PEMBANGUNAN DAFTAR STOPWORD MENGGUNAKAN PENDEKATAN TERM-BASED RANDOM SAMPLING PADA KATEGORISASI DOKUMEN BERBAHASA INGGRIS

Anju Vikhers S S[1], Moch. Arif Bijaksana[2], Yanuar Firdaus A.w.[3]

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

Abstrak

Kata-kata dalam suatu dokumen yang sering muncul tapi kurang berarti dalam proses kategorisasi disebut sebagai stopword. Untuk kata-kata yang dikategorisasikan ke dalam stopwords dianggap tidak memiliki kontribusi dalam proses kategorisasi, seharusnya dihapus sewaktu pengindeksan sebelum proses kategorisasi dilakukan. Bagaimanapun, penggunaan satu daftar stopword untuk koleksi dokumen yang berbeda-beda bisa mengurangi performansi dari peng-kategorisasian

Pada tugas akhir ini digunakan pendekatan Term-Based Random Sampling menghasilkan daftar stopword secara otomatis untuk dokumen yang diberikan. Pendekatan ini, menentukan seberapa besar informasi yang dimiliki suatu kata (term). Dengan ini akan bisa ditentukan suatu daftar stopword secara otomatis. Dalam tugas akhir ini digunakan koleksi dokumen Reuter. Untuk daftar stopword yang dihasilkan akan dievaluasi dengan melakukan kategorisasi pada dokumen yang menggunakan daftar stopword yang dihasilkan.

Pendekatan ini juga nanti akan dievaluasi dengan membandingkan hasil performansi kategorisasi yang dihasilkan dengan pre-proses pembuangan daftar stopword menggunakan daftar stopword yang dihasilkan dengan menggunakan pendekatan ini, hasil performansi kategorisasi menggunakan daftar stopword Salton and Buckley I dan Salton and Buckley II, Google stopword, defalutl English Stopword dan hasil performansi kategorisasi tanpa menggunakan pre-proses pembuangan stopword.

Dari hasil evaluasi yang dilakukan, daftar stopword yang lebih efektif bisa diperoleh dengan menggunakan pendekatan Term-Based Random Sampling. Dengan akurasi pengkategorisasi sebesar 88.24%.

Kata Kunci : Kategorisasi, term, stopword

Abstract

Words in a document that frequently occuring but meaningless in categorization are called as stopwords. Words that categorize as stopwords do not contribute for categorization, they should be remove during indexing before categorization process. However, using a single fixed stopword list across diffrent document collection could be decrimental the performansi of categorization. On this final project, Term-Based Random sampling is used as approach to provide stopword list automatically for document collection that is processed. This approach, define how informative a term is. So, it's enable us to define a stopword list automatically. The stopword list that is produced will evaluate by categorization step for document that use this stopword list.

This approach will evaluate by comparing the result of categorization that produce by use preprocessing using stopword list that produce by this approach with the performans using stopword Salton and Buckley I, Salton and Buckley II, Google's stopword, default English Stopword, and also camparing with categorization's performansi without using stopword. From the evaluation, better stopword list is coming from stopword list that produce by using Term-Based Random Sampling Approach. The accuration is 88,24%.

Keywords : Categorization, term, stopword

1. Pendahuluan

1.1 Latar belakang

Adanya berita elektronik berbahasa Inggris merupakan salah satu penyajian informasi pada *digital library* yang mempunyai peranan sangat penting terutama dalam meningkatkan kegunaan dari sekumpulan dokumen. Dokumen-dokumen tersebut tentunya mengandung berbagai informasi yang berharga yang dapat dimanfaatkan oleh banyak pihak. Suatu dokumen merupakan sumber informasi yang memiliki nilai yang berharga. Sebenarnya secara eksplisit, bisa saja mengetahui informasi pada suatu dokumen dengan mudah, salah satu contohnya dengan melihat judul yang ada pada bagian dokumen tersebut. Namun dokumen-dokumen yang ada mempunyai nilai informasi yang bersifat implisit dan pada umumnya tidak pernah diperhatikan secara seksama. Sedangkan informasi inilah yang sebenarnya merupakan pengetahuan yang mempunyai informasi sangat berguna yang perlu kita peroleh.

Text mining merupakan bagian dari data mining yang memfokuskan bidangnya pada pengambilan informasi berharga dari basis data yang berupa *text*. Kategorisasi adalah salah satu teknik dari *text mining* yang bertujuan untuk menentukan topik dari suatu artikel atau *text document* berdasarkan atribut kelasnya, kategorisasi bersifat *supervised*. Untuk memperoleh hasil kategorisasi yang baik maka diperlukan suatu preproses *stopwords removal* yaitu penghapusan kata-kata yang sering muncul tapi tidak memiliki kontribusi terhadap informasi suatu dokumen.

Dengan terus bertambahnya jumlah dokumen pada *digital library*, pengkategorian secara manual tentu saja akan menjadi suatu masalah baru untuk pengguna. Hal tersebut akan memakan banyak waktu dan menimbulkan kejenuhan. Oleh karena itu, akan digunakan teknik *text mining* yaitu kategorisasi untuk mengkategorikan dokumen berbahasa Inggris. Sebelum proses kategorisasi dilakukan, perlu dilakukan proses *stopwords removal* terlebih dahulu pada dokumen. Maka diperlukan daftar *stopword* yang baik yang digunakan pada proses *stopword removal* ini, agar hasil pengkategorisasian yang dihasilkan bisa lebih baik.

Pada tugas akhir ini akan dibahas tentang pembangunan daftar stopword yang akan digunakan dalam proses *stopword removal* yang akan mempengaruhi hasil kategorisasi. Salah satu pendekatan yang digunakan untuk memperoleh daftar stopword adalah pendekatan *Term Based Random Sampling*. Pendekatan *Term-Based Random Sampling* ini bisa mengukur seberapa pentingnya suatu *term* dalam suatu dokumen yaitu dengan menghitung bobot. Bobot dari suatu *term* bisa dihitung dengan menggunakan *Kullback-Leibler Divergence Measure*. Semakin kecil bobot suatu *term* maka semakin cocok *term* tersebut disebut sebagai *stopword*. Dengan menggunakan pendekatan *Term-Based Random Sampling* ini maka bisa diperoleh suatu daftar *stopword* yang lebih efektif.

1.2 Perumusan masalah

Berdasarkan latar belakang, permasalahan yang dijadikan objek penelitian dan pengembangan tugas akhir ini adalah sebagai berikut :

1. Bagaimana menghasilkan daftar stopword pada dokumen yang digunakan dengan menggunakan pendekatan *Term-Based Random Sampling* untuk kategorisasi dokumen berbahasa Inggris.
2. Bagaimana mengaplikasikan pre-proses dengan *stopwords removal* dengan daftar *stopword* yang diperoleh dengan pendekatan *Term-Based Random Sampling* dan daftar *stopword* bahasa Inggris yang sudah didefenisikan.
3. Bagaimana mengukur dan menganalisis performansi kategorisasi setelah dilakukan *stopwords removal* dengan *Term-Basd Random Sampling* untuk kategorisasi dokumen berbahasa Inggris berdasarkan F-Measure..

1.3 Tujuan

Dalam tugas akhir ini, diharapkan tercapai hal-hal berikut :

1. Menghasilkan daftar *stopword* pada dokumen yang digunakan dengan menggunakan pendekatan *Term-Based Random Sampling* untuk kategorisasi dokumen berbahasa Inggris.
2. Mengaplikasikan pre-proses dengan *stopwords removal* dengan daftar *stopword* yang diperoleh dengan pendekatan Term-Based Random Sampling dan daftar *stopword* bahasa Inggris yang sudah didefenisikan.
3. Mengukur dan menganalisis performansi kategorisasi setelah dilakukan *stopwords removal* untuk kategorisasi dokumen berbahasa Inggris berdasarkan F-Measure.

1.4 Batasan masalah

Untuk memfokuskan penulisan tugas akhir ini, penulis mengambil studi kasus kategorisasi untuk berita berbahasa Inggris. Masalah yang akan dibahas memiliki batasan-batasan sebagai berikut:

1. Metode kategorisasi yang digunakan adalah Naïve Bayes dan J48 yang ada pada *Software* WEKA.
2. Dokumen yang digunakan adalah dokumen berbahasa Inggris yang bersifat offline yaitu dokumen Reuters dengan jumlah 200 dokumen.
3. Dokumen yang akan digunakan dilakukan proses *stemming* terlebih dahulu.

1.5 Metodologi penyelesaian masalah

Metodologi penyelesaian masalah yang digunakan pada Tugas Akhir ini adalah:

- **Studi Literatur**
  a. Pencarian referensi
     Mencari referensi yang layak dan berhubungan dengan topik tugas akhir, memahami dan memperlajari tentang *text mining, stopword, text categorization* dari berbagai jurnal, buku, Internet dan referensi lainnya yang mendukung.

- **Pengumpulan Data**
  Melakukan pengumpulan artikel berita yang berbentuk file txt yang akan digunakan sebagai data *collection*.

- **Perancangan perangkat lunak yang meliputi :**
  a. Implementasi perangkat lunak
     Mengimplementasikan perancangan menjadi perangkat lunak. Proses implementasi perangkat lunak dilakukan berdasarkan dari proses analisis dan perancangan yang telah dibangun.
  b. Pengujian
     Memeriksa error handling yang ada pada perangkat lunak yang dibangun misalnya kesalahan perhitungan, kesalahan dalam penginputan data, human error dan lain sebagainya.

- **Analisis hasil**
  Melakukan analisis terhadap hasil daftar *stopword*, kemudian melihat hasil kategorisasi dari dokumen atau artikel web berita dengan mengevaluasi hasil keakuratan pengkategorian dokumen. Proses pengukuran keakuratan dari hasil kategorisasi dengan menggunakan konfusion matrik yang ada pada tools WEKA.

- **Pengambilan Kesimpulan dan Pembuatan Laporan**
  Pembuatan laporan Tugas Akhir yang mendokumentasikan tahap-tahap kegiatan dan hasil dalam Tugas Akhir ini.

5. Kesimpulan dan Saran

5.1 Kesimpulan

1. Pendekatan term based random sampling mampu menghasilkan daftar *stopword* dapat meningkatkan hasil kategorisasi pada datamining berdasarkan hasil uji coba perhitungan F-Measure dan akurasi pada dataset Reuter dengan menggunakan parameter nilai Y, X dan L yaitu 25:40:400 untuk kategorisasi menggunakan Naïve Bayes.
2. Penambahan jumlah sampel dokumen, bisa membantu dalam menghasilkan daftar *stopword* yang cocok untuk kategorisasi.
3. Daftar *stopword* yang dihasilkan Term Based Random Sampling membantu meningkatkan performansi kategorisasi untuk algoritma-algoritma kategorisasi tertentu saja seperti Naïve Bayes.
4. Banyakanya jumlah *stopword* tidak menjamin hasil kategorisasi yang baik, tapi pemilihan term yang tepat untuk dimasukkan kedalam daftar *stopword* yang akan memberikan performansi kategorisasi yang lebih efektif.

5.2 Saran

1. Proses pembangunan daftar *stopword* dapat dikembangkan secara online, sehingga proses pembangunan daftar *stopword* dapat dilakukan secara otomatis.
2. Pembangunan daftar *stopword* bisa dikembangkan tidak hanya melihat tingkat kepentingan suatu kata hanya dari pembobotan dengan kullback leibler divergence measure tapi bisa juga dengan meneliti fungsi dari kata tersebut, apakah sama pentingnya informasi yang diberikan dari suatu kata benda dan kata kerja. Dan apakah harus membuang kata benda dan kata kerja yang sering muncul sebagai *stopwords*.
3. Penambahan jumlah sampel dokumen dalam pembangunan daftar *stopword* akan menghasilkan dokumen sampel yang lebih baik yang menggambarkan distribusi term-term yang ada dalam dokumen sehingga bisa menghasilkan daftar *stopword* yang lebih efektif.

6. Daftar Pustaka

[1] C. Fox. Lexical analysis and stoplists. In Information Retrieval - Data Structures & Algorithms, pages 102{130. Prentice-Hall, 1992.

[2] D. Hawking. Overview of the TREC2002. In Proceedings of the Nineth Text REtrieval Conference (TREC 9), pages 87-94, Gaithersburg, MD, 2000.

[3] E. M. Voorhees. Overview of TREC2002. In Proceedings of the Eleventh Text REtrieval conference (TREC2002), pages 1{16, Gaithersburg, MD, 2002.

[4] G. Amati and C. J. van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. ACM Transactions on Information Systems (TOIS), 20(4):357{389, 2002.

[5] Jiawei, Han. and Michelle, Kamber.(2001). *Data Mining: Concepts and Techniques*. San Francisco : Morgan Kaufmann Publisher.

[6] H. Zipf. Human Behaviours and the Principle of Least Effort. Addison-Wesley, Cambridge, MA, 1949.

[7] Effective Technique for Indonesian Text Retrieval. Jelita Asian B.comp. Sc.(Hons), School of Computer Science and Information Technology, Science, engineering, and Technology Portfolio, RMIT University, Melbourne, Victoria, Australia. 2007.

[8] Http://www.coffeecup.com/help/articles/default-english-stopwords/?PHPSESSID=188516aae0ea25b13fe018e2faf238ce

[9] Http://www.ranks.nl/resources/stopwords.html

[10] Rachel TszWai Lo, Ben He, Iadh Ounis. Automatically Building a Stopword List for an Information Retrieval System. Department of Computing Science University of Glasgow 17 Lilybank Gardens Glasgow, UK

[11] R. K. Belew. Finding Out About. Cambridge University Press, 2000.

[12] S. Charkrabarti. Mining the Web: Discovering knowledge from hypertext. Morgan Kaufmann, 2003.

[13] T. M. Cover and J. A. Thomas. Elements of Information Theory. John Wiley, 1991.

[14] W. B. Croft. Combining approaches to information retrieval. In Advances in Information Retrieval -Recent Research from the Center for Intelligent Information, pages 1{28. Kluwer Academic Publishers, 2000.

[15] W. Francis. Frequency Analysis of English Usage: Lexicon and Grammar. Houghton Mi²in, 1982.