ISSN 2355-3286

Klasifikasi Artikel Berita Online Sederhana dengan Menggunakan Struktur Kategori Wikipedia

William Aprilius
Program Studi Teknik Informatika, Universitas Multimedia Nusantara, Tangerang, Indonesia
william.aprilius@yahoo.com

Diterima 08 April 2014
Disetujui 16 Juni 2014

***Abstract*—The growth of information and communication technology makes the electronic news portal as a source of information. It makes the increasing numbers of online news articles that need to be classified. The classification is done to facilitate the users to access news. This paper proposes a simple method of classification of online news articles into categories. This method uses Wikipedia Bahasa Indonesia as a source of external knowledge and consists of 6 steps. In general, this method works by exploiting the structure of categories in Wikipedia, then check for the existence of entities of a news article in Wikipedia articles. This paper is an early stage of the research to be conducted and the proposed method has not been implemented. This makes the researchers have not been able to draw conclusions with regard to the method proposed.**

***Index Terms*—news classification, text classification, Wikipedia**

I. PENDAHULUAN

Informasi menjadi suatu hal yang dibutuhkan seiring dengan perkembangan teknologi informasi dan komunikasi. Salah satu sumber informasi tersebut adalah portal berita elektronik. Hal ini ditunjukkan dengan masuknya portal berita, seperti detik.com[1] dan kompas.com[2] dalam 15 *website* dengan lalu lintas tertinggi di Indonesia menurut Alexa[3].

Suatu portal berita elektronik mengklasifikasi artikel-artikel berita ke dalam kategori dan memiliki karakteristik untuk selalu bersifat *up-to-date*. Oleh karena itu, akan terdapat banyak dokumen elektronik yang perlu diklasifikasi setiap harinya [1]. Pengklasifikasian dilakukan untuk mempermudah pengguna mengakses dokumen tersebut. Dengan demikian, proses pengklasifikasian tersebut menjadi sulit untuk dilakukan dan memerlukan waktu.

Proses pengklasifikasian artikel berita dalam kategori telah menjadi pembahasan dalam beberapa penelitian seperti pada [1, 2]. Fokus utama dalam melakukan pengklasifikasian tersebut adalah bagaimana mengimplementasi suatu algoritma atau metode tertentu yang menghasilkan klasifikasi yang tepat dan memiliki performa yang baik.

Walaupun telah banyak algoritma yang terbukti dapat melakukan pengklasifikasian dokumen teks, dalam pelaksanaannya masih ditemukan beberapa masalah. Hal ini menjadikan metode yang diajukan bersifat terlalu kompleks jika ingin mengatasi seluruh masalah tersebut [2].

Tujuan dari penulisan *paper* ini adalah mengajukan metode pengklasifikasian artikel berita ke dalam suatu kategori. Metode yang diajukan menggunakan sumber pengetahuan eksternal untuk mengklasifikasi dokumen teks, yaitu Wikipedia. Hal yang ingin dicapai adalah suatu metode yang memiliki presisi yang baik dan tidak bersifat kompleks (*simplicity*).

Bagian selanjutnya dari *paper* ini diorganisasi sebagai berikut. Bab II menjelaskan penelitian terkait pengklasifikasian dokumen teks dan hal-hal terkait lainnya, seperti ekstraksi topik, *keyphrase* dan *headlines*. Bab III menjelaskan secara singkat mengenai jenis pengaturan klasifikasi dokumen. Bab IV menjelaskan mengenai metode pengklasifikasian yang diajukan. Simpulan dan saran pengembangan disajikan pada bab V.

II. KLASIFIKASI DOKUMEN

Pengklasifikasian dokumen teks merupakan cara mengelompokkan dokumen teks ke dalam beberapa kategori yang telah didefinisikan sebelumnya. Hal ini umumnya mengacu pada kategorisasi teks. Salah satu contoh dari kategorisasi teks adalah mengelompokkan artikel berita dalam kategori tertentu.

Terdapat tiga macam pengaturan kategorisasi teks, yaitu secara *binary*, *multi-class*, dan *multi-label* [3]. Pada pengaturan *binary*, dokumen teks hanya dikelompokkan menjadi tepat salah satu dari dua kategori. Sebagai contoh, dokumen teks hanya

1 http://www.detik.com
2 http://www.kompas.com
3 http://www.alexa.com/topsites/countries/ID diakses pada 26 Maret 2014

dikelompokkan apakah menyangkut hal mengenai olahraga atau tidak, atau mengenai teknologi atau tidak. Pada pengaturan *multi-class*, sebuah dokumen teks masuk dalam tepat salah satu kategori dari beberapa kategori. Pada pengaturan *multi-label*, sebuah dokumen dapat berada di dalam lebih dari satu kategori atau bahkan tidak dalam kategori manapun.

Pada *paper* ini, klasifikasi artikel berita dilakukan dengan pengaturan *multi-class*. Dengan demikian, setiap artikel berita akan berada dalam tepat sebuah kategori.

III. PENELITIAN TERKAIT

Proses pengklasifikasian artikel berita *online* pada [2] dilakukan dengan menggunakan suatu *framework*, bernama *Mutually Beneficial Learning* (MBL), yang mengintegrasi dua tahap pembelajaran. Tahap pertama adalah *clustering* untuk menemukan struktur lokal yang valid dan tahap kedua adalah tahap *classification*. Kedua tahap ini dilakukan secara iteratif sampai suatu kondisi terminasi dipenuhi. Hasil pengujian menunjukkan bahwa MBL secara signifikan lebih baik dari metode Naive Bayes dan SVM (*Support Vector Machines*).

Dalam *paper* yang sama, juga diajukan pengklasifikasian artikel berita *online* dengan memanfaatkan URL *pattern*. Namun, hal ini bersifat kurang cocok jika diterapkan pada artikel berbahasa Indonesia karena terdapat ketidakkonsistenan penggunaan bahasa pada beberapa *website* portal berita dalam menentukan URL *pattern*-nya. Sebagai contoh, detik.com yang menggunakan bahasa Inggris untuk kategori ekonomi (*finance*) sedangkan kompas.com menggunakan bahasa Indonesia untuk kategori yang sama (ekonomi).

Proses pengklasifikasian artikel berita *online* berbahasa Indonesia juga telah dilakukan seperti pada [1] dengan menggunakan metode Naive Bayes. Proses yang dilakukan juga meliputi *stemming* dan pembobotan kata. Hasil yang didapatkan adalah pengklasifikasian dengan persentase rata-rata *recall* dan presisi yang tinggi.

Suatu model ekstraksi topik dapat dimanfaatkan untuk mendukung hal-hal yang berkaitan dengan pengklasifikasian dokumen teks, seperti pada [3]. Hal ini dikarenakan, pada dasarnya proses pengklasifikasian juga melakukan ekstraksi informasi dari dokumen teks, sama seperti ekstraksi topik. Ekstraksi topik pada artikel berita *online* telah dapat dilakukan secara otomatis [4]. Selain itu, penelitian dengan pembahasan ekstraksi topik juga terdapat pada [5, 6].

Pada rujukan [4], ekstraksi topik dilakukan dengan menggunakan suatu teknik *Topic Detection and Tracking* (TDT) dan teori *aging*. Secara singkat, sistem bekerja dengan mengekstrak kata kunci dari sebuah artikel berita, memilah kata kunci tersebut dengan informasi topik dan menggabungkan kata kunci terpilih menjadi frasa. Frasa dengan peringkat tertinggi terpilih sebagai *topic keyphrase*. Hasil pengujian menunjukkan sistem dapat bekerja efektif.

Pada rujukan [5], ekstraksi topik dilakukan dengan menggunakan teknik berbasis *graph* yang menghasilkan suatu himpunan kata, yang mana himpunan kata tersebut memiliki kecenderungan untuk muncul pada suatu *corpus*. Kemudian, himpunan kata tersebut digabungkan untuk menghasilkan suatu topik. Jangkauan keluasan topik diatur melalui suatu parameter, yang dapat dilakukan oleh pengguna melalui suatu antarmuka. Hal ini menjadikan pengguna tidak perlu mengetahui perhitungan matematis untuk mengatur nilai parameter.

Terdapat perbedaan antara teks pada halaman web dan teks pada dokumen tradisional, yaitu halaman web memiliki hierarki dan antarhalaman terdapat hubungan, sedangkan pada dokumen teks tradisional, masing-masing dokumen bersifat independen. Oleh karena itu, [6] mengajukan suatu teknik ekstraksi *keyphrase* untuk menyimpulkan isi dari halaman web dalam suatu hierarki topik.

TitleFinder [7] menerapkan metode untuk mengekstrak *headlines* dari suatu halaman web berita secara *unsupervised*. Pendekatan yang dibangun berbasis konten dan independen terhadap domain dan bahasa. Algoritma yang diajukan berdasarkan sebuah motivasi bahwa terdapat kemiripan yang besar antara kata-kata pada *headline* dan *title element*. Oleh karena itu, dilakukan perbandingan tingkat kesamaan antara bagian *title element* dan segmen teks pada artikel dengan menggunakan empat jenis *similarity*. Empat jenis *similarity* tersebut berbasis pada *cosine similarity* dan *overlap scoring similarity*.

Masih pada *paper* yang sama, hasil pengujian menunjukkan metode yang diterapkan memiliki kinerja yang efektif dan efisien. Namun, penggunaan TitleFinder [7] untuk mengekstrak *headline* suatu artikel berita, tidak dapat memberikan hasil yang optimal apabila teks pada *title element* tidak berhubungan dengan isi artikel berita.

Penggunaan sumber pengetahuan eksternal untuk mendukung proses pengklasifikasian dokumen teks telah menjadi fokus beberapa penelitian, seperti pada [8, 9, 10]. Pada [8, 9] metode yang diajukan memanfaatkan Wikipedia, sedangkan pada [10] memanfaatkan fitur WordNet. Kemudian, [11] juga telah melakukan pendekatan untuk memperbaiki pengklasifikasian teks di Wikipedia. Hal ini dilakukan untuk mendukung pengklasifikasian dokumen teks yang menggunakan Wikipedia sebagai sumber pengetahuan eksternal.

Pada rujukan [8], dokumen teks diklasifikasi ke dalam kategori geografi atau berdasarkan tempat yang berhubungan dengan isi dari dokumen teks tersebut. Hal ini dilakukan dengan memanfaatkan Wikipedia untuk meningkatkan presisi dari hasil klasifikasi.

Pada rujukan [9], Wikipedia digunakan untuk membangun tesaurus, yang kemudian digunakan untuk memberikan informasi semantik pada representasi dokumen. Hal ini dilakukan untuk mengatasi kekurangan pendekatan tradisional BOW (*Bag of Words*) dalam merepresentasikan dokumen teks. Hasil pengujian menunjukkan bahwa terdapat perbaikan dalam akurasi hasil klasifikasi dokumen teks.

Dalam mengklasifikasikan artikel berita berbahasa Indonesia, fitur WordNet tidak dapat digunakan. Hal ini dikarenakan WordNet tidak menyediakan fitur untuk bahasa Indonesia. Selain itu, penggunaan fitur WordNet dalam pengklasifikasian dokumen teks tidak meningkatkan akurasi secara signifikan terhadap metode pengklasifikasian teks yang umum, seperti Naive Bayes dan SVM (*Support Vector Machines*) [10].

IV. METODE YANG DIAJUKAN

Metode pengklasifikasian artikel berita yang diajukan dalam *paper* ini menggunakan Wikipedia bahasa Indonesia sebagai sumber pengetahuan eksternal. Wikipedia merupakan ensiklopedia elektronik terbesar di dunia saat ini [9]. Selain itu, Wikipedia memiliki fitur-fitur yang menjadikannya suatu ontologi yang berpotensi untuk memperkaya representasi semantik suatu teks [12]. Hal ini menjadikan Wikipedia banyak digunakan untuk mendukung proses klasifikasi teks [8, 9] dan *text clustering* [12].

Metode yang diajukan terdiri dari 6 tahap, yaitu sebagai berikut.

A. Pengolahan Kategori dalam Wikipedia

Wikipedia telah memiliki struktur kategori yang dapat digunakan sebagai dasar kategori untuk mengklasifikasi artikel berita. Namun, pada metode ini, tidak seluruh struktur kategori dalam Wikipedia digunakan dalam proses pengklasifikasian. Hal ini karena tidak seluruh struktur kategori yang terdapat pada Wikipedia bersesuaian dengan kategori dasar yang ditetapkan untuk mengklasifikasikan artikel berita.

Setelah menentukan struktur kategori yang akan digunakan, dilakukan penyederhanaan halaman artikel Wikipedia yang berada dalam struktur tersebut. Pada artikel Wikipedia, terdapat beberapa bagian, seperti judul, paragraf, gambar, dan tabel. Pada metode ini, hanya bagian judul dan paragraf (teks dan *link*) yang akan digunakan. Hal ini dilakukan untuk menyederhanakan proses dan meningkatkan performa.

Berdasarkan penjelasan tersebut, diketahui bahwa tahap ini terdiri dari 2 subtahap, yaitu membentuk struktur kategori berdasarkan kategori pada Wikipedia dan menyederhanakan halaman artikel Wikipedia. Subtahap pertama dilakukan secara *supervised* dan sifatnya adalah reduksi struktur kategori, sehingga dihasilkan struktur kategori yang sesuai dan sederhana. Subtahap kedua dilakukan secara *unsupervised* dengan mengolah *tag-tag* HTML atau XML pada dokumen artikel Wikipedia.

B. Ekstraksi Entitas

Ekstraksi entitas dilakukan pada artikel berita yang akan diklasifikasi. Entitas merupakan representasi dari kata penting dalam artikel berita tersebut. Hal ini dilakukan dengan cara sebagai berikut.

- Membuang kata kerja dan kata sifat.
- Melakukan pembobotan *tf (term frequency)* untuk kata-kata yang tersisa dalam artikel berita.
- Memilih *n* kata atau frasa dengan bobot tertinggi. Kata atau frasa tersebut merupakan entitas.

Tantangan utama dalam tahap ini adalah bagaimana mengekstrak entitas yang tidak menghilangkan makna frasa. Dengan demikian, proses yang dilakukan tidak sekadar mengolah kata per kata.

C. Pencarian Artikel Wikipedia yang Bersesuaian

Setelah mendapatkan daftar entitas dari artikel berita, tahap selanjutnya adalah menemukan artikel Wikipedia yang bersesuaian dengan setiap entitas. Artikel Wikipedia disebut bersesuaian jika memiliki judul yang bersesuaian dengan entitas, baik bersesuaian secara tepat (*exactly match*) maupun bersesuaian secara parsial (*partialy match*).

D. Evaluasi Nilai Kemiripan

Seluruh artikel Wikipedia yang bersesuaian, yang dihasilkan dari tahap sebelumnya, akan dievaluasi untuk mengukur nilai kemiripannya dengan artikel berita yang akan diklasifikasi. Hal ini dilakukan dengan mengadopsi metode yang dilakukan TitleFinder [7] untuk menentukan *headline*. Namun, pengukuran dilakukan hanya dengan menggunakan *cosine similarity* berdasarkan skema pembobotan *tf-idf*. Adapun yang menjadi *query* adalah daftar dari entitas. Kemudian, pengukuran nilai kemiripan dilakukan antara *query* dan isi dari artikel Wikipedia yang bersesuaian. Dengan demikian, setiap artikel Wikipedia tersebut akan memiliki nilai kemiripan dengan entitas dari artikel berita.

E. Penentuan Kandidat Kategori

Kandidat kategori ditentukan dengan menghitung jumlah nilai kemiripan dari setiap artikel Wikipedia yang bersesuaian dengan entitas dalam suatu kategori. Makin besar jumlah nilai kemiripan artikel Wikipedia dalam suatu kategori, makin besar peluang bahwa suatu artikel berita masuk dalam kategori tersebut. Persamaan (1) berikut digunakan untuk menghitung peluang dari suatu artikel berita untuk masuk dalam

kategori ke-j, yaitu p_j, dimana $\sum_{i=1}^{N} d_i$ adalah jumlah nilai kemiripan dari artikel Wikipedia yang bersesuaian ke-i dalam kategori ke-j, *N* adalah jumlah artikel Wikipedia yang bersesuaian dalam kategori ke-j, dan *S* adalah jumlah nilai kemiripan dari seluruh artikel Wikipedia yang bersesuaian dengan entitas.

$$p_j = \frac{\sum_{i=1}^{N} d_i}{S} \quad (1)$$

F. Pemilihan Kategori

Pada tahap ini, peluang suatu artikel berita untuk masuk dalam suatu kategori (p_j) digunakan untuk menentukan kategori dari artikel berita tersebut. Kategori dengan peluang terbesar akan dipilih menjadi kategori dari artikel berita.

V. SIMPULAN

Metode yang diajukan menggunakan Wikipedia bahasa Indonesia sebagai sumber pengetahuan eksternal untuk mengklasifikasi artikel berita. Wikipedia digunakan karena dapat memperkaya representasi semantik dari suatu teks.

Metode yang diajukan terdiri dari 6 tahap, yaitu pengolahan kategori dalam Wikipedia, ekstraksi entitas dari artikel berita, pencarian artikel Wikipedia yang bersesuaian, evaluasi nilai kemiripan, penentuan kandidat kategori, dan pemilihan kategori. Secara singkat, metode ini bekerja dengan memanfaatkan struktur kategori yang telah dimiliki oleh Wikipedia, kemudian memeriksa keberadaan entitas artikel berita dalam artikel Wikipedia yang terdapat dalam struktur kategori tersebut.

Penulisan *paper* ini merupakan tahap awal dari penelitian yang akan dilakukan dan metode yang diajukan belum diimplementasi. Hal ini menjadikan peneliti belum dapat menarik simpulan berkaitan dengan metode yang diajukan. Namun, peneliti berhipotesis bahwa metode yang diajukan memiliki kelebihan, yaitu sifatnya yang sederhana (*simple*) dalam implementasi dan pengklasifikasian yang dapat dilakukan secara otomatis, serta tidak membutuhkan tahap pembelajaran. Akan tetapi, peneliti juga berhipotesis bahwa metode ini sangat bergantung pada struktur kategori yang dimiliki oleh Wikipedia.

Daftar Pustaka

[1] Ami D. Asy'arie dan Adi W. Pribadi, "Automatic News Articles Classification In Indonesian Language By Using Naive Bayes Classifier Method", di dalam *iiWAS '09 Proceedings of the 11th International Conference on Information Integration and Web-based Applications & Services*, New York, USA, 2009, hal. 658-662.

[2] Lei Wu, Zhiwei Li, Mingjing Li, Wei-Ying Ma, dan Nenghai Yu, "Mutualy Beneficial Learning with Application to On-line News Classification", di dalam *PIKM '07 Proceedings of the ACM first Ph.D. workshop in CIKM*, New York, USA, 2007, hal. 85-92.

[3] Davide Magatti, Fabio Stella, dan Marco Faini, "A Software System for Topic Extraction and Document Classification", di dalam *WI-IAT '09 Proceedings of the 2009 IEEE/WIC/ ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, Washington, USA, 2009, hal. 283-286.

[4] Canhui Wang, Min Zhang, Liyun Ru, dan Shaoping Ma, "An Automatic Online News Topic Keyphrase Extraction System", di dalam *WI-IAT '08 Proceedings of the 2008 IEEE/ WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, Washington, USA, 2008, hal. 214-219.

[5] Ajitesh Srivastava, Axel J. Soto, dan Evangelos Milios, "A Graph-based Topic Extraction Method Enabling Simple Interactive Customization", di dalam *DocEng '13 Proceedings of the 2013 ACM symposium on Document engineering*, New York, USA, 2013, hal. 71-80.

[6] Nan Liu dan Christopher C. Yang, "Keyphrase Extraction for Labeling A Website Topic Hierarchy", di dalam *ICEC '09 Proceedings of the 11th International Conference on Electronic Commerce*, New York, USA, 2009, hal. 81-88.

[7] Hadi Mohammadzadeh, Thomas Gottron, Franz Schweiggert, dan Gerhard Heyer, "TitleFinder: Extracting the Headline of News Web Pages based on Cosine Similarity and Overlap Scoring Similarity", di dalam *WIDM '12 Proceedings of the twelfth International Workshop on Web Information and Data Management*, New York, USA, 2012, hal. 65-72.

[8] Rafael Odon de Alencar, Clodoveau Augusto Davis, Jr., dan Marcos André Gonçalves, "Geographical Classification of Documents using Evidence from Wikipedia", di dalam *GIR '10 Proceedings of the 6th Workshop on Geographic Information Retrieval*, New York, USA, 2010, Article No. 12.

[9] Pu Wang dan Carlotta Domeniconi, "Building Semantic Kernels for Text Classification using Wikipedia", di dalam *KDD '08 Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, USA, 2008, hal. 713-721.

[10] Trevor Mansuy dan Robert J. Hilderman, "A Characterization of WordNet Features in Boolean Models for Text Classification", di dalam *AusDM '06 Proceedings of the fifth Australasian conference on Data mining and analystics - Volume 61*, Darlinghurst, Australia, 2006, hal. 103-109.

[11] Zeno Gantner dan Lars Schmidt-Thieme, "Automatic Content-based Categorization of Wikipedia Articles", di dalam *People's Web '09 Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources*, Stroudsburg, USA, 2009, hal. 32-37.

[12] Xiaohua Hu, Xiaodan Zhang, Caimei Lu, E. K. Park, dan Xiaohua Zhou, "Exploiting Wikipedia as External Knowledge for Document Clustering", di dalam *KDD '09 Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, USA, 2009, hal. 389-396.