PENGGUNAAN ALGORITMA CLUSTERING BASED ON FREQUENT WORD SEQUENCES (CFWS) DALAM PENGELOMPOKKAN ARTIKEL BERBAHASA INDONESIA

Mauliza[1], Imelda Atastina[2], Finny Pensiska .[3]

[1]Teknik Informatika, Fakultas Teknik Informatika, Universitas Telkom

---

Abstrak

Sebagai bahasa yang kaya dengan kosa kata, bahasa Indonesia memiliki banyak kata yang berbeda namun memiliki arti sama (sinonim). Hal ini dapat menyebabkan banyak berita masuk ke dalam kelompok atau kategori yang tidak relevan dengan isi beritanya. Untuk itu diperlukan cara untuk mengolah data untuk mendapatkan manfaat dari data itu, salah satu cara untuk mengolah berita tersebut yaitu data mining.Dalam data mining terdapat salah satu metode yang sering digunakan yaitu clustering. Clustering merupakan pengelompokkan objek berdasarkan karakteristiknya. Pengelompokkan berita dapat menggunakan metode clustering dengan tujuan untuk mengelompokkan artikel berita sesuai dengan topik beritanya.

Dalam tugas akhir ini mengimplementasikan suatu metode clustering, yaitu algoritma Clustering based on Frequent Word Sequences (CFWS) pada artikel berita berbahasa Indonesia. CFWS merupakan algoritma yang mempresentasikan dokumennya dengan menggunakan kata-kata yang paling sering muncul secara berurutan pada setiap dokumen. Dengan menggunakan algoritma ini dapat mengurangi dimensi dari setiap dokumen secara signifikan sehingga proses clustering menjadi lebih efisien. Pengujian dilakukan untuk melihat kualitas hasil cluster berdasarkan metode pengukuran akurasi F-measure.

Berdasarkan pengujian yang sudah dilakukan, algoritma CFWS dapat menghasilkan hasil kualitas hasil cluster yang baik. Selain itu algoritma CFWS dapat menghasilkan hasil cluster yang baik untuk dataset dengan topik yang berdekatan maupun topik yang sangat berbeda.

Kata Kunci : data mining, clustering, CFWS, F-measure

Abstract

As a rich with vocabulary language, Indonesian language has many words with the same meaning (synonym). This can cause news report being grouped in a non relevant category with the news' content. Therefore, a method to to process data is needed for getting te benefit from that data. One of the method used to process news is data mining. In data mining, there is a method that is used often, which is clustering. Clustering is the grouping of object according to its characteristic. The news grouping can use the clustering methode with the purpose to group a news article appropriate with its news topic.

In this final assignment, a clustering method is implemented, which is the Clustering based on Frequent Word Sequences (CFWS) algorithm on Indonesian language news article. CFWS is an algorithm that represents documents by using the most frequent word sequncesthat appear inthe document. By using this algorithm, the dimension of the document can be reduced significantly so the clustering process can be more efficient. THe testing was done to see the quality of the final cluster according to the accuracy calculation with F-Measure.

According to the test that have been done, CFWS algorithm produce a good quality of cluster. Beside that, the CFWS algorithm can produce a good cluster for the data set with a similar topic and different topic.

Keywords : data mining, clustering, CFWS, F-measure

---

1. Pendahuluan

1.1 Latar belakang masalah

Berita merupakan salah satu bentuk penyebaran informasi yang sangat tepat pada sasaran.Saat ini, berita tidak hanya disebarkan lewat televisi,radio,maupun media cetak. Akan tetapi, berita juga ikut tersebar melalui internet. Banyak situs di internet yang berlomba merilis berita terbaru, sehingga perputaran berita akan sangat cepat. Bahkan hanya dalam hitungan detik berita baru bermunculan dan berita sebelumnya akhirnya tenggelam.

Sebagai bahasa yang kaya dengan kosa kata, bahasa Indonesia memiliki banyak kata yang berbeda namun memiliki arti sama (sinonim). Hal ini dapat menyebabkan banyak berita masuk ke dalam kelompok atau kategori yang tidak relevan dengan isi beritanya. Kasus pengelompokkan berita ini dapat dilakukan secara manual jika update berita tidak terlalu sering. Namun, pengelompokkan berita secara manual ini akan menjadi tidak efektif lagi untuk kasus berita yang akan di update berjumlah sangat banyak. Sebagai contoh, pengguna harus memilih atau memasukkan satu persatu berita sesuai dengan kategorinya. Tentu saja hal ini akan membutuhkan waktu yang lama jika berita yang harus dimasukkan berjumlah banyak. Untuk itu diperlukan perangkat lunak yang dapat mengelompokkan berita secara otomatis dan akurat. Teknik dokumen *clustering* dapat menjadi suatu alternatif dalam pengelompokkan berita. Teknik document *clustering* yang standar pada umumnya menggunakan representasi vektor sebagai representasi dokumennya atau biasa disebut *vector space model.* Pada representasi vektor ini setiap dokumen di representasikan sebagai vektor dari jumlah kemunculan kata pada dokumen tersebut. Oleh sebab itu, walaupun satu kata hanya muncul satu kali dalam suatu dokumen maka kata tersebut tetap akan menjadi satu dimensi dalam *vector space model*. Permasalahan utama dalam teknik *clustering* dengan menggunakan *vector space model* adalah besarya dimensi untuk setiap dokumen sehingga dibutuhkan suatu basis data yang sangat besar ukurannya. Contoh algoritma yang menggunakan teknik ini *yaitu K-Means, Average-link, dan Scater/Gather* [11].

Permasalahan utama dalam *vector space model* inilah yang memunculkan suatu teknik baru yang tidak menggunakan *vector space model* dalam representasi dokumennya yaitu teknik *sequence of words* [1]. *Sequence of words* sering digunakan untuk merepresentasikan dokumen yaitu dengan menganggap dokumen sebagai sekumpulan kata-kata yang terurut sehingga arti semantik dari kata-kata tersebut akan tetap terjaga. Dengan memiliki dokumen yang tetap terjaga maka informasi yang terkandung di dalam dokumen akan lebih mudah didapatkan. Salah satu algoritma yang menggunakan teknik *sequence of words* adalah *Clustering based on Frequent Word Sequences (CFWS)* [5]. Algoritma CFWS merepresentasikan

dokumennya dengan menggunakan kata-kata yang paling sering muncul secara berurutan pada setiap dokumen. Dengan menggunakan algoritma ini dapat mengurangi dimensi dari setiap dokumen secara signifikan sehingga proses *clustering* akan menjadi lebih efisien[1].

Dalam tugas akhir ini dibuat suatu implementasi *clustering* untuk mengelompokkan berita sesuai dengan kata-kata yang paling sering muncul secara berurutan dari berita-berita tersebut. Data yang digunakan adalah berita dari koran berbahasa indonesia. Hasil dari *clustering* dari algoritma CFWS akan diukur tingkat akurasi nya dengan menggunakan sebuah metode yaitu metode *F-measure*. *F-measure* merupakan suatu metode pengukuran akurasi berdasarkan dengan nilai *precision* dan *recall*.

Diharapkan dengan adanya program ini, pengelompokkan berita dapat dilakukan lebih akurat.

1.2 Perumusan masalah

Tugas akhir ini membahas tentang implementasi *clustering* dengan menggunakan algoritma *Clustering based on Frequent Word Sequences (CFWS)*. Dalam tugas akhir ini terdapat beberapa perumusan masalah, antara lain:

1. Bagaimana mengimplementasikan *clustering* dengan menggunakan algoritma *Clustering based on Frequent Word Sequences (CFWS)* untuk melakukan segmentasi artikel berita berbahasa Indonesia .
2. Bagaimana mengelompokkan dan menggabungkan kandidat *cluster* berdasarkan teknik *sequence of words.*
3. Bagaimana mengelompokkan berita dengan tepat dan memiliki akurasi yang tinggi

.

1.3 Batasan masalah

Adapun yang menjadi batasan masalah dari penyusunan tugas akhir ini adalah:

1. Berita yang digunakan merupakan berita berbahasa Indonesia
2. Berita yang digunakan tidak diambil secara langsung melalui web,melainkan melalui database.
3. Algoritma yang akan digunakan adalah *algoritma Clustering based on Frequent Word Sequences (CFWS)*.

1.4 Tujuan

Tujuan pembuatan tugas akhir ini adalah sebagai berikut :

1. Membuat perangkat lunak yang dapat mengelompokkan berita berbahasa Indonesia dengan *algoritma Clustering Based on Frequent Sequences* (CFWS).

2. Mengevaluasi dan menganalisis hasil klasterisasi algoritma CFWS dengan berdasarkan nilai akurasinya dengan menggunakan *metode F-measure*.

1.5 Metodologi penyelesaian masalah

1. Studi Literatur

Mencari dan mengumpulkan beberapa referensi yang berkaitan dengan Data Mining khususnya *clustering* . Melakukan pendalaman materi, identifikasi masalah.Kemudian mempelajari dasar teori dan literatur-literatur yang relevan dengan data mining, text mining, dokumen clustering,teknik-teknik *clustering* khususnya algoritma CFWS,dan *sequential patterns.*

2. Pengumpulan data

Melakukan pengumpulan data sampel,berupa berita dari koran berbahasa indonesia.

3. Implementasi perangkat lunak

Mengimplemenasikan sistem perangkat lunak yang telah ditentukan kedalam bahasa pemrograman untuk menghasilkan suatu program yang dapat menganalisis berdasarkan perumusan masalah yang telah diuraikan diatas.

4. Testing dan Analisa Hasil

Pengujian dilakukan terhadap sistem yang telah dibangun pada tahap implementasi.

5. Pengambilan kesimpulan dan penyusunan laporan

Membuat kesimpulan dari hasil analisis yang telah dibuat, serta mendokumentasikan hasil perancangan, implementasi, pengujian, dan analisis ke dalam suatu bentuk laporan .

6. Perbaikan

Perbaikan dilakukan terhadap kesalahan-kesalahan yang mungkin terjadi pada perangkat lunak, laporan, maupun dokumentasi teknis.

1.6 Sistematika Penulisan

Bab 1 Pendahuluan, dimana bab ini menguraikan tugas akhir ini secara umum, meliputi latar belakang masalah, perumusan masalah, batasan masalah, tujuan, dan metodologi penyelesaian masalah.

Bab 2 Landasan Teori, dimana bab ini membahas mengenai uraian teori yang berhubungan dengan *text mining*. *Document clustering*, CFWS, tahapan CFWS, *suffix tree*, *k-mismatch*, Apriori, dan *Association Rule Mining.*

Bab 3 Analisis Perancangan dan Implementasi, dimana bab ini berisi analisis kebutuhan dari system dan masalah-masalah yang ada di dalamnya. Hasil

analisis ini dituangkan ke dalam suatu sistem pemodelan yang berorientasi objek. Dari tahap analisis kemudian dilanjutkan ke tahap perancangan dan implementasi.

Bab 4 Pengujian dan Analisis Hasil Percobaan, dimana bab ini membahas mengenai pengujian hasil implementasi yang telah dilakukan pada bab sebelumnya. Pengujian dilakukan dengan melakukan pengujian terhadap nilai *minimum support*, k dan *threshold*

Bab 5 Kesimpulan dan Saran, dimana bab ini berisi kesimpulan dari penulisan Tugas Akhir ini dan saran-saran yang diperlukan untuk pengembangan lebih lanjut.

5. Kesimpulan dan Saran

5.1 Kesimpulan

Dari hasil pengujian dan analisa yang dikerjakan dalam tugas akhir ini dapat diperoleh kesimpulan sebagai berikut:

1. Penentuan nilai *minimum support* pada *association rule mining* merupakan bagian terpenting dalam pembentukan *frequent 2-word sequences* yang merupakan inputan awal dari proses *clustering* dalam algoritma CFWS.
2. Nilai *minimum support* berpengaruh terhadap nilai akurasi, pada saat nilai *minimum support* lebih kecil, akan menghasilkan *frequent 2-word sequences* yang lebih besar dari yang seharusnya dan mengakibatkan hasil akurasi *cluster* yang tidak maksimal, sementara pada saat nilai *minimum support* yang terlalu besar akan mengakibatkan *frequent 2-word sequences* tidak dapat dihasilkan maka tidak ada *cluster* yang dihasilkan.
3. Nilai k mempengaruhi hasil *cluster* dan akurasi *cluster,* pada saat nilai k cukup besar maka jumlah *cluster* yang dihasilkan lebih sedikit dan sebaliknya apabila nilai k terlalu kecil nilai *cluster* yang dihasilkan lebih besar hal ini dikarenakan nilai k menunjukkan banyaknya jumlah kata yang dapat ditolerir untuk dimasukkan ke dalam satu *cluster.*
4. Nilai *threshold* tidak terlalu mempengaruhi nilai akurasi tetapi berpengaruh pada jumlah *cluster* yang dihasilkan.Nilai *threshold* yang terlalu kecil akan menyebabkan jumlah *cluster* menjadi lebih banyak dan nilai *threshold* yang terlalu besar menyebabkan jumlah *cluster* yang dihasilkan lebih sedikit.

5.2 Saran

Saran-saran untuk pengembangan tahap selanjutnya antara lain:

1. Dilakukan perbandingan dengan menggunakan proses stemming pada tahap preprocesing dengan tanpa menggunakan stemming.
2. Diperlukan pengembangan metode yang disertai dengan makna dari dokumen.
3. Penggunaan jumlah data yang lebih besar agar diperoleh sifat yang lebih general.

Daftar Pustaka

[1] **Li, Yanjun , B.S** . 2007. *High Performance Text Doument Custering* . Wright State University School Graduate Studies.

[2] **Beil, Florian, Martin Ester,Xiaowei Xu**. 2002. *Frequent Term Base Text Clustering*. ACM.

[3] **Han, Jiawei and Kamber, Micheline.** 2001. Chapter 1. *Data Mining: Concepts and Techniques*. Intelligent Database Systems Research Lab School of Computing Science Simon Fraser University, Canada.

[4] **Gusfield, Dan**. 1997. *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge,UK:Cambridge University Press.

[5] **Witten, Ian H**. 2003. *Text Mining*. University of Waikato, Coumputer Science, Nw Zealand.

[6] **Nong Ye.** 2003. *The Handbook of Data Mining*. New Jersey : Lawrence Erlbaum Associates, Inc., Publishers

[7] **Beil, Florian, Martin Ester, Xiaowei xu**. 2002. *Frequent Term-Based Text Clustering*. ACM.

[8] **Agrawal, Rakesh, Ramakrishnan Srikant.** *1995. Mining Sequential Pattrens.* IBM Research Center.

[9] **Ahonen-Myka, Helena.** 2005. *Mining All Maximal Frequent Word Sequences in a set of sentences.* ACM.

[10] **Tanoto, Andri**. 2007 . Pemanfaatan Sequential Patterns Dan Pengembangan CFWS Unttuk *Document Custering* . Undergraduate Theses from JBPTITBPP. Institut Teknologi Bandung. Indonesia.Available at:

http://digilib.itb.ac.id./gdl.php?op=read&id=jbptitbpp-gdl-andritanot-28941.

Di download pada tanggal 27 Maret 2010.

[11] **Zamir, Oren, Oren Etzioni.** 1998. *Web Document Clustering:A Feasibility Demonstration*. ACM

[12] **Taherizadeh S, Moghadam N**. 2009. *Integrating Web Content Mining Into Web Usage Mining for Finding Patterns and Predicting Users' Behaviors*. International Journal of Information Science and Management. Available at : http://www.ijism.ir/?_action=article&au=60&_au=S.++Taherizadeh