2.3 SIMILARITY

Cosine Similarity

Cosine similarity measures the similarity between two vectors by taking the cosine of the angle the two vectors make in their dot product space. If the angle is zero, their similarity is one, the larger the angle is, the smaller their similarity. The measure is independent of vector length (the two vectors can even be of different length), which makes it a commonly used measure for high-dimensional spaces.[14]

$$sim(\mathbf{x}_a, \mathbf{x}_b) = cos(\theta) = \frac{\mathbf{x}_a \cdot \mathbf{x}_b}{\|\mathbf{x}_a\|\|\mathbf{x}_b\|} = \frac{\sum_{i=1}^{d} \mathrm{x}_a^i \times \mathrm{x}_b^i}{\sqrt{\sum_{i=1}^{d} (\mathrm{x}_a^i)^2} \times \sqrt{\sum_{i=1}^{d} (\mathrm{x}_b^i)^2}} \quad (1)$$

Jaccard Similarity

Jaccard similarity measures the similarity between two nominal attributes by taking the intersection of both and divide it by their union. In terms of the above definitions this gives [14];

$$J = \frac{A_{11}}{A_{01} + A_{10} + A_{11}} \quad (2)$$

A11 = total number of binary values where both vectors have the value 1.
A01 = total number of binary values where first vector has value 1, other has value 0.
A10 = total number of binary values where first vector has value 0, other has value 1.
A00 = total number of binary values where both vectors have the value 0

3. RESULT AND DISCUSSION

In this study, divided into several sections including preprocessing, calculation of similarity and clustering methods SNN. Chronology of practice at the process of grouping the working title is described in Figure 2.

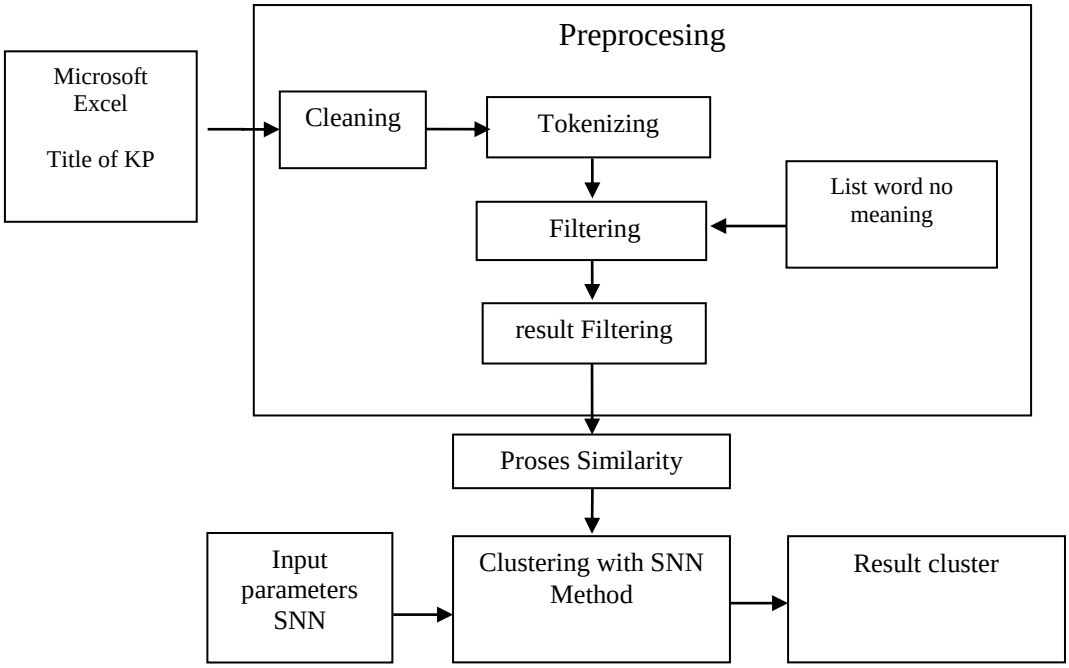

Figure 2. Title grouping process flow of practical work

1. Preprocesing

In preprocessing, there are several processes that is dry, tokenizing and filtering. In which the filtering process has been reserved for words that are not important in the database.

Table 1
Title Practical Work List

ID	Judul KP
KP1	sistem infomasi service komputer berbasis web
KP2	Media pembelajaran untuk kelas 5 berbasis web
KP3	pelatihan macromedia flash untuk guru di SMP gunung kidul
KP4	sistem informasi managemen apotik
KP5	Pembuatan web sekolah SMP bantul
KP6	Pembuatan web jurnal informatika
KP7	Pelatihan ms Office 2007 bagi guru SD Glagah Sari
KP8	Pembuatan website penjualan pada toko baju Ka-shop
KP9	Instalasi Jaringan Komputer di SMP N 2 Pleret
KP10	Pelatihan Desain Web Menggunakan Macromedia Dreamweaver 2004MX

2) Proses tokenizing

In this tokenizing limited to titles that have a word count 12, for the title more than 20 words will be deleted.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Table 2.
List Token Word

ID	K1	K2	K3	K4	K5	K6	K7
KP1	System	Informasi	Service	komputer	berbasis	web	
KP2	Media	pembelajaran	Untuk	Kelas	5	berbasis	web

3). Filtering
Filtering (Eliminate the word is not important / conjunctive)

Table 3
List Word Result filtering

ID	K1	K2	K3	K4	K5	K6
KP1	sistem	Informasi	service	komputer	berbasis	Web
KP2	media	pembelajaran	kelas	5	berbasis	Web

2. *Similarity Process*

The process of calculating the value of similarity is the closeness of each title with another title. This similarity used in the calculation formula and the cosine similarity Jaccard's similarity.

a. Cosine similarity

In calculating the similarity using the cosine similarity calculation done for one title with another title. Example of calculating a similarity to the title and the title of the two as follows:

$$Similarity\ 12 = \frac{1.0+1.0+1.0+1.0+1.1+1.1+0.1+0.1+0.1+0.1}{\sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2}\ .\ \sqrt{1^2+1^2+1^2+1^2+1^2+1^2+1^2}} = 0.33$$

Similarity value calculation is done until the last title. Ahir result of the calculation so that this similarity is the matrix n. n where n is the number of data.

Table 4
Result calculating cosine similarity for 10 data

Title	1	2	3	4	5	6	7	8	9	10
1	1.00	0.33	0.00	0.40	0.17	0.00	0.00	0.00	0.15	0.14
2	0.33	1.00	0.00	0.00	0.16	0.00	0.00	0.00	0.00	0.14
3	0.00	0.00	1.00	0.00	0.16	0.00	0.27	0.00	0.14	0.27
4	0.40	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.17	0.16	0.16	0.00	1.00	0.20	0.00	0.17	0.15	0.17
6	0.00	0.00	0.00	0.00	0.20	1.00	0.00	0.20	0.00	0.18
7	0.00	0.00	0.27	0.00	0.00	0.00	1.00	0.00	0.00	0.13
8	0.00	0.00	0.00	0.00	0.17	0.20	0.00	1.00	0.00	0.00
9	0.15	0.00	0.14	0.00	0.15	0.00	0.00	0.00	1.00	0.00
10	0.14	0.14	0.27	0.00	0.17	0.18	0.13	0.00	0.00	1.00

b. jaccard similarity

In calculating the similarity using the jaccard similarity calculation done for one title with another title. Example of calculating a similarity to the title and the title of the two as follows:

$$Similarity\ judul\ 1\ judul\ 2 = \frac{2}{6+6+2} = 0.5$$

Similarity value calculation is done until the last title. Ahir result of the calculation so that this similarity is the matrix n. n where n is the number of data.

Table 5
Result calculating jaccard similarity for 10 data

Title	1	2	3	4	5	6	7	8	9	10
1	1.00	0.14	0.00	0.17	0.08	0.09	0.00	0.00	0.07	0.07
2	0.14	1.00	0.00	0.00	0.08	0.09	0.00	0.00	0.00	0.07
3	0.00	0.00	1.00	0.00	0.07	0.00	0.12	0.00	0.07	0.12
4	0.17	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.08	0.08	0.07	0.00	1.00	0.09	0.00	0.08	0.07	0.07
6	0.09	0.09	0.00	0.00	0.09	1.00	0.00	0.09	0.00	0.08
7	0.00	0.00	0.12	0.00	0.00	0.00	1.00	0.00	0.00	0.06
8	0.00	0.00	0.00	0.00	0.08	0.09	0.00	1.00	0.00	0.00
9	0.07	0.00	0.07	0.00	0.07	0.00	0.00	0.00	1.00	0.00
10	0.07	0.07	0.12	0.00	0.07	0.08	0.06	0.00	0.00	1.00

c. Combine cosinus similarity dan jaccard similarity

$$Similarity\ judul\ 1\&2\ cosinus\ dan\ judul\ 1\&2 jaccard\ = \frac{0.33+0.50}{2} = 0.40$$

Results of the combined value of cosine similarity and Jaccard similarity if the data is a number 10 data is as follows:

Table 6
Result calculating combine cosine similarity dan jaccard similarity for 10 data

Judul	1	2	3	4	5	6	7	8	9	10
1	1.00	0.24	0.00	0.28	0.12	0.05	0.00	0.00	0.11	0.11
2	0.24	1.00	0.00	0.00	0.12	0.05	0.00	0.00	0.00	0.11
3	0.00	0.00	1.00	0.00	0.12	0.00	0.19	0.00	0.10	0.19
4	0.28	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
5	0.12	0.12	0.12	0.00	1.00	0.15	0.00	0.12	0.11	0.12
6	0.05	0.05	0.00	0.00	0.15	1.00	0.00	0.15	0.00	0.13
7	0.00	0.00	0.19	0.00	0.00	0.00	1.00	0.00	0.00	0.09
8	0.00	0.00	0.00	0.00	0.12	0.15	0.00	1.00	0.00	0.00
9	0.11	0.00	0.10	0.00	0.11	0.00	0.00	0.00	1.00	0.00
10	0.11	0.11	0.19	0.00	0.12	0.13	0.09	0.00	0.00	1.00

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

3. Clustering with SNN

In this grouping process, is the result of a continuation of the process of similarity. By entering a few parameters SNN then grouping titles will be done. In the formation of clustering methods SNN there are several parameters predetermined value k (specify how many titles the nearest neighbor of a title that will be included in the list of proximity based on the similarity between the two titles), the value of Eps (keywords that are set up to use together in 2 title (the same number of words)), and the mint value (threshold minimum number of titles to form clusters).

a. k = 5, Eps = 1 dan MinT = 2

In the case of this cluster is formed by the number of nearest neighbors, namely 5, keywords used simultaneously in the two titles is 1 word and the minimum number for forming each cluster are the two titles.

Table 7
Result of the first experiment

Cluster	Data Cluster
1	KP2, KP1, KP10
2	KP5, KP6
3	KP9, KP4
4	KP8, KP3

In Table 7 shows that the resulting four clusters, where each cluster on average there are two titles practical work.

b. k = 5, Eps =2 dan MinT = 2

In the case of this cluster is formed by the number of nearest neighbors, namely 5, keywords used simultaneously under two headings are two words and the minimum number for forming each cluster are the two titles.

Table 8
Results of the second experiment

Cluster	Data Cluster
1	KP2, KP4
2	KP5, KP6

In Table 8 shows that the resulting two clusters, where each cluster on average there are two titles practical work. This is because some of the other titles that do not meet the value eps is 2 so is not included in the cluster

4. CONCLUSION

In this study it can be concluded that:

1. Results of cosine similarity has the highest value in comparison with Jaccard similarity and the joint between Cosine and Jaccard similarity.

2. From the grouping hasi use SNN eps parameters greatly affect the formation of clusters. The larger the value, the cluster formed eps will be less.

REFERENCES

[1] B. Santosa, *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis*. Yogyakarta: Graha Ilmu, 2007.

[2] C. Triawati, M. A. Bijaksana, N. Indrawati, and W. A. Saputro, "PEMODELAN BERBASIS KONSEP UNTUK KATEGORISASI ARTIKEL BERITA," in *Seminar Nasional Aplikasi Teknologi Informasi (SNATI)*, 2009, vol. 2009, no. Snati, pp. 48–53.

[3] N. W. S. Saraswati, "Text Mining dengan Metode Naïve Bayes Classifier dan Support Vector Machines untuk Sentiment Analysis," Universitas Udayana Denpasar, 2011.

[4] R. F. Zainal and A. Djunaidy, "ALGORITMA SHARED NEAREST NEIGHBOR BERBASIS DATA SHRINKING," pp. 1–8.

[5] S. Jiang, G. Pang, W. Meiling, and K. Limin, "An Improved K-Nearest-Neighbor Algoritm for Text Categorization," *Expert Syst. with Appl.*, vol. 39.1, pp. 1503–1509, 2012.

[6] C. Li and L. Nan, "A Novel Text Clustering Algorithm," *Energy Procedia*, vol. 13, pp. 3583–3588, 2011.

[7] R. Handoyo, S. M. Nasution, P. Studi, S. Komputer, S. Linkage, and S. Coefficient, "Perbandingan Metode Clustering Mengggunakan metode Single Linkage dan K-Means Pada Pengelompokkan Dokumen," *JSM STMIK Mikroskil*, vol. 15, no. 2, pp. 73–82, 2014.

[8] T. Alfina, B. Santosa, J. T. Industri, and F. T. Industri, "Analisa Perbandingan Metode Hierarchical Clustering, K-Means dan Gabugan Keduanya dalam Membentuk Cluster Data (Studi Kasus : Problem Kerja Praktek Jurusan Teknik Industri ITS)," *J. Tek. POMITS*, vol. 1, no. 1, pp. 1–5, 2012.

[9] X. He, S. Zhang, and Y. Liu, "An Adaptive Spectral Clustering Algorithm Based on the Importance of Shared Nearest Neighbors," *Algorithms*, vol. 8, pp. 177–189, 2015.

[10] C. D. Manning, P. Raghavan, and H. Schutze, *introduction to Information Retrieval*. Camridge: Cambridge University Press, 2008.

[11] W. Junjie, *Advances in K- Means Clustering : a Data Mining Thinking*. Springer Science & Business Media, 2012.

[12] W. Xilon and K. Isua, *The Top Ten Algorithms in Data Mining*. London: CRC Press Tailor & Francis Group, 2009.

[13] V. Gupta, L. C. Science, and G. S. Lehal, "A Survey of Text Mining Techniques and Applications," vol. 1, no. 1, pp. 60–76, 2009.

[14] C. Plattel, "Distributed and Incremental Clustering using Shared Nearest Neighbours," Utrecht University, 2014.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)