IMPLEMENTASI ALGORITMA AMBIGUITY MEASURE FEATURE SELECTION PADA KATEGORISASI DOKUMEN TEKS BAHASA INDONESIA
IMPLEMENTATION AMBIGUITY MEASURE FEATURE SELECTION ALGORITHM ON CATEGORIZATION OF INDONESIAN TEXT DOCUMENT

Taajul Arifin[1], Shaufiah[2], Kemas Rahmat Saleh Wiharja[3]

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

Abstrak

Perkembangan teknologi internet sangat cepat membuat jumlah informasi berupa dokumen teks semakin banyak, oleh karena itu diperlukan suatu metode yang memudahkan pembaca untuk mencari informasi melalui proses kategorisasi. Namun tingginya dimensi data dapat mengganggu performansi hasil kategorisasi. Oleh karena itu dibutuhkan pemilihan feature yang berpengaruh besar terhadap kategorisasi yaitu feature selection. Ada beberapa algoritma dalam feature selection salah satunya yaitu Ambiguity Measure. Pada tugas akhir ini mengimplementasikan algoritma feature selection yaitu AM (Ambiguity Measure), dilakukan analisis hasil feature selection dengan menggunakan nilai threshold untuk memilih feature-feature yang berpengaruh terhadap proses kategorisasi. Kemudian diamati nilai precision dan recall menggunakan algoritma naïve bayess yang terdapat pada tools WEKA. Setelah dilakukan percobaan dengan menetapkan nilai threshold untuk pemilihan feature, menunjukan semakin tinggi nilai threshold yang ditetapkan maka jumlah feature yang dipilih oleh sistem semakin sedikit, namun performansi hasil kategorisasi meningkat. Performansi kategorisasi mencapai nilai tertinggi ketika ditetapkan threshold 0.95. Kemudian dilakukan perbandingan akurasi antara dataset sebelum dilakukan feature selection dan dataset setelah dilakukan feature selection, menunjukan akurasi yang dihasilkan setelah dilakukan feature selection lebih tinggi daripada dataset sebelum dilakukan feature selection.

Kata Kunci : feature selection, kategorisasi teks, Ambiguity Measure

Abstract

The development of Internet technology is very fast to make the amount of information of text documents is increasing. Therefore, the method is needed to find information though the categorization process. But the high-dimensional data can interfere with the performance results of categorization. Therefore feature selection is needed greatly affect categorization is feature selection. There are several algorithms in feature selection, one of which is ambiguity Measure. In this final report implements AM (ambiguity Measure) feature selection algorithm, to analyze the results of feature selection using a threshold value for selecting these features that influence the categorization process. Then the observed value of precision and recall using naïve bayess algorithm contained in the Weka tool. After doing the experiment by setting the threshold value for feature selection show the higher set threshold value then number of features selected by the system decrease, but the performance of the categorization increases. Categorization performance reaches the highest value when the specified threshold with 0.95. Then compare the accuracy of the dataset prior to feature selection and feature datasets after selection, the results yielded accuracy after feature selection is higher than the dataset prior to feature selection.

Keywords : feature selection, text categorization, ambiguity measure

1. PENDAHULUAN

1.1. Latar belakang masalah

Di era *globalisasi* sekarang ini, kebanyakan teks digital yang terdapat di World Wide Web (WWW) tidak teratur, sehingga pengguna Internet tidak dapat memanfaatkan informasi yang terkandung dalam data dari sebuah situs web. Contoh data teks digital yang terdapat dalam situs web yaitu : artikel berita, email, data pasien, dan lain sebagainya. Salah satu solusi untuk mengatur dan memanfaatkan informasi yang terkandung dalam data teks digital yaitu dengan melakukan proses kategorisasi, yang merupakan salah satu fungsionalitas dari *data mining*, tujuannya yaitu untuk mengetahui informasi berharga yang belum diketahui sebelumnya dari sekumpulan besar data.

Salah satu masalah dalam kategorisasi teks yaitu tingginya dimensi dari *feature space* data. Hal ini dapat mengganggu efektifitas dari hasil kategorisasi itu sendiri. Oleh karena itu, untuk mengurangi tingginya dimensi data harus dilakukan pemilihan terhadap beberapa atribut yang dapat berpengaruh besar terhadap hasil kategorisasi, yaitu *feature selection*.

Perhitungan nilai *feature* pada proses *feature selection* membutuhkan sebuah *measurement function*. Saat ini terdapat banyak *measurement function* dalam proses *feature selection* yang dapat digunakan, diantaranya: *CHI* (x^2), *Information Gain*, Improve Gini Index yang bertujuan untuk mendapatkan performansi yang bagus dari proses kategorisasi dokumen. Kelemahan dari *measurement function* yang ada seperti *CHI* (x^2), *Information Gain* yaitu memberikan *score* yang tinggi terhadap *term* jika term tersebut terdapat di lebihdari satu kategori , sedangkan dalam *Improve Gini Index* jika *term-term* yang memiliki frekuensi sama di dua atau lebih kategori, diberikan nilai yang lebih tinggi, sehingga *term* tersebut tidak dapat menunjuk kategori yang ada atau sering disebut *term* yang ambigu (*ambiguous term*)[7]. Masalah ini, dapat diatasi menggunakan algoritma *Ambiguity Measure feature selection. Ambiguity Measure* merupakan salah satu *measurement function* pada proses *feature selection* yang memberikan nilai yang tinggi terhadap *term* jika *term-term* tersebut hanya terdapat dalam satu kategori, sebaliknya apabila *term-term* yang terdapat dalam beberapa kategori dengan frekuensi *term* yang hampir sama atau frekuensi dari *term* tidak dominan dalam suatu kategori yang ada, maka akan diberikan nilai yang rendah.

Berdasarkan pertimbangan diatas, Tugas Akhir ini akan meneliti *measurement function Ambiguity Measure Feature Selection*, serta pengaruhnya terhadap performansi kategorisasi. Dengan mengimplementasikan *measurement function Ambiguity Measure* pada preprocessing data diharapkan *feature* yang terpilih dapat merupakan *feature* yang menunjuk pada suatu kategori tertentu, sehingga akan menghasilkan performansi kategorisasi yang baik.

1.2. Perumusan Masalah

Permasalahan yang dijadikan objek penelitian dalam tugas akhir ini yaitu :

Pengaruh algoritma *Ambiguity Measure feature selection* terhadap performansi kategorisasi dokumen berdasarkan parameter *precision* dan *recall*.

1.3. Tujuan

Tujuan yang ingin dicapai dalam penelitian ini adalah :

1. Mengimplementasikan algoritma *Ambiguity Measure* pada tahap *preprocessing* data.
2. Menganalisis hasil *feature selection Ambiguity Measure* dengan penetapan nilai *threshold* untuk memilih *feature*, dan pengaruh *Ambiguity Measure feature selection* terhadap performansi kategorisasi berdasarkan *precision* dan *recall.*

1.4. Batasan masalah

Dalam implementasi tugas akhir ini dibatasi oleh beberapa hal, sebagai berikut:

1. Dataset yang digunakan adalah artikel berita berbahasa Indonesia yang didapatkan dari web dan bersifat *offline* dan disimpan dalam file berekstensi .txt.
2. Proses pemilihan *feature* hanya dilakukan dengan cara penetapan nilai *threshold*.
3. Dokumen teks bahasa Indonesia yang digunakan didapat dari hasil riset research group Laboratorium Data Mining Centre (DMC).
4. Proses kategorisasi dilakukan dengan menggunakan Naïve Bayess yang terdapat pada *tools data mining*.
5. *Feature* hanya berupa kata bukan frase.

1.5. Metodologi PenyelesaianMasalah

Metodologi yang dilakukan untuk menyelesaikan permasalahan adalah sebagai berikut:

a. Studi litelatur

Melakukan pencarian informasi dan pembelajaran khususnya mengenai algoritma *Ambiguity Measure feature selection* dan Naive Bayess yang nantinya digunakan sebagai referensi tugas akhir.

b. Pengumpulan data

Melakukan pencarian dan pengumpulan data, data yang akan digunakan berupa artikel berita bahasa Indonesia yang diambil dari lab DMC (*Data Mining Center*).

c. Analisa kebutuhan dan implementasi

Tahap ini dilakukan dengan menganalisa dan merancang kebutuhan perangkat lunak yang dibangun.Sedangkan implementasi dilakukan terhadap hasil analisa dan perancangan kebutuhan perangkat lunak.

d. Pengujian

Pada tahap ini dilakukan pengujian dan analisa terhadap perangkat lunak yang telah dibangun menggunakan dataset yang telah disediakan, kemudian dilakukan kategorisasi menggunakan *tools data mining* untuk mengukur performansi yang dihasilkan.

e. Kesimpulan dan penyusunan laporan.

1.6. Sistematika Penulisan

Laporan Tugas Akhir ini disusun dengan sistematika sebagai berikut :

BAB I PENDAHULUAN

Pada bab ini dibahas mengenai latar belakang, perumusan, batasan, dan tujuan penelitian, metode pendekatan yang dipakai serta sistematika penulisan laporan.

BAB II LANDASAN TEORI

Menjelaskan teori-teori yang relevan dengan masalah yang diteliti, yaitu: dasar teori *feature selection*, *measurement function* yang ada pada *feature selection*, dan kategorisasi naive bayes.

BAB III ANALISIS DAN PERANCANGAN

Menjelaskan mengenai perancangan sistem, spesifikasi kebutuhan sistem, dan perancangan subsistem.

BAB IV EVALUASI HASIL

Menjelaskan tentang struktur program aplikasi, pengujian program dan hasil yang dicapai

BAB V KESIMPULAN DAN SARAN

Bab ini berisi tentang kesimpulan yang diambil dari hasil penelitian serta saran-saran untuk pengembangan lebih lanjut.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Dari analisis dan pengujian pada bab sebelumnya dalam tugas akhir ini, maka dapat disimpulkan bahwa :

1. *Feature selection* dapat meningkatkan performansi kategorisasi dengan artian bahwa setelah dilakukan *feature selection* menghasilkan jumlah feature yang lebih sedikit dengan nilai *precision, recall* dan akurasi yang lebih baik dari hasil kategorisasi menggunakan Naïve Bayes sebelum dilakukan proses *feature selection*. Nilai performansi yang paling baik terdapat pada threshold 0.95.
2. Nilai probabilitas yang dihasilkan *measurement function Ambiguity Measure* dipengaruhi oleh jumlah kategori.

5.2 Saran

Saran yang bisa diberikan penulis untuk perkembangan sistem ini yaitu:

1. Dikembangkan aplikasi yang dapat mengkategorisasikan dokumen secara online.
2. Mengganti algoritma pengukuran performansi kategorisasi dengan algoritma kategorisasi yang lain, misalnya SVM, K-NN, Naïve Bayes Multinomial.

DAFTAR PUSTAKA

[1] Christoper D. Manning, Prabhakar Raghavan, and Hinrich Schutze. 2008. **IR Online Book**. Cambridge University Press

[2] Dunja Mladenic and Marco Grobelnik. **Feature Selection for Classification based on Text Hierarchy**. J. Stefan Institute. www.citeseer.ist.psu.edu/605859.html [7 Feabruary 2010]

[3] Dunja Mladenic. **Slide : Dimensionality Reduction by Feature Selection in Machine Learning**.

[4] Guyon I, Elisseeff A**. An Introduction to Variable and Feature Selectio***n*. Journal of Machine Learning Research 3 (2003) 1157-1182.

[5] Ian H. Written and Eibe Frank. 2005. **Data Mining : Practical Machine Learning Tools and Techniques 2nd edition**. San Francisco : Morgan Kaufmann Publisher

[6] Mengle, S., & Goharian, N. (2008). **Using Ambiguity Measure Feature Selection Algorithm for Support Vector Machine Classifier.** In Proceedings of the 2008 ACM Symposium on Applied Computing (916–920).

[7] Mengle S, Goharian N.2009. **Ambiguity Measure Feature-Selection Algorithm**. Journal Of The American Society For Information Science And Technology, 60(5):1037–1050.

[8] Mark A. Hall and Llioyd A. Smith. **Feature Subset Selection : A Correlation Based Filter Approach**. University of Wakaito. www.cs.waikato.ac.nz/~ml/publications/1997/Hall-LSmith97.pdf [25 February 2010]

[9] Ronen Feldman and James Sanger. 2006. **Text Mining Handbook**. Cambridge University Press

[10] Santosa B.2007. **Data mining Teknik Pemanfaatan Data untuk Keperluan Bisnis**. Yogyakarta. Graha Ilmu

[11] Tala. Fadillah Z. **A Study Stemming Effect on Information Retrieval**. Netherland :Universiteit van Amsterdam. www.illc.uva.nl/Publications/ResearchReports/MoL-2003-02.text.pdf [20 April 2010]

[12] Wenqian Shang, Houkuan Huang, Haibin Zhu, Yongmin Lin, Youli Qu, and Zhihai Wang. 2007. **A Novel Feature Selection Algorithm for Text Categorization.** www.npissing.edu/haibin_zhu [7 April 2010]

[13] www.en.wikipedia.org/wiki/Naïve_Bayes_classifier [26 Mei 2010]

[14] Yiming Yang and Jan O. Pederson. **A Comparative Study on Feature Selection in Text Categorization**. In Proceeding of the 14th International Conference in Machine Learning. Nashville, USA, pp. 412-420. www.jsbi.org/journal/GIW02/GIW02F006.pdf [7 April 2010]