KATEGORISASI DOKUMEN BERITA BERBAHASA INDONESIA DENGAN FEATURE PROJECTION INDONESIAN NEWS DOCUMENT CATEGORIZATION WITH FEATURE PROJECTION

Yunita Rahmawati[1], Moch. Arif Bijaksana[2], Kiki Maulana[3]

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

Abstrak

Pesatnya penggunaan Internet sebagai sarana penyebarluasan informasi mengakibatkan tersedianya informasi dalam jumlah yang banyak, beraneka ragam dan pada umumnya tidak terstruktur. Hal ini mendorong kebutuhan untuk mengelola informasi guna mendapatkan pengetahuan yang bermanfaat. Salah satu komponen penting dalam pengelolaan informasi adalah kategorisasi teks yang dapat membantu pengguna dalam mendapatkan informasi berdasarkan kategori tertentu secara cepat dan akurat.

tertentu secara cepat dan akurat. Pada tugas akhir ini digunakan Feature Projection dan k-Nearest Neighbor Feature Projection (k-NNFP) untuk mengklasifikasikan dokumen berita berbahasa Indonesia. Dari perangkat lunak yang dihasilkan, diukur performansi dan waktu klasifikasi dari kedua classifier pada berbagai data set dan dibandingkan dengan k-Nearest Neighbor.

Feature Projection dan k-Nearest Neighbor Feature Projection (k-NNFP) memiliki performansi yang tinggi dan bervariasi pada berbagai data set namun selisih yang terjadi hanya sedikit. Dari hasil penelitian pada tugas akhir ini juga diketahui bahwa Feature Projection memiliki waktu klasifikasi yang lebih cepat jika dibandingkan dengan k-NNFP dan k-NN. Hal ini menunjukkan Feature Projection dapat menjadi sebuah classifier yang berguna dalam proses kategorisasi teks yang membutuhkan kecepatan dan performansi yang tinggi.

Kata Kunci : Feature Projection, k-Nearest Neighbor Feature Projection , k-

Abstract

The rapid usage of internet as an information spreading medium causes information availability in the large amount, various, and commonly unstructured. This lead to the need of information management to get useful knowledge. One of important components in information management is text categorization that can help user to get information based on certain category fast and accurately.

In this final project will be used Feature Projection and k-Nearest Neighbor Feature Projection (k-NNFP) to classify Indonesian news document. From resulted software, it will be measured performance and classification time from both classifier in many data sets and compared to k-Nearest Neighbor.

Feature Projection and k-Nearest Neighbor Feature Projection (k-NNFP) have high performance and varied in many data sets but it have little difference. From the research in this final project also can be known that Feature Projection has faster classification time compared to k-NNFP and k-NN. This is demonstrated that Feature Projection can be a useful classifier in text categorization process that needs high performance and high speed.

Keywords : Feature Projection, k-Nearest Neighbor Feature Projection , k-Nearest

1. Pendahuluan

1.1 Latar belakang

Internet atau lebih tepatnya WWW(World Wide Web) merupakan sarana yang paling umum digunakan untuk menyebarluaskan informasi. Pesatnya penggunaan Internet mengakibatkan tersedianya informasi dalam jumlah yang sangat besar, beraneka ragam, dan pada umumnya tidak terstruktur. Hal ini mendorong meningkatnya kebutuhan untuk menemukan dan mengelola informasi dengan baik sehingga dihasilkan pengetahuan yang berguna.

Salah satu komponen yang penting dalam pengelolaan informasi adalah pengkategorisasian teks. Tugas utama dari pengkategorisasian teks ini adalah mengklasifikasikan dokumen ke dalam kategori tertentu yang telah didefinisikan sebelumnya. Hal ini dapat membantu pengguna dalam menemukan informasi yang dibutuhkan berdasarkan topik tertentu dengan cepat dan akurat.

Terdapat beberapa *classifier* yang dapat digunakan untuk mengkategorisasikan teks diantaranya yaitu *k-NN Feature Projection* dan *Feature Projection.* Ide utama dari kedua *classifier* ini berasal dari *k-Nearest Neighbor (k-NN). k-NN* merupakan *classifier* yang sering digunakan pada kategorisasi teks karena memiliki performansi yang tinggi, algoritma yang sederhana dan mudah untuk dipelajari. Tetapi *k-NN* memiliki kelemahan utama yaitu waktu yang dibutuhkan untuk mengklasifikasikan dokumen sangat lambat[5]. Hal ini karena *k-NN* harus menghitung nilai kesamaan antara setiap dokumen uji dengan keseluruhan data latih untuk mendapatkan sejumlah *k* tetangga terdekat dari dokumen uji tersebut.

Pada tugas akhir ini digunakan *k-NN Feature Projection dan Feature Projection* sebagai *classifier* untuk melakukan pengkategorisasian dokumen berita berbahasa Indonesia. Berdasarkan referensi 5, proses yang dilakukan pada kedua *classifier* ini yaitu menormalisasikan setiap data latih sebagai sebuah unit *vector*. Dokumen *vector* ini ditransformasikan ke dalam *feature projections* dan disimpan pada masing-masing dimensi *feature*-nya. Hasil dari *feature projection* pada sebuah *feature* dapat dilihat sebagai sekumpulan bobot dari dokumen untuk *feature* tersebut. Pada *k-NN Feature Projection* klasifikasi awal pada sebuah *feature* dilakukan berdasarkan pada sejumlah k tetangga terdekat dari data uji pada *feature* tersebut[2]. Sedangkan pada *Feature Projection* masing-masing *feature* dari data latih dilakukan *voting* awal untuk sebuah proses klasifikasi[5]. Klasifikasi akhir dari sebuah dokumen uji ditentukan oleh sebuah *majority voting* dari setiap *voting* awal yang dilakukan pada data latih untuk masing-masing *feature*.

Pada *k-NN Feature Projection* dan *Feature Projection*, setiap *feature* secara individual berkontribusi pada proses klasifikasi sehingga kedua *classifier* ini tidak terpengaruh oleh adanya *feature* yang tidak relevan. Selain itu pada *k-NNFP* perhitungan similarity hanya dilakukan pada masing-masing feature saja. Sedangkan pada *Feature Projection* pengklasifikasian dokumen uji hanya membutuhkan perhitungan yang sederhana yaitu berdasarkan pada hasil voting dari masing-masing *feature* untuk masing-masing kategori. Oleh karena itu kedua *classifier* ini mengatasi kekurangan dari *k-NN* dalam hal waktu klasifikasi dan

mampu mempertahankan poin kekuatan dari *k-NN* yaitu performansi yang tinggi[2].

1.2 Perumusan masalah

Pada tugas akhir ini masalah yang akan diselesaikan yaitu:

1. Bagaimana cara untuk mengatasi keberagaman nilai *document frequency* untuk masing-masing feature.
2. Bagaimana mengkategorisasikan dokumen berita berbahasa Indonesia dengan *Feature Projection.*
3. Bagaimana performansi dari penerapan *Feature Projection* pada pengkategorisasian dokumen berita berbahasa Indonesia.

Batasan masalah pada tugas akhir ini yaitu:

1. Dataset yang akan digunakan merupakan artikel berita berbahasa Indonesia yang bersifat *offline*.
2. Dataset yang digunakan adalah data yang telah berlabel.
3. Dokumen dikategorikan ke dalam *single label*.
4. Tidak menangani *preprocessing* data.
5. Klasifikasi dokumen berdasarkan pada kata tunggal bukan berdasarkan pada frasa.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan *Feature Projection* dan *k-Nearest Neighbor (k-NNFP)* untuk mengklasifikasikan dokumen berita berbahasa Indonesia.
2. Mengevaluasi perbandingan performansi yang berupa nilai *precision, recall, f-measure* dan *micro average f-measure* serta waktu klasifikasi dari *Feature Projection* dan *k-Nearest Neighbor Feature Projection (k-NNFP)* serta dibandingkan dengan *k-Nearest Neighbour (k-NN)* .

1.4 Metodologi penyelesaian masalah

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

1. Studi literatur.
   Mencari, mengumpulkan informasi, memahami dan mempelajari konsep kategorisasi teks dengan menggunakan *Feature Projection* serta konsep *text mining* melalui literatur berupa makalah, buku, atau jurnal yang berhubungan dengan *text mining, text categorization* dan *Feature Projection*.
2. Pencarian dan pengumpulan data.
   Mengumpulkan data berupa dokumen berita berbahasa Indonesia yang dibutuhkan untuk keperluan proses implementasi dan pengujian metode yang digunakan.
3. Analisis kebutuhan dan perancangan aplikasi yang akan dibangun.

Tahapan ini dilakukan dengan menganalisa kebutuhan perangkat lunak dan merancang perangkat lunak untuk pengkategorisasian dokumen berita berbahasa Indonesia dengan *Feature Projection* dan *k-Nearest Neighbor (k-NNFP)*

4. Implementasi dan Pengujian
   Mengimplementasikan hasil analisis dan perancangan perangkat lunak serta melakukan pengujian terhadap dataset yang akan diklasifikasikan untuk mengukur performansi dari *classifier*.
5. Analisa hasil pengujian dan penarikan kesimpulan.
   Perangkat lunak yang dihasilkan dievaluasi berdasarkan data yang diperoleh dari hasil pengujian.
6. Penyusunan laporan tugas akhir.
   Pembuatan laporan tugas akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam tugas akhir ini.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Nilai *precision, recall,* dan *f-measure* dari *Feature Projection* dan *k-NN Feature Projection* pada data set tanpa *noise* dan dengan *noise* bervariasi namun tidak ada selisih yang signifikan jika dibandingkan dengan *k-NN*. Ketiga *classififer* memiliki *precision, recall,* dan *f-measure* yang tinggi yaitu di atas 0,88. Hal ini menunjukkan bahwa Feature Projection dan *k-NNFP* memiliki performansi berdasarkan nilai *precision, recall* dan *f-measure* yang tinggi dan sebanding dengan *k-NN*.
2. Terdapat anomali nilai *precision* pada ketegori internasional untuk *Feature Projection* dimana data set dengan *noise* memiliki nilai *precision* yang lebih tinggi dibandingkan dengan data set tanpa *noise.* Data *noise* pada data latih mengandung *feature-feature* yang memberikan kontribusi pada kategori internasional. Hal ini mempengaruhi pembangunan model pada tahap *training* dimana hasil voting masing-masing *feature* untuk kategori internasional semakin tinggi dan akurat yang mengakibatkan nilai *precision* untuk kategori internasional lebih baik pada data set yang mengandung *noise.*
3. *k-NNFP* memiliki nilai *micro-average f-measure* yang paling tinggi dibandingkan dengan *Feature Projection* dan *k-NN* pada seluruh data set yang digunakan disebabkan oleh pemilihan tetangga terdekat dari dokumen uji berdasarkan pada masing-masing *feature* secara terpisah sehingga mengurangi efek adanya *noise* pada data latih. Hal ini sangat efektif untuk meningkatkan keefektifan dari *k-NNFP* dalam mengklasifikasikan dokumen uji.
4. *Feature Projection* memiliki waktu klasifikasi yang paling cepat jika dibandingkan dengan *k-NNFP* dan *k-NN* karena pengklasifikasian dokumen uji hanya membutuhkan perhitungan yang sederhana yaitu berdasarkan pada hasil voting dari masing-masing *feature* untuk masing-masing kategori.

5.2 Saran

1. Perlu dilakukan penelitian lebih lanjut untuk mengatasi masalah pada *Feature Projection* jika data set yang digunakan sebagai data latih memiliki jumlah dokumen pada masing-masing kategori yang tidak sebanding. Penyebab masalah ini adalah kategori yang lebih banyak dokumennya memiliki kandidat voting yang lebih banyak jika dibandingkan dengan kategori yang ukuranya lebih kecil. Hal ini dapat diatasi dengan memodifikasi nilai voting utama sehingga kategori yang lebih besar tidak mendominasi proses klasifikasi.
2. Perlu dilakukan penelitian lebih lanjut untuk meningkatkan performansi dari *Feature Projection* dengan memperhatikan adanya *co-occurrence frequency* antara dua buah *feature* yaitu jumlah kemunculan bersama antara dua buah *feature* dalam data set.

3. Perlu dikembangkan aplikasi yang dapat mengkategorisasikan dokumen berita secara online.

Referensi

[1] Adiwijaya. 2006. *Text Mining and Knowledge Discovery*. Kolokium bersama Komunitas Data Mining Indonesia & Softcomputing Indonesia.

[2] Akkus, Aynur dan Guvenir H.A. 1996. *K-Nearest Neighbour Classificaton on Feature Projections*. ICML 1996. Didownload pada tanggal 14 Februari 2008 di www.citeseer.ist.psu.edu/87838.html

[3] Fahmi, Ismail. 2004. *Examining Learning Algorithms for Text Classification in Digital Libraries*. Groningen, Netherland : University of Groningen. Didownload pada tanggal 14 Februari 2008 di www.let.rug.nl/alfa/scripties/IsmailFahmi.pdf

[4] Feldman, Ronen dan James Sanger. 2007. *The Text Mining Handbook*. New York: Cambridge University Press.

[5] Ko, Youngjoong dan Jungyun Seo. 2002. *Text Categorization Using Feature Projections*. International Conference on Computational Linguistics ke-19. Didownload pada tanggal 14 Februari 2008 di www.aclweb.org/anthology-new/C/C02/C02-1074.pdf

[6] Ko, Youngjoong dan Jungyun Seo. 2004. *Text Ccategorization Using Unlabeled Data*. Sogang Unoiversity Korea. Didownload pada tanggal 14 Februari 2008 di http://portal.acm.org/citation.cfm?id=985318

[7] Manning, Christopher D (dkk.). 2007. *An Introduction to Information Retrieval*. Cambridge University Press. Didownload tanggal 14 Februari 2008 di www.nlp.stanford.edu/IR-book/pdf/irbookprint.pdf

[8] Salton, Gerard dan Christopher Buckley. 1988. *Term Weighting Approaches in Automatic Text Retrieval*. Information Processing and Management. Didownload pada tanggal 20 Februari 2008 di http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.101.9086

[9] Yang, Yiming. 1999. *An Evaluation on Statistical Approaches to Text Categorization*. Jounal of Information Retrieval :67-88. Didownload tanggal 5 Maret 2008 di http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.50.8921

[10] Yang, Yiming(dkk). 2002. *A Study of Approaches to Hypertext Categorization*. Journal of Intelligent Information System, volume 18, number 2. Didownload tanggal 25 Juni 2008 di….

[11] _____. 2008. *Precision and Recall*. Didownload tanggal 25 Juni 2008 di http://en.wikipedia.org/wiki/Precision_and_recall