IMPLEMENTASI DAN ANALISIS RANDOM-WALK TERM WEIGHTING UNTUK KLASIFIKASI ARTIKEL BERITA BERBAHASA INDONESIA

Rizqi Meirina Fadilla[1], Yanuar Firdaus A.w.[2], Zk. Abdurahman Baizal[3]

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

Abstrak

Term frequency telah lama digunakan sebagai metode pembobotan term dalam dokumen teks. Metode ini mengasumsikan bahwa setiap term memiliki nilai kepentingan yang sebanding dengan jumlah kemunculannya pada dokumen. Hal ini menjadi kelemahan utama dari term frequency sehingga mengakibatkan term menjadi independen dan mengabaikan keterhubungan yang mungkin ada antar term dalam dokumen. Kelemahan term frequency ini dapat diatasi dengan mengimplementasikan metode random-walk. Metode ini direpresentasikan dengan algoritma perankingan berbasis graf yang diterapkan dalam graf tekstual yang dapat mengintegrasikan dependensi antar term dan konteks sekitarnya.

Pada Tugas Akhir ini dibahas pembobotan term dengan menggunakan metode term frequency dan random-walk pada dataset artikel berita berbahasa Indonesia. Ada dua skema pembobotan yang akan digunakan yaitu skema tf – rw dan skema tf.idf – rw.idf. Lalu, dataset ini akan diklasifikasikan dengan menggunakan klasifier pada tools Weka. Analisa performansi hasil klasifikasi dilakukan dengan menggunakan nilai akurasi dan macro-average f-measure.

Hasil percobaan menunjukan bahwa metode random-walk memberikan performansi yang lebih baik dari metode term frequency khususnya pada skema tf.idf - rw.idf.

Kata Kunci : klasifikasi, term weighting, term frequency, random-walk, akurasi,

Abstract

Term frequency has been long used as a method of term weighting in text document. The method assumes that every term has importance value which is proportional to its frequency on document. It is the main weakness of term frequency which causes term becomes independent and disregards any dependencies that may exist between terms in the text. The problem of term frequency can be solved by applying the method of random-walk term weighting. The method is represented by graph-based ranking algorithm which is applied in textual graph that is able to integrates the dependencies of a term and its surrounding context.

The final project researches term weighting using the methods of term frequency and random-walk toward Indonesian news articles. There are two weighting schemes which will be used, tf – rw scheme and tf.idf – rw.idf scheme. Then, the datasets will be classified using Weka. The performance analysis of classification result is done by using the accuracy value and macro-average Fmeasure.

The experiments show that random-walk give better performance than term frequency especially on tf.idf - rw.idf scheme.

Keywords : classification, term weighting, term frequency, random-walk,

1. Pendahuluan

1.1 Latar belakang masalah

Saat ini informasi artikel berbahasa Indonesia berbasiskan web semakin banyak jumlahnya. Hal ini menyebabkan diperlukannya suatu kategorisasi terhadap artikel tersebut untuk memudahkan pembaca dalam mencari topik berita yang mereka inginkan. Salah satu cara yang dapat dilakukan sebagai solusi untuk masalah ini adalah dengan menggunakan proses klasifikasi teks dalam *text mining.*

Informasi yang akan digali pada *text mining* memiliki struktur sembarang. Oleh karena itu, diperlukan proses pengubahan bentuk menjadi data yang terstruktur sesuai kebutuhannya untuk pengolahan lebih lanjut. Proses ini sering disebut *text preprocessing* yang terdiri dari tahap *case folding*, *tokenization*, dan *filtering*. Setelah data menjadi data terstruktur, lalu dilakukan proses *term weighting* untuk memberikan bobot pada setiap *term* yang ditemukan pada sekumpulan dokumen teks. Bobot ini menyatakan kepentingan/kontribusi *term* terhadap suatu dokumen.

Term frequency telah lama digunakan sebagai metode pembobotan *term* pada dokumen teks [5]. Dalam metode ini, tiap *term* diasumsikan memiliki nilai kepentingan yang sebanding dengan jumlah kemunculan *term* tersebut pada dokumen teks. Hal inilah yang menjadi kelemahan utama dari *term frequency* sehingga mengakibatkan *term* menjadi independen dan mengabaikan ketergantungan yang mungkin ada antar *term* dalam dokumen. Hal ini mungkin efektif untuk mengambil *term-term* yang relevan dalam konteks lokal, tapi tidak dalam konteks global dimana keberadaan suatu *term* berpengaruh pada teks secara keseluruhan.

Kelemahan metode *term frequency* diatas dapat diatasi dengan mengimplementasikan metode pembobotan *random-walk* [11]. Metode ini direpresentasikan dengan algoritma perankingan berbasis graf yang diterapkan dalam graf tekstual yang dapat mengintegrasikan keterhubungan antar *term* dan konteks sekitarnya. Pada perankingan graf tekstual, teks direpresentasikan menjadi sebuah graf. *Vertex/node* pada graf tekstual adalah unit teks yang akan diranking, yaitu berupa *term-term* dalam dokumen. *Edge/link* dalam graf menunjukkan keterhubungan yang bermakna antar *vertex/node*.

Penelitian akan dilakukan dengan menerapkan metode *random walk* dan *term frequency* pada beberapa dataset artikel berita berbahasa Indonesia. Ada dua skema yang akan digunakan, yaitu skema *tf - rw* sebagai skema pertama dan *tf.idf - rw.idf* sebagai skema kedua. Setelah itu dilakukan proses klasifikasi dokumen dengan menggunakan klasifier pada *tools* Weka. Analisa difokuskan pada pengaruh *random walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure.*

1.2 Perumusan masalah

Objek penelitian pada Tugas Akhir ini adalah implementasi metode *random-walk* dan *term frequency* untuk menghitung bobot term dalam suatu dokumen. Akan tetapi, tugas akhir ini diutamakan kepada analisis karakteristik *random-walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure*.

Adapun batasan masalah yang dipakai pada Tugas Akhir ini adalah :

1. Dataset yang digunakan adalah artikel berita berbahasa Indonesia yang diperoleh dari web, bersifat *offline* dan disimpan dalam *file* berekstensi .txt.
2. Tidak melakukan *stemming* pada tahap *text preprocessing* dengan pertimbangan bahwa penggunaan *stemming* tidak terlalu berpengaruh pada proses pembobotan.
3. Proses klasifikasi yang akan dilakukan pada dokumen dengan menggunakan klasifier pada *tools data mining* yaitu Weka.

1.3 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Mengimplementasikan *random-walk* dan *term frequency* untuk menghitung bobot *term* dalam suatu dokumen.
2. Melakukan klasifikasi dokumen terhadap dataset yang telah dilakukan pembobotan *random-walk* dan *term frequency*.
3. Menganalisis pengaruh *random-walk* dan *term frequency* terhadap performansi klasifier berdasarkan nilai *akurasi* dan *macro-average F-measure*. Analisis tersebut dilakukan pada skema *tf - rw* dan skema *tf.idf - rw.idf*.

1.4 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah dalam Tugas Akhir ini adalah :

1. Studi literatur
   Mencari referensi dan sumber-sumber yang berhubungan dengan permasalahan yang ada seperti *text mining*, *text classification*, *term weighting*, dll.
2. Pencarian dan pengumpulan data
   Data yang akan digunakan berupa artikel berita berbahasa Indonesia yang diambil dari web.
3. Analisis kebutuhan dan implementasi sistem
   Analisis kebutuhan dilakukan dengan pembuatan sistem kebutuhan perangkat lunak. Melakukan implementasi sistem dengan membangun perangkat lunak sesuai dengan perancangan yang telah dilakukan.
4. Pengujian sistem dan analisa hasil
   Pengujian dilakukan terhadap beberapa metode pembobotan *term* berdasarkan parameter yang telah didefinisikan di awal.
5. Pengambilan keputusan dan penyusunan Tugas Akhir
   Melaporkan semua yang telah dilakukan selama implementasi dan pengujian dalam penulisan Tugas Akhir ini.

5. Penutup

5.1 Kesimpulan

Dari hasil analisis dan pengujian pada bab sebelumnya dalam tugas akhir ini, maka didapatkan kesimpulan :

1. Pada skema *rw - tf* metode *random-walk* dapat meningkatkan performansi hasil klasifikasi pada klasifier SVM, RBFN, dan KNN. Sementara pada klasifier Naïve Bayes metode *random-walk* belum dapat memberikan hasil performansi yang lebih unggul dibandingkan dengan pembobotan *term frequency*.
2. Pada skema *rw.idf - tf.idf* metode *random-walk* memiliki performansi hasil klasifikasi berupa *akurasi* dan *macro-average F-measure* yang lebih baik dibandingkan dengan metode *term frequency* untuk klasifier Naive Bayes, SVM, RBF Network, dan KNN.
3. Berdasarkan nilai *akurasi* dan *macro-average F-measure*, klasifier SVM, RBF Network, dan KNN dapat meningkatkan performansinya jika pembobotan *term* dilakukan dengan metode *random-walk*.
4. Metode *random-walk* dapat mengatasi kelemahan *term frequency* yang tidak dapat mengambil *term-term* yang relevan dalam konteks global dimana keberadaan suatu *term* berpengaruh pada teks secara keseluruhan khususnya pada skema *tf.idf – rw.idf*.

5.2 Saran

1. Perlu melakukan perbandingan antara penggunaan *stemming* dan non-*stemming* pada tahap *text preprocessing* untuk mengetahui apakah penggunaan *stemming* tersebut dapat mengubah performansi hasil klasifikasi.
2. Perlu melakukan perbandingan hasil klasifikasi dengan klasifier lainnya untuk pembobotan *random walk* dan *term frequency*.

Daftar Pustaka

[1] Adiwijaya, Igg. (2006). *Text Mining dan Knowledge Discovery*. Komunitas *Data mining* Indonesia & Soft-omputing Indonesia.

[2] F. Debole and F. Sebastiani. *Supervised Term Weighting for Automated text categorization*. In SAC '03: Proceedings of the 2003 ACM symposium on Apllied computing, pages 784-788, New York, NY, USA, 2003. ACM Press

[3] Francois Fouss, Alain Pirotte, Jean-Michel Renders, and Marco Saerens. 2007. *Random-Walk Computation of Similarities between Nodes of a Graph with Application to Collaborative Recommendation*. IEEE Transaction on Knowledge ang Data Engineering.

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

[5] M. Lan, C. Tan, H. Low, and S. Sungy. A comprehensive comparative study on term weighting schemes for text categorization with support vector machines. In Proceedings of the 14th international conference on World Wide Web, pages 1032–1033, 2005.

[6] Mihalcea, Rada and Paul Tarau. 2004. *TextRank: Bringing Order into Texts*. Department of Computer Science and Engineering University of North Texas USA.

[7] Nikolaos Nanas, Victoria Uren, and Anne De Roeck. *A Comparative Study of Term Weighting Methods for Information Filtering*. Department of Computing ang Mathematics Milton Keynes, U.K.

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

[9] S. Brin and L. Page. 1998. *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Computer Networks and ISDN Systems, 30(1-7).

[10] Salton, G. (1989). Automatic Text Processing: *The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-wesley, Reading, Pennsylvania.

[11] Samer Hassan, Rada Mihalcea and Carmen Banea. 2004. *Random-Walk Term Weighting for Improved Text Classification*. Department of Computer Science University of North Texas.

[12] Tan, Pang-ning, Michael Steinbach, dan Vipin Kumar. 2006. *Introduction to Data mining*. Pearson education, Inc.

[13] Tokunaga, Takenobu. *and* Iwayama, Makoto. 1994. *Text Categorization based on Weighted Inverse Document Frequency*. Department of Computer Science. Tokyo Institute of Technology. Japan.

[14] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[15] www.en.wikipedia.org/wiki/Naïve_Bayes_classifier [25 Oktober 2009]

[16] www.en.wikipedia.org/wiki/Radial_Based_Function_Network [25 Oktober 2009]