Ontology Learning pada Teks Tidak Terstruktur

Rajif Agung Yunmar[1], Hartanto Tantriawan[2], Andika Setiawan[3]

Program Studi Teknik Informatika Institut Teknologi Sumatera
Jl. Terusan Ryacudu Lampung Selatan, 35365, Indonesia.
[1] rajif@if.itera.ac.id
[2]hartanto.tantriawan@if.itera.ac.id
[3]andika.setiawan@if.itera.ac.id

Intisari — **Informasi yang tersebar pada berbagai sumber di internet banyak ditujukan hanya untuk manusia saja. Sementara itu, muncul kebutuhan agar informasi tersebut tidak hanya bisa dibaca dan dipahami oleh manusia saja, tetapi juga oleh mesin. Informasi dalam format yang dapat dipahami oleh mesin dapat digunakan untuk berbagai keperluan, misalnya: menjadi basis pengetahuan untuk penalaran, sharing knowledge antar mesin, semantic search, visualisasi informasi, dsb. Ontology learning adalah metode yang dapat mengekstrak informasi dari teks tidak terstruktur pada suatu dokumen atau halaman web untuk kemudian diubah menjadi basis pengetahuan dalam format yang dapat dipahami oleh mesin, yaitu dalam bentuk ontologi. Metode tersebut terdiri dari beberapa tahapan, yaitu: preprocessing, ekstraksi konsep, ekstraksi relasi, dan evaluasi. Preprocessing menyiapkan korpus uji sehingga siap untuk masuk kedalam metode ekstraksi konsep, yang menggunakan algoritma entropy concept extraction, pada bagian ekstraksi relasi digunakan algoritma subcat relation extraction, sedangkan pada bagian evaluasi ontologi menggunakan metode expert evaluation. Hasil akhir menunjukkan akurasi pada nilai 89.84% untuk ekstraksi konsep, 93.02% untuk ekstraksi relasi, dengan kepercayaan terhadap ekstraksi relasi pada prosentase 71.15%.**
Kata kunci — **ontology learning, entropy concept extraction, subcat relation extraction.**

Abstract — **Information that spread on various sources on the internet intended only for humans. Meanwhile, there is a need that the information can understood and read not only by humans, but also by machines. Information in a format that can be understood by machines can be used for the various purposes, for example: being a knowledge base for reasoning, sharing knowledge between machines, semantic search, information visualization, etc. Ontology learning is a method that can extract the information from unstructured text on a document or web page, and then convert it into a knowledge base in a format that can be understood by machines, namely in the ontology form. The method consists of several stages, namely: preprocessing, concept extraction, relationship extraction, and evaluation. Preprocessing prepare the text to be ready to proses by concept extration method that use entropy concept extraction algorithm, in the relation extraction stage the subcat relation extraction algorithm is used, while the ontology evaluation section uses the expert evaluation method. The final results show accuracy of concept extraction is 89.84%, 93.02% for relation extraction, with confidence in relation extraction at a percentage of 71.15%.**
Keywords — **ontology learning, entropy concept extraction, subcat relation extraction.**

I. PENDAHULUAN

Dewasa ini teknologi terkait web dan internet berkembang dengan pesatnya. Hal tersebut membantu informasi untuk dapat tersebar luas dengan cepat. Informasi yang terdapat pada berbagai sumber di internet disajikan dengan mengutamakan aspek presentasi, yaitu bagaimana informasi tersebut disajikan secara menarik dan mudah dimengerti oleh manusia. Informasi tersebut biasaya ditulis dalam teks tidak terstruktur dan dalam bentuk bahasa alami. Suatu informasi pada suatu halaman web dengan informasi pada halaman web lain kadang tidak saling terhubung dalam hal konsep pengetahuan. Hal ini membuat hadirnya informasi terkadang kontradiktif antara satu sumber dengan sumber lainnya.

Sementara itu muncul kebutuhan yang lebih besar, yaitu bagaimana agar informasi yang ada di internet tersebut selain dapat dimengerti oleh manusia, juga dapat dipahami oleh mesin. Informasi dalam bentuk yang dapat dimengerti oleh mesin dapat digunakan untuk berbagai keperluan, misalnya: menjadi basis pengetahuan untuk suatu penalaran, *sharing knowledge* antar mesin, semantic search, visualisasi informasi, dan lain sebagainya. Permasalahan tersebut dapat diselesaikan dengan teknologi semantic web. Teknologi semantic web diperkenalkan oleh Tim Berners Lee dengan tujuan membuat informasi dan pengetahuan pada web menjadi saling terhubung pada skala global, dan yang paling penting yaitu dapat dimengerti dan diproses oleh mesin [1], [2].

Ontologi memainkan peran penting dalam mengimplementasikan konsep semantic web. Ontologi merupakan model dari dunia nyata pada domain tertentu yang memuat hubungan antar konsep, baik secara hirarki maupun relasi [3]. Ontologi sebagai basis pengetahuan mampu merepresentasikan informasi pengetahuan berdasarkan konsep semantik dari sebuah objek, properti, dan relasi antar objek yang terjadi pada domain tertentu [4]. Penggunaan ontologi akan menghasilkan sistem yang mana pengetahuan didalamnya tidak hanya bisa dipahami oleh manusia saja, tapi juga dapat dipahami oleh mesin. Dengan demikian informasi diberikan oleh sistem yang menggunakan basis pengetahuan ontologi dapat lebih tepat dan relevan [5].

Namun demikian, dengan segala kelebihan ontologi yang disebutkan diatas, terdapat masalah dalam hal proses pengembangannya, yaitu membutuhkan banyak sumberdaya. Pengembangan ontologi secara manual membutuhkan ahli yang memahami pengetahuan pada domain tersebut. Saat ini terdapat lebih dari 60 miliar halaman web [6], tidak dapat dibayangkan berapa banyak sumberdaya yang harus disediakan apabila secara manual diinginkan untuk membangun pengetahuan dari semua halaman web tersebut kedalam bentuk yang dapat dimengerti oleh mesin. Pengembangan ontologi dalam skala yang lebih besar dengan cara manual memungkinkan hasil yang tidak akurat, proses pengembangannya lama dan membosankan, serta mahal dari segi biaya [7].

Sulitnya pengembangan ontologi pada skala global inilah yang menjadi *bottleneck* dari implementasi semantic web dewasa ini [8]. Kendala yang dihadapi dalam pengembangan ontologi secara manual kemudian membuat pengembangan ontologi yang meminimalkan campur tangan manusia menjadi hal yang penting. Menjadi tren penelitian saat ini yaitu pengembangan ontologi dilakukan secara otomatis, menjadi lebih menantang apabila ontologi dibangun dari sumber teks yang tidak terstruktur seperti yang ada pada halaman web. Bidang penelitian ini disebut dengan istilah ontology learning.

Pada implementasinya, ontology learning memiliki beberapa tahapan, yaitu: preprocessing, concept extraction, relation extraction, dan ontology evaluation. Terdapat dua arus besar pada tahapan term extraction, yaitu menggunakan pendekatan pengolahan bahasa alami, dan pendekatan menggunakan statistik. Aplikasi dari kedua pendekatan tersebut asing-masing pendekatan memiliki kelebihan dan kekurangan [9]. Pada penelitian ini digunakan pengolahan bahasa pada tahapan *preprocessing*. Hasil pengolahan bahasa pada *preprocessing* digunakan pada tahapan ekstraksi konsep dan ekstraksi relasi yang keduanya menggunakan prinsip-prinsip statistik.

Pada tahapan ontology evaluation, peneliti ini melakukan evaluasi pada dua tempat, yaitu pada tahapan ekstraksi konsep, dan pada

tahapan ekstraksi relasi. Dengan dua kali evaluasi, diharapkan ontologi yang dihasillkan dapat diukur dengan lebih baik.

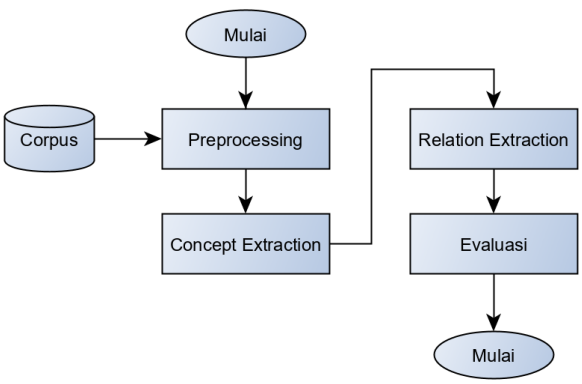
II. METODE

A. *Korpus*

Teks tidak terstruktur yang digunakan sebagai bahan dan sumber data penelitian ini berasal dari halaman-halaman web yang terdapat pada website bertema pariwisata Indonesia, khususnya Yogyakarta, yaitu: *www.indonesia-tourism.com*. Korpus terdiri dari 48 artikel pariwisata dengan total kata sejumlah 18915 item. Pemilihan tema pariwisata diharapkan dapat berdampak positif bagi sektor pariwisata dan membuat penelitian ini menjadi berkelanjutan. Hasil ontologi yang didapatkan dari penelitian ini dapat menjadi basis pengetahuan bagi perangkat lunak yang dikembangkan oleh peneliti lainnya, misalnya Question Answering System dengan basis pengetahuan ontologi.

B. *Rancangan Sistem*

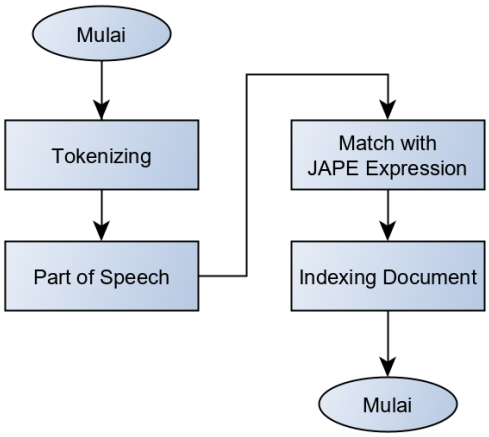
Perancangan sistem menggambarkan detail dari setiap tahapan yang akan dilakukan didalam sistem. Perancangan ini bertujuan untuk mengetahui bagaimana alur kerja daripada penelitian yang dilakukan. Sehingga dapat menjadi panduan dalam implementasi dan pengembangan sistem selanjutnya. Secara umum rancangan sistem pada penelitian ini dapat dilihat pada Gbr. 1. Tahapan yang akan dilalui diantaranya, yaitu: *preprocessing*, *concept extraction*, *relation extraction*, dan evaluasi. Pada penelitian ini, korpus yang dibangun akan digunakan pada tahapan *preprocessing*.



Gbr. 1 Rancangan sistem secara umum.

1) Preprocessing

Pada tahapan preprocessing digunakan Java Annotation Patterns Engine (JAPE) untuk melakukan pengolahan awal terhadap teks. Proses yang dilakukan diantaranya adalah tokenisasi dengan cara memecah kalimat menjadi kata/*term*, melakukan proses penandaan dengan Part of Speech (POS). Pada tahapan POS, setiap kata/term akan diberikan label, misalnya: *subject*, *object*, *predicate*, *noun*, *proper noun, verb,* dan lain sebagainya. Proses POS akan berguna bagi tahapan selanjutnya, yaitu pada tahapan ekstraksi konsep, dan ekstraksi relasi. Proses selanjutnya yang dilakukan pada tahapan *preprocessing* yaitu melakukan pencocokan hasil POS dengan JAPE Expression. Terakhir, *preprocessing* akan membuat indeks dalam bentuk *annotation set* untuk setiap dokumen korpus. Langkah-langkah pada tahapan *preprocessing* dapat dilihat secara detail melalui Gbr. 2.



Gbr. 2 Tahapan proprocessing.

2) Ekstraksi Konsep

Pada tahapan ini akan diekstrak kata/*term* yang dapat menjadi konsep dari korpus yang ada. Dalam ontologi konsep ini dapat disebut juga dengan istilah kelas. Metode ekstraksi yang digunakan dalam penelitian ini adalah *Entropy Concept Extraction*. Algoritma ini memiliki kelebihan tersendiri, karena tidak terikat dengan batasan algoritma pada bahasa tertentu, juga tidak memerlukan proses-proses terkait pengolahan bahasa. Algoritma ini mengekstrak konsep pada kata/*term* berjenis *noun*, yang dihitung melalui menggunakan Persamaan 1 [10].

$$Entropy\ term\ t = P(t) * \log(P(\ t)) \quad \textbf{(1)}$$

$$P(t) = \frac{Frekuensi\ t}{Kumulatif\ frekuensi\ semua\ term} \quad \textbf{(2)}$$

dimana *t* adalah sebuah kata/*term*, *P(t)* adalah nilai probabilitas dari *t*. Nilai probabilitas *t* didapatkan dari perhitungan Persamaan 2. Konsep diekstrak dengan melihat relevansi *term t* terhadap nilai *entropy* dari suatu *term t*. Semakin besar nilai *entropy term t*, semakin relevan ia untuk dijadikan sebuah konsep.

3) Ekstraksi Relasi

Ontologi tersusun dari term-term yang saling terhubung dengan predikat tertentu. Dalam ontologi learning, predikat tersebut bisa didapatkan dari tahapan ekstraksi relasi. Algoritma *Subcat Relation Extraction* digunakan untuk mengekstrak relasi. Algoritma tersebut menggunakan pencocokan pola sintaksis. Sebuah relasi diidentifikasi berdasarkan aturan pola sintaksis sebagaimana terlihat pada Tabel 1 [11].

Tabel 1. Pola aturan identifikasi relasi.

No	Aturan	Pola
1	Transitive	Subject + Verb + Object
2	Intransitive plus preposition phrase-complement	Subject + Verb + Preposition + Object
3	Trasitive + preposition phrase-complement	Subject + Verb + Object + Preposition + Prepositional object

Tabel 2. Contoh identifikasi relasi.

No	Aturan	Kalimat dan Hasil
1	Transitive	**Kalimat**: A man kick the ball **Hasil**: kick(man, ball)
2	Intransitive plus preposition phrase-complement	**Kalimat**: Barcelona depend on Messi **Hasil**: depend_on(barcelona, messi)
3	Trasitive + preposition phrase-complement	**Kalimat**: John buy the Boba using Gopay **Hasil**: 1. buy(john,boba) 2. buy_using(john,Gopay)

Pencocokan pola sintaksis untuk menemukan relasi dari sebuah kalimat dapat diberikan contoh seperti pada Tabel 2.

Relasi yang didapatkan menggunakan aturan pola yang ada pada Tabel 1 akan diekstrak dengan mengukur tingkat kepercayaan (*confidence*). Perhitungan tingkat kepercayaan dapat diperoleh melalui Persamaan 3 [11].

$$Confidence = PD * PR \quad \textbf{(3)}$$

dimana PD adalah probabilitas domain, dan PR adalah probabilitas *range*. Nilai dari PD dan PR dapat dihitung melalui Persamaan 4, dan Persamaan 5.

$$PD = \frac{Frekuensi\ domain\ untuk\ relasi\ R}{Frekuensi\ relasi\ R} \quad \textbf{(4)}$$

$$PR = \frac{Frekuensi\ range\ untuk\ relasi\ R}{Frekuensi\ relasi\ R} \quad \textbf{(5)}$$

Istilah domain dan relasi pada Persamaan 4 dan Persamaan 5 mengacu kepada objek-objek pada sebuah relasi R. Jika terdapat relasi R, maka domain dituliskan dengan fungsi R(domain,range). Jika merujuk kepada contoh pada Tabel 2.2, misalnya: kick(man,ball), maka *man* adalah *domain*, sedangkan *ball* adalah *range*.

Algoritma *Subcat Relation Extraction* akan mengekstrak relasi dengan dengan *domain* dan *range* yang paling sering muncul dalam relasi tersebut. Sebagai contoh: relasi R(A1,B1) muncul dua kali, resi R(A2,B2) muncul satu kali, R(A3,B3) muncul tiga kali, dan R(A1,B2) muncul tiga kali, maka dapat dihitung frekuensi A1, A2, A3, B1, B2, B3 dalam relasi R sebagai berikut:

- Frekuensi A1 = 2 + 3 = 5
- Frekuensi A2 = 1
- Frekuensi A3 = 3
- Frekuensi B1 = 2
- Frekuensi B2 = 1 + 3 = 4
- Frekuensi B3 = 3

Maka, dalam kasus ini A1 menjadi *domain* yang memiliki kemunculan paling sering, sedangkan B2 menjadi *range* yang muncul paling sering. Maka, algoritma *Subcat Relation Extraction* akan mengekstrak relasi R(A1,B2).

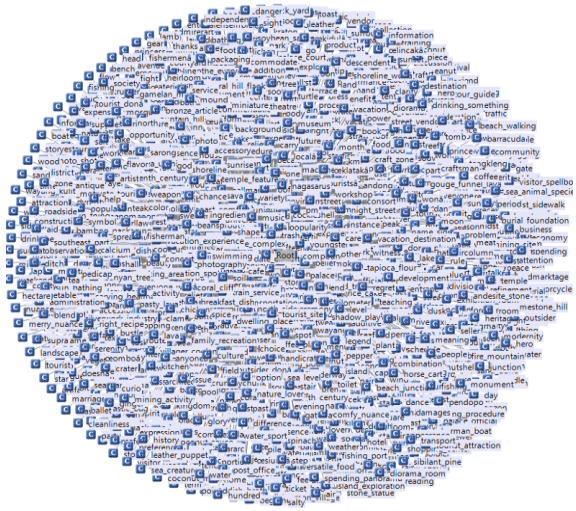
4) Evaluation

Tahapan ekstraksi konsep dan ekstraksi relasi, masing-masing akan menghasilkan term-term konsep, dan relasi yang menghubungkan term-term konsep tersebut. Evaluasi dilakukan pada setiap item yang dihasilkan dengan memberikan nilai 1 untuk item yang bernilai benar, dan nilai 0 untuk item term yang kurang tepat. Proses evaluasi dilakukan oleh ahli yang memiliki pengatahuan pada bidang pariwisata. Akurasi untuk item yang diekstraksi dapat dihitung menggunakan Persamaan 6.

$$Akurasi = \frac{Item\ bernilai\ benar}{Jumlah\ keseluruhan\ item} * 100 \quad \textbf{(6)}$$

III. IMPLEMENTASI

Implementasi metode penelitian memanfaatkan tool Text2Onto untuk melakukan ekstraksi terhadap korpus yang disusun. Pada tool tersebut telah terdapat modul *preprocessing*, ekstraksi konsep, dan ekstraksi relasi. Text2Onto mendapatkan input berupa korpus pada domain pariwisata Indonesia berbahasa Inggris. Hasil eksekusi metode pada Text2Onto menghasilkan term-term konsep dan relasi yang menghubungkan term. Hasil eksekusi tserebut dapat dieksport dalam bentuk file ontologi dengan format RDF. Hasil eksport dalam format tersebut dapat divisualisasikan seperti terlihat pada Gbr. 3.



Gbr. 3 Visualisasi ontologi hasil ekstraksi.

IV. HASIL DAN PEMBAHASAN

Ekstraksi konsep dengan algoritma *Entropy Concept Extraction* menghasilkan 965 term konsep. Evaluasi yang dilakukan terdapat 98 item konsep bernilai salah. Sementara ekstraksi relasi menggunakan algoritma *Subcat Relation Extraction* menghasilkan 41 relasi. Dari evaluasi yang dilakukan, terdapat 3 item relasi bernilai salah. Nilai kepercayaan terhadap ekstraksi relasi berdasarkan Persamaan 3 ada pada nilai presentase 71.15%. Tabel 3 menunjukkan akurasi ekstraksi konsep dan ekstraksi relasi apabila dihitung dengan menggunakan Persamaan 6.

Tabel 3. Akurasi hasil ekstrasi.

No	Metode	Hasil		Akurasi
		Item Total	Benar	
1	Ekstraksi konsep	965	867	89.84%
2	Ekstraksi relasi	41	38	93.02%

V. KESIMPULAN

Nilai prosentase akurasi yang tinggi pada tahapan ekstraksi konsep dan tahapan ekstraksi relasi seperti yang terlihat pada bagian pembahasan menunjukkan bahwa metode ontology learning yang diimplementasikan menggunakan Text2Onto memiliki tingkat keberhasilan yang tinggi. Salah satu kunci utama keberhasilannya terletak pada *preprocessing*, dimana setiap kata/*term* yang ada dalam korpus dipisahkan berdasarkan jenisnya. Hal ini menyebabkan tahapan selanjutnya lebih mudah melakukan ekstraksi berdasarkan jenis kata/*term* tertentu.

UCAPAN TERIMA KASIH

Penelitian ini dibiayai oleh PNBP Institut Teknologi Sumatera (ITERA) melalui Program Hibah Mandiri Institut Teknologi Sumatera 2019 dengan nomor kontrak B/318/IT9.C1/PT.01.03/ 2019.

REFERENSI

[1] J. Jensen, "A systematic literature review of the use of Semantic Web technologies in formal education," *Br. J. Educ. Technol.*, vol. 00, no. 00, 2017, doi: 10.1111/bjet.12570.

[2] A. Elnaggar, "The Semantic Web," 2015.

[3] R. Samo, Y. Anistyasari, and R. Fitri, *Semantic Search. Pencarian Berdasarkan Konten*. Yogyakarta: Penerbit ANDI, 2012.

[4] Azhari, Subanar, R. Wardoyo, and S. Hartati, "Model Representasi Informasi Dan Pengetahuan Untuk Proyek-Proyek Perusahaan Dengan," *J. Ilm. Teknol. Inf.*, vol. 7, pp. 85–92, 2008.

[5] J. Sun and L. Wang, "Research on E-commerce Data Management Based on Semantic Web," in *14th International Conference on High Performance Computing and Communications*, 2014, p. 925.

[6] M. de Kunder, "The size of the World Wide Web (The Internet)," 2019. [Online]. Available: https://www.worldwidewebsize.com/.

[7] I. Imam, A. Hamouda, and H. A. A. Khalek, "An Ontology-based Summarization System for Arabic Documents (OSSAD)," *Int. J. Comput. Appl.*, vol. 74, no. 17, pp. 38–43, 2013.

[8] L. Drumond and R. Girardi, "A survey of ontology learning procedures," in *CEUR Workshop Proceedings*, 2008, vol. 427.

[9] M. N. Asim, M. Wasim, M. U. G. Khan, W. Mahmood, and H. M. Abbasi, "A survey of ontology learning techniques and applications," *Database*, vol. 2018, no. 2018, pp. 1–24, 2018, doi: 10.1093/database/bay101.

[10] Ryadyo, "Concept and Instance Extraction Algorithms in Text2Onto," 2012. [Online]. Available: https://ryadyo.wordpress.com/2012/10/03/concept-and-instance-extraction-algorithms-in-text2onto/. [Accessed: 21-Jan-2020].

[11] Ryadyo, "SubcatRelationExtraction Algorithm," 2012. [Online]. Available: https://ryadyo.wordpress.com/2012/10/09/subcatrelationextraction-algorithm/. [Accessed: 21-Jan-2020].