IMPLEMENTASI DAN ANALISIS PENGARUH AFFIX REMOVAL STEMMING TERHADAP CLUSTERING STUDI KASUS CLUSTERING TERJEMAHAN AYAT-AYAT AL-QUR'AN TENTANG PERMASALAHAN AKIDAH

Mohammad Shobri[1], Arie Ardiyanti Suryani[2], Yanuar Firdaus A.w.[3]

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

Abstrak

Clustering merupakan pengelompokkan data tanpa proses label kedalam kelompok – kelompok atau cluster-cluster, yang mempunyai nilai parameter kualitas cluster yang dibentuk (purity cluster), dan sebelum melakukan membentukan cluster-cluster didalam clustering dilakukan proses pengelolaan data menjadi kumpulan teks yang disebut dengan text mining. Pada text mining ada dua jenis kata yaitu kelompok kata yang berimbuhan dan kelompok kata yang sudah hilang imbuhannya. Untuk melakukan proses membentukan kata dasar dari kata yang berimbuhan dibutuhkan suatu proses yang disebut stemming. Algoritma affix removal stemming merupakan salah satu algoritma stemming yang bisa diterapkan dalam data teks bahasa Indonesia karena mampu menstemming data teks dengan nilai akurasi mencapai 90%. Pada studi kasus clustering teks terjemahan ayat-ayat Al-Qur'an, dibuktikan bahwa pengaruh penerapan stemming pada text mining, dapat mengurangi jumlah data teks yang akan diproses pada clustering dan mampu memberikan nilai purity cluster yang lebih baik dari pada clustering yang tidak menerapkan stemming.

Kata Kunci : Clustering, Cluster, Text Mining, Affix Removal Stemming, Purity Cluster, Al-Qur'an.

Abstract

Clustering is the grouping of data without the process of label into groups - groups or clusters, that have a quality parameter values formed clusters (cluster purity), and before doing to create clusters, in the clustering have been doing the data management process into a collection of text is called text mining . In text mining there are two types of words : word groups that have affix and group of words that have been lost the affix. To perform basic word processing, required a process called stemming. Affix removal stemming algorithm is one stemming algorithm that can be applied in the Indonesian language text data, as text data can stemming with values reaching 90% accuracy. In the case study translated text clustering verses the Qur'an, proved that the effect of stemming on the text mining application, able to reduce the amount of text data will be processed on clustering and cluster purity capable of delivering value better than the clustering does not apply stemming.

Keywords : Clustering, Cluster, Text Mining, Affix Removal Stemming, Purity Cluster, Al-Qur'an.

1. PENDAHULUAN

1.1 Latar Belakang Masalah

Dunia Internet sangat diminati oleh banyak kalangan masyarakat. Banyak diantaranya yang mencari solusi, tutorial, pembahasan, atau diskusi lewat browsing di Internet. Latar belakang penulis mengambil judul tugas akhir ini adalah sulitnya mengelompokkan terjemahan ayat-ayat Al - Qur'an pada kehidupan sehari – hari, dikarenakan bentuk Al – Qur'an yang konvensional sulit untuk dipelajari. Ayat – ayat dalam Al – Qur'an disusun tidak berdasarkan sebuah permasalahan dan sifatnya yang berpencar, menyebabkan pencarian ayat – ayat berdasarkan kelompok tertentu membutuhkan waktu lama. Dengan begitu, penulis berencana membuat sebuah sistem untuk mengelompokkan kata-kata yang ada pada terjemahan ayat-ayat Al-Qur'an yang berhubungan dengan akidah. Sehingga sistem tersebut dapat mengelompokkan dan penganalisa pengaruh stemming pada clustering terjemahan ayat – ayat Al – Qur'an sebagai referensi dan solusi.

Dengan masalah diatas, maka digunakan sebuah proses yang disebut *text mining*. Salah satu proses *text mining* adalah *stemming*. *Stemming* digunakan untuk mengganti bentuk dari suatu kata menjadi kata dasar dari kata tersebut yang sesuai dengan struktur morfologi bahasa Indonesia yang baik dan benar. Pada algoritma *stemming* teks bahasa Indonesia yang ada selama ini mempunyai beberapa kendala, salah satunya adalah pengenalan affix serapan (Imbuhan Asing). Oleh karena itu, penulis mencoba mengembangkan algoritma Porter, sehingga algoritma affix removal *stemming* ini mampu menyelesaikan beberapa permasalahan tersebut.

Pengelompokan informasi yang berkaitan dengan suatu kejadian tentu sulit dilakukan, bila hanya mengandalkan query biasa. Sebab pemilihan query yang kurang spesifik akan berakibat membanjirnya dokumen-dokumen yang tidak relevan. Oleh karena itu, penulis menggunakan proses *clustering* dalam pengelompokannya, dengan menggunakan *Algoritma Hierarchical Clustering* karena mudah diimplementasikan dan menghasilkan hirarki tree sehingga mempermudah pada proses analisa pengaruh *stemming* terhadap *clustering*. Pada tugas akhir ini, penulis melakukan analisa terhadap proses yang menggunakan *stemming* dan tidak menggunakan *stemming*, sehingga diharapkan menghasilkan kesimpulan analisa pengaruh *stemming* pada *clustering* dalam hal kualitas *cluster* (*Purity*) dan akurasi *stemming*nya.

1.2 Perumusan Masalah

Berdasarkan uraian diatas, maka permasalahan yang muncul dan yang menjadi objek penelitian pada Tugas Akhir ini ialah:

a. Bagaimana menerapkan algoritma *affix removal stemming* pada *clustering* yang menggunakan *Algoritma Hierarchical Clustering?*

b. Merancang dan pembangun perangkat lunak untuk penerapan algoritma *affix removal stemming* pada *clustering*.
c. Bagaimana menganalisis akurasi dan kualitas cluster (*Purity*) pengaruh algoritma *affix removal stemming* pada *clustering*?

Batasan masalah agar tidak meluasnya materi pembahasan dalam tugas akhir ini ialah:

a. Terjemahan ayat-ayat Al-Qur'an yang diinputkan pada ayat yang berkaitan dengan akidah dalam teks bahasa Indonesia yang baku.
b. Bahasa pemrograman yang digunakan adalah Visual Basic 6.0 (VB6) sebagai design *interface* dan *database* yang digunakan adalah Microsoft Office Access (MOA).
c. Pada proses *clustering*nya, menggunakan *Algoritma Hierarchical Clustering*, dengan kondisi berhenti proses clusteringnya ketika iterasi memiliki satu cluster.

1.3 Tujuan

Tujuan yang hendak dicapai dalam pembuatan Tugas Akhir ini yaitu : mengimplementasikan algoritma *affix removal stemming* dan *Algoritma Hierarchical Clustering* pada suatu aplikasi simulasi *Clustering and Stemming Sistem (CSS)* pada terjemahan ayat-ayat Al-Quran tentang permasalahan akidah.

1.4 Metode Penyelesaian Masalah

Metodologi yang digunakan untuk menyelesaikan masalah dalam Tugas Akhir ini ialah:

1. Studi Literatur
   Mempelajari sumber-sumber pustaka yang ada, yang dapat dijadikan referensi mengenai *text mining* khususnya algoritma *stemming*, proses *clustering* dan proses perbandingan kualitas cluster dan akurasi serta sumber-sumber lain yang terkait untuk menunjang penyelesaian tugas akhir ini. Sumber-sumber pustaka dapat berupa paper, buku, Al-Quran digital, maupun halaman web.

2. Analisis dan Desain
   Tahap ini meliputi analisis kebutuhan serta penyelesaian masalah untuk merancang perangkat lunak yang mempunyai fungsionalitas proses *stemming* dan *clustering* terjemahan ayat-ayat *Al-Quran*.

3. Implementasi Sistem
   Tahap ini meliputi pembangunan perangkat lunak yang telah dirancang pada tahap sebelumnya. Pembangunan perangkat lunak berbasiskan aplikasi desktop yaitu menggunakan aplikasi desktop dengan bahada pemprograman Visual Basic (VB6).

4. Analisis dan Pengujian
   Melakukan pengujian perangkat lunak yang telah dikembangkan, dan kemudian menganalisis hasil performansi yang didapatkan. Pengukuran performansi adalah kualitas cluster (*Purity*) yang dihasilkan. Tujuan pengujian adalah untuk mengetahui hasil penerapan algoritma *affix removal stemming* pada *clustering*, maka juga dilakukan analisa akurasi *stemming*nya.

5. Penyusunan Laporan
   Hasil penelitian akan disusun menjadi suatu laporan yang meliputi aspek-aspek dalam penelitian yaitu teori, perancangan dan implementasinya, serta membuat kesimpulan dari hasil penelitian tersebut.

1.5 Sistematika Penulisan

Sistematika dari Penulisan Tugas Akhir ini adalah sebagai berikut :

BAB I PENDAHULUAN

Bab ini membahas mengenai latar belakang pembuatan tugas akhir ini, rumusan masalah yang akan di analisis, ruang lingkup masalah yang ada pada tugas akhir ini, tujuan dari pembuatan tugas akhir ini, metodologi pemecahan masalah serta sistematika penulisan dokumentasi.

BAB II LANDASAN TEORI

Bab ini terdiri dari teori-teori yang digunakan dalam mendukung dalam penyelesaian tugas akhir ini, dalam hal ini adalah penerapan *text mining* dan *information retrieval* yaitu algoritma *stemming* pada bahasa Indonesia dan *Algoritma Hierarchical Clustering* pada *clustering* terjemahan ayat – ayat Al Qur'an tentang permasalahan akidah.

BAB III PERANCANGAN PERANGKAT LUNAK

Bab ini berisi pengumpulan data analisis dan perancangan perangkat lunak yang terdiri dari perancangan struktur data, perancangan modul dan *interface*.

BAB IV IMPLEMENTASI DAN PENGUJIAN

Berisi tentang implementasi detil dari implementasi dan analisa pengaruh *affix removal stemming* terhadap *clustering* dengan studi kasus *clustering* terjemahan ayat – ayat Al-Qur'an. Pengujian terhadap sistem juga dibahas pada bab ini.

BAB V PENUTUP

Berisi mengenai kesimpulan dan saran-saran yang dapat diambil oleh penulis dari keseluruhan sistem yang telah dibuat untuk pengembangan tugas akhir ini.

Telkom University

5. KESIMPULAN DAN SARAN

Pada bab ini akan diuraikan hal yang dapat disimpulkan dari pelaksanaan Tugas Akhir ini. Selain itu diuraikan pula beberapa saran yang dapat digunakan dalam pengembangan Tugas Akhir di masa mendatang.

5.1 Kesimpulan

Berdasarkan hasil analisis dan pengujian perangkat lunak yang dilakukan dalam tugas akhir ini dapat diambil beberapa kesimpulan, yaitu:

a. *Clustering and Stemming Stemming* yang dibangun, mampu melakukan fungsionalitas yang ada, yaitu melakukan proses stemming, proses pembobotan, proses clustering dan proses analisa dengan tingkat keberhasilan mencapai 99%.
b. Algoritma affix removal stemming dapat diterapkan pada teks bahasa Indonesia karena mampu memberikan nilai akurasi mencapai 93%.
c. Penerapan algoritma hierarchical clustering pada studi kasus clustering terjemahan ayat-ayat Al-Qur'an kurang memberikan nilai purity cluster yang kurang bagus karena nilai purity pada pengujian dihasilkan nilai purity antara 0,1 sampai 0,21.
d. Penggunaan stemming pada clustering memberikan dampak positif sebagai berikut : data teks yang diproses lebih sedikit dan iterasi yang dihasilkan lebih sedikit sehingga mempercepat proses clustering, serta cenderung dapat meningkatkan nilai purity cluster.

5.2 Saran

Untuk pengembangan Tugas Akhir di masa mendatang, penulis menyarankan hal-hal sebagai berikut:

a. Algoritma stemming lebih disempurnakan dengan melibatkan imbuhan bahasa Indonesia yang tidak baku, yaitu dengan cara membuat database yang membedakan jenis kata tersebut seperti kata kerja, kata sifat, kata benda, dan kata keterangan. Sehingga diharapkan algoritma stemming dapat menstemming imbuhan bahasa Indonesia yang tidak baku tersebut.
b. Pada clustering terjemahan ayat-ayat Al-Qur'an gunakan algoritma selain algoritma hierarchical clustering, sehingga diharapkan mampu memberikan hasil cluster yang lebih baik.

DAFTAR PUSTAKA

[1] Agus Zainal Arifin dan Ari Novan Setiono, 7 Mei 2002, Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering, Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember.

[2] Wikipedia, 23 Desember 2009, Stemming, <URL: http://en.wikipedia.org/wiki/Stemming>.

[3] Wikipedia, 23 Desember 2009, Wikipedia:Pedoman ejaan dan penulisan kata, <URL: http://id.wikipedia.org/wiki/Wikipedia:Pedoman_ejaan>.

[4] F.Z.Tala, *A Study of Stemming Effects on InformationRetrieval in Bahasa Indonesia*, master thesis at Institute for Logic, Language and Computation, Universiteit van Amsterdam.

[5] *NLP Stanford, 10 Februari, Agglomerative Clustering, <http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html#26381>*

[6] Vendy. Wordpress, 10 Februari, Stemming *<http://vendy.wordpress.com/2007/12/24/stemming-kata-berimbuhan-bahasa-indonesia/>*

[7] Weiss, Indurkhya, Zhang, Damerau, Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer: 2005.

[8] Harlian Ch, Milkha. "Text Mining.pdf". University of Texas : 2006. Didownload pada tanggal 20 Agustus 2009.

[9] Polajnar, Tamara. "Intro to Text Mining.pdf". University of Glasgow Departement of Computer Science Bioinformatics Research Centre : 2008.

[10] Hadi, Syamsul. Humaniora Volume XV No. 2/2003 : "Perubahan Fonologi Kata-Kata Serapan dari Bahasa Arab dalam Bahasa Indonesia". Yogyakarta : 2003.

[11] Tim penyusun. Kamus Bahasa Indonesia. Pusat Bahasa Departemen Pendidikan Nasional, Jakarta : 2008.

[12] Dam, Nikolaos van. Kata Serapan Arab Dalam Bahasa Indonesia. *Republika*, 2 July 2009.

[13] Anwar, Rosihon. Ulum Al-Qur'an.Pustaka Setia : "Munasabah Al-Qur'an". Bandung : 2007.

[14] Wibiosono, Yudi dan Masayu Leylia Khodra. Jurnal KNS : "Clustering Berita Berbahasa Indonesia". Bandung : 2005.

[15] Matteucc, 21 Novemver 2009, clustering : hierarchical clustering. <URL: http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html dan http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html>

[16] Stanford, 20 November 2009, clustering : AGNES. <URL: http://nlp.stanford.edu/IR-book/html/htmledition/hierarchical-agglomerative-clustering-1.html#26381>

[17] Kios Project, 20 November 2009, download file "AlQuranDigital21.exe". <URL: http://www.alqurandigital.com/download.htm>

[18] CSE, 21 November 2009. Purity Cluster, <URL: www.cse.iitm.ac.in/~cs672/purity.pdf>
[19] 20 November 2009. Grammer: Prefix, Suffix, dan Confix. <URL: http://www.bahasakita.com/>
[20] Agus Zainal Arifin dan Ari Novan Setiono, 7 Mei 2002, Klasifikasi Dokumen Berita Kejadian Berbahasa Indonesia dengan Algoritma Single Pass Clustering, Proceeding of Seminar on Intelligent Technology and Its Applications (SITIA), Teknik Elektro, Institut Teknologi Sepuluh Nopember.
[21] Wikipedia, 23 Desember 2007, Stemming, <URL: http://en.wikipedia.org/wiki/Stemming>.
[22] Wibisono, Yudi dan Masayu Leylia khodra. Jurnal : "Clustering Berita Berbahasa Indonesia".
[23] Aditya Vendy Pradana Blog, 12 Maret 2010, Stemming, <URL: http://vendy.wordpress.com/2007/12/24/stemming-kata-berimbuhan-bahasa-indonesia/>