ISSN 2089-6131 (Print)
ISSN 2443-1311 (Online)
DOI 10.22146/teknosains.86111
https://jurnal.ugm.ac.id/teknosains

JURNAL TEKNOSAINS

VOLUME 6 No. 2, 22 Juni 2017 Halaman 59-138

KLASIFIKASI KATEGORI DOKUMEN BERITA BERBAHASA INDONESIA DENGAN METODE KATEGORISASI MULTI-LABEL BERBASIS *DOMAIN-SPECIFIC ONTOLOGY*

Pangestu Widodo, Janur Adi Putra,* dan *Suwanto Afiadi
Jurusan Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember (ITS)
Email: pangestu.widodo15@mhs.if.its.ac.id

ABSTRACT

A news document often related to more than one category, necessary for utilization the method of categorization that is not only fast but also able to classify a news into many categories. Many methods can be used to categorize the news documents, one of which is an ontology. Ontology approach in the categorization of a document is based on the similarity of news features in documents with features that exist in the ontology. The use of ontologies in categorization that just based on the occurance of the term in calculating the relevance of the document, led to the emergence of many other features that are actually very relevant is undetectable. This paper proposed a new method for categorizing news documents are related with many categories, the method is based on a specific domain ontology and for document relevance calculation is not only based on the occurrence of the term but also take into account the relationships between terms that are formed. Tests performed on the Indonesian language news document with two categories: sports and technology. The trial results show the value of the average accuracy is high, that the sports category was 93,85% and the technology category is 96,32%.

Keywords: ***Categorization; Domain-specific; Multi-label; News document; Ontology.***

ABSTRAK

Sebuah dokumen berita seringkali terkait lebih dari satu kategori, untuk itu diperlukan pemanfaatan metode kategorisasi yang tidak hanya cepat tetapi juga dapat mengelompokkan sebuah berita kedalam banyak kategori. Banyak metode yang dapat digunakan untuk mengkategorisasi dokumen berita, salah satunya adalah ontologi. Pendekatan ontologi dalam kategorisasi sebuah dokumen berita didasarkan pada kemiripan fitur yang ada di dokumen dengan fitur yang ada di ontologi. Penggunaan ontologi dalam kategorisasi yang hanya didasarkan pada kemunculan term dalam menghitung relevansi dokumen menyebabkan banyak kemunculan fitur lain yang sebenarnya sangat terkait menjadi tidak terdeteksi. Dalam paper ini diusulkan metode baru untuk kategorisasi dokumen berita yang terkait dengan banyak kategori, metode ini berbasis domain spesifik ontologi yang perhitungan relevansi dokumen terhadap ontologinya tidak hanya didasarkan pada kemunculan term tetapi juga memperhitungkan relasi antar term yang terbentuk. Uji coba dilakukan pada dokumen berita berbahasa indonesia dengan 2 kategori yaitu olahraga dan teknologi. Hasil uji coba menunjukkan nilai rata-rata akurasi yang cukup tinggi yaitu kategori olahraga adalah 93,85% sedangkan pada kategori teknologi adalah 96,32%.

Kata Kunci: ***Dokumen berita; Domain-spesifik; Kategorisasi; Multi-label; Ontologi.***

PENGANTAR

Berita telah menjadi kebutuhan pokok manusia seiring dengan berkembangnya teknologi dan internet. Perkembangan teknologi dan internet ini menyebabkan proses pendistribusian informasi pada berita beralih dari cara penyampaian era media cetak menuju era digital. Berita yang disajikan dalam bentuk teks pada media digital, biasanya dikelompokkan berdasarkan isinya seperti berita olahraga, ekonomi, sains, dan lain sebagainya. Isi dari berita adalah berupa teks, sedangkan berita pada media digital tersebut disimpan dan dikelompokkan sesuai dengan kelompoknya. Permasalahan yang muncul adalah penggunaan media digital dalam penyampaian informasi menyebabkan jumlah berita digital yang dirilis oleh portal berita tiap harinya menjadi sangat banyak. Hal ini berdampak pada ketersediaan berita yang jumlahnya sangat melimpah. Berdasar uraian tersebut dibutuhkan metode pengorganisasian yang baik dan cepat untuk memudahkan pengambilan informasi dari berita yang melimpah tersebut. Informasi ini yang nantinya akan digunakan sebagai dasar dari pengelompokan berita yang ada. Teks merupakan data yang tidak berstruktur dan jika teks tidak diorganisasikan maka proses pengambilan informasi akan membutuhkan proses yang lama. Proses pengorganisasian teks tersebut adalah *Text Mining*. Salah satu kegunaan dari *text mining* adalah pengklasifikasian dan pengorganisasian dokumen berdasarkan isinya atau disebut *text categorization* (Ben-Dov dkk., 2001).

Berita memiliki aliran yang dinamis dimana informasi yang terkandung didalamnya memungkinkan sebuah informasi baru yang tidak ada dalam dokumen sebelumnya. Berita juga seringkali terkait dengan lebih dari satu kategori (*multi-label*). Multi-label selalu terkait dengan data yang ambigu, dimana setiap satu objek berita merupakan anggota dari sejumlah kelas kategori (*label*) yang berbeda. Hal tersebut tentunya menambah tingkat kesulitan dalam memprediksi kategori dari sebuah berita, untuk itu diperlukan metode *text categorization* yang juga dapat mengkategorikan sebuah dokumen berita kedalam banyak kategori yang sesuai atau biasa disebut *multi-label text categorization.*

Saat ini, metode atau algoritma yang sangat terkenal dalam *text categorization* adalah *machine learning* (Sebastiani dkk., 200). Beberapa algoritma *machine learning* yang dapat digunakan dalam *text categorization* diantaranya adalah algoritma K-Nearest Neighbor, Naïve Bayes Classifier, dan ID3. Implementasi dari algoritma-algoritma tersebut pada *text categorization* berdasar pada kemuculan kata atau morfologi kata. Konsep dari klasifikasi teks dengan menggunakan algoritma-algoritma tersebut adalah memasukkan teks baru yang belum diketahui kategorinya ke dalam kategori dengan melakukan pelatihan terhadap sekumpulan teks yang telah diketahui kategorinya. Proses pelatihan tersebut adalah menentukan kemiripan antara teks uji dengan setiap teks latih. Teks uji dan teks latih dikatakan mirip bila ada sekumpulan term yang muncul pada kedua dokumen tersebut. Term yang muncul tersebut adalah yang memiliki huruf penyusun yang sama. Semakin banyak term yang sama maka semakin mirip pula kedua teks tersebut. Proses penentuan kemiripan teks ini memiliki kelemahan karena apabila terdapat teks uji yang memiliki term yang berbeda dari term pada teks latih padahal kedua term tersebut memiliki makna yang sama maka kedua teks tersebut tidak dapat dikatakan mirip. Hal ini memungkinkan teks uji tersebut akan dikelompokkan ke dalam kategori yang berbeda dari kategori teks latih tersebut.

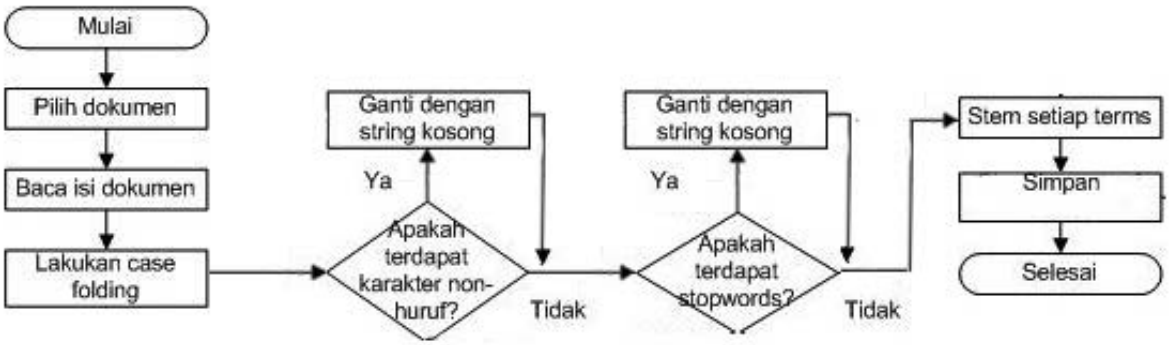
Alternatif algoritma *text-categorization* yang dapat mengatasi masalah tersebut adalah algoritma yang menggunakan pendekatan ontologi. Ontologi adalah deskripsi formal tentang suatu konsep secara eksplisit dalam sebuah *domain* dan properti dari setiap konsep beserta dengan batasannya (Noy dkk., 2001). Penggunaan pendekatan ontologi dalam kategorisasi sebuah dokumen berita berdasarkan kemiripan fitur yang ada di dokumen dengan fitur yang ada di ontologi. Pendekatan *knowledge engineering* disebut *rule base* karena pendekatan ini memanfaatkan keahlian

manusia (*human expert*) untuk membuat aturan-aturan (*rules*) secara manual melalui proses pemahaman pada sebuah *domain* penelitian (Milton, 2003). Penggunaan ontologi dalam kategorisasi yang hanya didasarkan pada kemunculan term dalam menghitung relevansi dokumen menyebabkan banyak kemunculan fitur lain yang sebenarnya sangat terkait menjadi tidak terdeteksi.

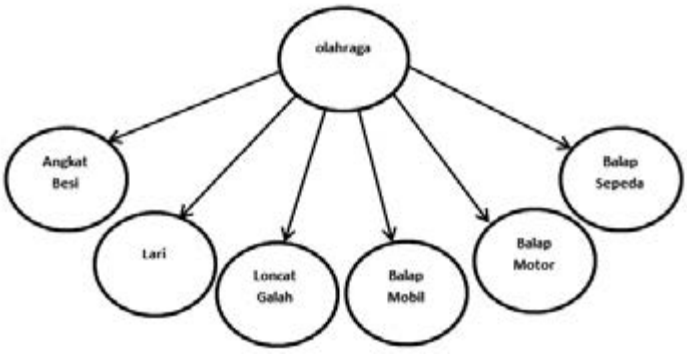
Berdasarkan permasalahan-permasalahan yang disebutkan, diusulkan sebuah metode baru untuk kategorisasi dokumen berita yang terkait dengan banyak kategori, metode ini berbasis domain spesifik ontologi yang perhitungan relevansi dokumen terhadap ontologinya tidak hanya didasarkan pada kemunculan term tetapi juga memperhitungkan relasi antarterm yang terbentuk. *Domain* spesifik ontologi dapat digunakan untuk melakukan kategorisasi dokumen berita dalam penelitian ini karena bersifat unik dan memiliki struktur hierarkis. Selain itu, sebuah model *domain* spesifik ontologi dapat menghilangkan makna ambigu, sehingga dapat menangani masalah yang muncul pada bahasa alami dimana sebuah kata memiliki lebih dari satu makna atau arti bergantung pada konteks kalimatnya. Metode usulan ini dalam pencarian kemunculan term ontologinya tidak lagi terpaku pada frase penuh yang tebentuk. Metode ini nantinya akan diuji untuk mengategorisasi dokumen berita Berbahasa Indonesia dengan kategori yang berbeda-beda.

Metode

Pada bab ini akan dibahas metode penelitian yang akan dilakukan mulai persiapan data, *preprocessing* data, metode usulan berupa diagram alur fase *training* dan diagram alur fase *testing* serta metode evaluasi yang digunakan.



Gambar 1
Tahapan persiapan dokumen berita (preprocessing)



Persiapan Dokumen Berita

Proses persiapan dokumen berita dilakukan terlebih dahulu sebelum dilakukan *text categorization*. Proses persiapan dokumen berita sebagaimana terlihat pada Gambar 1 meliputi proses *case folding*, *filtering*, pembuangan *stopwords* (*stopping*). Tujuan dari proses persiapan dokumen teks adalah untuk menghilangkan karakter-karakter selain huruf, menyeragamkan kata, mengurangi volume kosakata, dan mengubah kata kedalam bentuk aslinya.

Identifikasi Kata

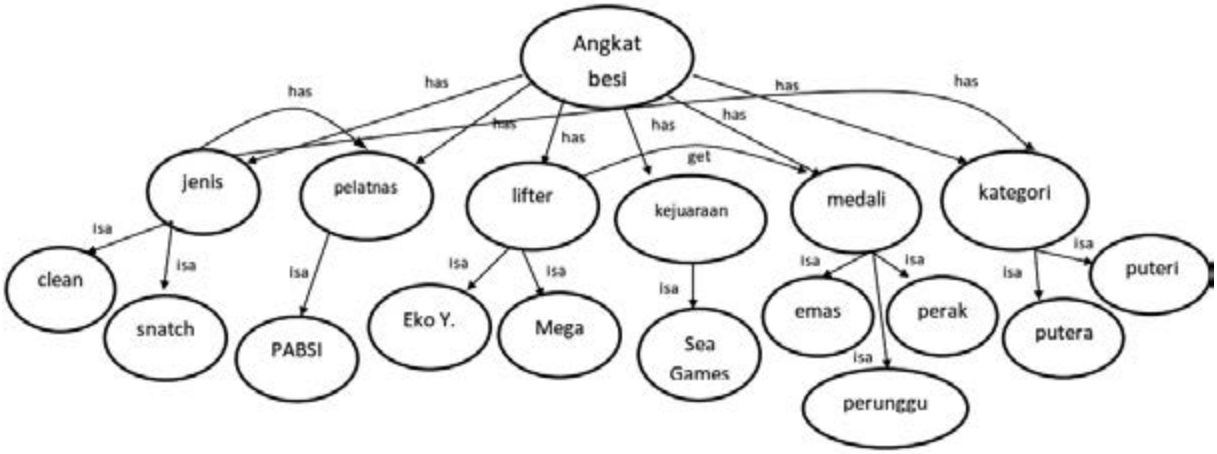
Proses identifikasi kata merupakan proses penghilangan angka, tanda baca dan konversi huruf kapital dan huruf kecil. Secara garis besar proses ini dapat dibagi menjadi dua proses, yaitu *filtering* dan *case folding*. Proses *filtering* merupakan proses yang berguna untuk menghilangkan karakter - karakter non-huruf seperti angka, tanda baca dan simbol, sedangkan dalam proses case folding variasi huruf harus diseragamkan (menjadi huruf kecil saja). Karakter selain huruf dihilangkan dan dianggap sebagai *delimiter* (Baeza-Yates dkk., 1999).

Stopping

Proses pembuangan *stopwords* merupakan proses yang dilakukan setelah proses identifikasi kata. *Stopwords* adalah kata-kata yang sering muncul dan tidak dipakai di dalam pemrosesan bahasa alami (Baeza-Yates dkk., 1999). *Stopwords* dapat berupa kata depan, kata penghubung, dan kata pengganti. Contoh *stopwords* dalam bahasa Indonesia adalah "yang", "ini", "dari", dan "di". Ukuran kata dalam sebuah dokumen teks menjadi berkurang setelah dilakukan proses pembuangan *stopwords* sehingga hanya kata-kata yang penting terdapat dalam sebuah dokumen teks dan diharapkan memiliki bobot yang tinggi.

Ontologi

Setelah proses persiapan dokumen hal selanjutnya adalah pembentukan *domain* spesifik ontologi. Ontologi adalah sebuah deskripsi formal tentang sebuah konsep secara eksplisit dalam sebuah *domain* dari setiap konsep beserta dengan batasannya (Noy dkk.,2001). Sebuah konsep di ontologi dapat memiliki objek (*instances*). Secara teknis, ontologi direpresentasikan dalam bentuk *class*, *property*, *facet*, dan *instances*. *Class* menerangkan konsep atau makna dari suatu *domain*. *Class* adalah kumpulan dari elemen dengan sifat yang sama. Sebuah *class* bisa memiliki *sub class* yang menerangkan konsep yang lebih spesifik. *Property* merepresentasikan hubungan diantara dua individu. *Property* menghubungkan individu dari *domain* tertentu dengan individu dari *range* tertentu. Ada tiga jenis *property*, yaitu *object property*, *data type property* dan *annotation property*. *Object property* menghubungkan suatu individu dengan individu lain. *Object property* terdiri dari empat tipe, yaitu *inverse property*, *functional property*, *transitive property*, dan *symmetric property*. *Data type property* menghubungkan sebuah individu ke sebuah tipe data pada *Resource Description Framework* (RDF) literal atau pada *Extensible Markup Language* (XML). *Annotation property* digunakan untuk menambah informasi (*metadata*) kelas, individu dan object/data *type property*. *Facet* digunakan untuk merepresentasikan informasi atau batasan tentang *property*. Ada dua jenis *facet*, yaitu *cardinality* dan *value type*. *Cardinality facet* merepresentasikan nilai eksak yang bisa digunakan untuk slot pada suatu kelas tertentu. *Cardinality facet* dapat bernilai *single* dan *multiple cardinality*. *Value type* menggambarkan tipe nilai yang dapat memenuhi *property*, seperti *string, number, boolean, dan enumerated*. Contoh pendekatan ontologi yang akan digunakan dalam penelitian ini terlihat pada Gambar 2 dan Gambar 3. Gambar 2 merupakan representasi ontologi dari kategori (*Class*) olahraga, sedangkan Gambar 3 merupakan representasi dari subclass (*domain*) dari Class olahraga yaitu angkat besi. Ontologi dapat digunakan untuk melakukan klasifikasi dokumen teks pada penelitian ini karena ontologi bersifat unik dan memiliki struktur hierarkis. Selain itu, sebuah model ontologi dapat menghilangkan makna ambigu sehingga dapat menanggulangi masalah yang



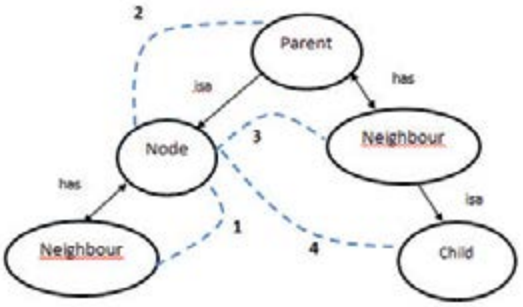
Gambar 3
Representasi domain angkat besi pada ontologi Olahraga

muncul pada bahasa alami di mana sebuah kata memiliki lebih dari satu makna atau arti bergantung pada konteks kalimatnya.

Konsep atau class merepresentasikan term atau kata dalam *domain* yang spesifik. Fitur (*instance*) atau dalam paper ini disebut dengan *node* merepresentasikan individu dari sebuah kelas. Relasi atau property merepresentasikan hubungan di antara konsep. Ada dua relasi yang digunakan dalam penelitian ini, yaitu relasi yang spesialisasi dan *non*-spesialisasi.

Metode Klasifikasi Usulan

Metode usulan yang diajukan tediri dari dua fase utama yakni fase *training* dan fase *testing*. Fase *training* atau fase pembelajaran merupakan fase yang digunakan untuk mencari nilai *threshold* yang optimal pada tiap *domain*nya. Nilai *threshold* yang optimal selanjutnya digunakan untuk mencari nilai relevansi suatu dokumen pada *domain* ontologinya atau pada kategorinya, sedangkan fase *testing* digunakan untuk mendapatkan nilai akurasi atau tingkat kebenaran dalam klasifikasi metode usulan ini. Dalam fase *training* semua dokumen berita akan dicocokkan dengan semua *domain* ontologi yang telah dibuat, hal ini digunakan untuk mencari nilai relevansi tiap dokumen terhadap semua *domain* ontologi.



Gambar 4
Kriteria *node* tetangga

Step awal yang dilakukan dalam fase *training* atau pembelajaran adalah menghitung nilai relevansi tiap dokumen terhadap semua *domain* dalam ontologi yang telah dibuat. Nilai relevansi dihitung dari kemunculan *node* dan jarak antar*node* yang berelasi yang muncul didalam dokumen. Dua buah *node* dihitung sebagai sebuah relasi apabila *node* tersebut memiliki relasi dengan *node* tetangga, sebuah node disebut mempunyai relasi dengan node tetangga apabila memenuhi salah satu kriteria seperti pada Gambar 4, kriteria tersebut sebagai berikut:

Relasi 1, relasi antara *node* dan *neighbour*
Relasi 2, relasi antara *node* dengan *parent*nya
Relasi 3, relasi antara *node* dengan *neighbour* pada *parent*nya

Relasi 4, relasi antara *node* dengan *child* dari *neighbour* pada *parent*nya

Jarak kemunculan *node* didapat dari selisih jarak kemunculan *node* pertama dengan *node* tetangga dimana *node* tersebut membentuk sebuah relasi baik itu relasi spesialisasi atau relasi non spesialisasi, kemudian jarak tersebut akan dikalikan dengan bobot (w). Nilai relevansi tiap *domain* ini disimpan dalam database sebagai acuan untuk mencari nilai *threshold* yang optimal.

Setelah didapat nilai relevansi ditiap *domain* maka langkah selanjutnya adalah mencari *threshold* yang optimal. *Threshold* optimal didapat dengan cara meniru algoritma *binary search* untuk mencari *threshold* dengan akurasi tertinggi dengan interval dari nilai relevansi minimum adalah 0 hingga nilai relevansi maksimum. Diagram alur fase *training* metode usulan ditunjukkan pada Gambar 5, sedangkan diagram alur pencarian *treshold* optimal ditunjukkan Gambar 6.

Gambar 5
Diagram alur fase training metode usulan

Ontologi direpresentasikan dengan kumpulan *node* yang saling terhubung dan membentuk sebuah jejaring. Untuk sebuah *node N*, himpunan kemunculan (*O*) *node* tersebut pada *domain* (Ω) dalam sebuah dokumen (*d*) dirumuskan pada persamaan 1.

$$O_{d\Omega}(N)=\{P_{d1}, P_{d2}, ...P_{dk}\} \quad (1)$$

P adalah indeks kemunculan *node N* tersebut didalam dokumen (*d*) dan *k* adalah banyak elemen dalam himpunan (*O*). Untuk menghitung relevansi (*S*) sebuah dokumen (*d*) terhadap sebuah *domain* (*Ω*) maka diusulkan persamaan 2.

$$S(d,\Omega)=\Phi(d,\Omega) + \Psi(d,\Omega)$$

$$S(d,\upsilon)=\phi(d,\upsilon) + \Psi(d,\Omega) \quad (2)$$

ф Φ merupakan jumlah kemunculan *node* tanpa relasi pada *domain* Ω yang terdapat dalam sebuah dokumen (*d*), sedangkan Ψ merupakan total score untuk seluruh relasi antar *node* yang terbentuk pada *domain* Ω yang muncul pada dokumen (*d*). Nilai dari Φ ditunjukkan pada persamaan 3, Dimana *j* adalah jumlah *node* yang muncul pada dokumen (*d*), T adalah jumlah kata dalam dokumen (*d*) dan k(N) adalah jumlah frekuensi kemunculan *node* (N).

$$\Phi(d,\Omega)=\frac{j\times\sum_{i=1}^{j}k(N_i)}{T} \quad (3)$$

Jika terdapat dua *node M* dan *node N*yang memiliki relasi dalam *domain* (Ω) yang muncul pada dokumen (d) dan himpunan kemunculan (*O*) *node M* (persamaan 4) dan himpunan kemunculan (*O*) *node N* (persamaan 5), Maka *score* untuk relasi dua *node* tersebut pada *domain* (Ω) terhadap dokumen (*d*) dapat dipresentasikan pada persamaan 6.

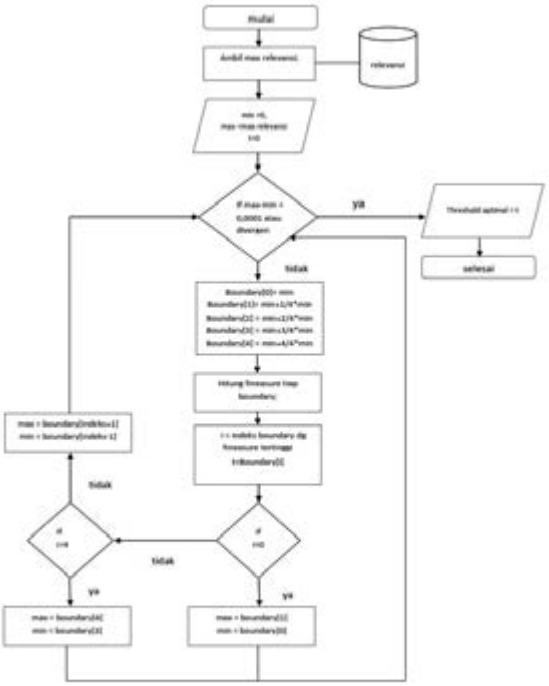
$$O_{d\Omega}(M)=\{P_{d1}, P_{d2}, ...P_{dy}\} \quad (4)$$

$$O_{d\Omega}(N)=\{P_{d1}, P_{d2}, ...P_{dx}\} \quad (5)$$

$$\partial(d,M,N,\Omega)=\sum_{i=1}^{x}\sum_{j=1}^{y}\frac{1}{P_{Mj}-P_N}\times W, P_{Mj} > P_{Ni} \quad (6)$$

∂ merupakan *score* relasi antar dua *node* pada *domain* (Ω) yang muncul pada dokumen (*d*), dimana $P_{Mj}>P_{Ni}$ dan *W* adalah bobot. Untuk menghitung total *score* (Ψ) seluruh relasi antar *node* yang terbentuk pada *domain* Ω yang muncul pada dokumen (*d*) maka dapat digunakan persamaan 7, dimana T adalah jumlah kata dalam dokumen (*d*) dan R adalah jumlah relasi yang muncul.

$$\Psi(d,\Omega)=\frac{R\times\sum_{i=1}^{R}\partial(d,M,N,\Omega)}{T} \quad (7)$$



Gambar 6
Diagram alur pencarian *threshold* optimal

Jika relasi antar dua *node M* dan *node N* yang terbentuk adalah relasi spesialisasi maka *score* relasi (∂) antar dua *node* pada *domain* (Ω) yang muncul pada dokumen (*d*) yang dihitung hanya *score* relasi *node M* terhadap *node N* saja yaitu ∂ (*d,M,N,Ω*), sedangkan jika

node M dan *node N* membentuk relasi non spesialisasi maka *score* relasi (∂) antar dua *node* pada *domain* (Ω) yang muncul pada dokumen (*d*) yang dihitung adalah score relasi *node M* terhadap *node* N yaitu ∂ (*d,m,n,Ω*) ∂ (*d,M,N,Ω*) seperti pada persamaan 6 dan score relasi *node N* terhadap *node M* yaitu ∂ (*d,M,N,Ω*) seperti pada persamaan 8.

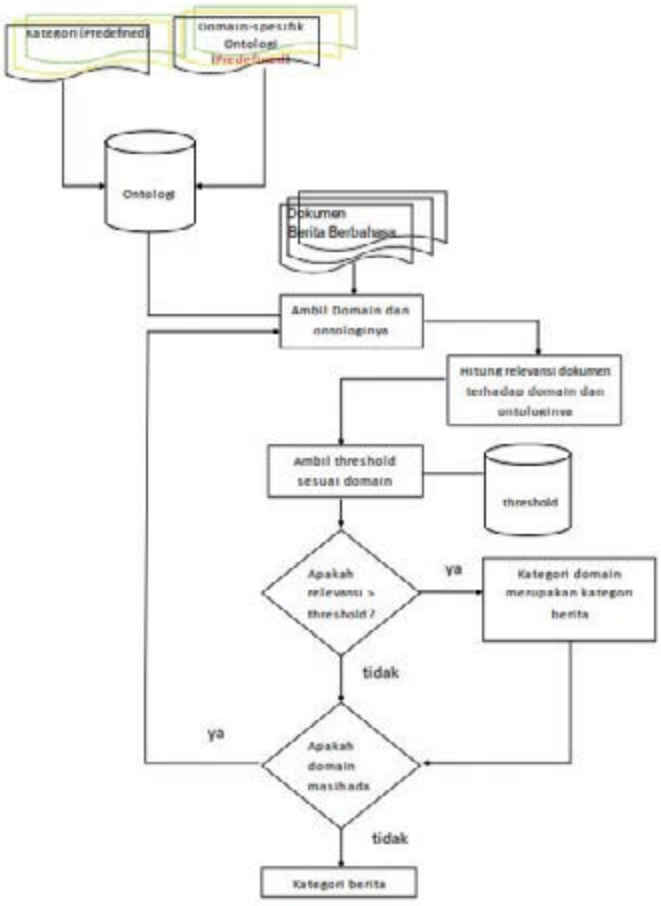
$$\partial(d,N,M,\Omega)=\sum_{i=1}^{y}\sum_{j=1}^{x}\frac{1}{P_{Ni}-P_{mj}}\times W, P_{Nj}>P_{mi} \quad (8)$$

Sehingga Untuk menghitung total score seluruh relasi antar *node* yang terbentuk pada *domain* (Ω) yang muncul pada dokumen (*d*) maka dapat digunakan persamaan 9.

$$\Psi(d,\Omega)=\frac{R\times\sum_{i=1}^{R}\partial(d,M,N,\Omega)+\partial(d,N,M,\Omega)}{T} \quad (9)$$

Perhitungan relevansi dilakukan pada dokumen (*d*) terhadap seluruh *domain* dalam ontologi yang telah dibuat. Setelah didapat nilai relevansi (S) sebuah dokumen (*d*) terhadap seluruh *domain* (Ω) maka langkah selanjutnya adalah mencari *threshold* yang optimal. *Threshold* optimal didapat dengan cara meniru algoritma *binary search* dimana mencari *threshold* dengan akurasi tertinggi dengan interval dari nilai minimum nilai relevansi adalah 0 hingga nilai maksimal relevansi *score*, diagram alur pencarian *treshold* optimal ditunjukkan Gambar 6.

Dalam fase *testing* step yang dilalui hampir sama dengan step-step pada fase *training*, yakni perhitungan kemunculan *node* dan jarak antar *node* yang berelasi serta *scoring*. Setelah nilai relevansi dokumen terhadap *domain* didapat maka akan dikomparasi dengan *treshold* hasil *training* yang telah dilakukan.



Gambar 7
Diagram alur fase *testing*

Apabila nilai relevansi lebih besar dari *treshold* yang ada maka kategori dari *domain* tersebut merupakan kategori dari dokumen berita yang diklasifikasikan. Tahap perhitungan nilai relevansi dilakukan pada setiap *domain* sehingga dapat diketahui apakah sebuah dokumen hanya terikat pada satu *domain* saja atau lebih.

Jika sebuah dokumen terikat lebih dari satu *domain* pada kategori yang berbeda maka dokumen tersebut memiliki multi kategori atau biasa disebut *multi-label.* Diagram alur pada fase *testing* metode klasifikasi usulan ditunjukkan pada Gambar 7.

Metode Evaluasi Hasil Klasifikasi

Pelaksanaan evaluasi uji coba menggunakan rumus *precision*, *recall*, *F-Measure* dan *Accuration* dengan pendekatan dokumen yang di*retrieve* dan relevan seperti pada Tabel 1. Tabel tersebut menunjukkan beberapa *item* yang diperlukan untk mengukur performa *classifier*. *Item-item* tersebut akan digunakan untuk menghitung *Precision*, *Recall*, *F-Measure* dan *Accuration* dengan rumus sebagai berikut:

Precision (P) = TP/(TP+FP)
Recall (R) = TP/(TP+FN)
F-Measure (F1) = 2*P*R/(P+R)
Accuration (A) = (TP+TN)/(TP+FP+FN+FN)

Tabel 1
Retrieve dan *Relevant*

	Relevant	**Not Relevant**
Retrieved	TP	FP
Not Retrieved	FN	TN

HASIL DAN PEMBAHASAN

Data yang digunakan untuk menguji aplikasi ini adalah dokumen berita untuk *training*, dokumen berita untuk *testing* dan *domain*-spesfik ontologi yang telah dibuat. Karakteristik dan jumlah data dijelaskan sebagai berikut:

Data *Domain* Spesifik Ontologi

Data ontologi berupa kumpulan kata yang terkait dengan *domain* tertentu. *Domain-domain* ini merupakan sub class dari kategori. Kata-kata yang merepresentasikan *domain* tersebut disebut *node*, *node* satu dengan *node* yang lain dalam *domain* yang sama akan memiliki sebuah relasi. Relasi yang digunakan dalam penelitian ini adalah relasi spesialisasi dan nonspesialisasi.

Kategori yang digunakan dalam percobaan ini adalah kategori olahraga dan teknologi, kategori olahraga memiliki 6 *domain* dan kategori teknologi memiliki 5 *domain*. Total *node* yang ada pada kategori olahraga adalah 300 *node* sedangkan pada kategori teknologi adalah 250 *node*. Relasi yang terbentuk 650 relasi pada kategori olahraga dan 340 pada kategori teknologi. Detail dari *domain* pada kategori olahraga dapat dilihat pada Tabel 2, sedangkan *domain-domain* yang terkait dengan kategori teknologi dapat dilihat pada Tabel 3.

Tabel 2
Domain kategori Olahraga.

Domain	**Jumlah *Node***	**Jumlah Relasi**
Angkat Besi	50	99
Lari	50	115
Loncat Galah	50	145
Balap Mobil	50	79
Balap Motor	50	100
Balap Sepeda	50	112
Total	**300**	**650**

Tabel 3
Domain kategori Teknologi.

Domain	**Jumlah *Node***	**Jumlah Relasi**
Internet	50	45
Inovasi	50	44
Gadget	50	105
Komputer	50	100
Aplikasi Komputer	50	46
Total	**250**	**340**

Data Dokumen Berita

Data berupa corpus berita online berbahasa Indonesia yang didapatkan dari berbagai situs berita *online*. Berita diunduh berdasar kategori yang telah ditetapkan. Kategori primitif dalam uji coba berguna untuk mengevaluasi hasil klasifikasi. Dokumen berita diambil berdasar

kategori yang telah ditentukan yaitu kategori Olahraga dan kategori Teknologi. Dokumen berita yang terkait untuk kedua kategori juga diikutkan dalam ujicoba hal ini dilakukan untuk mengetahui keakuratan pada klasifikasi yang *multilabel*.

Terdapat data yang digunakan sebagai data pelatihan aplikasi. Data tersebut disebut data *training* dan memiliki karakter yang sama dengan data uji atau *testing*, hanya saja dalam pembuatan *corpus*, data tersebut telah dilabeli kategori sesuai dengan kategori yang diberikan oleh situs berita tersebut. Antara sebuah kategori dengan kategori lainnya memiliki jumlah dokumen uji yang berbeda-beda. Jumlah dokumen berita untuk setiap kategori dapat dilihat pada Tabel 4. Data yang digunakan untuk pelatihan atau *training* adalah 80% dari total dokumen sedangkan data uji atau *testing* adalah 20% dari seluruh dokumen. Pada uji coba yang telah dilakukan pada program ditemukan bahwa nilai bobot (w) berpengaruh pada akurasi dan *fmeasure* (F1) sebuah dokumen pada suatu *domain* sehingga untuk mendapatkan akurasi dan *fmeasure* (F1) terbaik atau untuk mendapatkan sebuah *domain* yang paling relevan terhadap sebuah dokumen dilakukan dengan mengubah-ubah nilai bobot (w) tersebut, selanjutnya jika terdapat nilai akurasi atau *fmeasure* (F1) terbaik dengan nilai yang sama pada bobot (w) yang berbeda maka nilai bobot yang dipilih adalah bobot yang nilainya lebih kecil.

Tabel 4
Dokumen Berita.

Kategori	Jumlah Dokumen
Olahraga	120
Teknologi	180
Total	**300**

Dari hasil uji coba, didapatkan hasil rata–rata nilai akurasi untuk kategori olahraga adalah 93, 85% sedangkan *fmeasure* (F1) untuk kategori olahraga adalah 74, 74%. Dengan tingkat akurasi tertinggi adalah 100% dengan bobot (w) = 15 yaitu *domain* balap mobil sedangkan yang terendah adalah 86, 84% yaitu *domain* loncat galah dengan bobot (w) = 5. *fmeasure* (F1) tertinggi adalah *domain* balap mobil dengan nilai 100% pada bobot (w) = 15, dan yang terendah adalah 54, 54% pada *domain* loncat galah pada bobot (w) = 5. *Domain* Angkat Besi mempunyai nilai akurasi dan *fmeasure* (F1) terbaik pada bobot (w) 5, 8, 10, 13, 15, 18, , 20, 23, dan 25, sehingga diambil nilai bobot (w) = 5 sebagai nilai yang paling kecil. Detail hasil ujicoba pada kategori olahraga dapat dilihat pada Tabel 5.

Tabel 5
Hasil Kategori Olahraga.

Domain	**Akurasi Terbaik**		**F1 Terbaik**	
	Akurasi (%)	W	F1 (%)	W
Angkat Besi	97, 36	5	85, 71	5
Lari	97, 36	5	80, 00	5
Loncat Galah	86, 84	5	54, 54	5
Balap Mobil	**100, 00**	15	**100, 00**	15
Balap Motor	92, 10	5	66, 67	5
Balap Sepeda	89, 47	8	61, 54	18
Rata - rata	**93, 85**		74, 74	

Untuk rata–rata nilai akurasi untuk kategori teknologi adalah 96, 32% sedangkan *fmeasure* (F1) untuk kategori teknologi adalah 78, 96%. Dengan tingkat akurasi tertinggi adalah 100% dengan bobot (w) = 20 yaitu *domain* komputer sedangkan yang terendah adalah 92, 11% yaitu *domain* inovasi dengan bobot (w) = 5. *fmeasure* (F1) tertinggi adalah *domain* komputer dengan nilai 100% pada bobot (w) = 20, dan yang terendah adalah 40% pada *domain* inovasi pada bobot (w) = 5. *Domain* komputer mempunyai nilai akurasi terbaik pada bobot (w) 20, 23, dan 25, sehingga diambil nilai bobot (w) = 20 sebagai nilai yang paling kecil. Pada *fmeasure* (F1) terbaik untuk *domain* komputer juga mempunyai nilai bobot (w) = 20, sebagai nilai yang paling kecil. Detail hasil ujicoba pada kategori teknologi dapat dilihat pada Tabel 6.

Tabel 6
Akurasi Kategori Teknologi.

Domain	**Akurasi Terbaik**		**F1 Terbaik**	
	Akurasi (%)	**W**	**F1 (%)**	**W**
Internet	94, 74	13	75, 00	13
Inovasi	92, 11	5	40, 00	5
Gadget	97, 37	23	90, 91	23
Komputer	**100, 00**	20	**100, 00**	20
Aplikasi Komputer	97, 37	5	88, 89	5
Rata - rata	**96, 32**		78, 96	

Berdasarkan uji coba yang telah dilakukan didapatkan hasil bahwa pada kategori olahraga nilai akurasi dan *fmeasure* yang terbaik adalah pada *domain* balap mobil, sedangkan nilai akurasi dan *fmeasure* terendah adalah pada *domain* loncat galah. Akurasi dan *fmeasure* tertinggi pada kategori teknologi terdapat pada *domain* komputer sedangkan akurasi dan *fmeasure* terendah terdapat pada *domain* inovasi. Pada kategori olahraga jumlah relasi *node* terbanyak pada *domain* loncat galah, sedangkan *domain* loncat galah memiliki nilai akurasi dan *fmeasure* terendah pada kategori olahraga. Sebaliknya pada kategori olahraga *domain* balap mobil memiliki jumlah relasi *node* yang paling sedikit tetapi justru memiliki nilai akurasi dan *fmeasure* tertinggi.

Pada kategori teknologi hal tersebut tidak terjadi, jumlah relasi *node* yang memiliki nilai tinggi pada kategori ini adalah *domain* komputer dan gadget. Pada hasil uji coba nilai akurasi dan *fmeasure* tertinggi juga teradapat pada *domain* komputer dan gadget, begitu juga jumlah relasi *node* yang memiliki jumlah terendah adalah *domain* inovasi dan pada hasil uji coba nilai akurasi dan *fmeasurenya* juga merupakan yang terendah. Sehingga dapat dilihat bahwa tidak ada keterkaitan antara jumlah relasi *node* dengan nilai akurasi dan *fmeasure* yang diperoleh saat uji coba pada setiap *domain,* namun nilai akurasi dan *fmeasure* yang diperoleh sangat terkait dengan relasi antar *node* ontologi yang muncul pada dokumen hal ini dikarenakan hubungan kemunculan dan nilai relevansi dokumen adalah berbanding lurus. Berdasarkan uji coba yang telah dilakukan ditemukan bahwa untuk mencapai nilai akurasi dan *fmeasure* yang tinggi maka diperlukan pemilihan nilai *threshold* dan bobot (w). Mengubah-ubah nilai bobot dilakukan pada ujicoba untuk melihat perubahan yang terjadi pada tingkat akurasi dan *fmeasure* dimana didapatkan fakta bahwa setiap domain memiliki nilai bobot optimal yang berbeda-beda untuk mencapai tingkat akurasi dan nilai *fmeasure* yang tinggi hal ini dikarenakan nilai bobot (w) yang sangat mempengaruhi pada perhitungan nilai relevansi dokumen terhadap sebuah domain pada fase *training* dan *testing*.

SIMPULAN

Metode yang diusulkan terbukti mampu melakukan kategorisasi dengan sangat baik dimana hal ini dapat dilihat pada haasil uji coba yang dilakukan pada dokumen berita untuk kategori olahraga dan kategori teknologi. Akurasi yang dapat dicapai metode usulan pada kategori olahraga adalah 93, 85% sedangkan pada kategori teknologi adalah 96, 32%. Nilai *fmeasure* (F1) yang dapat dicapai algoritma klasifikasi pada kategori olahraga adalah 74, 74% sedangkan pada kategori teknologi adalah 78, 96%. Metode usulan juga mampu mengkategorisasi dokumen berita yang terkait lebih dari satu kategori (*multi label*). Berdasar ujicoba yang telah dilakukan juga didapat kesimpulan bahwa keakurasian dalam kategorisasi dokumen berita tidak berkaitan dengan jumlah relasi *node* ontologi yang dibuat, namun sangat terkait dengan relasi antar*node* ontologi yang muncul pada dokumen hal ini dikarenakan hubungan kemunculan dan nilai relevansi dokumen adalah berbanding lurus. *Threshold* pada metode usulan ini telah dapat dioptimasi pada fase *training* sedangkan bobot masih dilakukan secara manual, untuk itu diperlukan penelitian lebih lanjut terkait cara penentuan paramater bobot (w) yang optimal sehingga tingkat

keakurasian dalam mengkategorsasi dokumen dapat ditingkatkan. Penentuan bobot yang optimal pada fase *training* dapat menghasilkan nilai *treshold* yang optimal pula, sehingga dapat disimpulkan bahwa kedua parameter tersebut sangat berkaitan dan mempengaruhi tingkat keakurasian metode usulan. Metode usulan juga sangat bergantung pada ontologi yang telah terbentuk, untuk itu diperlukan kombinasi metode usulan dengan penggunaan database leksikal, semisal Wordnet sehingga referensi term pada ontologi yang terbentuk menjadi fleksibel dan lebih luas.

DAFTAR PUSTAKA

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). Modern information retrieval. New York: Addison Wesley.

Ben-Dov, M., dan Feldman R. 2001, Text Mining and Information Extraction, Chapter 38.

Husni. IR dan Klasifikasi. Diktat kuliah, Universitas Trunojoyo.

Milton, N. (2003). Knowledge Engineering. November 21,2015.http://www.epistemics.co.uk/Notes/61-0-0.htm

N.F. Noy, D.L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Knowledge Systems Laboratory (KSL) of Department of Computer Science Stanford, USA: Technical Report, KSL-01-05, 2001.

Noy, N.F., & McGuinness, D.L. (2001). Ontology Development 101: A guide to creating your first ontology. Knowledge Systems Laboratory (KSL) of Department of Computer Science Stanford, USA: Technical Report, KSL-01-05.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, Vol. 34 (1), 1-47.