Computer Engineering and Applications

Computer Engineering and Applications Vol. 6, No. 3, 2017

Text Mining for Internship TitlesClusteringUsing Shared Nearest-Neighbor Method

LisnaZahrotun

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

ABSTRACT

An Internship course becomes one of many compulsory subjects in Under graduate Program of Informatics Engineering in Ahmad Dahlan University, Yogyakarta.In the last few semesters, we found that some students were failed in taking this subject. After being identified, they were facing some obstacles such as determining the main theme for their job description. During this study, we proposed an application to classify the internship titles by using a technique in text mining called Shared Nearest-Neighbor and Cosine Similarity. From the result, we got values from the parameter K is 7, the epsilon value is 0.5, and the value of Mint t is 0.3 with 22 clusters and 0 outlier. These values presented that all data titles of internship activitiesareclassified into each cluster. 7 topics whichtook by majority of students are:1) Information Systems (7 titles);2) Instructional Media (5 titles);3)Archiving Applications (4 titles);4) Web Profile Implementation (3 titles); 5)Instructional Media for University Courses (3 titles); Multimedia (3 titles) and 6)Workshop & Training (3 titles).

Keywords: Shared Nearest-Neighbour, Cosine Similarity.

1. INTRODUCTION

An Internship becomes one of many compulsory subjects in Undergraduate Program of Informatics Engineering in Ahmad Dahlan University. In the last few semesters, we found that some students were failed in taking this subject. After being identified, they were facing some obstacles such as determining the main theme for their job description and having difficulty to find the workplace according to their expertise.

The implementation process of Internship course in Informatics Engineering Program should be managed well, so that every student can pass this course on time. Realizing that choosing its focus work becomes one of the important process, it is necessary to create groups for each themewhich been taken by students. Therefore, students can be assisted to decide which themes that match with their expertise, even since from the early semester.

To determine themes available for Internship course,we need to identify the Internship titles and themes which have been done before by applying some text mining techniques for creating the clusters. In order to solve this problem,we propose a study The Implementation of Internship Titles Classification Using Shared Nearest-Neighbor (SNN) Method.

2. LITERATURES REVIEW

Text mining becomes a step of text analysis processes performed automatically by computer machine to extract quality of information from a series of texts summarized in a document [1]. While text clustering is a process of unsupervised learning which clusters documents based on their similarity and place them into separate groups [2]. The grouping will make the document be visible in subtopics, ensuring that useful documents will not be ignored from the inquiry results.The basic clustering algorithm creates a topic vector for each document and measures the weight of how well the document corresponds to each cluster [3]. Cluster technology is able to support the management of information systems, which may contain thousands of documents [4]. Some researches in document clustering had been done through various methods, for example the use of K-Nearest Neighbour (K-NN) method for text categorization [5], data text clustering with Fuzzy C-Means [6], and documents clustering using K-Harmonic [7]. Classification using SNN method that has been done before was comparing Cosine Similarity, Jaccard and the combination of both, while Cosine Similarity had the best result [8]. The Shared Nearest Neighbor (SNN) approach is used for numerical data like measuring distance [9].

2.1 *Shared Nearest Neighbor Algorithm*

The Shared Nearest neighbor shared algorithm (SNN) is a density-based clustering algorithm that able to find an arbitrary set of shape, size and density,without specifying the number of clusters as parameters [10]. SNN algorithm it self requires 3 input parameters as follows:

1. k, the number of nearest neighbors.
2. e, radius, the threshold value of shared neighbors
3. MinT, the minimum amount of data for each cluster

The algorithmincludes these following steps [9]:

1. Create a distance matrix using a certain distance function and identify the nearest k-neighbor for each point.
2. For each of two points, calculate the similarities, which are given by the number of shared neighbors.
3. Set the SNN density at each point. The SNN density is given by the number of nearest neighbors who share Eps or more neighbors.
4. Identify the core points of the data set. Any point that has an equal orgreater SNN density than MinT is considered a core point.
5. Build the cluster from the core point. Two core points are allocated to the same cluster if they share Eps or more neighbors with each other.
6. Handle the noise point. Points are not classified as core points and those not within Eps from the core point are considered noise.
7. Set the remaining points in the group. All non-core and non-noise points assigned to the nearest cluster.

2.1 *Cosine Similarity*

This matrix calculates the cosine value of the angle between two vectors, where values used in this calculation are 1 and 0. The equation of cosine similarity is shown in equation 1. [11]

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

Where :
Xa = value in title 1
Xb = value in title 2
d = number of words in each title

3. METHODOLOGY

The stages of text mining done in this research are:

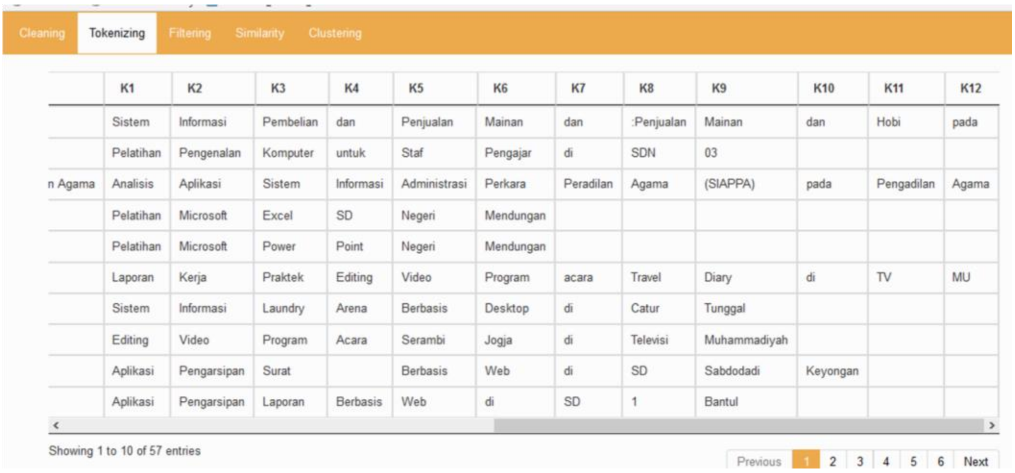
1. Preprocessing
   Preprocessing is the documents initial process prior to the grouping process, where the data cleaning process also takes place [12].
2. Tokenization
   Tokenization is the process of cutting a sentence into several parts of words [2].
3. Filtering
   In this research, filtering process is done by using stopword removal model. This is the procedure of removing words that are considered unimportant [2].
4. Calculation of cosines similarity
   The process is done by calculating the proximity distance between document titles, as the main object in this research. The distance will be measured between each title, resulting in the form of nxn matrix, where n is the number of titles.
5. SNN Clustering
   In the process of grouping with the SNN method, first we entered the value of k, epsilon and min t as parameters. K-NN table then will be generated, which is the list of proximityvalue between the titlesthat is limited by the value of k. By this K-NN value then will be formed into clusters using SNN method.
6. OutliersChecking
   Outliers or data are a collection of objects that are considered the most different than the overall data.
7. Knowledge Representation
   This stage will displayone or more found patterns and then being represented to the user.

4. RESULT AND DISCUSSION (EXPERIMENTS AND RESULTS?)

This research conducted several text mining processes, described as follows:

1. Preprocessing
   In the preprocessing, the data is filtered only to have 12 words at maximum. For instance, if a title is longer than 12 words, only the first 12 words are considered.

2. Tokenizing
   Tokenizing is the process of breaking a sentence into multiple tokens (e.g. word, phrase, etc). The example results of this process is shown in Figure 1.



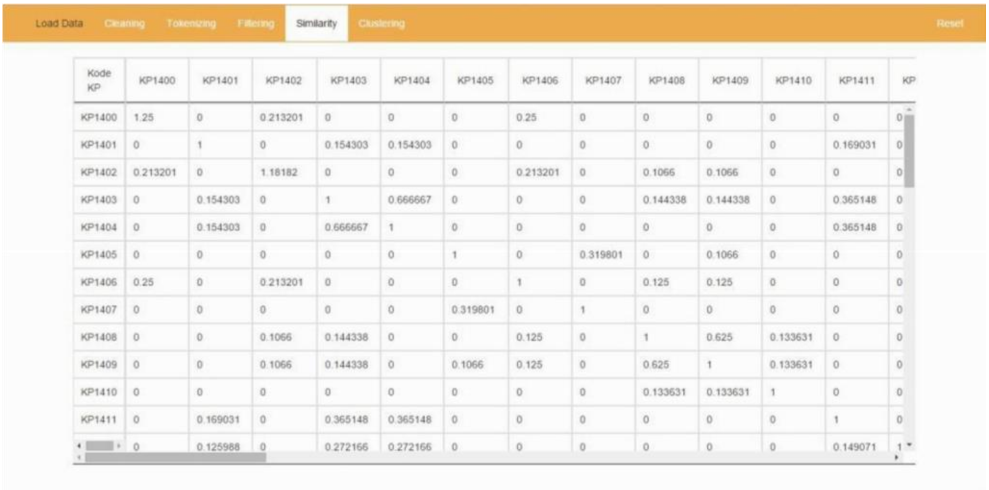
	K1	K2	K3	K4	K5	K6	K7	K8	K9	K10	K11	K12
	Sistem	Informasi	Pembelian	dan	Penjualan	Mainan	dan	:Penjualan	Mainan	dan	Hobi	pada
	Pelatihan	Pengenalan	Komputer	untuk	Staf	Pengajar	di	SDN	03			
n Agama	Analisis	Aplikasi	Sistem	Informasi	Administrasi	Perkara	Peradilan	Agama	(SIAPPA)	pada	Pengadilan	Agama
	Pelatihan	Microsoft	Excel	SD	Negeri	Mendungan						
	Pelatihan	Microsoft	Power	Point	Negeri	Mendungan						
	Laporan	Kerja	Praktek	Editing	Video	Program	acara	Travel	Diary	di	TV	MU
	Sistem	Informasi	Laundry	Arena	Berbasis	Desktop	di	Catur	Tunggal			
	Editing	Video	Program	Acara	Serambi	Jogja	di	Televisi	Muhammadiyah			
	Aplikasi	Pengarsipan	Surat		Berbasis	Web	di	SD	Sabdodadi	Keyongan		
	Aplikasi	Pengarsipan	Laporan	Berbasis	Web	di	SD	1	Bantul			

Showing 1 to 10 of 57 entries

Figure 1. Result of Tokenizing

3. Filtering
   In this process, we removed words that belong to the Indonesian stopwords from the resulting tokens. For instance, the word “di”, “ke”, “dari”, “dan”, “untuk”, “pada”, “atau” will be removed.

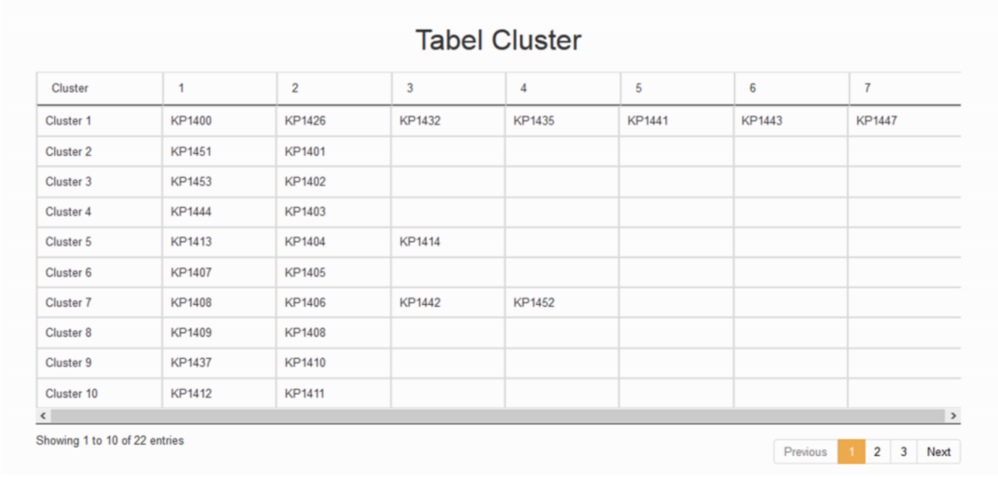
4. Cosine Similarity Calculation
   This process calculates similarity for every two titles. The result of this process is a matrix of NxN where N is the number of titles. The sample result is



Kode KP	KP1400	KP1401	KP1402	KP1403	KP1404	KP1405	KP1406	KP1407	KP1408	KP1409	KP1410	KP1411	KP
KP1400	1.25	0	0.213201	0	0	0	0.25	0	0	0	0	0	0
KP1401	0	1	0	0.154303	0.154303	0	0	0	0	0	0	0.169031	0
KP1402	0.213201	0	1.18182	0	0	0	0.213201	0	0.1066	0.1066	0	0	0
KP1403	0	0.154303	0	1	0.666667	0	0	0	0.144338	0.144338	0	0.365148	0
KP1404	0	0.154303	0	0.666667	1	0	0	0	0	0	0	0.365148	0
KP1405	0	0	0	0	0	1	0	0.319801	0	0.1066	0	0	0
KP1406	0.25	0	0.213201	0	0	0	1	0	0.125	0.125	0	0	0
KP1407	0	0	0	0	0	0.319801	0	1	0	0	0	0	0
KP1408	0	0	0.1066	0.144338	0	0	0.125	0	1	0.625	0.133631	0	0
KP1409	0	0	0.1066	0.144338	0	0.1066	0.125	0	0.625	1	0.133631	0	0
KP1410	0	0	0	0	0	0	0	0	0.133631	0.133631	1	0	0
KP1411	0	0.169031	0	0.365148	0.365148	0	0	0	0	0	0	1	0
	0	0.125988	0	0.272166	0.272166	0	0	0	0	0	0	0.149071	1

Figure 2. Result of cosine similiarity calculation

5. Clustering with SNN
   The clustering process in this research is done with SNN. The clustering result is shown in Figure 3.



Tabel Cluster

Cluster	1	2	3	4	5	6	7
Cluster 1	KP1400	KP1426	KP1432	KP1435	KP1441	KP1443	KP1447
Cluster 2	KP1451	KP1401					
Cluster 3	KP1453	KP1402					
Cluster 4	KP1444	KP1403					
Cluster 5	KP1413	KP1404	KP1414				
Cluster 6	KP1407	KP1405					
Cluster 7	KP1408	KP1406	KP1442	KP1452			
Cluster 8	KP1409	KP1408					
Cluster 9	KP1437	KP1410					
Cluster 10	KP1412	KP1411					

Showing 1 to 10 of 22 entries

Figure 3. Results Grouping using SNN method

6. Outlier detection
   We conducted several experiments to find the optimum clusters with minimum number of outliers. The result of this experiment is presented in Table 1.

Table 1. The results of experiments using the SNN method

k	E	Min t	Outlier
5	0.5	0.3	1
10	0.5	0.3	14
7	0.3	0.1	0

7. Knowledge representation
   The result of this research is that the best clusters are obtained using these combination of parameter values : k = 7, epsilon = 0.5, and mint = 0.3. From the clustering results we obtained 7 trending topics, they are: information system (7 data), instructional media (5 data), archiving software (4 data), developing web profile (3 data), instructional media for university course (3 data),multimedia (3 data), workshop and training (3 data). The detail of the result is presented in Figure 5.

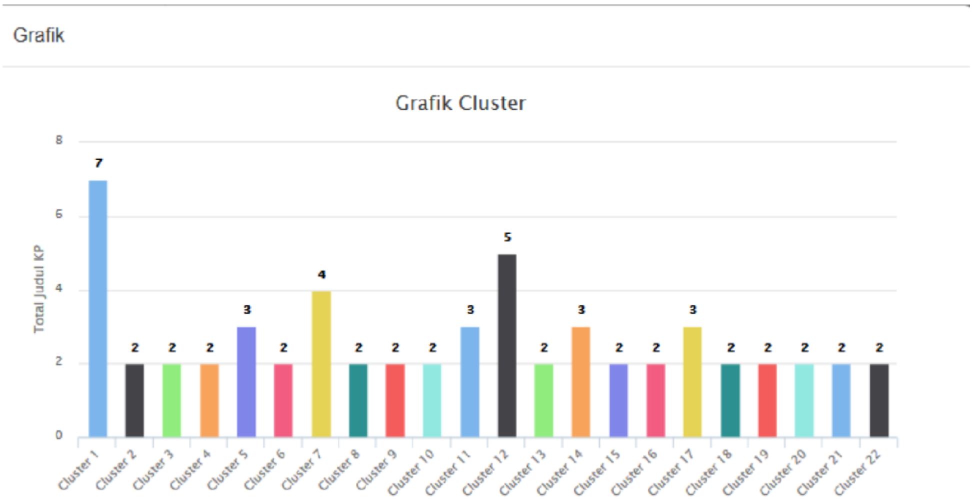


Figure 5. Cluster results with the smallest number of outliers

5. CONCLUSION

We have developed a text mining application that clusters the title of internship projects using Shared Nearest Neighbor and Cosine Similarity. The numbers of best clusters are found 22 with 0 outlier. This means that all the data get into the cluster. These clusters were obtained with the parameters k = 7, epsilon = 0.5, and minT= 0.3.

ACKNOWLEDGEMENT

This research has been supported by the research grant of KOPERTIS V.

REFERENCES

[1] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kauffman, 2006.

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

[3] M. Steinbach, K. George, and V. Kumar, "A Comparison of Document Clustering Techniques," Department of Computer Science and Engineering, University of Minnesota., pp. 1–20. [4] R. Janani and S. Vijayarani, "Text Mining Research: A Survey," Int. J. Innov. Res. Comput. Commun. Eng., vol. 4, no. 4, pp. 6564–6571, 2016.

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

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

[6] H. Emami, S. Dami, and H. Shirazi, "K-Harmonic Means Data Clustering With Imperialist Competitive Algorithm," U.P.B.Sci.Bull.,Series C,, vol. 77, no. 1, pp. 91–104, 2015.

[7] L. Zahrotun, "Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method," vol. 5, no. 1, pp. 11–18, 2016.

[8] B. N. Bai and A. M. Sowjanya, "An Incremental Shared Nearest Neighbour Clustering Approach For Numerical Data Using An Efficient Distance Measure," Int. J. Eng. Comput. Sci., vol. 4, no. 9, pp. 14192–14196, 2015.

[9] L. Ertoz, M. Steinbach, and V. Kumar, "A New Shared Nearest Neighbor

[10] Clustering Algorithm and its Applications," in Workshop on Clustering High Dimensional Data and its Applications at 2nd SIAM International Conference on Data Mining, 2002.

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

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