KLASIFIKASI TEKS DENGAN MENGGUNAKAN IMPROVED K-NEAREST NEIGHBOR ALGORITHM

Gema Megantara[1], Angelina Prima Kurniati[2], Arie Ardiyanti Suryani[3]

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

Abstrak

Klasifikasi merupakan proses mengelompokkan suatu data ke dalam kelompok data yang telah ditentukan berdasarkan tingkat kemiripannya. Klasifikasi ini pun dapat diterapkan dalam dokumen teks, dengan tujuan mempermudah penentuan seluruh dokumen dengan kategori tertentu. Terdapat berbagai cara untuk melakukan klasifikasi, salah satunya dengan menggunakan metode K-Nearest Neighbor. Metode K-Nearest Neighbor merupakan metode yang populer dalam klasifikasi, karena kemudahan dalam implementasinya.

Tetapi dibalik kemudahannya itu metode K-Nearest Neighbor memiliki kelemahan jika digunakan dalam dokumen yang memiliki distribusi yang tidak merata, karena saat nilai k yang digunakan semakin besar akan ada dominasi oleh kelas yang berukuran besar terhadap kelas yang berukuran kecil. Oleh karena itu digunakan metode Improved K-Nearest Neighbor untuk menanggulangi kelemahan tersebut.

Untuk mengevaluasi performansi dari K-Nearest Neighbor dan Improved KNearest Neighbor digunakan precision, recall, dan F1-Measure. Hasil yang didapat menunjukkan bahwa metode Improved K-Nearest Neighbor dapat menghilangkan efek dominasi dari kategori terbesar dalam berbagai jenis distribusi dokumen training.

Kata Kunci : klasifikasi, K-Nearest Neighbor, Improved K-Nearest Neighbor

Abstract

Classification is the process that grouping the data into the class based on similarity level. Classification can be also applied in text document, to make easier act of determining whole document with certain category. There is a various way to do the classification, one of them is with the K-Nearest Neighbor method. The K-Nearest Neighbor is a popular method in classification because of easy in implementation.

But, behind in the easiness, the K-Nearest Neighbor method has a weakness if it be used in a document that has uneven distribution, because when the k value more and more bigger will appear domination by the bigger class to the smaller class. Therefore the Improved K-Nearest Neighbor method has been used for to cope with the weakness.

Precision, recall, and F1-Measure are used for evaluating the performance from the K-Nearest Neighbor method and Improved K-Nearest Neighbor method. The result shows that the Improved K-Nearest Neighbor method can eliminate the domination effect from the largest category in various kind of document training distribution.

Keywords : classification, K-Nearest Neighbor, Improved K-Nearest Neighbor

1. Pendahuluan

1.1 Latar belakang

Klasifikasi merupakan proses mengelompokkan suatu data ke dalam kelompok data yang telah ditentukan berdasarkan tingkat kemiripannya. Klasifikasi ini pun dapat diterapkan dalam dokumen teks. Dengan tujuan mempermudah penentuan seluruh dokumen dengan kategori tertentu. Permasalahan ditemukan saat jumlah dokumen yang besar, tidak mungkin diklasifikasikan secara manual dengan dibaca satu persatu, maka harus dibuat sistem yang dapat mengklasifikasikan dokumen teks tersebut secara otomatis.

Metode pengklasifikasian teks yang umum digunakan adalah *K-Nearest Neighbor* regular [3]. Dan sudah banyak peneliti yang menemukan bahwa metode *K-Nearest Neighbor* regular memiliki performansi yang bagus dalam penelitiannya [2][4]. Ide dari *K-Nearest Neighbor* reguler untuk mengklasifikasikan dokumen baru, adalah dengan cara menemukan sejumlah k tetangga terdekat dari dokumen *training*, dan menggunakan kategori dari k tetangga terdekat untuk menentukan kategori dari data kandidat. Namun kelemahan *K-Nearest Neighbor* reguler memiliki kelemahan saat menentukan *class* dari data kandidat. Seperti *K-Nearest Neighbor* reguler akan salah menentukan *class* dari data kandidat saat k tetangga terdekat memiliki anggota *neighbor* lebih banyak yang memiliki nilai *similarity* yang kecil dari suatu *class* menjadi *class* pemenenang, sedangkan anggota *neighbor* yang memiliki nilai *similarity* lebih besar kalah dalam jumlah dimana yang seharusnya menjadi *class* dari data kandidat menjadi *class* yang kalah[1]. Untuk mengatasi kelemahan tersebut telah ditemukan metode *Improved K-Nearest Neighbor*[1]. Maka dari itu dalam tugas akhir ini akan digunakan metode *Improved K-Nearest Neighbor* untuk mengklasifikasikan dokumen teks.

Perbedaan algoritma *Improved K-Nearest Neighbor* dan algoritma *K-Nearest Neighbor* regular terletak pada jumlah *neighbor* yang digunakan [1]. Pada algortima *K-Nearest Neighbor* jumlah *neighbor* yang sama diterapkan pada setiap kasus walaupun jumlah anggota tiap *class* berbeda. Sedangkan pada algoritma *Improved K-Nearest Neighbor* jumlah *neighbor* yang digunakan adalah berbeda untuk *class* yang berbeda. Dengan kata lain pada algoritma *K-Nearest Neighbor* reguler, *nearest neighbor* yang digunakan adalah sejumlah K yang telah ditentukan. Sedangkan pada algoritma *Improved K-Nearest Neighbor*, hanya digunakan *top* n *nearest neighbor* yang mewakili dari tiap *class* yang ada.

1.2 Perumusan masalah

Untuk text classification dengan menggunakan algortima *Improved K-Nearest Neighbor*, terdapat beberapa masalah yang akan diselesaikan di Tugas Akhir ini, yaitu sebagai berikut:

1. Bagaimana mengubah suatu dokumen teks menjadi data yang siap untuk digunakan dalam proses klasifikasi?
2. Bagaimana membuat sistem *text classification* dengan menggunakan metode *Improved K-Nearest Neighbor*?

3. Bagaimana performansi *Improved K-Nearest Neighbor* dalam mengklasifikasikan dokumen teks dibandingkan dengan performansi *K-Nearest Neighbor Regular* berdasarkan confusion matrix dengan parameter F-measure?

Adapun batasan masalah Tugas Akhir ini adalah sebagai berikut :

1. *Document collection* yang digunakan adalah dokumen berbahasa Indonesia yang berasal dari situs www.okezone.com.
2. Dokumen yang diuji berupa dokumen teks berita dengan format *..txt*.

1.3 Tujuan

Tujuan yang ingin dicapai dalam penyusunan Tugas Akhir ini adalah sebagai berikut:

1. Menerapkan metode *Improved K-Nearest Neighbor* dalam klasifikasi dokumen teks.
2. Menganalisis hasil implementasi Algoritma *Improved K-Nearest Neighbor* dalam *text classification* dibandingkan dengan hasil implementasi Algoritma *K-Nearest Neighbor* regular dilihat dari *precision* dan *recall, F-measure,* dan standar deviasi.

1.4 Metodologi penyelesaian masalah

Metodologi yang digunakan dalam memecahkan masalah di atas adalah dengan menggunakan langkah-langkah berikut:

1. Studi literatur
   Pencarian referensi dan sumber-sumber yang berhubungan dengan text classification.
2. Pengumpulan data
   Mengumpulkan data *document collection* yang nantinya akan digunakan sebagai *data training* dan *data testing*.
3. Analisis dan perancangan sistem
   Melakukan analisis dan perancangan terhadap sistem yang dibangun, menganalisis metode yang akan digunakan untuk menyelesaikan permasalahan, termasuk menentukan bahasa pemrograman yang digunakan, arsitektur, fungsionalitas, dan antarmuka sistem. Menyiapkan data yang siap untuk dilakukan *data mining* dengan melewati proses *preprocessing*. Input sistem berupa data latih, data validasi, dan data uji. Data latih dan data validasi digunakan untuk membangun fungsi prediksi optimal sedangkan data uji digunakan untuk menguji akurasi sistem prediksi.
4. Implementasi dan pembangunan sistem
   Pada tahap ini, akan dilakukan implementasi sistem yang mampu mengklasifikasikan dokumen teks secara otomatis dengan menggunakan metode *Improved K-Nearest Neighbor* dan *K-Nearest Neighbor* regular.
5. Pengujian dan analisis
   Setelah sistem telah sempurna maka akan dianalisis hasil dari klasifikasi dan kestabilan dari sistem.
6. Pengambilan kesimpulan dan penyusunan laporan Tugas Akhir.

2. Dasar Teori

2.1 *Text Mining*

Text Mining memiliki definisi menambang data yang berupa teks dimana sumber data biasanya didapatkan dari dokumen, dan tujuannya adalah mencari kata-kata yang dapat mewakili isi dari dokumen shingga dapat dilakukan analisa keterhubungan antar dokumen[6].

Text Mining berbeda dengan *searching* pada dokumen seperti biasanya. *Searching* biasanya dilakukan untuk mencari sesuatu infromasi yang diinginkan, namun informasi itu sebelumnya sudah ada dalam dokumen. Sedangkan *Text Mining* adalah mencari informasi baru yang diinginkan dengan mengolah informasi yang ada. Informasi yang telah ada sebelumnya diproses dengan suatu cara khusus untuk menghasilkan informasi yang lebih berguna.

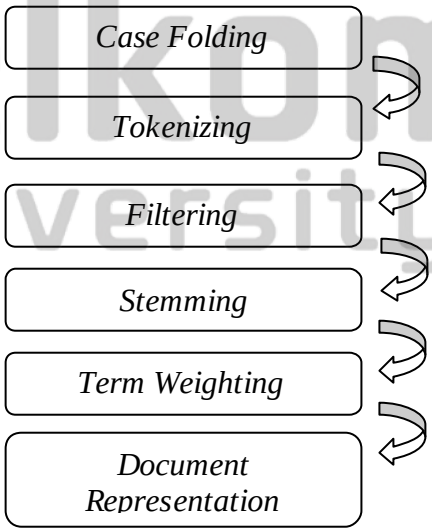
Dalam *Text Mining* terdapat dua tahapan yang akan dilakukan. Yang pertama adalah *Preprocessing*. *Preprocessing* bertujuan untuk mendapatkan data yang siap untuk dilakukan proses pada tahapan berikutnya, yang pada tugas akhir kali ini adalah proses klasifikasi.

2.2 *Preprocessing*

Dalam *preprocessing* dokumen yang ada sebelumnya akan dibuat menjadi dokumen yang terstruktur sehingga akan tercipta dokumen yang berkualitas untuk dilakukan proses klasifikasi. Tujuannya *preprocessing* dalam *text mining* adalah mentransformasi data ke suatu format yang prosesnya lebih mudah dan efektif untuk kebutuhan proses berikutnya, dengan indikator sebagai berikut :

a. Mendapatkan hasil yang lebih akurat.
b. Pengurangan waktu komputasi untuk *large scale problem*.
c. Membuat nilai menjadi lebih kecil tanpa merubah informasi yang dikandungnya.

Dalam Text Mining terdapat beberapa tahap dalam preprocessing [6], yaitu :



Gambar 2 - 1 Preprocessing

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Secara keseluruhan performa dalam melakukan klasifikasi metode *Improved K-Nearest Neighbor* mencapai hasil yang lebih baik dari metode *K-Nearest Neighbor* dalam berbagai kondisi. Baik itu diimplementasikan dalam distribusi dokumen merata, cukup rata, ataupun sangat timpang.
2. Metode *Improved K-Nearest Neighbor* memiliki kestabilan yang lebih baik daripada metode *K-Nearest Neighbor* regular dilihat dari standar deviasi yang dihasilkan.
3. Kriteria distribusi yang akan menimbulkan efek dominasi dari kategori terbesar adalah ketika kategori terkecil memiliki ukuran dokumen 1/5 dari dokumen terbesar dengan nilai *k* maksimal 1/2 dari dokumen terbesar, dan akan menyebabkan penurunan performa pada kedua metode baik *K-Nearest Neighbor* regular maupun *Improved K-Nearest Neighbor*.

5.2 Saran

Saran terhadap pengembangan terhadap tugas akhir ini adalah :

1. Metode *Improved K-Nearest* juga dapat juga digunakan dalam *document collection* yang bukan merupakan dokumen teks. Misalnya dokumen gambar, suara, dll. Asalkan semua jenis dokumen tersebut telah melalui *preprocessing* menjadi data numerik.
2. Dapat dicoba diterapkan pada sistem klasifikasi yang bersifat *online*.

Referensi

[1] Li Baoli. An Improved k-Nearest Neighbor Algorithm for Text Categorization.

[2] Yang Y. and Liu X., 1999. A Re-examination of Text Categorization Methods [A]. In: Proceedings of 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval [C]. 42-49.

[3] Manning C. D. and Schutze H., 1999. Foundations of Statistical Natural Language Processing [M]. Cambridge: MIT Press.

[4] Joachims T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features [A]. In: Proceedings of the European Conference on Machine Learning [C].

[5] Fadillah Z Tala. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia.

[6] Milkha Harlian Ch. Text Mining, Final Project. http://lecturer.eepis-its.edu/~iwanarif/kuliah/dm/6Text%20Mining.pdf, diakses tanggal 29 November 2009.

[7] Dr. E. Garcia. The Classic Vector Space Model. http://www.miislita.com/term-vector/term-vector-3.html, diakses tanggal 29 November 2009.

[8] Yiming Yang and Thorsten Joachims, 2008. Text Categorization. http://www.scholarpedia.org/article/Text_categorization, diakses tanggal 29 November 2009.

[9] Y Liao. Review of K-Nearest Neighbor Text Categorization Method. http://www.usenix.org/events/sec02/full_papers/liao/liao_html/node4.html , diakses tanggal 30 November 2009.

[10] Evaluation Method in Text Categorization. http://datamin.ubbcluj.ro/wiki/index.php/Evaluation_methods_in_text_categorization, diakses tanggal 14 Februari 2010.

[11] Rumus Standar Deviasi http://www.gealgeol.com/2009/04/23/rumus-standard-deviasi.html, diakses tanggal 5 Maret 2010.