JDMSI, Vol. 1, No. 5, 2020, 32-37

SURVEY UKURAN KESAMAAN SEMANTIC ANTAR KATA

Styawati[1)], Winda Yulita[2)], Ida Bagus Gede Sarasvananda[3)]

[123]Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Gadjah Mada
[123]Bulaksumur, Yogyakarta
Email: [1]styazps@gmai.com, [2]winda.yulita@gmail.com, [3]sarasvananda@gmail.com

Abstract

The measurement of the semantic similarities between words has been developed in recent years. Measures of semantic similarity between words are classify into three: distance based method, information content based method and distance and information content based method. Distance based methods are a more natural way of measuring the semantic similarities between words. This approach estimates the distance between nodes that match the concepts being compared. Information content based method uses a systematic knowledge base and statistical information of words from databases such as WordNet, Corpus and HowNet. The distance and information content based method is an combination of a distance-based approach to an information-based content approach.

Keyword: *semantic, distance based method, information based method, distance and information based method*

Abstrak

Pengukuran kesamaan semantic antara kata telah dikembangkan dalam beberapa tahun terakhir. Pengukuran kesamaan semantic antar kata dikelompokkan menjadi tiga yaitu *distance based method, information content based method* dan *distance and information content based method*. *Distance based method* adalah cara yang lebih alami untuk mengukur kesamaan semantic antara kata. Pendekatan ini memperkirakan jarak antara node yang sesuai dengan konsep yang dibandingkan. *Information content based method* menggunakan basis pengetahuan yang sistematis dan statistik informasi kata-kata dari database seperti WordNet, Corpus dan HowNet. *Distance and information content based method* merupakan penggabungan dari pendekatan berbasis jarak dengan pendekatan berbasis konten informasi.

Kata Kunci: *semantic, distance based method, information based method, distance and information based method*

1. Pendahuluan

Dalam beberapa tahun terakhir, pengukuran kesamaan semantic banyak digunakan pada semantic web dan *Natural Language Processing* (NLP). Pengukuran kesamaan semantic adalah pengukuran kesamaan antara term/konsep yang termasuk dalam sumber pengetahuan untuk melakukan estimasi nilai kesamaan [1]. Pengukuran kesamaan semantik adalah merupakan proses yang memerlukan keterlibatan beberapa disiplin ilmu, seperti bahasan, komputer, matematika logik dan domain yang bersangkutan. Langkah awal perhitungan kesamaan semantic adalah mengacu kepada kesamaan terminological atau kerap kali disebut label. Terminologi yang dimaksud dapat meliputi *class*, *property* hingga *instances* [2].

Kesamaan semantic (*semantic similarity*) dikelompokkan dalam beberapa jenis yaitu, *term similarity, sentence similarity, entity similarity, document similarity* dan *word similarity. Term similarity* adalah pemeriksaan kesamaan semantic dengan metode empiris yang menggunakan corpus dokumen web. *Sentence similarity* merupakan proses untuk mengetahui nilai tingkat kesamaan antara dua kalimat. Nilai bisa berupa numerik atau fuzzy. *Entity similarity* merupakan proses menentukan dua aspek yang sama sebagai sebuah entitas yang sama. Contohnya *Oil and Natural Gas Commission* merupakan entitas yang sama dengan ONGC. Document similarity diterapkan dengan pencarian dokumen di web. Tujuan dari proses *document similarity* adalah mengidentifikasi sejauh mana domain (dokumen) tersebut dalam wilayah tertentu yang sedang dipertimbangkan. *Word similarity* adalah proses penentuan tingkat kesamaan makna kata dengan bantuan *lexical resource* seperti WordNet [3].

Banyak metode pengukuran kesamaan semantic antara kata telah dikembangkan dalam beberapa tahun terakhir, seperti *distance based method* [4], *information content based method* [5], dan *distance and information content based method* [6]. *Distance based method* adalah cara yang lebih alami untuk mengukur kesamaan *semantic* antara kata. Pendekatan ini memperkirakan jarak antara *node* yang sesuai dengan konsep yang dibandingkan [7]. *Information content based method*

menggunakan basis pengetahuan yang sistematis dan statistik informasi kata-kata dari database seperti WordNet, Corpus dan HowNet [5]. *Distance and information content based method* merupakan penggabungan dari pendekatan berbasis jarak dengan pendekatan berbasis konten informasi [6].

Distance and information content based method menggunakan database dari WordNet, Corpus, dan HowNet telah banyak dilakukan, diantaranya Jiang dan Conrath [6] dan Li, et al. [8]. Namun, dari penelitian yang ada belum ditemukan pengukuran kesamaan *semantic* yang menggunakan gabungan dari database WordNet, Corpus, dan HowNet. WordNet adalah basis pengetahuan yang sistematis dan memiliki statistik informasi kata-kata, WordNet dapat digunakan sebagai sumber statistik tanpa menggunakan Corpus [17]. HowNet adalah basis pengetahuan *common-on-line* yang mengungkapkan hubungan antar konseptual dan atribut hubungan konsep-konsep yang berkonotasi dalam kamus bahasa Cina dan bahasa Inggris yang setara [9]. Corpus atau Corpora adalah sejenis "bank" bahasa yang berasal dari penggunaan bahasa dalam berbagai genre, ragam, dan bahan lisan maupun tertulis yang menjamin keragaman yang seluas-luasnya dan menghindari penggunaan bahasa yang sangat sempit seperti idiolek. Data tersebut disusun secara sistematis dan biasanya mudah diakses secara elektronis dengan komputer [10].

2. Klasifikasi Ukuran Kesamaan Semantic Antar Kata

Tabel 1 menunjukkan pengelompokkan metode yang digunakan untuk ukuran kesamaan semantic antar kata. Pengelompokan dibagi berdasarkan metode yang digunakan untuk mengukur kesamaan kata. Metode tersebut diantaranya yaitu: *Distance based method, information content based method*, dan *distance and information content based method* yang merupakan penggabungan dari *distance dan information content*.

3. Pembahasan

2.1. *Distance Based Method*

Rada et al. [4] menggunakan *knowledge based information retrieval* (KBIR) untuk fungsi kesamaan semantik sebagai mekanisme perbandingan antara kata-kata. Rada et al. mendefinisikan jarak konseptual antara dua simpul dalam jaringan semantik IS-A sebagai panjang jalur terpendek yang menghubungkan kedua simpul tersebut. Panjang minimum jalur yang menghubungkan dua konsep yang berisi kata-kata ditampung sebagai metrik untuk mengukur kesamaan semantik kata-kata. Penelitian ini menggunakan metode persamaan berbasis penghitungan tepi. Metode penghitungan tepi yang paling sederhana bekerja dengan baik untuk aplikasi spesifik domain dengan taksonomi yang sangat terbatas, seperti jaring semantic pada bidang medis. Namun, tidak memperhitungkan keseragaman dalam kepadatan link, dan tidak mempertimbangkan hubungan semantik lainnya seperti antonymy, holo-/ meronymy, dan lain-lain. Oleh karena itu, hal itu tidak berkinerja baik dalam ontologi semantik umum seperti WordNet.

Penelitian yang dilakukan oleh Richardson dan Smeaton [11] juga menggunakan *knowledge based information retrieval* (KBIR) untuk fungsi kesamaan semantik sebagai mekanisme perbandingan antara kata-kata seperti yang dilakukan oleh Rada et al. [4]. Penelitian ini menggunakan pendekatan terhadap IR berdasarkan perhitungan jarak semantik antara konsep atau kata dan menggunakan jarak kata ini untuk menghitung kemiripan antara kueri dan dokumen. Nilai jarak sangat bergantung pada hirarki jaringan yang dibangun secara subyektif. Hal ini dikarenakan tujuan utama perancangan WordNet bukan untuk tujuan perhitungan kesaaman semantic. Hasil penelitian ini sedikit mengecewakan karena memiliki nilai precision dan recall lebih rendah dari pendekatan tf*idf .

Wan and Angryk [12] menggunakan context vectors untuk membantu merepresentasikan arti kata dalam WordNet. Pengukuran kesamaan kata dalam penelitian ini menggabungkan konsep gloss information dengan semantic. Di Tahun yang sama Alvarez dan Lim [13] mengusulkan algoritma baru dari penelitian sebelumnya untuk menghitung kesamaan semantik antara dua kata yang disebut *Semantic Similarity Algorithm* (SSA). Penelitian ini memodelkan kesamaan semantik antara kata-kata sebagai grafik dan menggabungkan hubungan "sinonim", "is-a", dan "part-whole" ke dalam grafik. Untuk meningkatkan akurasi, Alvarez dan Lim mengintegrasikan skema pembobotan dan menilai tingkat persimpangan antar kata. Hasil penelitian yang dilakukan oleh Alvarez dan Lim menunjukkan korelasi antara hasil SSA dan peneliaian manusia sebesar 0.903. Korelasi SSA yang tinggi dengan penilaian manusia menunjukkan bahwa SSA akan mudah dalam menyelesaikan beberapa masalah pengumpulan data dan pencarian informasi.

Ahsaee and Naghibzadeh [14] menyajikan sebuah model baru yang merupakan perbaikan dari penelitan sebelumnya seperti yang dilakukan oleh Rada, et al. [4].

Tabel 1. Klasifikasi ukuran kesamaan semantic antar kata

Peneliti	Metode yang digunakan		
[4]	Distance based method	Information content based method	Distance and Information Content based method Jiang
[11]	Penelitian yang dilakukan oleh Rada, et al., menggunakan jarak antara simpul pada ontology hirarkis untuk mengukur kesamaan kata.		
[14]	- Dalam makalah ini menggunakan pendekatan terhadap IR berdasarkan perhitungan jarak semantik antara konsep atau kata dan menggunakan jarak kata ini untuk menghitung kemiripan antara kueri dan dokumen. - Nilai jarak sangat bergantung pada hirarki jaringan yang dibangun secara subyektif. Hal ini dikarenakan tujuan utama perancangan WordNet bukan untuk tujuan perhitungan kesaaman semantic.		
[12]	Dalam penelitian ini menggunakan Particle Swarm Optimization untuk memperbaiki pengukuran kesamaan kata menggunakan perhitungan jarak pada WordNet.		
[13]	Menggunakan context vectors untuk membantu merepresentasikan arti kata dalam WordNet		
[15]	Menggunakan *Semantic Similarity Algorithm* menghitung kesamaan semantikantara dua kata		
[16]		Melakukan evaluasai *Information Content* berdasarkan HowNet untuk mengetahui kesamaan semantik antara dua istilah yang sama dan kata aslinya	
[17]		Menetapkan ukuran kesamaan semantik dalam sebuah IS-A taksonomi berdasarkan *information content*. Evaluasi eksperimental dilakukan dengan menggunakan korpus	
[19]		Menggunakan cara konvensional dalam mengukur *Information Content.* Pengukuran IC menggunakan pengetahuan tentang struktur hierarkis dari sebuah ontologi seperti WordNet dengan statistik dalam teks yang berasal dari corpus	
[18]		Menggunakan HowNet sebagai basis pengetahuan yang mendasai Liuling DAI et al dalam melakukan penelitian	
[6]		Mempresentasikan secara singkat tentang berbagai metode untuk menghitung *information content* pada	

		konsep ontologis	
[8]			Kombinasi pendekatan *edge-based* dari skema perhitungan edge dan pendekatan *node-based* dari perhitungan *information content*
[19]			Metode kesamaan kata dengan *path length, depth,* dan *local density*
[9]			Metode kesamaan kata dengan *path length* dan *depth*
[19]			*HowNet-based similarity measurement*

Model baru yang disajikan untuk menghitung kesamaan semantik kata dalam WordNet 2.1 dengan Particle Swarm Optimization untuk memperbaiki pengukuran kesamaan kata menggunakan perhitungan jarak. Hasil penelitian menunjukkan bahwa penambahan parameter pada fungsi transfer rumus kesamaan memiliki pengaruh positif terhadap korelasi kesamaan.

2.2. *Information Content Based Method*

You et al. [20] menunjukkan pendekatan berbasis node dan menentukan kesamaan konseptual yang disebut metode berbasis *Infomation Content*. Evaluasi Information Content yang dilakukan oleh You Bin berdasarkan HowNet untuk mengetahui kesamaan semantik dua istilah atau kata aslinya. Berbeda dengan metode konvensional yaitu berdasarkan korpus dan WordNet. Seco et al. [17] mengusulkan bahwa WordNet dapat digunakan sebagai sumber statistik tanpa menggunakan korpus. Namun, WordNet adalah kamus bahasa Inggris, dan itu tidak bisa membantu dalam mengevaluasi kesamaan semantik antara kata-kata cina [17]. Terbukti dalam penelitian yang dilakukan You Bin dengan eksperimen menghitung IC berdasarkan HowNet, bukan menghitung berdasarkan jarak sememe, dapat mendekati nilai yang dilakukan oleh manusia dan HowNet dapat mengekspresikan kata-kata bahasa inggris dengan baik.

Penelitian lain yang dilakukan oleh Resnik [16] yaitu menetapkan ukuran kesamaan semantik dalam sebuah IS-A taksonomi berdasarkan gagasan dari *information content*. Pada penelitian ini evaluasi eksperimental dilakukan dengan menggunakan korpus yang dibangun secara independen, taksonomi yang dibangun secara independen dan human subject data yang ada sebelumnya, dan hasilnya menunjukkan bahwa ukuran kesamaan semantik dalam sebuah taksonomi jauh lebih baik daripada pendekatan yang lain. kesamaan semantik yang diukur dengan IC, terbukti bermanfaat dalam menyelesaikan 2 bentuk ambiguitas linguistik, yaitu sintaksis dan semantik.

Seco et al. [17] membahas cara konvensional mengukur Information Content of Word Senses dengan menggunakan pengetahuan tentang struktur hierarkis dari sebuah ontologi seperti WordNet dengan statistik dalam teks yang berasal dari corpus. Dalam tulisannya Nuno menyajikan intrinsik pengukuran IC dengan menggunakan struktur hirarkis. Sedangkan ekstrinsik ukuran IC menggunakan analisis corpus. Hasil yang diperoleh menggunakan nilai IC dalam rumus teorema informasi tampaknya telah mengungguli homolog yang memberi kesan bahwa asumsi awal mengenai struktur taksonomi WordNet adalah benar. Perlu dicatat bahwa nilai maksimal yang didapat, menggunakan formulasi Jiang dan Conrath, sangat dekat dengan apa yang diusulkan oleh Resnik [16] sebagai computational upper bound. Satu keuntungan utama dari pendekatan ini adalah bahwa nilai IC dalam rumus teorema informasi tidak bergantung pada analisis korpora. Dengan demikian kita menghindari masalah data yang tidak lengkap yang banyak terjadi pada pendekatan berbasis korpus.

Dai et al. [9] juga meneliti tentang algoritma untuk mengukur kesamaan semantik antara dua kata. Liung Dai menggunakan HowNet sebagai basis pengetahuan yang mendasarinya dalam melakukan penelitian. Penelitian ini menunjuakan konsep dari sebuah kata melalui sebuah grafik konsep sesuai dengan definisi HowNet. Pada penelitiannya dilakukan proses perhitungan kesamaan antar kosep dengan mengukur kesamaan antar grafik konsep. Hal tersebut dilakukan untuk menagani penyimpangan kata atau kata yang tidak baku dalam HowNet. Tesaurus diadopsi selama pengukuran kesamaan kata. Pada penelitiannya Liuling melakukan eksperimen pada pasangan kata dan mengambil rating kesamaan manusia sebagai baseline untuk mengevaluasi metode. Hasil percobaan menunjukkan bahwa algoritma yang diusulakan oleh Liuling et al dapat bersaing dengan karya berbasis WordNet.

Banu et al. [18] mempresentasikan secara singkat tentang berbagai metode untuk menghitung information content pada konsep ontologis. Ayesha menjelaskan delapan informasi berbasis konten berdasarkan ukuran kesamaan semantik. Pada penelitiannya kemudian Ayesha memilih dua tindakan terbaik di antara delapan pengukuran tersebut. Dua langkah tersebut adalah langkah yang dilakukan oleh Lin dan Jiang, dan Conrath, karena langkah tersebut menunjukkan tingkat korelasi yang lebih tinggi yaitu memiliki peringkat korelasi 1 atau 2. Banu et al. [18] mengusulkan pengukuran dan algoritma baru untuk menghitung kesamaan semantik. Pada dua dataset yang diteliti membuktikan bahwa

ukuran tersebut menunjukkan hasil yang lebih baik. Pengukuran baru yang diusulkan ini dapat diterapkan dalam kategori kesamaan semantik hibrida yaitu pengukuran berbasis tepi dan informasi.

2.3. *Distance and information content based method*

Sejak tahun 1989, banyak peneliti yang melakukan riset pengukuran semantic antara kata berdasarkan pendekatan. Salah seorang peneliti bernama Jiang dan Conrath [6] mengusulkan suatu pendekatan yang berbeda dari penelitian sebelumnya. Jiang dan Conrath mengajukan sebuah pendekatan untuk mengukur kesamaan *semantic* antara kata dan konsep berdasarkan *corpus static* dan *lexical taxonomy*. Pengukuran yang diusulkan dengan menggabungkan pendekatan *edge-based (distance)* dari skema penghitungan tepi dan disempurnakan dengan pendekatan *node-based (information content)* dari pengukuran *information content.* Data uji yang digunakan berupa *benchmark* yang berdasarkan penilaian manusia.

Pada tahun 2003, Li, et al. [8] melakukan riset pengukuran *semantic* antara kata menggunakan beberapa *information source,* seperti *structural semantic information* dari *lexical taxonomy* dan *information conten*t dari corpus.. Alasan Li, dkk. melakukan riset ini, karena setiap metode yang berbeda menggunakan *information source* yang berbeda juga, sehingga hasil penelitian berada pada tingkat kinerja yang berbeda, misalnya penelitian yang dilakukan oleh Jiang dan Conrath [6] berbeda dengan penelitian yang dilakukan Resnik [5]. Li, et al. mengusulkan pengukuran baru dengan mengkombinasikan path length dan depth of subsumer. Hasil penelitian menunjukkan bahwa pengukuran yang diusulkan lebih baik dari pada pengukuran sebelumnya.

Pada tahun 2006, Li, et al. [19] mengusulkan metode pengukuran antara kata yang diterapkan untuk menghitung similarity antara kalimat. Proses penghitungan antara kalimat harus menghitung *similarity* antara kata terlebih dahulu. Penghitungan *similarity* antara kata dilakukan dengan menghitung *path length* dan *depth of subsumer*.

Penelitian yang dilakukan oleh Dai et al. [9] berbeda dari penelitian sebelumnya. Perhitungan *Lexical semantic* pada penelitian sebelumnya menggunakan WordNet. Pada penelitian yang dilakukan oleh Dai dan Liu menggunkan HowNet sebagai pengganti WordNet. Selain itu, juga dihitung kesamaan antara konsep dengan mengukur kesamaan antara graph concept.

4. Simpulan dan Saran

Ukuran kesamaan semantik adalah proses yang memerlukan keterlibatan beberapa disiplin ilmu, seperti bahasa, komputer, matematika logik dan domain yang bersangkutan. Klasifikasi metode ukuran kesamaan semantic antar kata berupa *distance based method, information content based method*, dan *distance and information content based method*. Penelitian berikutnya diharapkan dapat menggabungkan WordNet, Corpus, dan HowNet sebagai database untuk pengukuran kesamaan semantik antar kata.

Daftar Pustaka

[1] T. Slimani, 2013. Description and Evaluation of Semantic Similarity Measures Approaches. International Journal of Computer Applications, 80(10), pp.25–33.

[2] L.Y. Banowosari, 2006. Tinjauan Similaritas Semantik Dalam Pemilihan Ontology pada Peer-To-Peer (P2P).

[3] M. Kathuria, C. K. Nagpal & N. Duhan, 2016. A Survey Of Semantic Similarity Measuring Techniques For Information Retrieval. 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp.3435–3440.

[4] R. Rada, H. Milli, E. Bicknell, M. Blettner, 1989. Development and Application of a Metric on Semantic Nets. IEEE Transactions on Systems, Man and Cybernetics, 19(1), pp.17–30.

[5] P. Resnik, 1995. Using Information Content to Evaluate Semantic Similarity in a Taxonomy. , 1.

[6] J.J. Jiang & D.W. Conrath, 1997. Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In Proceedings of International Conference on Research in Computational Linguistics.

[7] A. Maind, A. Deorankar & P. Chatur, 2012. Measurement of Semantic Similarity Between Words: a Survey. International Journal of Computer Science, Engineering & Informa;,, 2(6), p.p51.

[8] Y. Li., Z.A. Bandar, & D. McLean, 2003. An approach for measuring semantic similarity between words using multiple information sources. IEEE Transactions on Knowledge and Data Engineering, 15(4), pp.871–882.

[9] L. Dai, B. Liu, Y. Xia dan S. Wu, 2008. Measuring Semantic Similarity Between Words Using HowNet. Proceedings of the International Conference on Computer Science and Information Technology, ICCSIT 2008, pp.601–605.

[10] A. Adhikari, S. Singh, A.Dutta, B. Dutta, 2015. A novel Information Theoretic Approach for Finding Semantic Similarity in WordNet. TENCON 2015 - 2015 IEEE Region 10 Conference, pp.1–6.

[11] R. Richardson & Smeaton, A..1995. Using WordNet in a Knowledge-Based Approach to Information Retrieval.

[12] S. Wan & R. A. Angryk, 2007. Measuring semantic similarity using WordNet-based context vectors. Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics, pp.908–913.

[13] M.A. Alvarez,. & S. Lim, 2007. A Graph Modeling of Semantic Similarity between Words. International Conference on Semantic Computing, pp.355–362.

[14] M.G. Ahsaee, & M. Naghibzadeh, 2010. Using WordNet to determine semantic similarity of words. , pp.1019–1027.

[15] B. You, L. Xiao-ran, L. Ning dan Y. Yue-song, (2012) 'Using information content to evaluate semantic similarity on HowNet', Proceedings of the 2012 8th International Conference on Computational Intelligence and Security, CIS 2012, pp. 142–145. doi: 10.1109/CIS.2012.39.

[16] P. Resnik, 1999. Semantic Similarity i n a T axonomy: An Information-Based Measure and its Application to Problems of Ambiguity i n Natural Language. Journal of Artiicial Intelligence Research Submitted, 11(3398), pp.95–130.

[17] N. Seco, T. Veale & J. Hayes, 2004. An Intrinsic Information Content Metric for Semantic Similarity in WordNet. Ecai, 16(Ic), p.1089.

[18] A. Banu, S.S. Fatima, K.U.R. Khan, 2015. Information Content Based Semantic Similarity Measure for Concepts Subsumed By Multiple Concepts. International Journal Web Applications Volume, 7(3), pp.85–94.

[19] Y. Li, D. McLean, Z.A. Bandar, J. D. O'Shea & K. Crockett, 2006. Sentence Similarity Based on Semantic Nets and Corpus Statistics. IEEE Transactions on Knowledge and Data Engineering, 18(8), pp.1138–1150.

[20] B. You, X.R. Liu, N. Li, Y.S. Yan 2012. Using information content to evaluate semantic similarity on HowNet. Proceedings of the 2012 8th International Conference on Computational Intelligence and Security, CIS 2012, pp.142–145.