PERANCANGAN TEXT MINING PENGELOMPOKKAN PENELITIAN DOSEN MENGGUNAKAN METODE SHARED NEAREST NEIGHBOR DENGAN EUCLIDEAN SIMILARITY

Mushlihudin[1*], **Lisna Zahrotun**[2]
[1, 2] Program Studi Teknik Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan
Jl. Prof. Dr. Soepomo, Janturan, Yogyakarta 55164
[*]Email: mushlihudin@tif.uad.ac.id

Abstrak

Penelitian merupakan salah satu dari Tridarma dosen. Selain mengajar seorang dosen diwajibkan untuk melakukan penelitian guna mengembangkan ilmunya. Universitas Ahmad Dahlan (UAD) merupakan salah satu Universitas Swasta di Yogyakarta. Selama ini penelitian dosen UAD dikelola oleh Lembaga Penelitian dan Pengembangan atau yang sering disingkat dengan LPP. Penelitian dosen terdiri dari penelitian internal dan penelitian eksternal. Salah satu kendala dosen dalam penelitian adalah mencari pasangan yang tepat yang sesuai dengan bidang keilmuan. Padahal bidang keilmuan ini dapat dilakukan dengan dosen antar program studi. Jika para dosen mengetahui bidang minat dan riwayat dari penelitian-penelitian sebelumnya dosen lain tentu ini akan memudahkan dosen dalam berkolaborasi dengan dosen lain untuk melakukan penelitian. Tujuan penelitian ini adalah untuk membuat perancangan text mining dalam mengelompokkan judul penelitian dosen berdasarkan kemiripan antar judul penelitian. Metode yang digunakan dalam penelitian merupakan salah satu metode pengelompokkan dalam text mining yaitu Shared Nearest Neighbor dengan Euclidean Similarity". Luaran dari penelitian ini adalah rancangan aplikasi text mining dalam mengelompokkan judul-judul penelitian dosen yang memiliki kemiripan sehingga memudahkan para dosen untuk mencari relasi dalam penelitian berikutnya.

***Kata Kunci** : Text mining, Shared Nearest Neighbor, Euclidean Similarity*

1. PENDAHULUAN

Penelitian merupakan salah satu dari Tridarma dosen. Selain mengajar seorang dosen diwajibkan untuk melakukan penelitian guna mengembangkan ilmunya. Sedangkan Perguruan tinggi berkewajiban menyelenggarakan penelitian dan pengabdian kepada masyarakat disamping melaksanakan pendidikan sebagaimana diamanahkan oleh Undang- undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional Pasal 20. Sejalan dengan kewajiban tersebut, Undang-undang Nomor 12 Tahun 2012 tentang Pendidikan Tinggi Pasal 45 menegaskan bahwa penelitian di perguruan tinggi diarahkan untuk mengembangkan ilmu pengetahuan dan teknologi, serta meningkatkan kesejahteraan masyarakat dan daya saing bangsa.(Direktorat Riset dan Pengabdian Kepada Masyarakat 2016)

Universitas Ahmad Dahlan (UAD) merupakan salah satu Universitas Swasta di Yogyakarta. Selama ini penelitian dosen UAD dikelola oleh Lembaga Penelitian dan Pengembangan atau yang sering disingkat dengan LPP. Penelitian dosen terdiri dari penelitian internal dan penelitian eksternal. Penelitian internal merupakan penelitian yang diselengggarakan oleh UAD dalam mengembangkan ilmu dari dosen. Sedangkan penelitian dari eksternal merupakan penelitian dengan dana yang bersumber dari luar UAD diantaranya Koordinasi Perguruan Tinggi Swasta (Kopertis) dan Direktorat Riset dan Pengabdian kepada Masyarakat, Direktorat Jenderal Penguatan Riset dan Pengembangan Kementerian Riset, Teknologi, dan Pendidikan Tinggi (RISTEKDIKTI).

Penelitian internal maupun eksternal dibuka satu kali setiap tahunnya. Jumlah penelitian internal UAD setiap tahunnya bisa mencapai 150 judul penelitian, Untuk penelitian Kopertis 30 judul sedangkan untuk penelitian RISTEKDIKTI untuk tahun 2016 mencapai 56 judul penelitian. Dengan banyaknya penelitian dan skema-skema dari penelitian maka tentunya penelitian yang dihasilkan juga akan sangat beragam dan bervariasi. Salah satu kendala dosen dalam penelitian adalah mencari pasangan yang tepat yang sesuai dengan bidang keilmuan. Padahal bidang keilmuan ini dapat dilakukan dengan dosen antar program studi. Jika para dosen mengetahui

bidang minat dan riwayat dari penelitian-penelitian sebelumnya dosen lain tentu ini akan memudahkan dosen dalam berkolaborasi dengan dosen lain untuk melakukan penelitian.

Text mining merupakan salah satu teknik yang digunakan untuk menggali data yang tersembunyi dari data yang berbentuk *text*. Salah satu metode dalam *text mining* adalah *clustering*. Teknik *Clustering* sendiri merupakan teknik pengelompokkan yang banyak dipakai dalam *data mining*. Menurut Santosa (2007) tujuan utama dari metode *cluster* adalah pengelompokkan sejumlah data/ objek ke dalam *cluster* (group) sehingga dalam setiap *cluster* akan berisi data yang semirip mungkin. Ini berart objek dalam satu *cluster* sangat mirip satu sama lain dan berbeda dengan objek dalam *cluster-cluster* yang lain. Menurut Jarvis dan Patrick, (1973, dalam penelitian Zainal 2008) menyebutkan bahwa pendekatan *Shared Nearest Neighbor* (SNN) merupakan cara yang terbaik untuk mengelompokkan data dalam jumlah besar. Dengan mengunakan teknik *similarity* atau kesamaan, setelah ketetanggaan terdekat dari semua titik data telah ditentukan, maka nilai kesamaan yang baru diantara titik-titik data ditentukan dari jumlah ketetanggaan yang dimiliki secara bersama-sama. (Zainal & Djunaidy 2008). Penelitian lainnya yaiu penggunaan SNN untuk pengelompokkan data tiga dimensi clud computing (Wu et al. 2015). Penelitian lainnya penggunaan SNN dalam pengelompokkan spectral dilakukan oleh (He et al. 2015). Penelitian tentang kedekatan atau sering disebut sebagai similarity pernah dilakukan sebelumnya, dalam penelitian ini membandingkan beberapa similarity dan dari beberapa similarity dihasilkan *Euclidean similarity* yang memiliki kurasi paling baik.(Patidar et al. 2012). Penelitian tentang penggunaan SNN juga pernah dilakukan oleh (Zahrotun 2016). Dalam penelitian ini membandingkan dua metode *similarity* yaitu *cosine similarity* dan *jaccard similarity*.

Dengan demikian, dengan melihat dari penelitian-penelitian sebelumnya dan agar judul-judul penelitian dosen dari UAD dapat bermanfaat maka di buat Perancangan *Text Mining* Pengelompokkan Penelitian Dosen Menggunakan Metode *Shared Nearest Neighbor* dengan *Euclidean Similarity*. Dengan adanya penelitian ini diharapkan dapat membantu dalam pembuatan rancangan program yang dapat mengelompokkan judul-judul penelitian yang memiliki kemiripan sehingga memudahkan para dosen untuk mencari relasi dalam penelitian berikutnya.

2. METODOLOGI

2.1 Dasar Teori

1. *Shared Nearest Neighbour*

Algoritma *shared nearest neighbor* (SNN) pada proses pengklasteran dikenal sebagai cara untuk mengatasi masalah pengukuran jarak dalam data berdimensi tinggi (Jarvis dan Patrick, 1973) yang dikembangkan oleh (Gupta, 1999). (Zainal & Djunaidy 2008)

Algoritma SNN memerlukan 3 input parameter :

1). k , yakni jumlah daftar tetangga terdekat
2). e, yakni jari-jari(radius), nilai ambang ketetanggaan yang dimiliki secara bersama
3). MinT, yakni jumlah minimal data untuk setiap kluster

Langkah-langkah Algoritma Shared Nearest Neighbor

1) Hitung nilai kesamaan dari data set
2) Bentuk daftar k-tetangga terdekat masing-masing titik data untuk k data
3) Bentuk graph ketetanggaan dari daftar k tetangga terdekat
4) Temukan kepadatan untuk setiap data
5) Temukan titik-titik representatif
6) Bentuk cluster dari titik-titik representative tersebut

2. *Euclidean Similarity*

Jarak euclidean menentukan akar perbedaan persegi antara koordinat sepasang objek. Untuk jarak vektor x dan y (x, y) ditunjukkan dalam persamaan 1 (Patidar et al. 2012)

$$\text{Sim}(x, y) = d = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad 1)$$

Dimana x dan y adalah vektor n-dimensi

2.2 Pengumpulan Data

1. Pengumpulan data dilakukan dengan studi literatur dari berbagai macam buku, artikel, publikasi ilmiah untuk mempelajari:
   a. Mekanisme *Teks Mining* dalam mengelompokkan data.
   b. Mekanisme pengelompokan data menggunakan metode SNN
   c. Algoritma SNN dan Euclidean similarity.
2. Metode Wawancara
   Metode wawancara ini dilakukan dengan cara bertanya langsung terhadap nara sumber dalam hal ini wawancara dilakukan kepada bpk Wahyudin selaku karyawan yang menangani tentang pendataan penelitian dosen UAD.
3. Metode Observasi
   Metode observasi ini dilakukan dengan cara melihat proses pendataan judul penelitian dosen melalui website lpp.uad.ac.id dan juga pendataan judul penelitian dosen menggunakan *form hardcopy*.

2.3 Alur Proses Penelitian

Dalam penelitian ini dimulai melakukan tahap-tahap pada *Text Mining* yaitu *cleaning, tokenizing, filtering,*. Kemudian penghitungan similarity menggunakan *Euclidean similarity* dan pemahaman dari algoritma *Shared Nearest Neighbor* (SNN) dilanjutkan dengan perancangan *Text mining* pengelompokkan data judul penelitian dosen.

3. HASIL DAN PEMBAHASAN

Berdasarkan analisis sistem yang telah disebutkan diatas, rancangan sistem yang akan dibangun dalam penelitian terbagi menjadi beberapa bagian diantaranya preprocesing, penghitungan similarity, dan pengelompokkan menggunakan metode SNN.

3.1 Tahap-tahap *Text mining*

a. Data awal
   Data awal ini merupakan data penelitian dosen awal yang belum dilakukan perubahan apapun. Data awal ini ditunjukkan dalam Tabel 1.

Tabel 1. Data Awal

Judul Penelitian	Nama Peneliti	Tahun
Perancangan dan Implementasi Pembelajaran Perkalian dan Pembagian Bilangan untuk Sekolah Dasar Kelas 2	Agung Dwi Haryanto	2014
Analisis Perbandingan Metode Li dan Chan-Vese pada Proses Segmentasi Citra Digital	Rizki Mulasari	2014
Sistem Pakar untuk Mendiagnosa Penyakit Kambing Etawa Berbasis Web	Bagus Primatoro	2014
Sistem Pakar untuk Mendiagnosa Penyakit Kelinci Berbasis Web	Sulis Trianto	2014
Sistem Penentuan Keterkaitan Antar Skripsi Berdasarkan Keyword Seeking	Mulyadin	2013

b. Cleaning

Proses *cleaning* merupakan proses pembersihan data, dimana dalam proses ini data nama peneliti dan data tahun tidak dipakai sehingga kolom nama peneliti dan tahun di hapus. Selain itu dalam proses *cleaning* ini judul penelitian dibatasi maximal 30 kata, jika lebih dari 30 kata maka kata setelah ke 30 akan dihapus. Hasil dari proses *cleaning* ditunjukkan dalam Tabel 2

Tabel 2. Hasil Cleaning

Judul Penelitian
Perancangan dan Implementasi Pembelajaran Perkalian dan Pembagian Bilangan untuk Sekolah Dasar Kelas 2
Analisis Perbandingan Metode Li dan Chan-Vese pada Proses Segmentasi Citra Digital
Sistem Pakar untuk Mendiagnosa Penyakit Kambing Etawa Berbasis Web
Sistem Pakar untuk Mendiagnosa Penyakit Kelinci Berbasis Web
Sistem Penentuan Keterkaitan Antar Skripsi Berdasarkan Keyword Seeking

c. Tokenizing

Tokenizing merupakan proses memisah kalimat menjadi kata (Manning et al. 2008), sehingga judul penelitian yang sudah di lakukan proses *cleaning* akan dilakukan proses *tokenizing*. Hasil dari proses *tokenizing* ditunjukkan dalam Tabel 3.

Tabel 3. Hasil *Tokenizing*

Kata 1	Kata 2	Kata 3	Kata 4	Kata 5	Kata 6	Kata 7	Kata 8	Kata 9	Kata 10	Kata 11	Kata 12	Kata 13
Perancangan	Dan	Implementasi	Pembelajaran	Perkalian	Dan	Pembagian	Bilangan	Untuk	Sekolah	Dasar	Kelas	2
Anaisis	Perbandingan	Metode	Li	Dan	Chan-Vese	Pada	Proses	Segmentasi	Citra	Digital		
Sistem	Pakar	Untuk	Mendiagnosa	Penyakit	Kambing	Etawa	Berbasis	Web				
Sistem	Pakar	Untuk	Mendiagnosa	Penyakit	Kelinci	Berbasis	web					
Sistem	Penentuan	Keterkaitan	Antar	Skripsi	Berdasarkan	Keyword	Seeking					

d. *Filtering*

Salah satu proses *filtering* adalah *stopword removal*, *stopword removal* adalah menghapus kata-kata yang tidak penting (Manning et al. 2008). Kata yang tidak penting yang dihapus dalam prose ini adalah dan , di, ke ,dari, untuk, pada. Hasil proses filering ditujukkan dalam Tabel 4

Tabel 4. Hasil Filtering

Kata 1	Kata 2	Kata 3	Kata 4	Kata 5	Kata 6	Kata 7	Kata 8	Kata 9	Kata 10
Perancangan	Implementasi	Pembelajaran	Perkalian	Pembagian	Bilangan	Sekolah	Dasar	Kelas	2
Anaisis	Perbandingan	Metode	Li	Chan-Vese	Proses	Segmentasi	Citra	Digital	
Sistem	Pakar	Mendiagnosa	Penyakit	Kambing	Etawa	Berbasis	Web		
Sistem	Pakar	Mendiagnosa	Penyakit	Kelinci	Berbasis	web			

Sistem	Penentuan	Keterkaitan	Skripsi	Berdasarkan	Keyword	Seeking

e. *Euclidean Similarity*
Euclidean similarity adalah mencari nilai kesamaan antara satu judul dengan judul yang lain. Hasil dari penghitungan Euclidean similarity ditunjukkan dalam Tabel 5.

Tabel 5 Hasil penghitungan *Euclidean similarity*

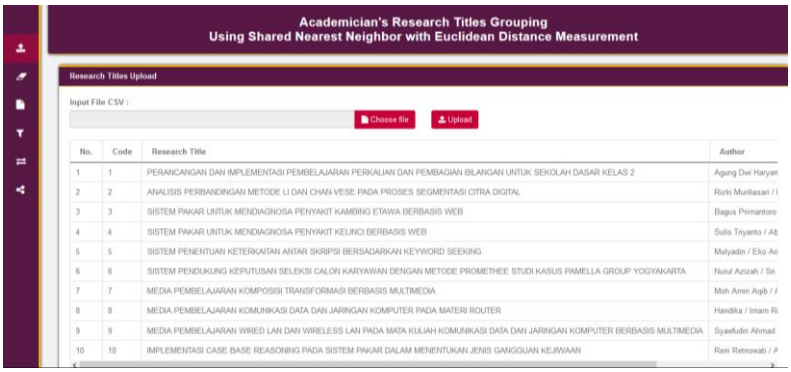
Judul	1	2	3	4	5
1	0	0.198778	0.696358	0.772469	0.653441
2	0.198778	0	0.708685	0.791712	0.584551
3	0.696358	0.708685	0	0.344153	0.400616
4	0.772469	0.791712	0.344153	0	0.456013
5	0.653441	0.584551	0.400616	0.456013	0

f. Pengelompokkan menggunakan *Shared Nearest Neighbor* (SNN)
Dalam pengelompokkan ini ditentikan nilai parameter awal terlebih dahulu yaitu nilai k yang merupakan nilai kedekatan, nilai e yang merupakan nilai *epsilon* dan nilai Min T yang merupakan jumlah minimal untuk setiap *cluster*. Dalam penelitian ini dibuat nilai k = 4, nilai e = 1 dan nilai Min T=2. Hasil dari pengelompokkan menggunakan SNN ditunjukkan dalam Tabel 6.

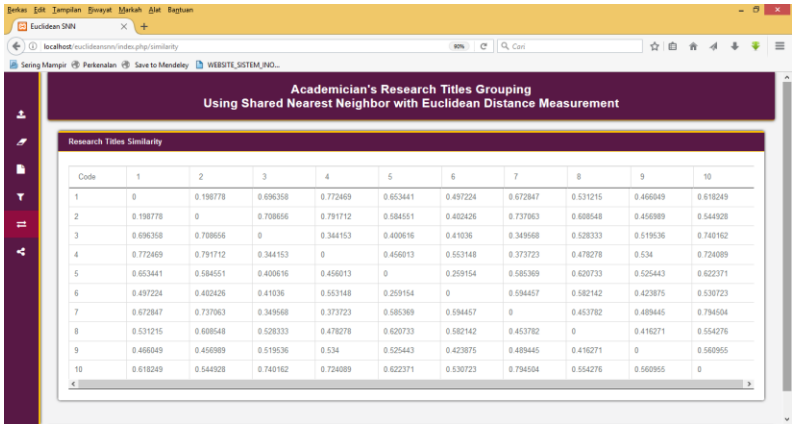
Tabel 6. Hasil *Cluster*

Cluster	Data *Cluster*
1	Judul 1, judul 2, judul 5
2	Judul 3
3	Judul 4
4	-

g. Perancangan *user interface*
Hasil dari proses tahap-tahap text mining diimplementasikan dalam perancangan *User interface text mining* pengelompokkan judul penelitian dosen. Gambar 1 menunjukkan data judul penelitian yang di ambil dari Microsoft excel yang akan dilakukan pengelompokan. Gambar 2. menunjukkan hasil dari penghitungan similarity antar judul dan Gambar 3. Menunjukkan hasil pengelompokkan yang dilakukan dengan metode SNN.



Gambar 1. Load data judul penelitian



Gambar 2. Hasil penghitungan Similarity

Gambar 3. Hasil Pengelompokkan Judul Penelitian

4. KESIMPULAN

Hasil dari penelitian ini adalah sebuah perancangan *user interface* untuk mengelompokkan data judul penelitian dosen menggunakan Metode *Shared Nearest Neighbor* dan *Euclidean similarity.* Hasil dari desaign user interface dapat langsung diaplikasikan dalam bahasa pemrograman yang terintegrasi dengan website dari Lembaga Pengembangan dan Penelitian (LPP) Universitas Ahmad Dahlan.

UCAPAN TERIMA KASIH

Penelitian ini telah didukung oleh hibah penelitian RISTEK DIKTI dengan skema Penelitian Dosen Pemula (PDP) tahu anggaran 2017

DAFTAR PUSTAKA

Direktorat Riset dan Pengabdian Kepada Masyarakat, D.J.P.R. dan P., 2016. Panduan Pelaksanaan Penelitian dan Pengabdian Kepada Masyarakat di Perguruan Tinggi Edisi X Tahun 2016. *www.risetdikti.go.id*.

He, X., Zhang, S. & Liu, Y., 2015. An Adaptive Spectral Clustering Algorithm Based on the Importance of Shared Nearest Neighbors. *algorithms*, 8, pp.177–189.

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

Patidar, A.K., Agrawal, J. & Mishra, N., 2012. Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach. *International journal of Computer application*, 40(16), pp.1–5.

Wu, F. et al., 2015. A Nearest Neighbor Searches (NNS) Algorithm for Fast Registration of 3D Point Clouds based on GPGPU. In *International Conferences on Intelligent Research and Mechatronics Engineering (ISRME)*. pp. 2153–2158.

Zahrotun, L., 2016. Comparison Jaccard similarity , Cosine Similarity and Combined Both of the Data Clustering With Shared Nearest Neighbor Method. , 5(1), pp.11–18.

Zainal, R.F. & Djunaidy, A., 2008. ALGORITMA SHARED NEAREST NEIGHBOR BERBASIS DATA SHRINKING. *JUTI*, 7, pp.1–8.