Pengkategorian Berita Online Secara Otomatis Menggunakan Metode PLSA

Nur Cahyo Wibowo [1], Dhian Satria Yudha Kartika [2], Septiyawan Rosetya Wardhana [3]

Program Studi Sistem Informasi Fakultas Ilmu Komputer, UPN Veteran Jawa Timur [1,2]
Program Studi Teknik Informatika Fakultas Teknologi Industri, Institut Teknologi Adhi Tama Surabaya[3]
agen2009@gmail.com

Abstrak

Pengguna internet di Indonesia berkembang pesat mampu mengubah gaya hidup masyarakat. Masyarakat memanfaatkan internet untuk mengakses informasi pada berita online. Berita online yang beragam jenis dan kategori membuat penyedia layanan berita online harus menyediakan informasi sesuai dengan permintaan pengguna. Pada paper ini menjelaskan proses pengkategorian berita online secara otomatis. Tujuan pengkategorian secara otomatis untuk mempermudah penyedia layanan dalam membuat sebuah berita. Proses pengkategorian berita secara otomatis menggunakan metode probabilistyc laten semantic analysis (PLSA). Dari 60 dokumen berita sebagai dataset yang diambil pada tiga berita online: kompas, sidomi, liputan6 menghasilkan 4 kelompok kategori (labels) berita. Sehingga dapat disimpulkan dengan metode PLSA mampu menghasilkan nilai presisi tertinggi sebesar 0,68 dengan iterasi sebanyak 300 kali.

***Kata kunci** : kategori berita, PLSA, otomatisasi kategori*

Abstracts

Internet users in Indonesia are rapidly grow that enable to change people lifestyles. People use the internet to access online news information. The various type and category of online news force the service providers to provide the information according to user demand. This paper describes the process of categorizing online news automatically. The purposes of automatic categorization are to ease the service provider in creating news. The automatic categorization process used probabilistyc lateness semantic analysis (PLSA) method. There are 60 news documents as a dataset that is taken from three online news: kompas, sidomi, and liputan6 which then generate 4 groups of category (label) news. It can be conclude that the PLSA method is able to produce the highest precision value of 0.68 with iteration as much as 300 times.

***Keywords**: news category, PLSA, categorical automation*

1. Pendahuluan

Perkembangan internet di dunia sangat pesat seiring kemajuan teknologi dan kebutuhan bagi masyarakat. Pada tahun 2014 kementerian komunikasi dan informatika republik indonesia menyebutkan pengguna inernet oleh penduduk indonesia sekitar 8,2 juta. Survey tersebut diperoleh data, mereka menggunakan internet untuk berkomunikasi dan mengakses informasi. Dibandingkan masyarakat dunia, Indonesia menduduki peringkat ke-8 dimana masyarakatnya menggunakan internet untuk aktivitas sehari-hari. Dan dari 80% pengguna internet masyarakat Indonesia memanfaatkannya untuk mengakses berita online [1].

Dalam penelitian yang lain, Rani menyebutkan pada tahun yang sama rata-rata masyarakat Indonesia menghabiskan waktu untuk mengakses internet selama 3 jama setiap harinya. Dan internet menjadikan media tradisional seolah-olah menjadikan pesaing baru dalam mendistribusikan berita dan informasi [2].

Perkembangan dan kemajuan dalam dunia informasi mampu dimanfaat oleh sebagian pendidik untuk bisa mendukung proses pendidikan. Perkembangan teknologi informasi diharapakan bisa meningkatkan mutu pendidikan. Misalnya seorang pengguna akan mendapatkan informasi diseluruh dunia. Termasuk mengakses informasi dan menggunakan buku pada perpustakaan di negara-negara seluruh dunia [3].

Potensi penggunaan internet mengubah gaya hidup masyarakat untuk mengakses segala informasi khususnya berita online. Selain mendukung pendidikan, banyak penyedia layanan informasi memanfaatkan internet untuk menyajikan berita. Pergeseran ini seiring kebutuhan masyarakat yang juga

Received February 8, 2018; Revised April 27, 2018; Accepted April 29, 2018

bergeser dari media tradisional ke media digital memanfaatkan internet. Dalam penelitian yang lain menyebutkan selamat datang di era digital [4].

Beberapa layanan berita yang bisa diakses misalnya http://kompas.com, http://liputan6.com, http://republika.co.id, http://detik.com dan masih banyak lagi portal yang bisa diakses. Beragam informasi yang disajikan oleh layanan berita menjadikan informasi sangat beragam. Beragamnya informasi yang ada di internet dimanfaatkan oleh penyedia layanan informasi untuk bisa mengelompokkan informasi berdasarkan kategori. Hal ini bertujuan untuk mempermudah pembaca mendapatkan banyak informasi. Kategori berita digunakan untuk mengelompokkan informasi berdasarkan jenis berita. Politik, kesehatan, olah raga, wanita, *life style*, *entertainment* dll. Sejalan dengan penelitian sebelumnya bahwa pengelompokkan mempermudah pencarian informasi berdasarkan kejadian atau topik berita [5].

Setiap berita mempunyai judul berita, isi berita, penulis, kata kunci, berita terkait dan informasi yang melekat pada berita. Beberapa penelitian sebelumnya memanfaatkan bagian-bagian berita untuk menghitung tingkat kesamaan (*similarity*) sebuah berita dari beberapa berita online dari media online yang lain [6]. Pada penelitian yang lain menyebutkan cara mengoptimalkan bagian-bagian berita untuk membangun relasi antara berita terkait atau bisa disebut *proximity* [7]. Penelitian pada berita online semakin berkembang bersamaan dengan banyaknya metode yang ditemukan untuk menggali kembali informasi yang dimiliki.

Pada penelitian sebelumnya, beberapa metode yang digunakan untuk mengelompokkan berita diantaranya algoritma suffix tree clustering untuk pengelompokkan dokumen berita yang memungkinkan untuk sharing topik berita antar dokumen [5], Implementasi algoritma Fuzzy C-Means untuk pengelompokkan berita pada twitter [8], Pengelompokkan Berita Indonesia Berdasarkan Histogram Kata Menggunakan *Self-Organizing Map* [9] dan proses clustering artikel berita bahasa indonesia menggunakan *unsupervised feature selection* [10].

Proses pengkategorian berita selama ini masih bersifat manual dibuat oleh penyedia layanan berita. Setelah berita selesai ditulis, baru ditambahkan *author* dan bagian tag (relasi) berita. Belum menggunakan pengkategorian secara otomatis. Proses pengkategorian secara manual sangat merepotkan dan membuat berita baru membutuhkan waktu cukup lama dibandingkan secara otomatis.

Pada penelitian ini akan dilakukan proses pengkategorian otomatis memanfaatkan berita online. Proses pengkategorian berita online dimulai dengan menggabungkan kata-kata yang mempunyai bobot lebih pada judul dan body (isi) berita. Proses setelah mengetahui bobot pada judul dan isi akan dilakukan penggabungan beberapa berita untuk membentuk berita baru sehingga terbentuk kategori baru.

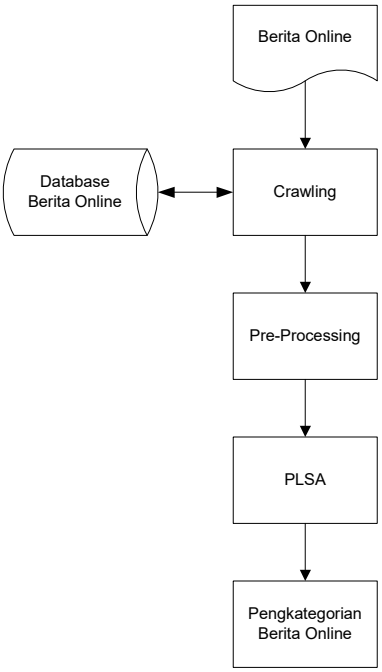
Dalam penelitian ini diusulkan proses pengkategorian secara otomatis pada berita menggunakan metode *probabilystic laten semantic analysis* (PLSA). Proses pengkategorian berita menggunakan PLSA pada penelitian ini digunakan untuk mengelompokkan (*cluster*) berita. Proses pengkategorian berita ini bertujuan mempermudah penyedia layanan berita membentuk berita baru. Sehingga kategori yang dihasilkan lebih beragam dan proses lebih cepat.

Hasil dari penelitian yang dilakukan akan terbit sebuah portal berita baru berdasarkan topik berita. Metode PLSA menjadi pelengkap dari metode sebelumnya yang perlah diusulkan. Portal berita yang sudah melalui proses penelitian akan dilihat nilai kesesuaiannya, menggunakan *precission, recall* sehingga mampu menunjukkan tingkat akurasi.

2. Metode Penelitian

Dalam penelitian ini dilakukan pengambilan dataset dari 3 berita online diantaranya http://sidomi.com, http://liputan6.com dan http://kompas.com. Berita online yang berhasil diambil sebanyak 60 dokumen berita yang akan dijadikan dataset.

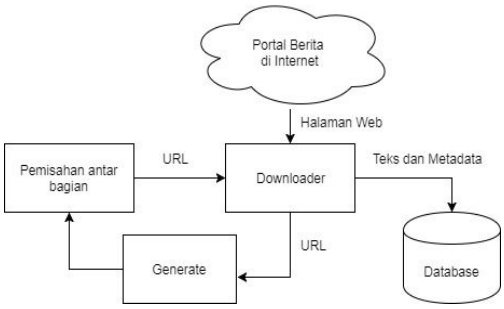
Masing-masing portal berita online mempunyai topik terdiri dari berita politik, berita hiburan, berita olah raga dll. Metode penelitian yang diusulkan seperti pada Gambar 1 berikut ini.



Gambar 1. Metode Penelitian

2.1 Crawling

Proses pengambilan dataset pada beberapa portal berita online dengan cara *crawling*. *Crawling* adalah proses algoritma yang mampu mengambil informasi pada website untuk disimpan pada database. Proses *crawling* akan mengambil seluruh informasi pada website mulai dari url, judul, penulis, waktu penulisan, isi (*body*) berita, berita terkait, gambar, *keyword* atau informasi lain yang ada pada website. Proses crawling seperti pada Gambar 2 berikut ini.

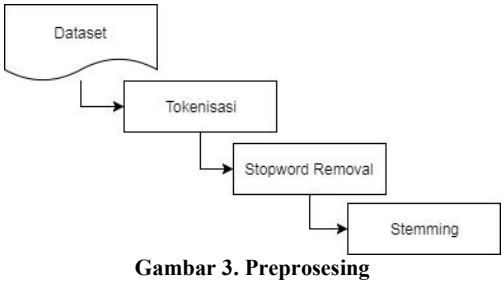


Gambar 2. Proses Crawling

2.2 Pre-Processing

Setelah proses *crawling*, didapatkan 60 berita dari beberapa portal berita online. Selanjutnya dataset dokumen berita online akan dilakukan *preprosesing*. *Preprosesing* sangat penting dilakukan sebelum tahap selanjutnya, berkaitan dengan normalisasi data.

Tahapan *preprocessing* yang pertama adalah tokenisasi. Tokenisasi bertujuan untuk menghilangkan spasi dan tanda baca pada dokumen. Langkah selanjutnya adalah melakukan *stopword removal* dengan menghilangkan kata (term) yang tidak memiliki nilai informasi. Contoh kata yang tidak mempunyai nilai informasi diantaranya yang, pada, suatu, ketika dll. Setelah didapatkan kata dari hasil *stopword removal*, kemudian dilakukan *stemming*.



Gambar 3. Preprosesing

Stemming bertujuan untuk mendapatkan kata dasar dari setiap kata. Dalam penelitian kali ini digunakan Sastrawi stemmer. Sastrawi dinilai mempunyai kualitas baik untuk proses *stemmer*, bisa diunduh pada http://sastrawi.github.io.

2.3 TF. IDF

Setelah dilakukan *preprosesing* dan dataset sudah mempunyai standar yang sama, proses selanjutnya adalah menghitung bobot kata pada masing-masing berita. Pembobotan kata yang digunakan menggunakan metode *Term Frequency* (TF). TF merupakan pembobotan dengan menghitung frekuensi kemunculan kata pada suatu dokumen [6]. Untuk setiap kata dalam dokumen akan dihitung bobot menggunakan persamaan 1. Kata setiap dokumen t_i akan dihitung dalam dokumen d_j kemudian dihitung nilai TF nya.

$$W_{TF}(t_i, d_j) = f\ (t_i, d_j) \tag{1}$$

Sedangkan IDF (*Inverse Document Frequency*) melakukan pendekatan dengan menganggap kata yang sering muncul pada satu dokumen, tapi jarang muncul pada seluruh data set akan diberikan nilai bobot yang lebih tinggi. Dalam sebuah *corpus* yang terdiri dari D dokumen terdapat $d_{(t_i)}$ dokumen yang mengandung kata i. Perhitungan IDF dari dokumen yang mengandung kata i dapat dilakukan dengan melihat persamaan

$$W_{IDF}(t_i) = 1 + log\left(\frac{D}{d_{(t_i)}}\right) \tag{2}$$

Perhitungan bobot TF.IDF dilakukan dengan melakukan perkalian antara persamaan 1 dengan 2 sehingga menghasilkan persamaan 3

$$W_{TF.IDF}(t_i, d_j) = f(t_i, d_j) \times \left(1 + log\left(\frac{D}{d_{(t_i)}}\right)\right) \tag{3}$$

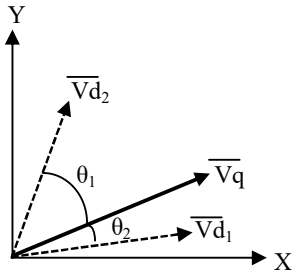
2.4 Similarity

Salah satu cara untuk mengetahui ukuran kesamaan teks adalah dengan *cosine similarity*. Ukuran ini menghitung nilai cosinus dari sudut antara dua vektor. Penggunaan *cosine similarity* pada *text matching* memiliki batasan sudut antara 0◦ dan 90◦. Hal ini dikarenakan pada text matching perhitungan kesamaan dokumen tidak dapat bernilai negatif.

Dimisalkan terdapat dua vektor dokumen d_1 dan d_2 vektor $query\ q$. *Cosine similarity* akan menghitung nilai θ dari setiap dokumen ke $query\ q$. Untuk setiap kata dalam dokumen yang memiliki bobot $W(t_i, d_j)$ dan setiap kata dalam query yang memiliki bobot $W(t_i, q)$ perhitungan *cosine similarity* dapat dilakukan dengan menerapkan rumus pada persamaan 4 berikut.

$$\cos(\theta) = \frac{d \cdot q}{\|d\| \times \|q\|} = \frac{\sum_{i=1}^{N} W(ti,q) \cdot W(ti,dj)}{\sqrt{\sum_{i=1}^{Nq} |Wq|^2} \cdot \sqrt{\sum_{i=1}^{N} |Wd|^2}} \tag{4}$$

Perhitungan similaritas antara dokumen dengan *query* menggunakan cosine similarity akan menghasilkan nilai antara 0 hingga 1. 0 menunjukan antara dokumen dengan *query* sama sekali tidak ada kesamaan, sedangkan 1 menunjukan bahwa dokumen dan query tersebut identik. Sedangkan untuk ilustrasi dari dasar *cosine similarity* dapat dilihat pada Gambar 4



Gambar 4. Representasi Cosine Similarity

2.5 PLSA

Topik modeling merupakan teknik yang dikembangkan untuk menghasilkan representasi dokumen berupa kata kunci kata kunci. Kata-kata kunci digunakan sebagai dalam proses pengindeksan serta pencarian dokumen untuk temu kembali sistem informasi [11].

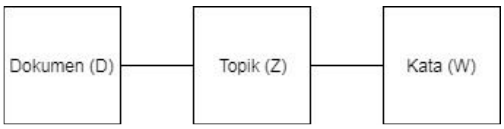
Probabilistic Latent Semantic Analysis (PLSA) muncul pada tahun 1999, menciptakan suatu teknik yang bernama *Latent Semantic Analysis* (LSA). LSA adalah sebuah teknik statistik terotomasi untuk menbandingkan kesamaan semantik dari beberapa kata atau beberapa dokumen.

PLSA melakukan klasifikasi dokumen teks. Dari sejumlah data training yang sudah disediakan, maka algoritma *Expectation Maximization* (EM) akan memproses data training. Proses training dilakukan sejumlah angka iterasi tertentu. Output algoritma EM merupakan model dari hasil training yang dilakukan PLSA. Proses yang sama akan dilakukan pada data testing.

Join probability pada PLSA antara dokumen (D) dan kata (W) bisa digambarkan pada persamaan berikut

$$P(d,w) = P(d)\,P(w|d).P(w|d) = \sum_{z\in Z} P(w|z)P(z|d)$$

Antara dokumen dan kata apabila ditarik garis maka diperoleh kata kunci (*keyword*). Kunci inilah yang disebut *aspect model*. *Aspect Model* didefinisikan sebagai sebuah variabel yang tidak terlihat *(latent variable)* dari sebuah dokumen. Variabel untuk memodelkan *aspect model* melibatkan asumsi *conditional independence (CI)*. Dokumen dan kata dalam satu kondisi *CI* dihubungkan dengan topik seperti pada Gambar 5 berikut ini.



Gambar 5. Hubungan antara Dokumen, Topik dan Kata

3. Hasil dan Pembahasan

Pada pengujian sistem ini dilakukan untuk mengetahui nilai perhitungan presisi dari hasil metode PLSA. Hasil perhitungan tingkat presisi metode PLSA ini diperoleh dari hasil pembagian antara jumlah hasil relevan yang dihasilkan oleh sistem dengan jumlah keseluruhan berita online yang terambil oleh sistem.

Berdasarkan hasil pengujian diatas menunjukkan bahwa semakin banyak iterasi yang dilakukan dalam metode PLSA maka terjadi peningkatan relevansi dokumen terhadap topik yang diujicobakan.

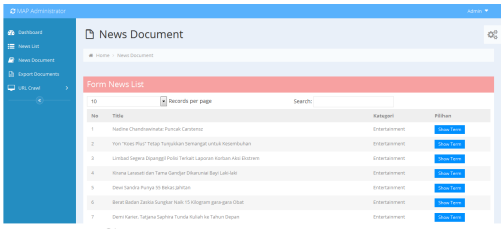
Namun pada beberapa kategori juga mengalami penurunan akibat penambahan iterasi. Hal tersebut dikarenakan tingkat keterkaitan term yang masih kurang dan tingkat keterkaitan antara dokumen dengan topik yang kecil.

3.1 Implementasi

Halaman *news list* admin digunakan untuk melakukan *crawling website* yang bersifat *list view*. Jadi dengan menggunakan fitur *news list* ini, keseluruhan berita online yang ditampilkan dalam list kategori akan terambil semua oleh sistem *crawling* yang sudah terpasang dalam sistem. Seperti pada Gambar 6.

Gambar 6. Halaman Crawling

Halaman export dokument admin digunakan untuk melakukan export dokumen judul berita dan dokumen isi berita yang telah melalui proses preprocessing ke dalam bentuk file txt. Tujuan dari penggunaan fitur ini adalah mempercepat proses analisa metode PLSA yang nantinya menggunakan Java. Seperti pada Gambar 7 berikut.

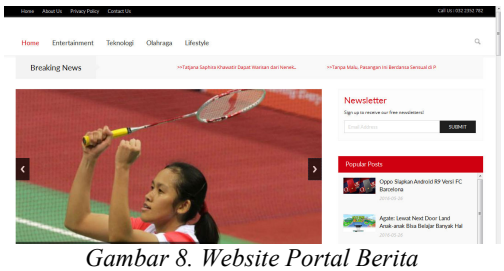


Gambar 7. Export Dokumen

Halaman URL Crawl pada sistem digunakan untuk melakukan *crawling* pada dokuemn berita yang terdapat pada website dengan sifat *single page*. Beberapa halaman website yang termasuk dalam URL *Crawl* ini adalah kompas dan liputan6. Seperti pada Gambar 8.

Gambar 8. URL Crawl

Antarmuka pengguna merupakan halaman website yang berbentuk portal berita. Halaman ini berfungsi untuk mempermudah pengguna dalam mengakses maupun membaca berita secara online. Website portal berita ini berisi halaman depan, halaman kategori, halaman pencarian dan halaman detail berita. Seperti pada Gambar 8.



Gambar 8. Website Portal Berita

3.2 Hasil Pengujian Sistem

Pada pengujian ini digunakan 60 dokumen berita online dan 4 topik meliputi *entertainment*, teknologi, olahraga, dan *lifestyle*. Hasil pengujian pada Tabel 1 berikut:

Tabel 1: Tabel Hasil Pengujian

Topik	Hasil Pengujian (Iterasi)		
	100x	200x	300x
Entertainment	0,43	0,62	0,68
Teknologi	0,51	0,54	0,46
Olah Raga	0,38	0,45	0,56
Lifestyle	0,46	0,56	0,63

4. Kesimpulan

Berdasarkan hasil penelitian yang sudah dilakukan, didapatkan beberapa hasil kesimpulan diantaranya :

1. Sistem portal berita yang memanfaatkan crawler sebagai fitur untuk mengumpulkan berita online secara otomatis berhasil dirancang dan dibangun.
2. Proses pengkategorian berita otomatis dengan menggunakan metode PLSA berhasil diterapkan dalam sistem ini.
3. Berdasarkan hasil pengujian yang telah dilakukan, nilai presisi tertinggi yang diperoleh adalah 0.68.

Ucapan Terima Kasih

Terima kasih kepada seluruh pimpinan dan rekan kerja di Fakultas Ilmu Komputer, khususnya program studi Sistem Informasi. Atas kesempatan yang telah diberikan untuk bisa mempublikasikan penelitian ini. Semoga sedikit ilmu ini bermanfaat dan bisa mengembangkan karya ilmiah kedepannya. Kepada rekan perjuangan sekaligus kolega, mas Septiyawan Rosetya di kampus Institut Teknologi Adi Buana (ITATS) atas perjuangan di Pascasarjana Teknik Informatika Intitut Teknologi Sepuluh Nopember (ITS) Surabaya. Sehingga penelitian ini bisa tuntas dan terpublikasi.

Daftar Pustaka

[1] Kominfo, [online], diakses di https://kominfo.go.id/index.php/content/detail/3980/Kemkominfo%3A+Pengguna+Internet+di+Indonesia+Capai+82+Juta/0/berita_satker [8 Nopember 2017]]

[2] Rani, O., & Lestari, D. QUALITY NEWS DAN POPULAR NEWS SEBAGAI TREND PEMBERITAAN MEDIA ONLINE (Studi Deskriptif Kualitatif Trend Pemberitaan Quality News dan Popular News pada Media Online Nasional di Indonesia Periode 2016), 83–94. (2017).

[3] Gafar, A. Jurnal Ilmiah Universitas Batanghari Jambi Vol.8 No. 2 Juli 2008 Penggunaan Internet Sebagai Media Baru dalam Pembelajaran Abdoel Gafar 1. *Jurnal Ilmiah Universitas Batanghari Jambi*, *8*(2), 36–43. (2008).

[4] Setiawan, W. Era Digital dan Tantangannya. *Seminar Nasional Pendidikan 2017*, 1–9. (2017).

[5] Arifin, A. Z., Darwanto, R., Navastara, D. A., & Ciptaningtyas, H. T. KLASIFIKASI ONLINE DOKUMEN BERITA DENGAN MENGGUNAKAN ALGORITMA SUFFIX TREE CLUSTERING. *Seminar Sistem Informasi Indonesia*. (2008).

[6] Salton, Gerard. Buckley, C. Term Weighting Approaches in Automatic Text Retrieval. (1988).

[7] Leal, P. Using proximity to compute semantic relatedness in RDF graphs. *ComSIS*, *10 no 4*, 1727–1746. (2013). https://doi.org/10.2298/CSIS121130060L

[8] Putri, E. N. (UIN S. G. J. IMPLEMENTASI ALGORITMA FUZZY C-MEANS UNTUK PENGELOMPOKKAN BERITA PADA TWITTER Oleh. *Skripsi*. (2017).

[9] Ambarwati. Winarko, edi (Ugm, F. Pengelompokan Berita Indonesia Berdasarkan Histogram Kata Menggunakan Self-Organizing Map, *8*(1). (2014).

[10] Langgeni, D. P., Baizal, Z. K. A., & W, Y. F. A. CLUSTERING ARTIKEL BERITA BERBAHASA INDONESIA, *2010*(semnasIF), 1–10. (2010).

[11] Suhartono, D. Probabilistic Latent Semantic Analysis (PLSA) untuk Klasifikasi Dokumen Teks Berbahasa Indonesia. *Technical Report Program Studi Doktor Ilmu Komputer Fakultas*. (2014).