ALGORITMA KELIP-KELIP DENGAN PENAMBAH BAIKAN PENGEKSTRAK FITUR UNTUK PEMBELAJARAN TAKSONOMI DARI TEKS MELAYU

TRI BASUKI KURNIAWAN

UNIVERSITI KEBANGSAAN MALAYSIA

ALGORITMA KELIP-KELIP DENGAN PENAMBAH BAIKAN PENGEKSTRAK FITUR UNTUK PEMBELAJARAN TAKSONOMI DARI TEKS MELAYU

TRI BASUKI KURNIAWAN

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEHI
IJAZAH DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2018

PENGAKUAN

Saya akui karya ini adalah hasil kerja saya sendiri kecuali nukilan dan ringkasan yang tiap-tiap satunya telah saya jelaskan sumbernya.

08 July 2018

TRI BASUKI KURNIAWAN
P54418

PENGHARGAAN

Pertama dan yang utama sekali saya panjatkan rasa syukur kehadirat Allah SWT, yang dengan karuniaNya dapat terselesaikan tesis ini.

Ucapan terima kasih yang tak terhingga disampaikan kepada Assoc Prof. Dr. Mohd Zakree Ahmad Nazri, sebagai penyelia, suadara dan sahabat yang telah memberikan bantuan berupa dukungan dan motivasi yang tak dapat diungkapkan dengan kata-kata. Juga kepada Prof. Dr. Abdul Razak Hamdan, atas dukungan dan bantuannya yang tak terhitung, mula dari awal saya datang ke UKM lagi.

Juga kepada semua sahabat dan kawan seperjuangan yang tak dapat disebutkan namanya satu-persatu disini.

Terakhir, ungkapan syukur kepada keluarga besar saya yang dengan sabar telah melayani, menemani, menghibur dan menyemangati perjalan panjang ini. Insya Allah kita akan sampai pada satu perhentian sementara, yang akan memberikan kita waktu untuk sedikit bernapas lega. Namun perjalanan masih akan panjang, terutama kepada anak-anakku. Dengan cinta kasih sayang dari Ananda lah ‘abi’ dapat terus bertahan.

ABSTRAK

Taksonomi adalah menyusun konsep suatu domain pengetahuan, Membina dan menyelenggara taksonomi secara manual adalah tugas membosankan dan memakan masa. Banyak taksonomi buatan telah dibina berasaskan kedua-dua domain terbuka berikut contohnya, WordNet dan domain khusus (mis., MeSH, untuk bidang perubatan). Namun, pengetahuan sentiasa berubah dan berkembang. Akibatnya, walaupun untuk membangunkan taksonomi domain khusus, taksonomi hasil buatan tangan (manual) tidak dapat dielakkan menjadi punca kekurangan liputan, dan mahal untuk terus dikemas kini. Ini telahmenarik minat penyelidikan dalam mempelajari teks taksonomi secara automatik dari teks. Tesis ini mencadangkan algoritma kelip-kelip yang tidak dikawal selia untuk mempelajari taksonomi secara automatik dari awal dengan menganalisis korpus teks Melayu yang diberikan. Pendekatan penyelidikan ini direka unutk menangani masalah kejarangan data, sehingga dapat secara efektif mendorong taksonomi bahkan dari korpora Melayu kecil. Tesis ini menyumbang tiga sumbangan penting. Pertama, ia melakukan penaakulan untuk membina taksonomi berdasarkan kaedah algoritma kelip-kelip sebagai kaedah pengelompokan berhierarki dan hipotesis distribusi Harris, yang dapat menangkap hubungan di antara konsep. Kedua, tesis ini mencadangkan algoritma berasaskan kelip-kelip untuk membina hubungan taksonomi. Dan ketiga, membina pelombong teks berdasarkan prinsip pindah pembelajaran, Penguraian Nilai Tunggal dan Pengindeksan Semantik Pendam. Penilaian empirik terhadap tiga teks Melayu menunjukkan kegunaan pendekatan yang dicadangkan. Secara empiris, kami membandingkan model yang dicadangkan bernama AKK-PD dan AKK-PT kepada satu set pendekatan terkini ke atas tiga domain yang berbeza iaitu Fiqh, Teknologi Malumat dan Biokimia. Keputusan menunjukan bahwa AKK-PD dan AKK-PT adalah lebih baik dari pendekatan yang dibandingkan pada korpora yang diuji. Di samping itu, kajian kami menunjukkan bahwa (i) algoritma pengelompokan berhierarki berasaskan Kelip-kelip meningkatkan ukuran-F Tindanan Taksonomi (F_{TO}) dan (ii) strategi perlombongan teks kami meningkat ketepatan leksikal dan ketepatan taksonomi.

FIREFLY ALGORITHM WITH IMPROVED FEATURE EXTRACTION FOR TAXONOMY LEARNING FROM MALAY TEXTS

ABSTRACT

Taxonomies hierarchically organize concepts in a domain. Building and maintaining them by hand is a tedious and time-consuming task. Many handcrafted taxonomies have been built that capture both open-domain (e.g., WordNet) and domain-specific (e.g., MeSH, for the medical domain) knowledge. Yet, our knowledge is constantly evolving and expanding. Consequently, even domain-specific, handcrafted taxonomies inevitably lack coverage, and ara expensive to keep up-to-date. This thesis proposes a new unsupervised Firefly algorithm for automatically learning taxonomy from scratch by analyzing a given Malay text corpus. Our approach is designed to deal with data sparseness, so it can effectively induce taxonomies even from small Malay corpora. The thesis makes three important contributions. First, it performs inference based on a hybridized hierarchical clustering and Harris distribu-tion hypothesis, which can capture links among concepts. Second, this thesis proposed a new firefly based algoritm to construct is-a relationship (taxonomy). And third, a text mining based on transfer learning principle, Singular Value Decomposition and Latent Semantic Indexing. An empirical evaluation on three Malay text demonstrates the utility of our proposed approach. Empirically, we compare our proposed model named AKK-PD and AKK-PT to a set of state-of-the-art approaches on three different domain which are Fiqh (Islamic jurisprudence), Information Technology (IT) and Biochemical. We found that AKK-PD and AKK-PT outperforms them on the tested corpora. Additional, our study shows that (i) our hierarchical clustering algorithm based on Firefly increases the F-measure Taxonomy Overlaps (F_{TO}) and (iii) our text mining strategy increases both lexical recall and taxonomic overlap precision.

KANDUNGAN

Halaman

SENARAI JADUAL

SENARAI ILUSTRASI

SENARAI SINGKATAN

Penggunaan dalam Bahasa Melayu

AKK	Algoritma Kelip-kelip
AKK-PD	Algoritma Kelip-kelip Pembagi Dua
AKK-PT	Algoritma Kelip-kelip Pengelompokan Taksonomi
PDSC	Pembagi Dua Sama Caraballo
PTGS	Perlombongan Teks berasaskan Google dan SVD
PTLSI	Perlombongan Teks berasaskan LSI
PTDT	Pembelajaran Taksonomi daripada Teks

Penggunaan dalam Bahasa Inggeris

HO	Hypermym Oracle
HAC	Hierachical Algomerative Clustering
GAHC	Guided Algomerative Hierarchical Clustering
FCA	Formal Concept Analysis
PSO	Particle Swarm Optimization
NLP	Natural Language Processing
GTM	Google Text Miner
AIS	Artificial Immune System
SVD	Single Value Decomposition
LSI	Latent Semantic Indexing

BAB I

PENDAHULUAN

1.1 PENGENALAN

Antara tugas penting dalam pembangunan Web Semantik adalah mewakil maklumat dalam bentuk yang berstruktur agar komputer boleh "memahami" maksud suatu perkataan dari kandungan halaman Web dan seterusnya menyelesaikan masalah yang kompleks seperti penaakulan atau pun penaabiran. Ontologi adalah satu daripada komponen utama Web Semantik. Ontologi menyediakan perbendaharaan kata suatu domain berkepentingan. Ontologi juga menerangkan sifat suatu istilah atau perkataan supaya mesin, aplikasi atau perkhidmatan boleh berkongsi maklumat dan pengetahuan dengan berkesan, sekaligus memastikan keantarakendalian antara mesin.

Walau bagaimanapun, kemajuan ke arah Web Semantik diperlahankan oleh masalah kesesakan pemerolehan pengetahuan. Pemodelan ontologi secara manual, digambarkan ramai penyelidik seperti Alesso dan Smith (2005) sebagai satu proses yang berintensifkan buruh, membosankan, kompleks, memakan masa dan mahal. Maka ramai penyelidik cuba memeroleh ontologi secara automatik seperti yang diusahakan oleh Gulla dan Brasethvik (2008), Yeh dan Sie (2006) dan Cimiano (2006). Pemerolehan ontologi dari pelbagai sumber seperti kamus, tesauri, wordnet dan teks seperti laman Web tidak berstruktur, buku, majalah dan jurnal. Teks merupakan sumber pemerolehan ontologi terbesar di dunia.

Antara komponen asas ontologi adalah taksonomi. Taksanomi juga dikenali sebagai hierarki konsep atau tesauri. Skop tesis ini diterhadkan kepada memeroleh taksonomi secara automatik dari teks bahasa Melayu kerana penggunaan teks Melayu sebagai sumber pemerolehan taksonomi masih kurang digunakan dalam kajian.

Kebanyakan penyelidik telah berusaha untuk membangunkan taksonomi secara automatik dari teks dengan menggunakan kaedah pembelajaran mesin, analisis statistik dan pemprosesan bahasa tabie. Proses pemerolehan pengetahuan automatik untuk mewujudkan taksonomi dipanggil pembelajaran taksonomi. Zakree (2011) telah menakrifkan usaha untuk memeroleh taksonomi secara automatik atau pun separa automatik dari teks sebagai pembelajaran taksonomi daripada teks (PTDT).

PTDT bahasa Inggeris bukanlah suatu bidang penyelidikan yang baharu. Tetapi PTDT Melayu secara relatifnya masih baharu dan yang terkini adalah kajian Zakree (2011). Kajian beliau telah menggunakan pelbagai fitur atau ciri serta pola bahasa Inggeris seperti pola Heaest (1992) yang boleh digunakan untuk mengekstrak taksonomi dari teks Melayu. Beliau menggunakan kaedah pembelajaran mesin berasaskan sistem imun buatan untuk memeroleh pengetahuan. Namun, Zakree (2011) membuat kesimpulan bahawa terdapat banyak ruang penambahbaikan untuk penyelidikan PTDT Melayu kerana kesukaran memeroleh alat pemprosesan Bahasa tabie Bahasa Melayu yang ampuh dan percuma untuk digunakan. Oleh yang demikian, tesis ini adalah berkenaan reka bentuk, pelaksanaan dan penilaian kaedah pengelompokan konsep berasaskan algoritma kelip-kelip untuk PTDT Melayu.

1.2 LATAR BELAKANG MASALAH

Web Semantik bergantung kepada ketersediaan sumber semantik yang luas iaitu ontologi. PTDT adalah antara inovasi yang dicadangkan ke arah penciptaan ontologi secara semi atau automatik sepenuhnya. PTDT adalah berkenaan penggunaan perbendaharaan kata yang diekstrak daripada teks dan perkataan tersebut dilombong menggunakan teknik perlombongan data seperti pengelompokan data.

Menurut Camino (2006), usaha memeroleh pengetahuan tersirat yang terkandung di dalam teks adalah satu cabaran yang besar. PTDT asasnya merupakan perlombongan teks termaju yang bertujuan untuk mengenal pasti struktur pengetahuan dalam bentuk taksonomi atau hierarki. Setiap perkataan yang diperolehi daripada teks mempunyai tahap (hierarki) yang berbeza-beza. Kaedah lazim adalah dengan mengeksploitasi pola-sintaktik lexico seperti 'komputer adalah mesin'. Pola ini menunjukkan hubungan taksonomi atau pola-sintaktik lexico antara komputer dan

mesin. Komputer adalah hyponim manakala mesin adalah hypernim. Walau bagaimanapun Cimiano (2006) mendapati, jika sesuatu teks bergenre teknikal dan khusus, maka semakin kurang pola tersebut digunakan. Penggunaan alat pemprosesan bahasa tabie (*Natural Language Processing*, NLP) digunakan dengan meluas oleh saintis barat dalam PTDT Bahasa Inggeris oleh kerana alatan NLP Inggeris termaju boleh diperolehi dengan mudah. Walau bagaimanapun, tidak bagi PTDT Melayu.

Menurut Zakree (2011), alatan NLP Melayu masih belum matang kerana kebanyakan alatan tersebut diuji pada korpus genre tertentu sahaja. Walaupun bahasa Melayu mempunyai struktur tatabahasa subjek-kata_kerja–objek seperti Bahasa Inggeris, tatabahasa Melayu adalah berbeza dari bahasa Inggeris (Azhar 1988). Oleh yang demikian, kajian NLP Melayu berkembang perlahan kerana memerlukan penyelidikan fundamental yang berbeza dengan NLP Inggeris. Kesannya ialah tiada alatan NLP Melayu umum yang boleh digunakan dengan berkesan pada apa sahaja genre seperti teks al-Quran, penulisan teknikal dan novel remaja. Malahan, alatan NLP yang dikatakan berkesan tidak dapat digunakan kerana sukar mendapat kebenaran untuk menggunakanya. Oleh yang demikian, penyelidikan PTDT Melayu sangat jarang ditemui dalam kesusasteraan penyelidikan kecuali hasil penyelidikan Zakree (2008, 2009, 2010 dan 2011) serta Saidah (2008). Kedua-dua penyelidik ini tidak bergantung kepada alat NLP Melayu sepenuhnya kerana menggunakan kaedah pengelompokan konsep berasaskan hipotesis pengagihan Harris (1954).

Hipotesis Harris (1954) menganggap bahawa perkataan dikatakan sebagai serupa atau sinonim atau memiliki maksud (semantik) yang sama dengan merujuk kepada sejauh mana perkataan berkenaan berkongsi konteks linguistik yang serupa. Dalam erti kata lain, hipotesis Harris menyatakan bahawa kesepadanan atau makna suatu perkataan boleh diagak dengan melihat kepada cara penggunaan perkataan berkenaan di dalam teks. Kaedah PTDT berasaskan hipotesis Harris (1954) dilihat sebagai perlombongan teks (Cimiano, 2006; Zakree, 2011) kerana tugas ini melibatkan pengekstrakan data (perkataan) dari teks sebelum diproses (atau dilombong) untuk mendapatkan pola yang tidak diketahui.

Perlombongan teks ini dianggap oleh Cimiano (2006) sebagai satu proses kejuruteraan terbalik kerana ia merupakan satu tugas membina semula model dunia penulis dokumen yang digunakan untuk memeroleh taksonomi. Tugas ini adalah kompleks dan mencabar kerana dua sebab utama. Pertama, terdapat hanya sebahagian kecil pengetahuan domain penulis tersurat di dalam teks. Kedua, domain pengetahuan dalam teks jarang disebut dengan jelas, kecuali di dalam kamus. Oleh yang demikian, pembelajaran taksonomi dari teks itu boleh dilihat sebagai bekerja dengan 'simbol yang tidak diketahui' dan mengalami masalah kejarangan data yang serius.

Pelbagai percubaan untuk mengatasi kejarangan data telah dibuat seperti yang dapat dilihat di Zakree (2011), Buitelaar et al. (2003), Cimiano et al. (2004a; 2004b), Reinberger dan Spyns (2005) dan Blohm dan Cimiano (2007). Cimiano (2005) sebagai contoh, memperkenalkan algoritma Pengelompokan Agglomerat Hierarki Berpandu (*Guided Algomerative Hierachical Clustering*, GAHC) dan Analisis Konsep formal (*Formal Concept Analysis*, FCA). GAHC mengeksploitasi WordNet dan *Hypernym Oracle* (HO) untuk membimbing proses kelompok dalam usaha untuk mewujudkan kelompok yang munasabah walaupun tanpa data yang cukup. HO mengandungi pasangan hypernim/hyponim. GAHC menghasilkan keputusan yang lebih baik berbanding dengan lain-lain teknik tanpa pengawasan seperti HAC (*Hierarchical Algomerative Clustering*) atau FCA (Cimiano, 2006). Teknik pelicinan digunakan dalam usaha untuk mengatasi kejarangan data dengan menggunakan kebarangkalian bersyarat untuk menapis kata yang diekstrak sebelum persamaan antara perkataan diukur menggunakan metrik kosinus. Zakree (2011) telah membina HO bahasa Melayu dalam kajiannya. Walaubagaimana pun, pendekatan yang berdasarkan corak sepadan seperti yang digunakan dalam GAHC memperoleh ukuran perolehan kembali (recall) yang sangat rendah. Keputusan ini diperolehi disebabkan pola-sintaktik lexico adalah sangat jarang berlaku terutamanya apabila genre teks adalah teknikal (Cimiano, 2006).

Selain daripada membina HO, WordNet merupakan sumber yang digunakan oleh penyelidik barat dalam membina taksonomi dan ontologi. Tetapi menurut Lim dan Hussein (2006), WordNet mempunyai masalah kerana sering takrifan atau hierarki suatu perkataan di WordNet hanyalah takrifan umum manakala semantiknya

dalam konteks domain yang lebih spesifik sering diabaikan. Zakree (2011) telah menguji WordNet dengan teliti dan beliau mendapati WordNet tidak sesuai digunakan dalam masalah PTDT Melayu. Ini adalah kerana terjemahan dilakukan tidak semata-mata sebagai satu tindakan pemindahan linguistik, tetapi ia juga melibatkan interaksi budaya dan pemindahan budaya (Elkateb et al., 2006). Sebagai contoh, perkataan 'nabi wanita' wujud di WordNet yang sudah pasti tidak diterima oleh budaya Melayu dan penganut Islam di Malaysia khususnya. Selain itu, teknik melicinkan digunakan oleh Cimiano (2005) tidak meningkatkan hasil pendekatan berasaskan FCA (Ryu et al., 2006). Kesimpulannya, PTDT Melayu tidak boleh bergantung kepada WordNet untuk mencari hypernim dan hyponim, melainkan untuk domain yang spesifik yang tidak punya perbezaan dari aspek budaya dan kepercayaan.

Sebagai alternatif lain, Fortuna et al. (2006) menggunakan Pengindeksan Semantik Terpendam (*Latent Semantic Indexing*, LSI) dan algoritma pengelompokan K-means untuk membangunkan ontologi secara semi-automatik. Fortuna et al. (2006) mencadangkan teknik perlombongan teks untuk menemui topik dalam korpus, berasaskan kaedah LSI dan K-Means. Mereka menggabungkan semua kaedah ini ke dalam sistem interaktif untuk membina topik ontologi. Kajian beliau masih separa-automatik dan satu peluang penyelidikan disini ialah untuk menjawab persoalan bagaimana mengurangkan keperluan untuk berinteraksi dengan pengguna untuk membangunkan ontologi (dalam tesis ini, taksonomi). Walau pun LSI telah lama diperkenalkan, ia masih digunakan dalam penyelidikan pembangunan ontologi misalnya oleh Rani et al. (2017).

Dalam kaedahnya, LSI menggunakan teknik dari algebra linear yang dikenali sebagai Penguraian Nilai Tunggal (*Singular Value Decomposition*, SVD) dan perwakilan bag-of-words yang mewakili suatu dokumen untuk mengekstrak perkataan dengan makna yang sama. Ini juga boleh dilihat sebagai pengekstrakan konsep semantik tersembunyi atau topik dari dokumen teks.

Pelbagai algoritma pengelompokan konsep telah digunakan untuk memeroleh taksonomi seperti Pengelompokan Aglomerat Berhierarki (HAC), algoritma memecahbelahkan seperti *Bisecting K-Mean*, Analisis Konsep Formal (FCA) dan

sistem imun buatan (AIS). FCA dan HAC telah menarik banyak perhatian sejak hasil pengelompokan boleh dibentangkan dalam bentuk struktur pokok atau kekisi. Kaedah berasaskan hipotesis pengagihan Harris (1954) ini berupaya memeroleh taksonomi daripada teks secara (semi) automatik. Walaupun kaedah ini telah menghasilkan keputusan yang baik, terdapat beberapa isu yang dihadapi oleh kaedah ini iaitu kejarangan data. Cimiano (2006) menyatakan bahawa sesetengah ciri-ciri kontekstual yang diekstrak dari teks adalah salah kerana kehilangan/ketidakcukupan data atau disebabkan kesilapan tatabahasa yang tidak disengajakan oleh pengarang. Oleh yang demikian, suatu perkataan akan dikelompokkan di dalam kelompok yang salah iaitu yang tidak sepadan dengan makna atau semantik sebenar. Malahan, pendekatan ini tidak berupaya melabel kelompok konsep yang diperolehi dengan nama konsep yang tepat.

Untuk mengatasi masalah ini, Zakree (2011) telah membuat kajian yang didorong oleh kepercayaan bahawa mana-mana sistem pembelajaran yang hendak digunakan di dalam PTDT Melayu mesti berupaya menyesuaikan (adaptasi) diri dengan masalah kejarangan data dan hingar. Sistem Imun Buatan (*Artificial Immune System*, AIS) telah menarik perhatian Zakree (2011) dan ramai penyelidik lain dalam berurusan dengan data yang tidak lengkap yang mengakibatkan masalah kejarangan data dan data hingar (iaitu data yang mengandungi ciri yang tidak betul). Kerja sebelumnya telah menunjukkan bahawa AIS, mempunyai ciri-ciri yang diperlukan untuk mengatasi masalah ini. Terdapat beberapa sifat AIS yang menjadi sumber inpirasi untuk menggunakan sistem imun untuk pembelajaran iaitu sifat ketegapan, kebolehpercayaan, pengiktirafan, kepelbagaian, ingatan, peraturan sendiri, dan pembelajaran (Dasgupta 1999).

Cabaran asas dalam pembelajaran taksonomi dari teks Melayu adalah pemilihan ciri-ciri (juga dipanggil fitur atau atribut) yang terbaik untuk mewakili konsep. Dengan andaian bahawa Bahasa Inggeris yang sedia ada mempunyai ciri-ciri sintaksis dan corak lexico-sintaktik yang boleh bekerja dalam PTDT Melayu, kaedah pengelompokan yang sedia ada dan pendekatan yang dicadangkan oleh Zakree (2011) boleh digunakan untuk PTDT Melayu. Zakree (2011) telah membangunkan tiga kaedah PTDT Melayu seperti berikut:

1) Hibridisasi antara GAHC dan aiNet yang dipanggil GCAINT (Pengelompokan Berpandu dan aiNet untuk Pembelajaran Taksonomi).
2) Penggunaan algorithm CLONALG dan Bi-Secting K-Mean untuk pembelajaran taksonomi yang dipanggil CLOSAT.
3) Pembangunan enjin pengekstrakan fitur berasaskan enjin Google untuk mengkayakan dan memperluaskan konteks setiap kata yang dipanggil GTM (*Google Text Miner*).

Percubaan pertama Zakree (2011) adalah dengan membangunkan GCAINT iaitu dengan menyediakan HO Melayu yang menyimpan pasangan hypernim-hyponim Melayu yang diperolehi dari teks Melayu. Apabila selesai memperoleh taksonomi berasaskan HO Melayu, barulah kaedah pengelompokan konsep berasaskan AIS digunakan untuk menghasilkan taksonomi. Taksonomi yang dihasilkan oleh AIS dan HO Melayu digabungkan. Kajian Zakree (2011) menggunakan set teks Melayu yang mengandungi tiga genre berbeza iaitu teks fekah, teks teknologi maklumat (IT) dan biokimia. Teks biokimia menghasilkan set data yang paling jarang di antara ketiga-tiga teks ini dan merupakan teks akademik yang teknikal. Seperti yang dibimbangi (Cimiano 2006), teks bergenre teknikal mempunyai pola sintaktik lexico yang paling sedikit sekali. Oleh itu, GCAINT berjaya menghasilkan keputusan yang lebih baik berbanding GAHC dan HAC pada teks teknologi maklumat dan fekah sahaja. Jelas, GCAINT tidak dapat menangani teks biokimia kerana masalah kejarangan data dan pasangan hypernim-hyponim tidak banyak ditemui.

Tesis ini berandaian bahawa HO Melayu mungkin menjadi penyebab rendahnya kualiti taksonomi kerana HO Melayu dihasilkan dari pola linguistik dan tiada data (fitur) yang dikutip dari teks untuk membolehkan taksonomi yang terbina berasaskan HO dan pengelompokan konsep dapat dihasilkan dengan baik. Walau bagaimanapun, Zakree (2011) telah membangunkan CLOSAT iaitu suatu kaedah yang dibangunkan tanpa bergantung kepada HO Melayu. CLOSAT memeroleh taksonomi dengan kualiti yang lebih baik untuk teks biokimia namun agak sama kualitinya dengan GCAINT bagi teks fekah dan Teknologi Maklumat. Kajian ini membuktikan bahawa penggunaan sistem imun buatan (AIS) dapat mengatasi masalah kejarangan data kerana operator pengklonan, mutasi dan pemilihan AIS. Secara umumnya, kualiti

taksonomi fekah dan teknologi maklumat yang dihasilkan oleh GCAINT adalah lebih baik berbanding CLOSAT.

Kekurangan utama pada CLOSAT adalah jumlah parameter yang begitu tinggi. Pada CLOSAT terdapat 8 parameter. Kualiti keputusan CLOSAT dan GCAINT menjadi lebih baik kerana parameternya telah ditala oleh algoritma Pengoptimuman Partikel Kerumunan (*Particle Swarm Optimization*, PSO). Justeru, kos pengkomputeran untuk menggunakan CLOSAT dan GCAINT adalah tinggi. Tesis ini berhasrat untuk membangunkan algoritma metaheuristic yang lebih 'santai' atau mudah digunakan untuk menghasilkan taksonomi dari teks Melayu.

Apabila GTM dibangunkan oleh Zakree (2011) untuk mengekspolitasi limpahan maklumat di Web untuk melengkapkan set data yang mengalami masalah kejarangan data. Walau bagaimanapun, setelah menggunakan set data yang diperkayakan oleh GTM pada GCAINT dan CLOSAT, kualiti ketiga-tiga taksonomi telah merudum jatuh. Masalah ini disebabkan oleh jumlah data (atribut) baru yang tidak diperlukan. GCAINT turut merosot kualitinya kerana GTM yang dibangunkan Zakree (2011) tidak mengekstrak hypernim dan hyponim untuk HO Melayu. Penggunaan GTM telah menghasilkan data yang 'terlebih' banyak bagi AIS untuk memprosesnya dan masalah ini dilihat bukan berpunca dari AIS tetapi kaedah pemilihan dan pengekstrakan fitur atau atribut (Zakree 2011) dan rekabentuk algoritma AIS sendiri. AIS sebelum ini telah dibuktikan berkesan untuk melakukan tugas pengelompokan, namun tidak pernah dimodelkan untuk melakukan tugas pengelompokan berhierarki. Oleh itu, Zakree (2011) telah menggunakan algoritma Bi-Secting K-Means untuk membina hierarki bagi setiap kelompok yang dihasilkan oleh AIS sebelum mencantumnya. Tesis ini berandaian bahawa ini adalah kelemahan utama algoritma berasaskan AIS untuk PTDT Melayu.

Kesimpulannya ialah kaedah pengelompokan konsep berasaskan AIS sangat baik untuk mengatasi masalah kejarangan data manakala kaedah HO adalah baik hanya jika terdapat banyak pola sintaktik lexico di dalam teks. Oleh yang demikian, sekali lagi teori Wolpert dan Macready (1997) iaitu *No Free Lunch Theorem* dilihat benar dalam kes ini. Mereka berdua menyatakan bahawa mana-mana dua algoritma

pengoptimuman adalah sama apabila purata prestasi kedua-dua algoritma tersebut diukur merentasi kesemua masalah yang mungkin. Secara ringkasnya, suatu algoritma mempunyai kekuatan pada masalah tertentu (seperti kejarangan data akut) dan suatu algoritma lain pula punya kelemahan pada masalah ini namun punya kekuatan untuk masalah lain. Zakree (2011) telah menjalankan kajian berdasarkan kepada andaian bahawa kaedah pembelajaran taksonomi sedia ada (bersumberkan penyelidikan teks barat) boleh digunakan untuk teks Melayu tetapi bagaimana sistem tersebut harus direalisasikan? Maka, kajian ini adalah untuk memperkukuhkan kaedah yang dipelopori Zakree (2011) untuk PTDT Melayu. Persoalannya masih sama iaitu bagaimana sistem cadangan tersebut harus direalisasikan?

Secara ringkasnya isu berikut yang menjadikan ini satu tugas yang sukar seperti yang diketengahkan oleh Cimiano (2005 & 2006) dan Zakree (2011) seperti berikut:

1) Kandungan teks lazimnya mengandungi hingar. Oleh itu ciri-ciri yang diekstrak mempunyai keberangkalian tinggi adalah salah, iaitu tidak semua ciri-ciri yang diekstrak adalah betul.
2) Tidak semua ciri-ciri yang diekstrak adalah 'relevan' dalam erti kata bahawa ciri-ciri yang yang diekstrak akan dapat membantu untuk membezakan antara objek yang berbeza.
3) Maklumat yang sempurna tidak akan dapat dipenuhi, iaitu koleksi teks tidak akan menjadi 'cukup besar' untuk mencari semua kejadian (data) yang mungkin (Zipf 1932); dan andaian
4) AIS tidak sesuai digunakan apabila suatu data tidak mengalami masalah kejarangan data (data tidak lengkap) atau keadaan yang set data diandaikan mencukupi atau hingar yang terlalu tinggi.
5) AIS mempunyai jumlah parameter yang banyak dan kos pengkomputeran menjadi tinggi apabila PSO digunakan untuk menala parameter.

1.3 PERNYATAAN MASALAH

Masalah PTDT Melayu boleh dinyatakan secara ringkas seperti berikut:

Cabarannya adalah untuk memeroleh taksonomi secara automatik dari set data yang diektrak dari teks Melayu. Set data yang diekstrak dari teks yang berbeza mengalami masalah kejarangan data dan hingar pada kadar yang berbeza. Oleh itu, suatu kaedah yang teguh yakni kaedah yang berupaya memeroleh taksonomi yang berkualiti tidak kira set data berkenaan mengalami masalah kejarangan data atau hingar data. Pada masa yang sama, kaedah cadangan berkenaan dapat mengatasi kelemahan kaedah pengelompokan konsep sedia ada. Kaedah cadangan mesti mampu menghasilkan taksonomi yang lebih baik dalam ukuran Ukuran-F Tindanan Taksonomik iaitu min diantara kejituan (precision) dan dapatan semula (recall).

Kaedah pengelompokan konsep sedia ada adalah berasaskan hipotesis pengagihan Harris (1968). Oleh itu, terdapat dua isu utama dalam PTDT yang perlu ditangani. Pertama, kaedah baru atau alternatif yang berupaya menangani masalah kejarangan dan hingar data yang ekstrem. Isu kedua ialah masalah pemilihan atribut yang melambakkan atribut yang tidak diperlukan.

Isu pertama dibangkitkan kerana fitur yang diekstrak bagi sesetengah kata atau konsep adalah salah yang menyebabkan hingar di dalam data. Manakala isu kejarangan data pula sentiasa menjadi isu dalam PTDT. Isu kedua ialah tentang bagaimana untuk menyediakan fitur bagi setiap kata dengan berkesan dan cekap.

Rangka kerja Pindah Pelajaran (PP) telah banyak mendapat perhatian penyelidik pembelajaran mesin khususnya dalam perlombongan teks. Penyelidik perlombongan teks berhadapan dengan masalah ketika melaksanakan tugas pembelajaran mesin dari teks disebabkan oleh kekurangan contoh (ciri/fitur) beranotasi yang berkualiti tinggi dan ‘cukup’ untuk membantu membina dan melatih model. Kerangka ini menawarkan penyelesaian yang menarik untuk masalah ini. Penyelidikan untuk tesis ini juga berhadapan dengan masalah yang sama.

Oleh kerana AIS merupakan sebuah algoritma metaheuristik dan berjaya mengatasi algoritma pengelompokan konsep yang lain, maka pencarian algoritma alternatif dalam kajian ini juga tertumpu kepada algoritma metaheuristik. Kajian literatur mendapati, algoritma kelip-kelip (AKK, *Firefly Algorithm*) yang secara relatifnya baru untuk tugas pengelompokan (Senthilnath et al. 2011), masih belum diuji dalam permasalahan PTDT dan pengelompokan berhierarki. Lebih-lebih lagi

AKK hanya memiliki dua parameter yang perlu ditala menjadikan AKK pilihan yang terbaik untuk dikaji.

Oleh yang demikian, adalah perlu untuk menjalankan kajian untuk menunjukkan bahawa algoritma AKK dan GTM yang dimajukan sesungguhnya dapat membuat perbezaan yang signifikan dalam membina model taksonomi yang lebih baik. Tesis ini cuba menjawab sama ada mekanisme AKK, dapat membina taksonomi yang lebih baik. "Lebih baik" itu bermaksud, ukuran F_{TO} yang lebih tinggi berbanding kaedah lain. Berikut adalah beberapa persoalan yang akan ditangani untuk menjawab pernyataan masalah di atas.

1) Model AKK hanya dimodelkan untuk tugas pengelompokan dan bukannya pengelompokan berhierarki atau pengelompokan konsep. Bagaimana AKK harus diubahsuai untuk PTDT Melayu?
2) Kaedah yang sedia (seperti *Artificial Immune System*, AIS) ada dapat menangani masalah kejarangan data namun lemah menangani hingar data, dapatkah AKK mengatasi kedua-dua masalah ini?
3) GTM dapat membantu mengatasi kejarangan data namun menyebabkan hingar data. Bagaimana GTM perlu diperbaiki?
4) Adakah pola lexico-sintaktik perlu dieksploitasi dalam taksonomi pembelajaran daripada teks Melayu dengan AKK?

Secara umumnya, persoalan kajian bolehlah dirumuskan sebagai berikut:

Bagaimana mengatasi masalah kejarangan dan hingar data dalam usaha memeroleh taksonomi dari teks Melayu yang berkualiti dengan mengunakan AKK dan GTM agar kualiti taksonomi yang diukur menggunakan F_{TO} (F-measure of taxonomic overlap) adalah lebih baik dari kaedah sedia ada?

Soalan penyelidikan ini adalah berasaskan kesan set data (yang diekstrak dari teks), set metrik dan teknik pemilihan dan pengekstrakan fitur. Berasaskan soalan-soalan di atas, maka berikut adalah hipotesis kajian ini:

1) Kaedah pembelajaran taksonomi berasaskan Algoritma Kelip-kelip (AKK) akan meningkatkan kualiti taksonomi berbanding kaedah lain yang dijadikan perbandingan.

2) Penggunaan kaedah Penguraian Nilai Tunggal (*Single Value Decomposition*, SVD) pada GTM dan pengekstrak fitur menggunakan kaedah Pengindeksan Semantik Terpendam (*Latent Semantic Indexing*, LSI) akan menghasilkan keputusan yang lebih baik berbanding kaedah lain.

1.4 MAKLAMAT PENYELIDIKAN

Tujuan kajian ini adalah untuk membangunkan kaedah pembelajaran taksonomi dari teks yang baharu berasaskan algoritma kelip-kelip dan meningkatkan keupayaan teknik pengekstrakan fitur dari teks Melayu untuk membentuk set data yang lebih baik berbanding teknik sedia ada agar kualiti taksonomi yang terhasil secara (separa) automatik adalah lebih baik. Matlamat penyelidikan ini adalah meningkatkan kualiti taksonomi yang terhasil dari penggunaan kaedah pembelajaran mesin untuk masalah TTDT Melayu untuk pembangunan ontologi secara (semi) automatik agar usaha pembangunan Web Semantik dapat dicapai dengan lebih pantas.

1.5 OBJEKTIF KAJIAN

Untuk mencapai tujuan dan matlamat penyelidikan ini, objektif-objektif yang perlu dicapai adalah seperti berikut:

1) Membangunkan model pembelajaran taksonomi hibrid berasaskan algoritma kelip-kelip dan kaedah pembahagi dua sama-caraballo.
2) Membangun model pembelajaran taksonomi berasaskan algoritma kelip-kelip untuk pengelompokan berhierarki
3) Membangunkan alat pelombong teks berasaskan konsep pindah pembelajaran (PP), antara Google Text Miner dan SVD.
4) Pemilihan fitur dari teks Melayu berasaskan kepada kaedah Pengindeks Semantik Terpendam (LSI) untuk membentuk matriks Harris.

1.6 SKOP PENELITIAN

Terdapat pelbagai pendekatan dan kaedah dari pelbagai spektrum bidang seperti analisi statistik, pembelajaran mesin dan pemrosesan bahasa tabie telah dicadangkan

untuk membina taksonomi secara (semi) automatik. Maedche dan Staab (2000) telah mengkategorikan PTDT kepada beberapa jenis, iaitu: i) pengesktrakan berasaskan pola; ii) seturuan peraturan; iii) pengelompokan konseptual; iv) pencantasan ontologi; dan v) pembelajaran konsep. Skop kajian ini tertumpu kepada pembangunan algoritma pengelompokan berhierarki berasaskan kardah metaheuristik yang boleh digunakan untuk pengelompokan konseptual atau pembangunan taksonomi. Kajian ini tertumpu kepada PTDT Melayu menggunakan algoritma diinspirasikan kelip-kelip (AKK) dan meningkatkan prestasi model berasaskan AKK ini dengan menghibrid AKK dengan kaerah pembahagi dua sama. Justeru, kajian ini dan metodologi tesis ini bertujuan memberikan sumbangan kepada pengkomputeran cerdas ataupun pengkomputeran lembut berbanding kejuruteran ontologi dan web semantik. Oleh yang demikian, kesan aplikasi taksonomi yang terhasil dalam kejuruteraan ontologi dan web semantik tidak diuji atau dikaji. Metodologi penyelidikan Cimiano (2006) adalah menjadi rujukan utama.

Kajian ini tidak melibatkan ujian mengenai kesan pemprosesan bahasa tabie Melayu kepada kualiti taksonomi tetapi mengkaji kesan kaedah pengekstrakan perlombongan teks berasaskan Google, LSI dan SVD kepada kualiti pengelompokan konseptual.

Kajian ini juga menguji keberkesanan pengekstrakan fitur berasaskan Google dan SVD ke atas algoritma yang dibangunkan oleh Zakree (2011). Hanya hubungan hypernim/hyponim (atau '*adalah'*) atau sub-konsep/super-konsep sahaja yang cuba diperolehi dari korpus. Hubungan yang selain daripada hypernim/hyponim adalah diluar skop kajian ini. Dalam usaha untuk mengekstrak ciri daripada teks, dua teknik akan digunakan dalam kajian ini. Teknik pergantungan sintaksis yang telah digunakan oleh Cimiano et al. (2006) dan pola lexico sintaktik yang diperkenalkan oleh Hearst (1992). Oleh yang demikian, skop kajian tesis ini boleh diringkaskan seperti berikut:

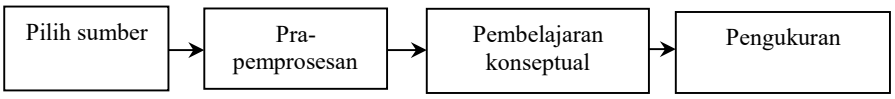
1) Korpus teks yang digunakan adalah korpus yang digunakan dalam kajian (Zakree 2011) yang terdiri daripada tiga domain berbeza iaitu Fekah, Biokimia, Teknologi Maklumat (IT).

2) Menggunakan SVD dalam meningkatkan keberkesanan pengekstrak fitur berasaskan Google untuk mengatasi masalah kejarangan data.
3) AKK akan dibandingkan dengan teknik pengelompokan berikut:
   a. Guided Agglomerative Hierarchical Clustering.
   b. Bi-Secting K-Means.
   c. Hierarchical Agglomerative Clustering: Single Linkage.
   d. Hierarchical Agglomerative Clustering: Average Linkage.
   e. Hierarchical Agglomerative Clustering: Complete Linkage.
   f. GCAINT
   g. CLOSAT

4) Kualiti taksonomi yang diperolehi akan diukur menggunakan ukuran yang diperkenalkan oleh (Cimiano 2006). Ukuran tersebut adalah ukuran Tindanan Taksonomik iaitu Kepersisan Leksikal (KL), Perolehan Kembali Leksikal (PKL), Leksikal F1 (F1), Kepersisan Tindanan Taksonomik (PTO), Perolehan Semula Tindanan Taksonomik (RTO) and Ukuran-F Kepersisan Tindanan Taksonomik (FTO).

1.7 RANGKA KERJA TEORITIKAL

Penyelidikan ini adalah berdasarkan rangka kerja penyelidikan yang digunakan oleh Maedche dan rakan-rakan (Kietz et al., 2000). Rangka kerja penyelidikan yang digambarkan dalam Rajah 1.1 terdiri daripada empat aktiviti seperti berikut:



Rajah 1.1 Rangka kerja pembelajaran taksonomi dari teks (PTDT)

Aktiviti 1 - Pilih sumber. Sumber adalah dokumen yang heterogen dalam format dan kandungannya. Dokumen tersebut boleh terdiri daripada teks domain atau dokumen teks generik. Dalam kajian ini, *World Wide Web* dan korpus umum digunakan.

Aktiviti 2 - Pra pemprosesan. Aktiviti utama yang dijalankan adalah untuk mengekstrak fitur dari teks menggunakan beberapa kaedah seperti pola leksiko sintaktik dan pembinaan ruang vektor.

Aktiviti 3. Pembelajaran Konsep dan Hubungan. Matlamatnya adalah untuk memperoleh konsep yang diekstrak daripada teks menggunakan pengekstrakan berasaskan pola dan pengelompokan konsep. Kajian ini hanya tertupu kepada memeroleh hubungan berbentuk sub kelas 'adalah' atau 'is-a.

Aktiviti 4 - Penilaian. Matlamat aktiviti ini adalah untuk mengukur kualiti taksonomi yang diperolehi secara automatik dengan membandingkan taksonomi yang diperoleh dengan taksonomi rujukan.

1.8 DEFINISI ISTILAH

1) Keteguhan atau Ketegapan atau Kekukuhan

   Suatu metod pengelompokan berhierarki adalah kukuh jika metod berkenaan boleh menghasilkan kelompok yang kualitinya boleh diterima walaupun mengalami masalah kejarangan data atau hingar.

2) Kejarangan Data

   Merujuk kepada situasi yang mana ciri-ciri kontekstual dari teks yang diekstrak adalah jarang dan tidak mencukupi untuk mengenal pasti persamaan antara segi. Oleh kerana ciri kontekstual diwakili dalam bentuk binari, kejarangan data juga merujuk kepada rentetan binari yang majoritinya adalah sifar.

3) Toleransi Hingar

   Keupayaan kaedah untuk mengenalpasti pola tanpa keperluan pengiktirafan mutlak sebagai kaedah toleran kepada hingar.

4) Parameter Vanilla

   Dalam teknologi maklumat, vanila adalah sejenis adjektif bermaksud biasa, asas atau standard. Dalam konteks kajian ini, parameter vanila merujuk kepada nilai yang biasa digunakan oleh penyelidik lain untuk mencari penyelesaian optimum dalam eksperimen mereka.

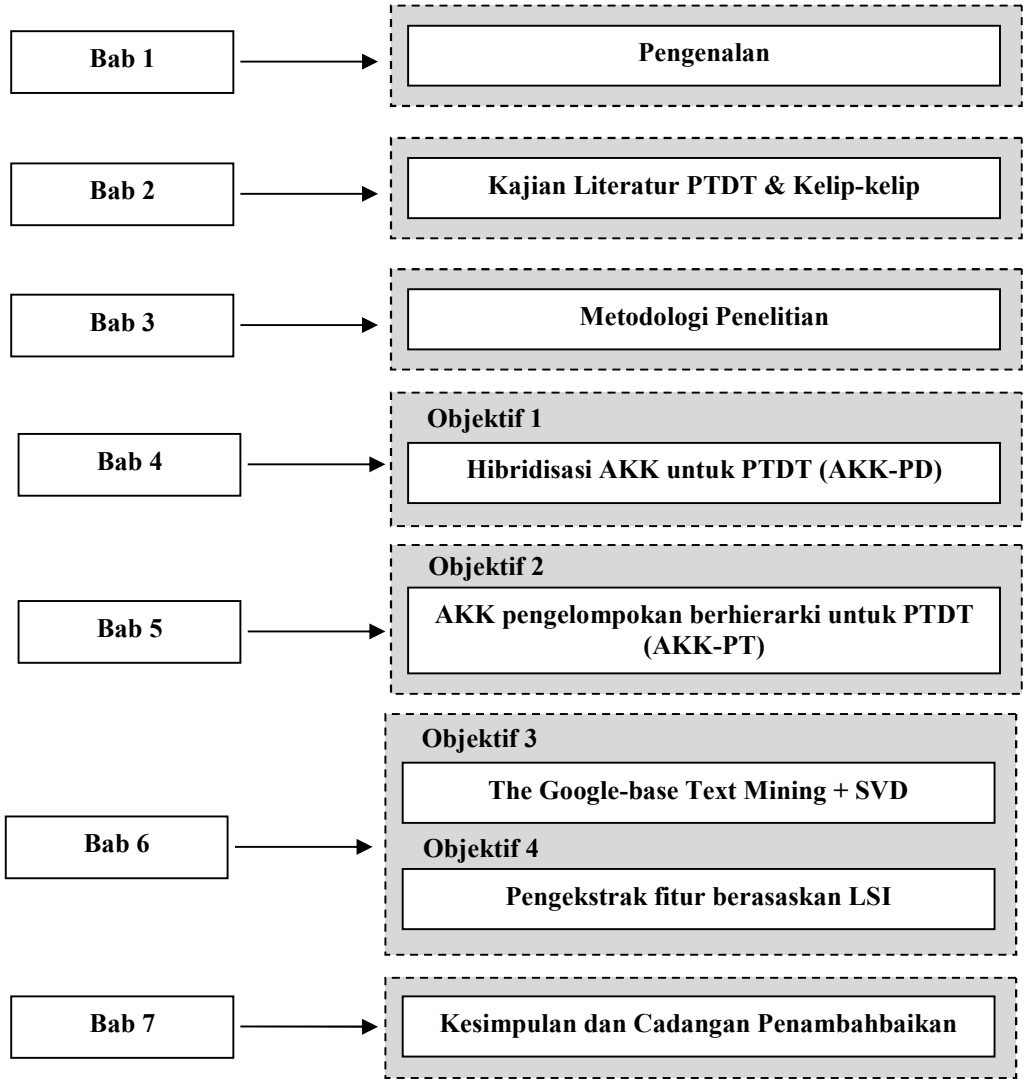
1.9 RINGKASAN DAN ORGANISASI TESIS

Bab ini telah membentangkan motivasi dan sumbangan kajian dengan mengkaji latar belakang masalah, serta menggariskan tujuan dan objektif kajian. Di samping itu, potensi sumbangan kajian juga telah diserlahkan. Tesis ini terdiri daripada enam bab seperti yang digambarkan dalam Rajah 1.2. Rajah 1.2 turut menunjukkan pemetaan objektif penyelidikan dan bab-bab dalam tesis.

Struktur tesis ini adalah seperti berikut.

1) Bab pertama adalah pengenalan kepada tesis seperti latar belakang masalah, pernyataan masalah, matlamat dan objektif kajian, skop kajian dan huraian ringkas metodologi penyelidikan.
2) Bab 2 adalah kajian literatur yang meliputi ontologi, taksonomi dan definisi penting istilah berkaitan. Bab ini membentangkan secara terperinci bidang pembelajaran taksonomi daripada teks, khususnya menerangkan penyelidikan terdahulu dan asas yang perlu untuk memahami teori dan konsep algoritma kelip-kelip (AKK). Bab ini turut berfungsi untuk menyediakan analisis dan hujah bahawa penggunaan AKK dan lain-lain sumbang kajian adalah penting kepada komuniti penyelidikan.
3) Bab 3 menerangkan pendekatan dan metodologi penyelidikan.
4) Bab 4 membincangkan model hibrid pembelajaran taksonomi berasaskan AKK dan kaedah pembahagi dua sama, metodologi dan hasil keputusan eksperimen.
5) Bab 5 membentangkan metod AKK bagi pengelompokan berhierarki, pelaksanaan dan hasil eksperimen PTDT Melayu menggunakan AKK yang dicadangkan.
6) Bab 6 membincangkan metodologi pengekstrakan fitur dari teks Melayu dengan menggunakan enjin carian Google Desktop dan SVD dan pengekstrakan fitur dari teks Melayu berasaskan pada Pengindeks Semantik Terpendam (LSI) untuk membentuk matrik Harris.
7) Akhirnya dalam Bab 7 membincangkan keputusan dan tafsiran taksonomi yang diperolehi. Bab ini diakhiri dengan menawarkan

beberapa pandangan tentang penggunaan sistem pintar untuk pembelajaran taksonomi dalam jangka panjang.



Rajah 1.2 Rangka tesis dan pemetaan objektif

BAB II

PEMBELAJARAN TAKSONOMI DARI TEKS

Bab ini membincangkan hasil tinjauan literatur ke atas domain kajian iaitu pembelajaran taksonomi dari teks menggunakan pembelajaran mesin (iaitu pembelajaran konseptual). Bahagian ini memberikan gambaran secara ringkas rangka kerja pembelajaran taksonomi seperti yang dibentangkan oleh Cimiano (2006). Selepas memperkenalkan domain kajian ini, bab ini akan memberi tumpuan kepada tiga bidang utama iaitu algoritma pembelajaran taksonomi dan kaedah pemprosesan ciri yang lebih berkesan dalam menangani isu kejarangan dan hingar data (ciri).

2.1 ONTOLOGI

Sebelum era kecerdasan buatan (AI) dimana perwakilan pengetahuan telah menjadi penting, ontologi lebih dikenali sebagai satu bidang falsafah atau sains mengenai kewujudan yang bertapak sejak abad 4 SM lagi. Perkataan ontologi berasal dari ontologia iaitu perkataan Yunani yang bermaksud "bercakap" (*-logia*) tentang "kewujudan" (onto). Ahli falsafah seperti Plato telah cuba untuk menjawab persoalan ontologi seperti *"Apa yang boleh dikatakan sebagai wujud?*" atau "*Bagaimana sifat (ciri) suatu objek adalah berkaitan dengan objek itu sendiri?*". Satu pendekatan dalam menjawab soalan falsafah seperti ini adalah dengan membahagikan entiti yang wujud kepada kumpulan-kumpulan berbeza yang dipanggil 'kategori'. Aristotle (384-322 SM) merupakan ahli falsafah yang memperkenalkan konsep *genus* dan *sub-spesies* (iaitu, perbezaan *superkonsep / subkonsep*) yang membentuk suatu ontologi dengan berlatarkan logik.

Manakala dalam bidang sains komputer moden, ontologi bukanlah satu kajian tentang kewujudan tetapi suatu perwakilan pengetahuan domain tertentu. Takrifan am Ontologi dalam komuniti kejuruteraan Ontologi ialah - spesifikasi formal suatu

pengkonsepan terkongsi" (Gruber 1993). Ontologi telah digunakan dalam pelbagai cara dan dalam pelbagai bidang seperti kecerdasan buatan, kejuruteraan sistem, informatik bioperubatan dan sains perpustakaan. Sebagai contoh, Ontologi atau 'ontologi gunaan' yang digunakan dalam bidang kecerdasan buatan biasanya digunakan untuk menyediakan ‘perbendaharaan kata yang piawai untuk satu set agen. Ontologi terus berkembang dengan kemunculan gagasan Web Semantik kerana ontologi merupakan satu daripada komponen utama Web Semantik. Ontologi Web Semantik digunakan untuk menerangkan spesifikasi bahasa yang diperlukan untuk membantu mesin secara berkesan dengan tujuan untuk berkongsi maklumat dan pengetahuan (Alesso dan Smith 2005). Ontologi membolehkan Web ‘faham’ dan ini akan membolehkan Web untuk memenuhi apa yang pengguna dan mesin perlukan. Shamsfard dan Barforoush (2003) telah memberi taktifan formal bagi Ontologi. Mereka mentakrifkan Ontologi sebagai $O = (C, R, A, Top)$ di mana:

1) C ialah set konsep yang tidak kosong.
2) R ialah set kesemua jenis hubungan antara dua atau lebih konsep yang saling berkaitan antara satu sama lain.
3) A adalah set aksiom
4) Top adalah konsep tahap paling atas dalam hierarki.

2.2 PEMBELAJARAN ONTOLOGI

Taksonomi adalah tulang belakang seni bina organisasi maklumat, merupakan sebahagian daripada sistem pengurusan kandungan atau pun ontologi berasaskan teknologi semantik. Walau bagaimanapun, memeroleh pengetahuan untuk membina taksonomi secara manual mengalami masalah cerutan pemerolehan pengetahuan yang kompleks. Pembangunan taksonomi dihalang oleh proses pembinaan yang memakan masa dan kos yang tinggi.

Dalam usaha untuk mengatasi masalah cerutan pemerolehan taksonomi ini, sebilangan besar aktiviti penyelidikan telah dilakukan untuk mengurangkan masalah cerutan pemerolehan dengan menggunakan pendekatan pembelajaran mesin, analisis statistik dan pemprosesan bahasa tabii (NLP). Proses pemerolehan pengetahuan secara automatik untuk penciptaan ontologi dikenali sebagai ***pembelajaran ontologi***

(Maedche dan Staab 2001). Borsje (2007) mendefinisikan pembelajaran ontologi sebagai tugas yang dikaitkan dengan mewujudkan dan mengekalkan domain ontologi spesifik. Istilah pembelajaran ontologi merujuk kepada usaha pemerolehan ontologi dari teks bahasa tabie (Wong et al. 2012).

Cimiano (2006) menganggap pembelajaran ontologi sebagai satu proses kejuruteraan terbalik (*reverse engineering*) kerana ia merupakan satu proses pembinaan semula model dunia pengarang daripada dokumen yang ditulis oleh pengarang berkenaan. Seorang pengarang menulis dokumen berdasarkan domain atau model dunianya yang berada di dalam fikiran yang membentuk kandungan teks. Tugas ini rumit kerana dua sebab. Pertama, hanya sebahagian kecil daripada pengetahuan domain pengarang yang wujud di dalam teks dalam proses rekreasi. Kedua, di dunia atau pengetahuan domain yang ditulis dalam teks jarang diterangkan dengan jelas seperti di dalam kamus, glosari atau thesauri.

Pembelajaran ontologi adalah berkenaan dengan penemuan ilmu pengetahuan dan dengan itu proses pembelajaran memerlukan input daripada mana-mana sumber untuk memeroleh konsep dan hubungan diantara konsep. Benz (2007) mengkelaskan tiga jenis input data untuk pembelajaran ontologi iaitu (i) data berstruktur iaitu data yang sudah disusun seperti skim pangkalan data, XML-DTDs, Rajah UML, dan sebagainya; (ii) data separa berstruktur seperti XML dan WordNet (Miller 1995); dan (iii) data tidak berstruktur seperti dokumen teks asli bahasa dan laman web berasaskan HTML. Sokongan (separa automatik) dalam pembinaan ontologi berdasarkan data tidak berstruktur dalam sumber-sumber teks secara automatik dirujuk sebagai pembelajaran ontologi dari teks (Cimianoc2006). Proses pembelajaran ontologi dari teks lazimnya melibatkan lapan tugas. Tidak kesemua lapan tugas berkenaan diaplikasi oleh kesemua sistem pembelajaran ontologi. Tugas pembelajaran ontologi dari teks adalah seperti berikut:

1) Pengesktrakan istilah domain.
2) Penemuan konsep.
3) Penjanaan hierarki konsep (taksonomi)
4) Pembelajaran hubungan selain hubungan taksonomi.
5) Penemuan petua.

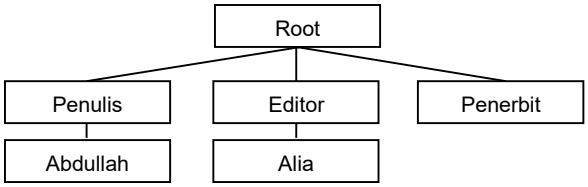
6) Pengisian ontologi.
7) Pengembangan hierarki konsep.
8) Pengesanan rangka dan peristiwa.

Tesis ini tertumpu kepada tugasan ke 3 sahaja iaitu penjanaan hierarki konsep atau taksonomi dari teks.

2.3 TAKSONOMI: SEJENIS ONTOLOGI RINGAN

Tesis ini mencadangkan kaedah pembelajaran taksonomi menggunakan algoritma diinspirasikan tabie. Justeru, adalah penting untuk memahami dan mentakrif taksonomi. Sekyen 2.1 telah membincangkan satu pendekatan untuk menjawab persoalan berontologi dengan membahagikan entiti ke dalam kumpulan yang berbeza, yang dipanggil ‘kategori’. Untuk mewujudkan kategori, ciri atau sifat yang serupa dan dimiliki oleh kebanyakan entiti lain perlu dikenalpasti. Ciri atau sifat yang serupa ini boleh digunakan untuk mengkelaskan entiti kepada kategori tertentu. Kaedah dan sains pengelasan ini dipanggil taksonomi.

Perkataan taksonomi berasal dari dua perkataan Yunani iaitu *‘taksi’* (bererti 'perintah', 'perkiraan') dan ‘*nomos’* ('undang-undang' atau 'sains'). Di samping itu, taksonomi juga ditakrifkan sebagai koleksi ‘Kosa Kata Kawalan’ yang disusun dalam struktur berhierarki (Ryu et al., 2006). Kini, taksonomi digunakan secara meluas untuk menyusun pengetahuan menggunakan hubungan ‘Adalah’ atau ‘Ialah’ (is-a), iaitu generalisasi/pengkhususan hubungan (Corcho dan Gomez-Perez, 2000). Rajah 2.1 menunjukkan contoh taksonomi. Ia terdiri daripada perbendaharaan kosa kata terkawal yang disusun ke dalam struktur hierarki.



Rajah 2.1 Satu contoh taksonomi

Rajah 2.1 menunjukkan taksonomi boleh digunakan untuk menentukan konsep *Pengarang*. Taksonomi di atas mempunyai struktur hierarki yang menakrifkan

hubungan ADALAH antara istilah manakala Abdullah dan Alia adalah contoh entiti lanjutan. Setiap istilah dalam Rajah 2.1 mempunyai hubungan induk/anak kepada istilah lain di dalam taksonomi (Batty et al. 2010). Dalam bidang linguistik, hubungan anak dipanggil hyponim. Hyponim adalah satu perkataan yang lebih khusus daripada hypernim. Misalnya, Rajah 2.1 menunjukkan yang Abdullah adalah hyponim kepada konsep Penulis dan Pengarang pula adalah hypernim kepada Abdullah.

Sebagai komponen penting dalam ontologi, taksonomi menyediakan satu model organisasi bagi suatu ontologi domain tertentu (Burgun dan Bodenreider 2001). Taksonomi domain adalah penting kerana ia adalah langkah pertama ke arah pengelasan yang berkesan, perolehan semula, perkongsian konsep, saling kendali di antara mesin dan komuniti web (Velardi et al. 2007).

Walau bagaimanapun, takrifan taksonomi masih berbeza mengikut perspektif akademik yang berbeza. Misalnya Studer et al. (1998) menganggap taksonomi sebagai ontologi penuh. Tetapi, Makrehci dan Kamel (2007) memanggil taksonomi sebagai draf ontologi yang merupakan elemen yang sangat penting dalam suatu ontologi. Manakala Lassila dan McGuinness (2001) mengkelaskan UNSPC, RosettaNet dan Yahoo! Directory yang merupakan taksonomi untuk menggelintar Web sebagai ontologi kerana mereka menyediakan konsep suatu domain yang telah mendapat persetujuan secara konsensus.

Swartout et al. (1997) menyatakan bahawa ontologi adalah satu set istilah dalam struktur berhierarki untuk menerangkan domain yang boleh digunakan sebagai rangka asas suatu pangkalan pengetahuan. Melz et al. (2006) pula memanggil taksonomi sebagai ontologi suatu domain yang spesifik. Menurut beliau, taksonomi menghubungkan istilah-istilah di dalam suatu domain secara berhierarki. Kadar spesifisiti suatu istilah menjadi lebih tinggi apabila kedudukannya di dalam hierarki menghampiri kepada nod penghujung pada rajah pepohon. Ia mendasari andaian bahawa istilah domain yang spesifik adalah hasil realisasi linguistik dari konsep yang spesifik dari suatu domain.

Walau bagaimanapun, komuniti kejuruteraan ontologi membezakan antara ontologi yang sifatnya sekadar taksonomi daripada ontologi yang memodelkan

domain dengan cara yang lebih mendalam sebagaimana Novacek (2005) telah mentakrifkan ontologi sebagai:

> *Suatu Ontologi terdiri daripada satu set nama entiti (istilah sub kelas, hubungan dan fungsi), dengan takrifan yang boleh dibaca oleh manusia, dan juga mungkin tmempunyai satu set aksiom yang mengekang tafsiran yang berbeza. (Novacek 2005).*

Sesetengah pihak membezakan antara ontologi dan taksonomi dengan mengkelaskan taksonomi sebagai 'ontologi ringan' dan ontologi sebagai 'ontology '. Alesso dan Smith (2005) dengan jelas membuat perbezaan antara ontologi dan taksonomi dengan menyatakan ontologi seperti berikut.

Ontologi=<taksonomi, petua inferens>
Taksonomi = <{kelas}, {hubungan}>

Cimiano (2006) menegaskan ungkapan taksonomi di atas dikenali sebagai hierarki konsep, yang merupakan tulang belakang ontologi dan beliau mentakrifkan taksonomi sebagai *kekisi separa-atas* $\leq_C$ $atas\, C$ *dengen elemen atas* $akar_C$ yang merupakan tulang belakang ontologi. Perbezaan utama antara ontologi ringan dan ontologi berat ialah ontologi berat mempunyai kekangan manakala ontologi ringan tidak mempunyai kekangan. Walau bagaimanapun, oleh kerana ontologi digunakan secara meluas untuk tujuan yang berbeza, Uschold (1998) memberikan takrifan yang lebih umum untuk mempopularkan ontologi kepada bidang lain. Beliau mentakrif ontologi sebagai:

> *Ontologi boleh terdiri dalam pelbagai bentuk, tetapi ia mesti mengandungi perbendaharaan istilah dan beberapa spesifikasi maknanya. Ia istilah sub takrifan dan petunjuk bagaimana konsep-konsep saling berkaitan yang secara kolektif membentuk struktur domain dan mengekang tafsiran yang mungkin berbeza daripada istilah yang berkenaan.*

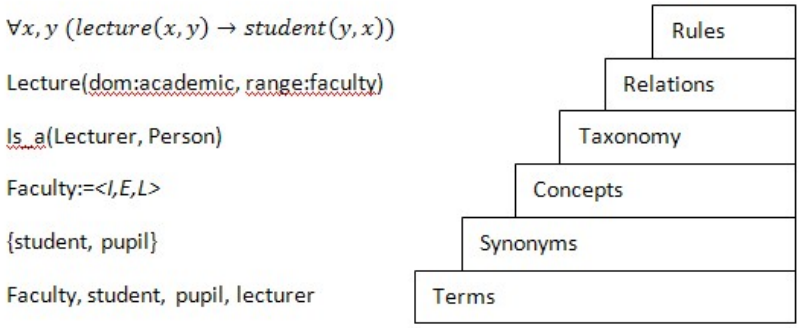
Seksyen ini telah mengumpulkan takrifan yang paling relevan untuk istilah ontologi dan taksonomi. Terdapat banyak definisi yang lain untuk istilah ini dan dapat ditemui dalam kesusasteraan. Walaupun terdapat beberapa perbezaan tentang bagaimana ontologi seharusnya ditakrifkan, namun terdapat konsensus yang tinggi kepada bagaimana ontologi harus digunakan (Gomez-Perez dan Manzano-macho 2004).

Matlamat utama tesis ini adalah untuk membina algoritma pembelajaran taksonomi. Taksonomi juga dikenali dalam komuniti kejuruteraan ontologi sebagai hierarki konsep atau ontologi ringan. Kesimpulan dari seksyen ini, takrifan utama yang diguna pakai dalam tesis ini sebagai panduan dan pengukuran adalah berasaskan kepada takrifan Swartout et al. (1997) seperti berikut:

> *Taksonomi yang juga dikenali sebagai ontologi ringan adalah satu hierarki berstruktur yang menetapkan spesifikasi konsepsual suatu domain.*

2.4 RANGKA PEMBELAJARAN TAKSONOMI DARI TEKS

Kaedah yang akan dibahas pada sub-bab ini adalah satu kaedah untuk mengekstrak komponen yang membentuk suatu ontologi daripada data tidak berstruktur. Bermakna pembelajaran itu tidak bergantung kepada mana-mana maklumat berstruktur untuk meningkatkan kualiti keputusan (Drumond dan Girardi 2008). Sehingga 2003, tiada rangka kerja standard pembelajaran taksonomi diperkenalkan, oleh itu Buitelaar et al. (2003) telah mencadangkan rangka kerja pembelajaran taksonomi dari teks seperti yang dipaparkan oleh Rajah 2.2. Rajah 2.2 menunjukkan contoh bagaimana maklumat dari domain universiti diwakilkan pada setiap lapisan. Lapisan yang dibangunkan oleh Buitelaar et al. (2003) ini memaparkan bahawa tugas pertama dalam membangunkan ontologi adalah dengan mengekstrak istilah atau kata nama seperti “fakulti’, “pelajar” dan “pensyarah”. Tugasan kedua adalah untuk mencari istilah yang sinonim misalnya “pelajar’ dan “mahasiswa”. Lapisan ketiga ialah pembentukan konsep.



Rajah 2.2 Lapisan tugas pembelajaran ontologi (Buitelaar et. al, 2003)

Cimiano (2006) berpendapat, terdapat pertindihan antara teknik yang digunakan untuk pengesanan sinonim dan pembentukan konsep. Dalam kes penemuan sinonim, perkataan yang serupa secara semantik dianggap sebagai calon yang berpotensi untuk sinonim tetapi calon itu juga menyediakan asas untuk mewujudkan konsep. Pembentukan konsep adalah satu isu yang sering dibahaskan kerana tidak begitu jelas apakah pengekstrakan konsep yang dimaksudkan (Cimiano 2006). Beberapa orang penyelidik seperti Hindle (1990), Lin dan Pantel (2002) dan Spyns dan Reinberger (2005) menganggap kelompok istilah-istilah yang berkaitan adalah konsep. Penyelidik lain pula telah menangani isu pembentukan konsep dari perspektif extensional seperti Evans (2003) dan Etzioni et al. (2004). Satu lagi pendekatan adalah untuk mempelajari konsep melalui perspektif *intension* seperti sistem OntoLearn (Velardi et al, 2005). Pada pandangan Cimiano (2006), konsep perlu terbina dari 3 komponen iaitu: i) i*ntention* konsep (I) ii) satu set tika konsep, iaitu *extension* (E) dan iii) satu set linguistik realisasi (L). Selepas terbentuknya konsep maka tugas selanjutnya adalah membina taksonomi. Rajah 2.2 menunjukkan kaedah perwakilan pengetahuan konsep pensyarah menggunakan predikat logik iaitu “adalah (pensyarah, orang)”.

Tugas pembelajaran taksonomi dari teks boleh dibahagikan kepada dua bahagian iaitu pengekstrakan konsep (istilah) dan mengorganisasi konsep yang telah diekstrak. Kebanyakan kajian terdahulu memberi tumpuan kepada pengekstrakan konsep sahaja seperti Ravichandran dan Hovy (2002); Liu et al. (2005), manakala Alfarone dan Davis (2015) memperkenalkan TAXIFY, untuk menangani kedua-dua tugas secara serentak.

Pembelajaran taksonomi bertujuan untuk membina sama ada taksonomi umum yang terbuka (taksonomi-terbuka), atau bidang taksonomi domain yang spesifik. Kedua-dua jenis taksonomi ini mempunyai cabarannya yang tersendiri. Pembelajaran taksonomi domain-terbuka seperti Ponzetto dan Strube (2011) dan Wu et al. (2012) berhadapan dengan cabaran seperti keperluan menganalisis korpus tekstual yang sangat besar dan mengatasi kesamaan (*ambiguity*) leksikal, tetapi boleh memanfaatkan lewahan leksikal yang besar dalam korpus yang besar. Sebaliknya, taksonomi domain spesifik lazimnya terhasil dengan mengeksploitasi korpus yang lebih kecil dan oleh itu tidak dapat mengeksploitasi ketumpatan data yang tinggi sekiranya korpus besar digunakan Alfarone & Davis (2015).

Tesis ini memberi tumpuan kepada penambahbaikan tugas Lapisan 1 (istilah) dan Lapisan ke-4 (taksonomi). Penyelidikan berkaitan yang paling relevan dengan skop dan fokus tesis ini adalah pembelajaran taksonomi domain-spesifik seperti Zakree (2011), Kozareva dan Hovy (2010) dan Velardi et al. (2013).

2.5 TUGAS LAPISAN 1: PENGEKSTRAKAN ISTILAH

Tugasan pertama dalam pembelajaran taksonomi ialah mengekstrak perkataan yang boleh menjadi calon istilah (konsep) beserta ciri-ciri yang menjadi konteks kepada istilah. Cimiano (2006) menyatakan konteks itu boleh digunakan sebagai asas untuk menentukan persamaan diantara istilah yang diekstrak. Oleh itu, adalah sangat penting untuk mewakili konteks setiap istilah daripada teks dengan mengekstrak ciri konteks istilah. Maksud yang betul bagi suatu perkataan hanya boleh ditentukan dengan menganalisis konteks yang mewakilinya. Kebanyakan penyelidikan bergantung kepada hipotesis pengagihan Harris untuk mereka bentuk teknik untuk mengenalpasti sinonim, konsep dan hubungan taksonomi. Harris (1954) menyatakan bahawa 'perkataan adalah sama sekiranya ia berkongsi konteks yang serupa'. Firth (1957) memperkenalkan sifat kebergantungan-konteks dengan idea 'situasi konteks'. Firth (1957) menyatakan bahawa 'anda akan mengetahui suatu perkataan melalui perkataan lain yang bersamanya' telah menjadi satu faktor penting dalam penyelidikan perlombongan teks, capaian maklumat dan pembelajaran ontologi. Oleh yang demikian, teori yang mendasari kajian ini adalah seperti berikut:

1) Kontekstual (pengagihan) hipotesis makna Harris (1954) dan Firth (1957) iaitu makna perkataan bergantung kepada penggunaannya dalam teks.
2) Hipotesis konteks persamaan semantik Miller dan Charles (1991), perkataan yang mempunyai persamaan konteks menggambarkan ia turut mempunyai persamaan semantik.

Terdapat beberapa penyelidikan telah dijalankan untuk memeriksa kesahihan hipotesis seperti Jiang dan Conrath (1997) dan Charles (2000). Siasatan empirikal mereka telah mengesahkan kesahihan hipotesis di atas. Data yang dikumpul oleh Charles (2000) mengesahkan dakwaan bahawa manusia mengikhtisar perwakilan konteks dari pengalaman beberapa konteks linguistik suatu perkataan. Penemuan ini menyokong hipotesis konteks makna. Grefenstette (1994) telah menunjukkan bahawa persamaan di dalam ruang vektor ada hubung kait dengan keberhubungan semantik perkataan.

2.5.1 Kebergantungan Sintaktik

Dalam pembelajaran taksonomi dari teks, istilah dan ciri-ciri sintaktik biasanya diekstrak daripada ayat yang telah terurai menggunakan pengurai ayat. Sebagai contoh, Cimiano (2005) mengekstrak ciri menggunakan kebergantungan sintaktik antara kata kerja yang terdapat dalam koleksi teks dan kepala subjek, objek dan pelengkap frasa kata depan. Untuk setiap kata nama yang muncul sebagai kepala argumen, kata kerja yang digunakan bersama kata nama dijadikan sebagai atribut (ciri).

Model tingkat n-perkataan juga digunakan untuk mengenalpasti kebergantuingan konteks. Menggunakan model ini, sejumlah *n* kata-kata ke kiri dan kanan perkataan sasaran dianggap sebagai ciri untuk menerangkan konteks istilah. Walau bagaimanapun, model ini hanya boleh mengekstrak konteks jarak pendek sahaja kerana *N* terbesar yang praktikal untuk digunakan adalah tiga (Zhou 2003). Oleh itu, adalah tidak jelas sama ada moel n-kata boleh menyelesaikan masalah makna ganda atau ketaksaan yang menyelubungi perkataan sasaran. Selain model tingkap n-kata, terdapat dua pendekatan lain untuk mengekstrak ciri konteks seperti

kebergantungan sintaktik (Cimiano et al, 2005a) dan pseudo-sintaktik (Grefenstette 1994).

Dalam pembelajaran taksonomi dari teks, pendekatan yang kerap digunakan untuk mengeluarkan ciri adalah untuk berdasarkan kebergantungan sintaktik daripada ayat yang telah diurai. Jadual 2.1 menunjukkan beberapa ciri sintaktik yang telah digunakan dalam pembelajaran ontologi dari teks. Biemann (2005) menyatakan bahawa kebanyakan pendekatan menggunakan hanya kata nama sebagai asas untuk membina ontologi dan tidak mengambil kira hubungan ontologi di antara kelas perkataan lain.

Jadual 2.1 Ciri-ciri sintaktik yang digunakan pada Kaedah yang berlainan

Pengarang	**Ciri-ciri**	**Teknik Pengelompokkan**	**Kaedah**
(Zakree 2011)	Kata kerja / PP-pelengkap, kata kerja / objek dan kata kerja / kebergantungan subjek	GAHC, CLOSAT dan CLONALG	Jarak persamaan
(Cimiano 2006)	Kata kerja / PP-pelengkap, kata kerja / objek dan kata kerja / kebergantungan subjek	FCA	Set-Teori
(Cimiano dan Staab 2005)	Pengelompokan Kata nama, corak Hearst, WordNet dan corak Web	Pengelompokkan berpandu	Jarak persamaan dan Pengekstrakan corak
(Sporleder 2002)	ciri-ciri morfologi	FCA	Set-Teori
(Osswald and Petersen 2002)	ciri-ciri morfologi	FCA	Set-Teori
(Bisson et al 2000)	Ketua / Perhubungan tatabahasa / Modifier	Pengelompokan konsep	Jarak persamaan
(Caraballo 1999)	Kebergantungan sintaktik: Kata hubung kata nama, membina positif dan corak Hearst	Pengelompokkan	Jarak Persamaan dan Pengekstrakan corak
(Faure dan Nedellec 1998)	bergantungan: kata kerja-hujah	iteratif Pengelompokan Dari Bawah ke atas	Jarak persamaan
(Pereira et al 1993)	Kebergantungan: Kata kerja-hujah Hubungan kata kerja-objek	Pengelompokkan atas bawah	Jarak persamaan
(Hindle 1990)	Predikat / Verb-hujah kebergantungan: Kata kerja / Subject dan Kata kerja / Objek		Jarak persamaan

Sumber:(Biemann 2005, Cimiano *et al.* 2004a) (Sanchez and Moeno 2005)

Cimiano (2006) pula memilih beberapa ciri berdasarkan kebergantungan sintatktik tertentu seperti frasa kata kerja/objek, kata kerja/subjek dan kata kerja/frasa preposisi (PP) pelengkap kebergantungan. Untuk setiap kata nama yang muncul sebagaikepal argumen, kata kerja yang sama digunakan sebagai ciri-ciri. Dalam fasa ini, alat NLP khas diperlukan untuk menjalankan carian pada teks yang telah diurai berdasarkan golongan sintaksis sebelum menukarkan kepala (kata nama) objek menjadi objek. Cimiano (2006) menggunakan alat yang dinamakan Tgrep untuk mendapatkan kebergantungan sintaktik. Menggunakan Rajah 2.2 sebagai contoh, pensyarah dan pelajar akan diekstrak sebagai istilah manakala perkataan kata kerja 'baca' akan diklasifikasikan sebagai sifat atau ciri yang akan digunakan sebagai perujuk kepada istilah.

Pendekatan Hindle (1990) mengambil kira kata nama yang muncul sebagai subjek dan objek bagi kata kerja. Faure dan Nedellec (1998) mengemukakan idea pengelompokan bawah-ke-atas berlelar untuk kata nama yang muncul dalam konteks yang sama. Pereira et al. (1993) mengemukakan pendekatan pengelompokan atas ke bawah untuk membina hierarki kata nama tidak berlabel. Merujuk kepada Jadual 2.1, Zakree (2011) menggunakan pendekatan Cimiano (2006) tanpa menggunakan alat NLP kerana tiada alat untuk mengekstrak kata nama berdasarkan kebergantungan sintaksis untuk Bahasa Melayu ketika itu. Zakree (2011) menggunakan ahli bahasa untuk mengekstrak kata nama dan ciri-cirinya yang terdiri daripada jenis kata kerja

2.5.2 Kebergantungan Pseudo-sintaksis

Satu lagi pendekatan untuk mendapatkan ciri konteks adalah dengan menggunakan heuristik mudah dan padanan pola yang bergantung kepada output penghurai cetek. Penghurai cetek yang juga dikenali sebagai penghurai sebahagian digunakan mengelompok kata-kata yang boleh dikenal pasti sebagai frasa sintaksis seperti frasa kata nama dalam ayat bahasa tabii. Walau bagaimanapun, sesetengah penghurai cetek boleh mengenal pasti struktur subjek-kata_kerja-objek yang digunakan oleh Reinberger dan Daelemans (2010) untuk membina kelas semantik kata nama. Kemudian, setiap ungkapan sintaksis akan dipadankan terhadap senarai ungkapan sintaktik. Padanan ungkapan biasa yang serupa akan diekstrak sebagai

sepasang objek-atribut (sifat/ciri). Berikut adalah contoh ungkapan sintaktik yang Cimiano gunakan untuk memeroleh ciri dari teks. Pasangan-sifat objek yang diekstrak dibentangkan dalam notasi predikat *(o),* di mana *a* adalah atribut dan *o* adalah objek.

i. Pengubah kata sifat (adjektif), contohnya *sebuah masjid yang indah* → *indah (masjid).*

ii. Frasa kata kerja dengan kata kerja *mempunyai*, contohnya. *setiap fakulti mempunyai dewan kuliah* →*mempunyai_dewan_kuliah (fakulti).*

Dalam usaha untuk membina taksonomi dari teks menggunakan pendekatan ini, konstruk kopula digunakan, contohnya *pensyarah adalah seorang saintis* → *saintis (pensyarah).* Kopula adalah kata kerja yang menyamakan (contoh: *seperti* dan *adalah*) yang menghubungkan subjek untuk melengkapkan ayat, contohnya objek atau adjektif.

2.5.3 Pengenalpastian Sinonim

Rajah 2.2 menunjukkan bahawa sebelum induksi taksonomi dilakukan, terdapat dua subtugasan yang perlu dilaksanakan iaitu pengenalpastian sinonim. Kebanyakan penyelidikan menangani pengenalan sinonim dalam syarat-syarat yang diekstrak oleh melalui algoritma pengelompokan dengan mengeksploitasi hipotesis pengagihan Harris. Sebagai contoh dengan menggunakan pendekatan itu, kedua-dua pelajar dan murid boleh dikenal pasti sebagai sinonim. Penemuan sinonim dalam kelompok kata nama (istilah) yang diekstrak membantu untuk mengelakkan konsep yang sama/seerti terbentuk dalam cabang yang berbeza dalam hierarki taksonomi. Masalah kesamaan semantik juga dapat dielakkan jika mewakili konsep yang sama, contohnya. pelajar, sarjana muda, murid, dan sebagainya. Sub-bab seterusnya akan menerangkan secara terperinci bagaimana hipotesis Harris digunakan dalam penemuan sinonim. Selain daripada pengelompokan, Baroni dan Bisi (2004) misalnya menggunakan maklumat statistik yang ditakrifkan di web untuk mengesan sinonim. Secara umum, kajian berkaitan penemuan sinonim telah lama dijalankan dan keputusan telah menunjukkan alat NLP ini telah mencapai prestasi setanding manusia

dalam pengenalpastian sinonim (Landauer dan Dumais 1997; Turney 2001). Selepas penemuan sinonim dan pembentukan konsep, taksonomi yang merupakan tulang belakang ontologi boleh diperolehi.

2.6 TUGAS LAPISAN 4: PEMBELAJARAN TAKSONOMI

Pembelajaran taksonomi merupakan objektif utama kajian ini dalam membangunkan algoritma pembelajaran taksonomi dari teks yang tidak berstruktur untuk membentuk lapisan ke-4. Berdasarkan kepada pengelasan yang dicadangkan oleh Gomez-Perez et al. (2005), Fortuna et al. (2005) dan Cimiano (2006), pembelajaran taksonomi dari teks boleh dibahagikan kepada tiga jenis iaitu i) Pendekatan linguistik; ii) Perlombongan teks dan iii) Pendekatan hibrid.

2.6.1 Pendekatan Linguistik

Pendekatan Linguistik adalah berdasarkan istilah yang dipetik dari Cimiano (2006). Pendekatan ini dinamakan linguistik kerana taksonomi diperolehi terus dari teks dengan mengeksploitasi analisis. Pendekatan ini bergantung kepada sumber lain untuk mengekstrak hubungan seperti WordNet atau Web seperti yang telah dicadangkan oleh Kietz et al. (2000). Kietz et al. (2000) memperkenalkan satu teknik yang boleh menghasilkan domain ontologi menggunakan ontologi teras seperti SENSUS atau WordNet.

Satu lagi jenis teknik pendekatan linguistik adalah dengan memadankan pola, contohnya pola Leksikal-sintaktik. Teknik pengekstrakan maklumat ini telah dicadangkan oleh Hearst (1992), Pasca (2004) dan Etzioni et al. (2004). Ini dilakukan dengan mencari pola ungkapan biasa dalam teks. Hearst (1992) mencadangkan agar hubungan *adalah* (atau hipernim) diekstrak dengan mencari padanan kepada pola semantik-terpendam. Beliau telah mereka bentuk pola hubungan hipernim/hiponim seperti berikut:

1) Frasa Nama (FN) seperti {FN, FN, …, (dan|atau)} FN
2) sebenarnya FN seperti FN {(dan|atau)} FN
3) FN {, FN} {,} atau lain-lain FN

4) FN {, FN) {,} atau lain-lain FN
5) FN {,} istilahsuk FN { dan|atau} FN
6) FN {,} terutamanya FN{dan|atau} NP

Antara kajian yang mebggunakan pola Hearst adalah Kozareva dan Hovy (2010) dan sumbernya adalah dari Web dan kemudian memangkas konsep konsep tidak relevan dari taksonomi tersebut dengan hanya memilih konsep yang memiliki laluan terpanjang. Velardi et al. (2013) mengekstrak hubungan *adalah* dari ayat-ayat takrifan seperti yang lazim ditemui dalam kamus daripada korpus yang diperolehi dari Web dan korpus bidang yang spesifik. Alfarone dan Davis (2015) mendapati ayat-ayat takrifan cenderung untuk mengekstrak s*upertypes* yang generik, menyebabkan terbentuknya rantaian konsep yang tidak relevan. Ini menyebabkan peratus ketepatan system menurun. Kozareva dan Hovy (2010) pula menggunakan strategi pemangkasan untuk mereka mengoptimumkan keseimbangan antara pengekalan laluan panjang dan memaksimumkan ketersambungan antara nod. Oleh itu, Alfarone dan Davis (2015) telah memperkenalkan kaedah yang dipanggil TAXIFY yang boleh mencerap pola yang terbentuk dan membiarkan laluan panjang wujud jika mematuhi syarat atau algoritma TAXIFY. Selain itu, TAXIFY berupaya mengesan hubungan *adalah* yang ralat ini meningkatkan kejituan sistem.

Kelemahan pengekstrakan taksonomi berasaskan pola ialah hubungan taksonomi *'adalah'* jarang nuncul dalam teks dan kebanyakan konsep dengan hubungan '*adalah'* tidak muncul dalam pola Hearst. Brunzel (2007) melaporkan bahawa pola Hearst jarang berlaku, dan frekuensi penemuan pola Hearst adalah rendah walaupun pada korpus dokumen yang besar. Kelemahan pendekatan itu adalah bahawa hubungan antara istilah yang dikenal pasti dalam 'bentuk perkataan' dan bukannya makna sebenar. Sebagai contoh, diberi sebaris ayat berikut: *'kawasan bersejarah di Bangi adalah sebuah kampung bernama Sri Putra'*. Kebanyakan pendekatan yang berdasarkan padanan pola leksikal-sintaktik akan memperolehi pengetahuan berikut (Dalam bentuk predikat kalkulus): adalah (Bangi, Kampung). Pengetahuan ini pastinya tidak betul, kerana ayat tersebut bermaksud "Kampung Sri Putra adalah sebuah kampung".

Kamus Mesin boleh dibaca (MRD) telah digunakan oleh Iwanska et al. (2000) dan Dolan et al. (1993) untuk mencari hubungan taksonomi. Kaedah ini dianggap oleh Cimiano (2006) sebagai satu daripada pendekatan linguistik dalam mencari hubungan hipernim/hiponim. Mereka bercadang untuk mengeksploitasi takrifan kamus standard. Kamus dan malah buku-buku teks akademik mengandungi pengetahuan yang jelas dalam bentuk definisi seperti "ikan paus adalah sejenis mamalia". Biasanya, kepala frasa nama pertama yang terdapat dalam menjelaskan maksud suatu perkataan di dalam kamus ialah hipernim. Beberapa orang penyelidik seperti Maedche et al. (2002) telah mengeksploitasi pola biasa untuk mengekstrak hubungan taksonomi dari teks. Velardi et al. (2005) menggunakan heuristik padanan-kepala untuk menentukam mana satu antara istilah adalah hipernim dan hiponim. Jika istilah pertama berada di depan istilah kedua, maka istilah pertama adalah hiponim dan istilah kedua adalah hipernim. Misalnya, merah adalah warna. Warna adalah hipernim kepada merah.

Cimiano et al. (2004b) melihat dua kelemahan utama dalam menggunakan MRD untuk pembelajaran ontologi. Pertama, pengetahuan yang diperolehi banyak bergantung kepada idiosinkrasi intrinsik yang berkaitan dengan takrifan yang diberikan. Setiap takrifan bergantung kepada keunikan suatu istilah dihuraikan. Jika tidak intrinsik, maka kebarangkalian berlaku ralat adalah tinggi. Kedua, tugas pembelajaran taksonomi biasanya berusaha untuk memperoleh pengetahuan domain-spesifik daripada domain tertentu, tetapi kamus lazimnya adalah dari domain sumber terbuka dan tidak terikat kepada suatu domain yang spesifik. Namun sehingga kini, tiada Kamus Melayu sumber terbuka yang boleh dibaca mesin boleh diperolehi dan digunakan dalam kajian.

Teknik ini sangat popular kerana penyelidik boleh menggunakan WordNet untuk mengekstrak hubungan dengan perkataan tertentu atau membandingkan keputusan mereka dengan WordNet. Walau bagaimana pun, Zakree (2011) menyatakan kelemahan utama menggunakan WordNet dalam membina taksonomi dalam konteks budaya dan agama masyarakat Melayu. Beliau memberi contoh gelintaran perkataan Nabi dan Rasul di WordNet. Kedua-duanya adalah konsep penting dalam Islam dan kajian ini kerana Fekah adalah antara teks yang digunakan untuk membina taksonomi dalam penyelidikan ini. Di WordNet, kedua-dua istilah

Nabi dan Rasul diiktiraf sebagai nabi. Walau bagaimanapun, kedua-dua Nabi dan Rasul sebenarnya tidak sinonim. Rasul adalah satu konsep yang berbeza kerana mereka mewarisi semua sifat-sifat Nabi dan beberapa sifat-sifat khas yang tidak dimiliki nabi. Misalnya, seorang Rasul membawa syariat baharu manakala Nabi hanya menyampaikan wahyu berdasarkan syariat yang dibawa oleh seorang Rasul (yang juga seorang Nabi).

WordNet berasaskan web turut memaparkan konsep asing mengenai nabi kepada umat Islam iaitu wujudnya Nabi wanita (iaitu *Prophetess)* sebagai hyponim bagi Nabi iaitu suatu konsep dari Rom kuno yang bermaksud seorang wanita yang dianggap sebagai seorang ahli nujum). Tukang ramal juga adalah hyponim bagi konsep Nabi. Oleh itu, Zakree (2011) membuat kesimpulan bahawa menggunakan semula konsep dari WordNet sebagai sumber pengetahuan tidak sesuai dalam konteks kajian ini dan hasilnya akan salah.

2.7 PERLOMBONGAN TEKS

Kaedah teknik pengelompokan tanpa selia untuk memeroleh taksonomi dari teks adalah berdasarkan kaedah persamaan. Biasanya, kaedah ini menerima pakai model ruang vektor. Sebelum menggunakan mana-mana teknik pengelompokan untuk memeroleh taksonomi, istilah yang diekstrak mesti diwakili sebagai satu kumpulan vektor ciri. Sebagai contoh, bit dalam vektor mewakili kehadiran ciri iaitu kata kerja yang muncul dalam konteks iaitu wujud bersama kata nama (istilah) dalam suatu frasa. Kata nama adalah calon bagi konsep. Adalah menjadi amalan bahawa kata nama mempunyai ciri konteks (iaitu kata kerja) yang diwakili secara binari. Jika wujud dalam konteks maka diwakili dengan 1 dan tiada dalam konteks diwakili dengan 0. Vektor ciri adalah ruang dimensi yang tinggi. Semua istilah adalah titik jarang di dalam ruang dimensi yang tinggi. Lebih banyak ciri diekstrak maka lebih jarang ruang vektor tersebut.

Salah satu kaedah pengelompokan seperti K-Min Pembahagi Dua Sama menggunakan ukuran persamaan untuk mengira persamaan pasangan demi pasangan untuk menentukan sama ada mereka mempunyai persamaan semantik untuk dikelompokkan. Apidianaki (2009) menerangkan ukuran jarak sebagai satu cara untuk

mengira persamaan daripada dua unsur. Ukuran persamaan yang terkenal dalam penyelidikan pembelajaran ontologi ialah kosinus sudut antara dua vektor yang dikenali sebagai pekali kosinus (Salton dan McGill 1983) atau vektor maklumat saling mereka. Kosinus ditakrifkan seperti berikut (Rujuk persamaan 2.1):

$$\cos(x, y) = \frac{x.y}{|x||y|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}} \quad \text{... (2.1)}$$

Pembangunan hierarki konsep atau taksomoni menggunakan pendekatan pembelajaran mesin merupakan satu pendekatan yang semakin popular. Pada tahun 1999, Sanderson dan Croft (1999) mencadangkan satu model statistik mudah yang dipanggil analisis kewujudan-bersama. Dua istilah atau konsep x dan y, y dikatakan sebagai sub kelas y jika syarat berikut berlaku: $P(x|y) \geq 0.8, P(y|x) < 1$. Algoritma yang menggunakan teknik analisis kewujudan-bersama pada ayat, paragraf atau dokumen yang sama boleh digunakan untuk mengenalpasti hubungan taksonomi. Schmitz (2006) menyatakan bahawa taksonomi yang dihasilkan menggunakan pendekatan ini agak bising kerana hanya 23% daripada pasangan (taksonomi) didapati benar mempunyai hubungan taksonomi, manakala 49% didapati jatuh ke dalam kategori yang lebih umum.

Kebanyakan kajian kaedah automatik untuk memeroleh taksonomi tertumpu kepada penggunaan teknik pengelompokan tanpa selia. Teknik pengelompokan tanpa selia ini mengekstrak hubungan taksonomi berdasarkan hipotesis pengagihan Harris (1968). Kaedah ini mengguna pakai model ruang vektor dan mewakili perkataan atau istilah sebagai vektor yang mengandungi ciri atau sifat yang berasal dari korpus tertentu. Apidianaki (2009) menggelar teknik ini sebagai ‘pengelompokan makna’ kerana sekumpulan unsur dikelompok berdasarkan persamaan semantik atribut (ciri) yang menunjukkan persamaan sifat penggagihan (atribut). Berdasarkan kepada 'pendekatan pengelompokan makna', ramai penyelidik seperti Zakree (2011), Neshati et al. (2009), Brewster et al. (2007), Widdows dan Dorow (2002), Faure dan Nedellec (1998), Bisson et al. (2000), Cimiano et al. (2004c), Cimiano dan Staab (2005) dan Cimiano et al. (2005b) telah membangunkan pelbagai teknik pnegelompokan tanpa selia untuk menjana taksonomi dari teks. Secara umum, pendekatan pengelompokan

biasanya digunakan untuk mencapai dua objektif iaitu pembentukan konsep dan pembentukan hierarki konsep. Ini adalah kerana hakikat itu, mereka mewujudkan kelompok atau kumpulan perkataan yang sama, yang boleh dianggap sebagai mewakili konsep sedikit, dan perintah selanjutnya-kelompok hierarki.

Cimiano dan Staab (2005) membangunkan Pengelompokan Hierarki Aglomerat Berpandu (GAHC) yang merupakan satu contoh bagaimana kaedah berasaskan pola digunakan untuk mengaruh hubungan taksonomi. GAHC meletakkan konsep yang memiliki persamaan yang tinggi dalam kelompok yang sama agar ia boleh dilabel dengan label yang sewajarnya.

Dalam konteks tesis ini, kajian tertumpu kepada pembangunan kaedah pengelompokan yang lebih berkesan sebagai algoritma pembelajaran tanpa selia untuk membangunkan taksonomi secara automatik. Antara sebab utama pendekatan tanpa pengawasan dipilih adalah kerana tiada data (korpus) Melayu bertag dengan jenis kata yang boleh diperolehi untuk pembelajaran mesin.

Pengelompokan secara umumnya adalah mencari kelompok atau kumpulan objek yang memiliki darjah persamaan yang tinggi daripada data. Oleh yang demikian, algoritma mengira darjah persamaan berdasarkan ciri konteks yang dimiliki oleh suatu istilah. Pendekatan ini adalah selaras dengan hipotesis pengagihan Harris (1954). Terdapat dua jenis pengelompokan iaitu yang pertama ialah pengelompokan agglomerat berhierarki dan yang kedua adalah pengelompokan pembahagi. Kaedah pengelompokan juga boleh dibahagikan kepada dua kelas. Pertama, kaedah yang berasaskan persamaan dan yang kedua yang berasaskan set-teori. Contoh kaedah berasaskan persamaan adalah K-Min Pembahagi Dua Sama manakala Analisis Konsep Formal (FCA) adalah satu contoh kaedah yang berdasarkan teori set.

2.7.1 Analisis Konsep Formal (FCA)

Analisis Konsep Formal (FCA) adalah kaedah yang digunakan untuk analisis data. FCA boleh digunakan untuk mengaruh hubungan tersirat antara objek menerusi set ciri yang dperoleh dari teks (Cimiano et al. 2005a). Data (ciri) yang diekstrak di struktur kepada unit yang merupakan abstrak formal kepada konsep-konsep yang

berada di dalam pemikiran manusia. Unit yang berstruktur ini memberi tafsiran yang boleh difahami dan bermakna. Untuk pengetahuan terperinci mengenai FCA, Ganter dan Wille (1999) telah menerbitkan makalah berkenaan FCA. Seksyen ini akan membincangkan konsep FCA yang diadaptasi dalam kebanyakan algoritma pembelajaran taksonomi.

Dalam mempelajari taksonomi dari teks menggunakan FCA, ciri bagi suatu konsep yang diekstrak berdasarkan sintaktik dari teks disusun dalam konteks formal. Konteks formal adalah 'jantung' bagi FCA. Ketga-tiga (G, M, I) dipanggil konteks formal jika G dan M adalah dalam suatu set dan mempunyai hubungan binari *adalah*. Unsur-unsur G dipanggil objek, M pula ciri-ciri (attribut) dan I pula mewakili hubungan insidens konteks antara objek dan ciri. Definisi ini adalah berdasarkan Ganter dan Wille (1999) di mana teori kekisi digunakan untuk analisis hierarki konsep. Pasangan (A, B) adalah satu konsep yang formal (G, M, I) jika dan hanya jika (A, B), di mana $A \subseteq G, B \subseteq M$, $A'=B$ and $B'=A$. Dalam erti kata lain, pasangan (A, B) adalah satu konsep yang formal jika set semua ciri objek A adalah sama dengan B dan A juga adalah set semua objek yang mempunyai semua ciri (fitur) = dalam B.

untuk $A \subseteq G, \quad A' := \{m \in M \mid \forall g \in A : (g,m) \in I\}$ ...(2.2)

dan untuk $B \subseteq M : B' := \{g \in G \mid \forall m \in B : (g,m) \in I\}$ ...(2.3)

Dalam erti kata lain, *A*' adalah set ciri yang lazim kepada semua objek dalam A dan B' ialah set objek yang mempunyai ciri dalam *B*. A dipanggil *extent* konsep formal *(A, B)* dan B adalah *intend* bagi A. Konsep konteks formal secara semula jadi menyusun hubungan subkonsep-superkonsep sebagaimana yang ditakrifkan oleh:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \; (\Leftrightarrow B_2 \subseteq B_1) \qquad \ldots (2.4)$$

Oleh itu, konsep formal sebahagiannya disusun dengan mengambil kira *extent* atau (yang bersamaan) balikan *intend* mereka. Hubungan $\leq$ adalah tertib pada S. Ganter dan Wille (1999) menyatakan bahawa (S(K), $\leq$) adalah kekisi yang lengkap dan kekisi ini dikenali sebagai kekisi konsep konteks K (G, M, I).

Satu fakta penting yang dipelajari dari liputan kesusasteraan ini adalah pendekatan set-teoritikal seperti FCA boleh bersaing dengan pendekatan berasaskan persamaan dari aspek kualiti kelajuan, kualiti hierarki yang dihasilkan, dan kebolehkesanan. Kualiti yang lebih baik menurut Cimiano et al. (2005a) adalah disebabkan oleh % *recall* (dapatan semula) yang lebih tinggi. Fakta yang menarik tentang pendekatan set-teoritikal adalah pendekatan ini adalah berasaskan kepada teori matematik Analisis Konsep Formal (FCA). Petersen (2001) menyatakan bahawa menggunakan FCA adalah satu idea yang inovatif di mana 'hubungan pewarisan' adalah berasaskan data (ciri). FCA diaplikasikan oleh oleh Maio et al. (2009), Quan et al. (2009), Obitko et al. (2004), Quan et al. (2004), Cimiano et al. (2005a) dan Zakree et al. (2008) dalam pembelajaran konsep hierarki.

Walau bagaimanapun, terdapat kelemahan apabila menggunakan FCA kerana FCA menghasilkan kekisi yang perlu diproses untuk mendapatkan taksonomi. Ini adalah kerana kekisi yang dihasilkan mengandungi kedua-dua istilah dab ciri (sifat-sifat konteks) yang sama. Oleh itu, ia adalah sukar untuk membezakan kedua-duanya.

2.7.2 Pengelompokan Berhierarki

Pengelompokan berhierarki adalah kaedah analisis kelompok dalam bidang perlombongan data dan statistik yang digunakan untuk menghasilkan hierarki kelompok. Terdapat dua strategi pengelompokan berhierarki yang digunakan seperti berikut:

1) Agglomerat: Pendekatan bawah-ke-atas ini bermula dengan pembentukan banyak kelompok secara berasingan, dan pasangan kelompok akan digabungkan apabila satu kelompok bergerak ke hierarki lebih tinggi.
2) Pembahagi: Pendekatan atas-ke-bawah ini bermula dengan satu kelompok dan kemudian dibahagikan secara rekursif.

Pengelompokan Berhierarki Agglomerat (HAC) adalah salah satu pendekatan yang telah digunakan oleh penyelidik terdahulu seperti Pereira et al. (1993), Caraballo (1999), Cimiano et al. (2004b), Cimiano dan Staab (2005) dan Kuo et al. (2006) untuk

memeroleh taksonomi. HAC telah digunakan untuk membina taksonomi kerana ia menghasilkan struktur pokok, yang dipanggil dendogram. Ia boleh digunakan untuk menggambarkan hierarki konsep atau taksonomi. HAC adalah contoh bagaimana kaedah berasaskan hipotesis Harris yang bergantung kepada persamaan antara objek digunakan untuk membentuk taksonomi.

Seperti FCA, algoritma pengelompokan berhierarki sama ada jenis agglomerat mahu pun pembahagi menggunakan ruang vektor untuk mewakilkan ciri konteks suatu istilah. Jarak persamaan antara semua pasangan kelompok akan dikira menggunakan ukuran persamaan seperti kosinus. Satu set objek (iaitu, kata nama) dan jarak persamaan antara mereka diberikan sebagai input. Pada HAC, setiap objek dikira sebagai kelompok berunsur tunggal. Algoritma jenis agglomerat bermula dengan setiap objek (sebagai kelompok tersendiri) yang bergabung dengan objek lain secara untuk membentuk satu kelompok. Algoritma tamat apabila satu kelompok besar yang mengandungi semua objek *D* telah terbentuk. Dalam algoritma ini, terdapat suatu prosedur yang dipanggil KelompokBaharu untuk menentukan bagaimana kelompok tunggal digabungkan di setiap peringkat. Terdapat pelbagai strategi untuk mengira jarak/persamaan antara dua kelompok yang berbeza seperti Pautan Tunggal, Pautan Purata dan Pautan Lengkap.

Pautan Tunggal Pautan tunggal ditakrifkan sebagai persamaan antara dua kelompok P dan Q. Dua kelompok digabungkan jika terdapat sekurang-kurangnya satu pasangan yang sama antara dua kelompok di atas nilai ambang yang telah ditetapkan (Cimiano et al. 2004b).

Pautan Purata Pautan purata atau juga dikenali sebagai purata-kumpulan adalah satu fungsi yang menentukan persamaan purata objek kedua-dua kelompok, (Cimiano et al. 2004b).

Pautan Lengkap Tidak seperti pautan sebelum ini, Pautan lengkap mengira persamaan yang paling rendah diantara dua kelompok.

Output HAC ialah dendogram yang membentuk kelompok hierarki perkataan yang berkaitan/serupa. Walau bagaimanapun, kelompok-kelompok masih tidak sesuai dilabel lagi untuk membentuk suatu taksonomi kerana ia menghasilkan struktur pokok. Ia perlu terus diproses untuk membentuk satu taksonomi. Untuk melabel kelompok yang dihasilkan dalam dendogram, Cimiano (2006) menggunakan kaedah

Caraballo (1999). Kaedah Caraballo menggunakan hypernim yang telah diperolehi dari pendekatan berasaskan pola untuk melabel setiap kelompok dalam dendogram. Hypernym paling kerap diambil sebagai label bagi kelompok dengan syarat bahawa ia adalah suatu hypernym yang sah untuk sekurang-kurangnya dua elemen dalam kelompok. Akhir sekali, hierarki dimampatkan dengan menanggalkan semua kelompok tanpa label. Oleh itu, taksonomi diwujudkan.

Pendekatan kedua dalam pengelompokan berhierarki yang digunakan dalam pembelajaran taksonomi adalah teknik pembahagi. Antara algoritma yang telah dicadangkan adalah Pembahagi Dua Sama K-Min (Steinbach et al. 2002). Algoritma ini bermula dengan suatu kelompok terbesar yang mengandungi istilah dan mula membentuk hubungan taksonomi dengan menggunakan langkah-langlah berikut:

1) Pilih kelompok terbesar untuk dibahagikan.
2) Cari (pisahkan) 2 sub-kelompok menggunakan algoritma K-min.
3) Ulangi langkah 2, ambil kelompok yang mempunyai persamaan tertinggi.
4) Ulangi langkah 1, 2 dan 3 sehingga bilangan kelompok yang diingini tercapai.

Algoritma bermula dengan satu set kelompok besar. Kemudian memilih kelompok terbesar untuk dibahagi. Jumlah kelompok yang perlu dihasilkan dari pembahagian kelompok besar ini bergantung kepada nilai paramaeter. Cimiano (2006) telah membuktikan bahawa pendekatan ini adalah yang paling efisien untuk membentuk taksonomi tetapi masih terdapat isu yang perlu ditangani iaitu pelabelan.

2.7.3 Algoritma Metaheuristik Diinspirasikan Alam

Menurut Siddique dan Adeli (2013), kecerdasan komputeran adalah satu set metodologi dan pendekatan komputeran yang diilhamkan oleh sifat dan tingkahlaku alam semula jadi untuk menangani masalah yang rumit yang mana pemodelan matematik atau tradisional tidak dapat digunakan untuk beberapa sebab seperti proses mungkin terlalu rumit untuk penaakulan berasaskan matematik kerana permasalahan mungkin mengandungi beberapa ketidakpastian atau bersifat stokastik.

Algoritma metaheuristik merujuk kepada prosedur aras-tinggi atau heuristik yang direka bentuk untuk mencari, menjana atau memilih heuristik (algoritma gelintaran) yang dapat menyediakan penyelesaian yang memadai untuk suatu permasalahan pengoptimuman (Leonora et al. 2009). Terdapat sejenis algoritma metaheuristik yang mendapat perhatian penyelidik dalam pembelajaran taksonomi iaitu dari jenis algoritma metaheuristik diinspirasikan-alam tabie seperti algoritma lebah koloni, algoritma berevolusi dan kecerdasan kerumunan.

Kecerdasan Kerumunan adalah disiplin kecerdasan buatan yang semakin popular (Blum & Li, 2008). Ia diilhamkan dari tingkah laku kolektif sosial kawanan semut, anai-anai, lebah, cacing, kawanan burung dan ikan. Walaupun kawanan ini terdiri daripada individu yang secara relatifnya tidak canggih, namun serangga dan haiwan berkenaan mempamerkan tingkah laku yang diselaras. Kawanan ini bertindak untuk mencapai matlamat yang diinginkan mengikut arahan yang dikoordinasi mengikut pola tertentu. Ini biasanya mengakibatkan sistem kehidupan mereka tersusun kerana setiap unit serangga dan binatang dalam kawanan memiliki sifat penyusunan-kendiri. Ini membentuk apa yang dipanggil sebagai ‘kecerdasan kolektif’ yang menjadi prinsip asas penyusunan kendiri dalamm sistem multi-agen. Tingkah laku yang diselaraskan ini dilakukan kerana interaksi antara individu, contohnya, ranai-anai dan cacing boleh membina sarang yang canggih, manakala semut dan lebah juga menggunakan kelakuan kolektif ketika mencari makanan. Biasanya, semut berinteraksi antara satu sama lain melalui laluan pheromone kimia untuk mencari laluan terpendek antara sarang dan sumber makanan mereka. koloni lebah, peranan pemberi maklumat dimainkan oleh lebah yang dpangil sebagai penjejak. Lebah penjejak bertanggungjawab untuk mencari sumber makanan. Di sini, komunikasi antara lebah dilakukan dengan apa yang disebut sebagai 'tarian waggle', di mana koloni lebah diarahkan oleh penjejak. Semasa makanan baru ini ditemukan, kesimbangan antara penjelajahan (koleksi baru maklumat) dan eksploitasi (penggunaan maklumat sedia ada) mesti dilakukan oleh koloni lebah.

Metaheuristik diinspirasikan alam telah digunakan oleh beberapa penyelidik dalam pembelajaran taksonomi seperti Zakree (2011) yang menggunakan sistem imun buatan sebagai asas kepada algoritma pembelajaran taksonomi manakala Yuan et al.

(2015) menggunakan pendekatan kecerdasan kerumunan iaitu menggunakan kaedah lebah koloni buatan untuk memeroleh hubungan bukan taksonomi.

Penyelidikan pembelajaran taksonomi terkini menggunakan kaedah metaheuristik adalah hasil penelitian Araujo et al. (2017) yang menggunakan algoritma evolusi tatabahasa untuk mengenal pasti hierarki konsep taksonomi dari Wikipedia. Setiap artikel di Wikipedia meliputi topik dan bersambung silang oleh hyperlink yang menghubungkan topik yang berkaitan. Taksonomi dan saling keberkaitannya kepada ontologi adalah sumber yang sangat berguna untuk banyak aplikasi kerana ia membolehkan carian semantik dan penalaran. Oleh itu, pengenalan automatik taksonomi yang terdiri daripada konsep-konsep yang berkaitan dengan laman Wikipedia yang berkaitan telah menarik banyak perhatian. Araujo et al. (2017) telah membangunkan sistem yang mengatur satu set konsep Wikipedia menjadi taksonomi. Teknik ini berdasarkan hubungan antara satu set ciri yang diekstrak dari kandungan halaman Wikipedia. Algoritma evolusi tatabahasa digunakan untuk mengetahui cara terbaik menggabungkan ciri-ciri yang dipertimbangkan dalam fungsi yang jelas. Fungsi-fungsi calon dinilai dengan menggunakan algoritma genetik untuk menghitung taksonomi optimum yang fungsi tersebut dapat sediakan sejumlah kes untuk pembelajaran. Ukuran prestasi dikira berdasarkan purata ketepatan yang diperoleh hasil pembandingan taksonomi rujukan dengan taksonomi yang dibangunkan secara automatik.

Walau bagaimanapun, algoritma metaheuristik masih jarang diteroka untuk membangunkan taksonomi. Antara algoritma diinspirasikan alam yang belum diterokai akan potensinya adalah algoritma kelip-kelip.

2.7.4 Kelip-Kelip

Menurut Lewis dan Cratsley (2008), serangga kelip-kelip atau kunang-kunang (Coleoptera: Lampyridae), adalah antara serangga yang paling ‘berkarisma’ dikalangan serangga, terutamanya cara kelip-kelip memikat yang telah memberi inspirasi kepada saintis. Kelip-kelip adalah sejenis serangga dari kumpulan kumbang. Menurut Fister et al. (2013), terdapat lebih 2000 spesis kelip-kelip di dunia dan spesis 'Pteroptyx Tener' adalah spesis yang terdapat di Kuala Selangor (Jusoh et al., 2013).

Keunikan yang ada pada kelip-kelip ini adalah pada ekornya, yang mengeluarkan kelipan cahaya. Serangga kelip-kelip lebih unik kerana cahaya yang dihasilkan oleh kerumunan serangga ini adalah serentak iaitu 3 kelipan sesaat.

Kelip-kelip hanya 6 mm panjang, hidup dalam iklim tropika di kawasan berpaya kerana sumber makanan utamanya adalah dari pokok Berembang. Pokok Berembang atau 'Sonneratia Caseolaris', sejenis pokok paya yang tumbuh liar adalah sebahagian dari ekosistem penting bagi kelip-kelip. Selain daripada itu, pokok Berembang juga penting sebagai habitat yang bertindak sebagai penapis untuk kotoran dan racun dan mengeluarkan air bersih untuk organisma dalam sungai. Jangkamasa hayat kelip-kelip ialah selama 2 hingga 3 bulan.

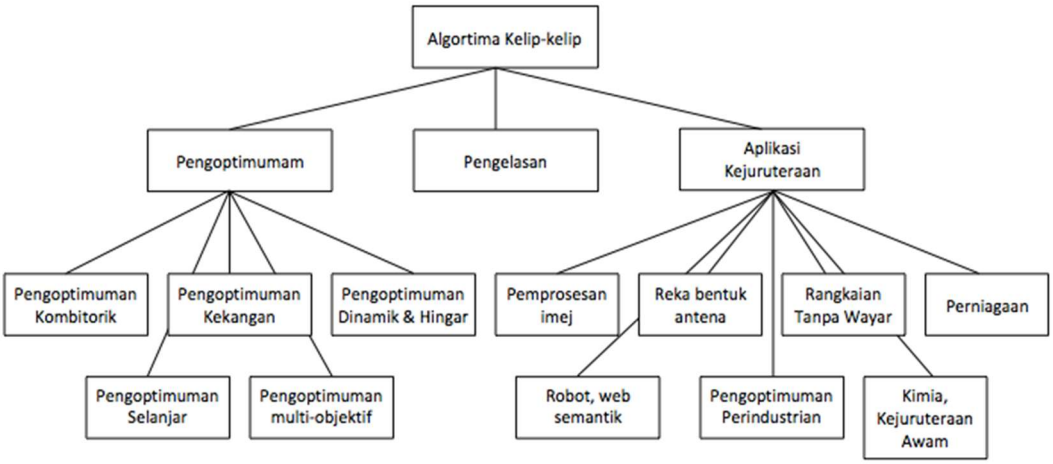
Kelipan yang dihasilkan oleh kelip-kelip adalah hasil proses biokimia iaitu biopendarcahaya *(bioluminescence)*. Organ yang menghasilkan tindak balas biopendarcahaya yang menghasilkan cahaya ialah lantera. Kelip-kelip jantan mengeluarkan cahaya yang lebih terang berbanding yang betina untuk menarik perhatian kelip-kelip betina. Kelip-kelip jantan berupaya mengawal biopendarcahaya untuk menghasilkan cahaya yang kuat dan berlainan (unik). Selain isyarat memikat, kelipan cahaya juga adalah untuk memberi amaran dan menghalau pemangsa. Namun ada juga kelip-kelip yang tidak mampu menghasilkan tindak balas biopendarcahaya. Kelip-kelip ini akan memikat pasangannya dengan menghasilkan feromon seperti semut. Lantera penghasil kelipan cahaya ini dimulakan oleh isyarat yang dikeluarkan dari sistem saraf pusat kelip-kelip.

Kebanyakan kelip-kelip bergantung kepada biopendarcahaya untuk memikat pasangannya. Lazimnya, kelip-kelip jantan akan mengeluarkan isyarat pertama untuk memikat kelip-kelip betina. Kelip-kelip betina akan membalas isyarat kelip-kelip jantan dengan menghasilkan kelipan cahaya yang berterusan. Kedua-dua kelip-kelip jantan dan betina menghasilkan isyarat cahaya yang pola kelipannya berbeza dan pemasaannya tepat untuk mengekod maklumat seperti identiti spesis dan jantina. Kelip-kelip betina tertarik kepada kelip-kelip jantan yang menghasilkan isyarat mengawan yang menunjukkan perbezaan sifat atau unik. Kebiasaannya, kelip-kelip betina tertarik kepada kelipan cahaya yang lebih terang.

Walaubagaimana pun, aras kecerahan cahaya atau keamatan cahaya berbeza-beza mengikut jarak sumber dan kelip-kelip betina tidak dapat membezakan di antara keamatan cahaya yang disebabkan oleh jarak atau kekuatan cahaya yang dihasilkan oleh lantera kelip-kelip jantan. Pancaran cahaya kelip-kelip amat mudah dilihat dan berkesan sebagai mekanisme pertahanan yang memberi amaran kepada pemangsa dan sekali gus mengelak kelip-kelip dari menjadi mangsa.

Terdapat dua ciri kecerdasan kerumunan iaitu i) organisasi swaatur dan ii) pembuatan keputusan tak terpusat. Kehidupan sosial kelip-kelip adalah sama seperti serangga lain yang hidup sekawan atau berkumpulan seperti lebah dan semut. Setiap kelip-kelip, seperti semut adalah individu berautonomi yang hidup dalam satu koloni yang harmoni. Keharmonian dicapai apabila setiap ekor kelip-kelip tidak hidup terpencil tetapi hidup bersosial iaitu saling berinteraksi dan berkomunikasi di antara satu sama lain.

Bagi kelip-kelip, kehidupan sosial atau kemasyarakatan ini bukan sahaja dalam aktiviti mencari makanan tetapi juga dalam pembiakan. Pembuatan keputusan secara kolektif berkait rapat dengan sifat dan tingkah laku memancarkan cahaya yang menjadi asas dan inspirasi kepada pembangunan algoritma kelip-kelip oleh saintis komputer bernama Yang (2010). Rajah 2.3 menunjukkan taksonomi aplikasi AKK dalam beberapa domain (Fister et al. 2013). Fister et al. (2013) telah menghasilkan hirarki aplikasi algoritma kelip-kelip menunjukkan bahawa AKK belum digunakan untuk tugas pengelompokan berhirarki namun Banati dan Bajaj (2013) telah melaporkan prestasi pengelompokan menggunakan AKK.



Rajah 2.3 Taksonomi Aplikasi AKK

Beberapa variasi algoritma kelip-kelip wujud dalam kesusasteraan. Fister et al. (2012) telah mencadangkan skim klasifikasi untuk mengklasifikasikan AKK kepada beberapa kategori dengan berasaskan kepada tetapan parameter mereka. Pengaturan tetapan parameter AKK ini adalah penting untuk mencapai prestasi yang baik yang perlu dipilih dengan berhati-hati. Secara umum, terdapat dua cara untuk menetapkan parameter algoritma dengan betul. Pertama adalah melalui penalaan nilai parameter sebelum algoritma dilarikan dan kaedah kedua adalah penalaan parameter dijalankan setelah AKK dilarikan. Penalaan parameter dijalankan setelah lengkap suatu lelaran. Selain daripada berdasarkan kaedah penetapan parameter, pengelasan yang digunakan oleh Fister et al. (2012) juga mengambil kira komponen dan ciri apakah yang terdapat algoritma AKK berkenaan. Berdasarkan kepada pengelasan Fister et al. (2013), Jadual 2.2 di bawah telah diisi dengan kajian berkaitan AKK.

Jadual 2.2 Aplikasi AKK dipelbagai bidang

Bidang	Rujukan
Pengoptimuman Perancangan	(Osaba et al 2016); (Tsuya et al. 2017);
Pengoptimuman Pengindustrian	(Singh et al 2017); (Ding & Liu 2017); (Teshome et al. 2017); (Apostolopoulos & Viachos 2011); (Yang et al. 2011); (Mauder et al. 2011); (Chatterjee et al. 2012); (Kazemzadeh 2011); (Aungkulonon 2011); (Rampriya 2010); (Chandrasekaran & Simon 2011); (Hu 2012); (Deckichi et al. 2012); (Roeva & Slavov 2012); (Roeva 2012); (Obedinia et al. 2012); (Dutta et al. 2011)
Pemprosesan Imej	(Su et al 2017); (Rodrigues et al. 2017); (Mary & Singh 2017); (Zhang & Wu 2012); (Horng & Jiang 2010a); (Horng 2012); (Hassanzadeh et al 2011a); (Hassanzadeh et al 2011b); (Mohd Noor et al. 2011); (Horng & Jiang 2010b); (Horng & Liou 2011)
Rekabentuk Antena	(Basu & Mahanti 2011); (Zaman & Matin 2012); (Chatterjee & Mahanti 2012); (Basu & Mahanti 2012)
Pengoptimuman Perniagaan	(Gomathi et al. 2017); (Yang et al. 2011); (Giannakouris et al. 2010)
Kejuruteraan Awam	(Erdal 2017); (Talatahari et al. 2012); (Gholizadeh & Barati 2012)
Robotik	(Hidalgo et al. 2017); (Jakimovski et al. 2010); (Severin & Rossmann 2012)
Web Semantik	(Bekkouche et al. 2017); (Pop et al. 2011); (Salomie et al. 2014)
Kejuruteraan Elektrik dan Tenaga Solar	(Ibrahim & Khatib 2017); (Wang et al. 2017)
Kejuruteraan Kimia	(Fateen et al. 2012)
Meteorologi	(Santos et al. 2013)
Rangkaian (Keselamatan)	(Adaniya et al. 2012)
Rangkaian Penderia Tanpa Wayar	(Breza & McCann 2008); (Sun et al. 2012); (Sarma & Gopi 2014)
Perlombongan Data (Algoritma Pengelompokan)	(Jain et al. 2017); (Senthilnath et al. 2011); (Hassanzadeh & Meybodi 2012); (Lei et al. 2016); (Nayak et al. 2014); (Mohammed et al. 2014);

Apa yang dapat disimpulkan dari Jadual 2.2 yang terhasil di atas adalah pembelajaran ontologi dan pembelajaran taksonomi masih belum diuji dengan AKK atau pun yang diinspirasikan oleh AKK.

2.7.5 Pendekatan Hibrid

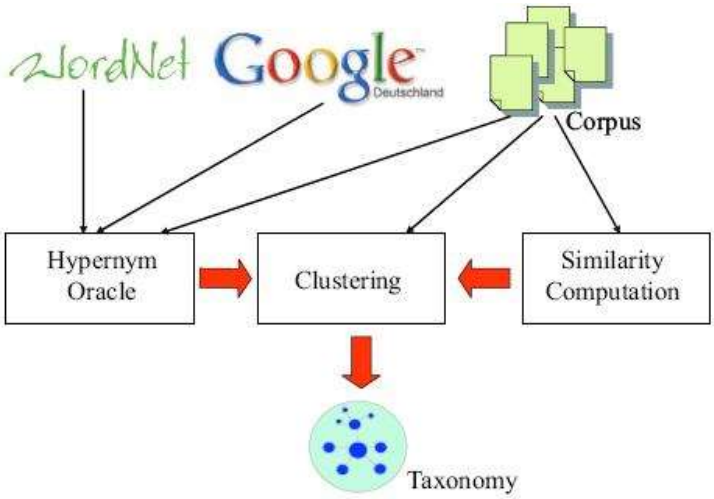
Dalam usaha untuk meningkatkan keberkesanan pembelajaran taksonomi, ramai penyelidik telah membangunkan pendekatan yang mengeksploitasi paradigma pembelajaran yang berbeza atau menggabungkan pendekatan atau teknik yang berbeza. Pendekatan hibrid dalam pembelajaran ontologi dapat dilihat dalam Jiang dan Tan (2010) dan Zakree (2011). Jiang dan Tan (2010) mencadangkan satu sistem

pembelajaran yang dikenali sebagai Concept-Relation-Concept Tuple-based ontology Learning (CRCTOL). CRCTOL dibangunkan untuk melombong ontologi secara automatik dari dokumen dari domain yang khusus. Perbezaan utama di antara CRCTOL dan sebelumnya ialah CRCTOL mengamalkan teknik penguraian teks penuh dan mengaplikasi gabungan kaedah statistik dan lexico-sintaktik. Mereka juga menggunakan algoritma statistik yang memetik konsep utama dari koleksi dokumen, algoritma yang menyelesaikan isu ketaksamaan makna konsep utama, algoritma berasaskan petua yang memetik hubungan antara konsep-konsep utama, dan algoritma memangkas hubungan tidak penting yang terhasil dari pembelajaran ontologi. Berbanding dengan alat lain seperti OntoLearn, Tex-to-Onto dan OntoBuilder, CRCTOL menghasilkan ontologi yang lebih padat serta tepat.

Stoica dan Hearst (2004) sebagai misal menggunakan WordNet untuk meningkatkan keberkesanan pendekatan analisis kejadian-bersama yang dicadangkan oleh Sanderson dan Croft (1999). Mereka menggunakan WordNet untuk mewujudkan kategori metadata yang menggambarkan kandungan kumpulan maklumat sasaran. Pertama, mereka akan memilih kata-kata tertentu sebagai perwakilan yang dapat mewakilkan topik yang terkandung dalam dokumen. Untuk setiap perkataan yang dipilih, mereka mendapatkan makna perkataan tersebut dari hypernym pertama yang ditemui dari WordNet. Kemudian, mereka membina taksonomi dengan mengumpulkan kesemua hypernym bagi setiap perkataan. Kemudian taksonomi tu dimampat. Satu lagi contoh yang menggabungkan paradigma yang berlainan dalam pembelajaran taksonomi ialah dengan menggunakan petua sekutuan (Maedche dan Staab 2000). Teknik ini mampu untuk menerokai hubungan selain hubungan taksonomi antara konsep dengan menggunakan taksonomi sebagai pengetahuan latar belakang dan statistik kejadian-bersama dalam teks.

Cimiano dan Staab (2005) telah mencadangkan pendekatan hibrid yang diberi nama Pengelompokan Hierarki Agglomerat Berpandu (GAHC). GAHC telah menunjukkan hasil yang baik dalam memperbaiki teknik pengelompokan berhieraki agglomerat dengan menghibrid HAC dengan pendekatan linguistik. Untuk mengatasi kelemahan kaedah berasaskan-hipotesis Harris, Cimiano dan Staab (2005) menggunakan WordNet dan Google untuk mendapatkan hubungan. Dalam kajian ini,

"Google desktop" digunakan untuk pemilihan ciri dari korpus yang disimpan di dalam komputer meja dan bukannya dari Web. Penyelidik sebelum ini menggunakan Google untuk mencari hypernim dari Web manakala WordNet digunakan untuk mencipta metadata dan juga untuk mencari hypernim. Pendekatan beliau adalah serupa dengan kaedah induksi taksonomi yang berasaskan persamaan teragih. Tetapi, perbezaan utama GAHC dari pengelompokan tanpa selia yang lain ialah GAHC mengeksploitasi hypernim yang diperolehi dari kaedha lain untuk untuk memandu proses pengelompokan. Dalam erti kata lain, algoritma pengelompokan tanpa selia menjalankan proses pengelompokan dengan 'petunjuk'. Petunjuk ini diperolehi daripada sumber-sumber lain seperti WordNet, World Wide Web serta pola lexico-sintaktik yang menunjukkan hubungan hyponim-hypernim seperti yang dicadangkan Hearst (1992). Kumpulan hubungan taksonomi yang terkumpul ini disimpan di dalam storan bernama Hypernym Oracle (HO).



Rajah 2.4 Pendekatan Pengelompokan Hierarki Berpandu Agglomerative
imej yang diguna pakai dari Cimiano dan Staab (2005)

Sebagai contoh, diberi dua perkataan yang dianggap sama kerana konteks yang diperolehi dari korpusnya adalah sama, maka GAHC akan merujuk kepada HO. Sekiranya kedua-dua perkataan memiliki hypernim yang sama maka kedua-dua perkataan tersebut akan menjadi anak kepada hypernim yang ditemui dalam HO. Rajah 2.4 berikut menggambarkan bagaimana GAHC beroperasi. Seperti pendekatan berasaskan hipotesis Harris yang lain, GAHC menggunakan kosinus untuk mengira jarak persamaan antara dua perkataan. Oleh itu, sebelum GAHC menggunakan

algoritma pengelompokan, HO perlu dibangunkan terlebih dahulu. GAHC menggunakan tiga sumber untuk membangunkan HO iaitu WordNet, padanan pola Hearst di dalam korpus dan padanan Hearst di Web.

Kekurangan bergantung kepada pola Hearst dan WordNet telah pun dibincangkan sebelum ini. Chen Li (2006) menyatakan bahawa pola Lexico-sintaktik jarang muncul dan kebanyakan penggunaan 'adalah (is_a)' tidak muncul seperti pola Hearst.

Teknik pengelompokan konsep diperkenalkan oleh Michalsky (1980) dan Faure & Poibeau (2000). Ia adalah agak serupa dengan kaedah berasaskan persamaan, tetapi tidak sama dari segi cara mereka mengukur jarak persamaan antara dua konsep. Gomez-Perez et al. (2005) memanggil teknik ini sebagai pengelompokan konseptual. Michalsky (1980) dan Faure & Poibeau (2000) menganggap bahawa dua konsep tergolong dalam kumpulan yang sama jika jarak semantik mereka adalah lebih rendah daripada nilai ambang yang telah ditentukan. Mereka mengira jarak antara konsep berdasarkan fungsi sintaksis untuk menentukan sama ada istilah adalah berkaitan dengan konsep dalam teks. Sebagai contoh:

Ayat 1: Ali mengembara dengan kereta api

Ayat 2: Ahmad mengembara dengan basikal.

Contoh yang dipinjam dari Gomez-Perez et al. (2005) menunjukkan bahawa jika kereta api dan basikal muncul di majlis yang sama sintaksis (contohnya, subjek dan kata kerja), maka konsep-konsep yang berkaitan dengan kereta api dan kereta dianggap secara semantik sebagai konsep yang sama maknanya dan perlu dikumpulkan.

Maedche dan Staab (2001) memperoleh konsep yang generik serta domain spesifik dengan menggunakan alat pemprosesan bahasa tabie (NLP) yang menggunakan pengekstrakan berasaskan pola dan pengelompokan konseptual. Walau bagaimanapun, kaedah ini menyatukan konsep yang dipeorlehi dengan ontologi teras (SENSUS, WordNet, dll.) menggunakan hubungan Sub-Kelas. Walau bagaimanapun, penggunaan semula ontologi teras atau sumber-sumber lain seperti WordNet dilihat tidak berdaya maju dalam pembelajaran konsep dari teks Melayu.

Rajah 2.4 menunjukkan Google menjadi sumber hubungan taksonomi bagi Cimiano & Staab (2005). Google telah dieksploitasi oleh ramai penyelidik kerana jumlah pengulangan maklumat boleh mewakili ukuran relevan. Sebab utama trend menggunakan web sebagai sumber maklumat dalam pemerolehan pengetahuan kerana banyak maklumat yang ada dalam Web dan frekuensi ulangan yang tinggi. Pendekatan ini diilhamkan oleh buku Surowiecki (2004) mengenai pengetahuan kolektif. Dia menyatakan bahawa pengetahuan kolektif dari khalayak adalah lebih baik daripada satu sumber pengetahuan. Soriano (2005) menyatakan bahawa "Sesetengah maklumat yang berulang terkandung dalam koleksi dokumen web membolehkan konteks dibina semula walaupun walaupun kehilangan sebahagian maklumat. Pemisahan yang boleh membantu untuk mendapatkan maklumat yang betul tetapi juga meningkatkan maklumat yang tidak relevan.

Faedah menggunakan laman web sebagai salah satu sumber maklumat yang boleh dilihat dalam kerja-kerja seperti Blohm dan Cimiano (2007), Lau et al. (2009) dan Makrehchi dan Kamel (2007) menggunakan Web untuk mendapatkan data dan maklumat mengenai data. Dalam usaha untuk mencari Web, banyak bergantung kepada Google kerana ia dianggap sebagai enjin carian yang paling popular di dunia (Wolf 2009). Chung et al. (2006), Ruenes (2007) dan Makhrehci dan Kamel (2007), misalnya, telah menggunakan API Google untuk mendapatkan maklumat dari Web. Mereka cuba untuk mengekstrak maklumat dari tujuh set data dengan menggunakan algoritma pola induksi terselia yang lemah. Walau bagaimanapun, keputusan eksperimen mereka menggunakan data di komputer meja telah menunjukkan bahawa kekurangan data berulang (frekuensi rendah) memerlukan data yang besar untuk latihan (pembelajaran mesin). Oleh itu, mereka menggunakan web untuk mengurangkan isu kejarangan data yang akan membawa kepada hasil yang lebih baik.

Sementara itu, Google Desktop juga telah mendapatkan populariti di kalangan penyelidik dalam pemerolehan pengetahuan. Trend baru muncul boleh dilihat di Erinjeri et al. (2008), Chau et al. (2008), Turnbull (2006) dan Zakree (2011). Erinjeri et al. (2008), sebagai contoh, menggunakan API Google Desktop untuk mengindeks pelayan fail mereka. Sebelum itu, 2.9 juta laporan teks turun dari sistem maklumat radiologi untuk fileserver. Kemudian, alat Google Desktop berasaskan perlombongan

digunakan untuk melombong laporan radiologi. Dengan menggunakan kuasa teknologi Google untuk pengguna akhir, ia telah meningkatkan produktiviti akademik radiologi dalam tugas-tugas klinikal dan penyelidikan.

Fortuna et al. (2005) membangunkan kaedah hibrid menggunakan algoritma pengelompokan K-Min dan Pendekatan Pengindeksan Semantik Terpendam (LSI) yang diperkenalkan oleh Deerwester et al. (1990). Pengindeksan Semantik Terpendam (LSI) adalah teknik untuk mengekstrak pengetahuan latar belakang ini dari dokumen teks. Ia menggunakan teknik dari algebra linear yang dipanggil Penguraian Nilai Singular (SVD) dan perwakilan *bag-of-words* untuk mengekstrak perkataan dengan makna yang sama. Fortuna et al. (2005) melihat penggunaan SVD dan Bag-of-words sebagai kaedah pengekstrakan konsep semantik tersembunyi atau topik dari dokumen teks. Walau bagaimanapun, Fortuna et al. (2005) hanya menggunakan algoritma K-Min dan LSI sebagai alat yang mencadangkan hubungan antara istilah (konsep).

Apabila pengguna alat yang dibangunkan oleh Fortuna et al. (2005) memilih topik, sistem secara automatik mencadangkan beberapa topik sebagai sub topik yang dipilih. Cadangan ini dilakukan oleh algoritma LSI atau k-min. Bilangan topik yang dicadangkan diselia oleh pengguna. Kemudian, pengguna memilih subtopik yang dirasakan munasabah dan sistem secara automatik menambahnya kepada ontologi dengan hubungan ’subtopik_kepada’ kepada topik yang dipilih. Pengguna juga boleh membuat keputusan untuk menggantikan topik yang dipilih dengan subtopik yang disyorkan. Hasil kerja Fortuna et al. (2005) memberi inspirasi kepada penyelidikan ini bahawa SVD dan algoritma pengelompokan boleh digunakan untuk mengenalpasti hubungan taksonomi kerana hubungan ‘subtopik kepada’ topik adalah seumpama hubungan ‘adalah’. Lebih-lebih lagi, penggunaan LSI mahu pun SVD belum teruji dalam mana-mana penyelidikan pembangunan taksonomi dari teks Melayu.

2.8 RINGKASAN DAN PERBINCANGAN

Dalam bab ini, kajian yang berkaitan dengan tesis ini telah dibincangkan dan gambaran keseluruhan mekanisme pembelajaran yang berbeza dan kajian lepas telah dibentangkan. Gambaran keseluruhan mengenai penyelidikan dalam bidang ini boleh

didapati di Drumond dan Girardi (2008), Zavitsanos et al. (2006) dan Gomez-Perez et al (2005).

Salah satu soalan yang paling asas dalam pembelajaran taksonomi dari teks-Melayu adalah pemilihan ciri terbaik untuk mewakili satu konsep. Zakree (2011) telah mengandaikan bahawa bahawa pemerolehan ciri sedia ada bersumberkan sintaksis Inggeris dan pola lexico-sintaktik boleh bekerja pada teks Melayu. Tesis beliau telah mendapat hasil yang cukup baik namun penulis mempercayai bahawa bahasa Melayu masih mempunyai ciri unik yang belum dikenal pasti oleh penyelidik sebelum ini untuk melengkapkan ciri sintaktik Melayu yang sedia ada dan lexico sintaktik-pola Inggeris yang 'diMelayukan' oleh Zakree (2011).

Pembelajaran taksonomi tanpa selia lazimnya dicirikan oleh hipotesis pengagihan Harris (1954). Dia mendakwa bahawa istilah itu serupa secara semantik setakat mana mereka berkongsi konteks sintaktik yang sama. Bagi tujuan ini, bagi setiap kata nama (konsep), ciri-ciri dari teks, contohnya pergantungan sintaktik, harus diekstrak dari teks. Beberapa kaedah pengelompokan berdasarkan pendekatan ini pengelompokan berhierarki aglomerat (iaitu, pautan tunggal, pautan purata dan pautan lengkap), algoritma yang membahagi seperti K-Min Pembahagi Dua Sama dan Konsep Analisis Formal (FCA). Walaupun terdapat usaha penyelidikan untuk memantapkan bidang ini, masih terdapat ruang yang luas untuk meningkatkan keberkesanan algoritma pembelajaran taksonomi terutamanya dalam menyelesaikan masalah kejarangan dan hingar data. Walaupun pendekatan Zakree (2011) telah disiapkan, terdapat beberapa masalah yang dikongsi bersama oleh pendekatan ini. Isu utama adalah mengenai kejarangan data. Pendekatan ini bergantung kepada kaedah penilaian persamaan semantik antara perkataan berdasarkan jumlah konteks linguistik yang dikongsi bersama. Cimiano (2006) menyatakan bahawa kadang-kadang, persamaan sintaktik boleh berlaku tanpa disengajakan sekali gus tidak sepadan dengan makna dalam dunia sebenar. Malah, pendekatan ini tidak berupaya untuk melabel kelompok yang dihasilkan. Selain itu, Cimiano (2006) turut memetik Ziph (1932) yang menyatakan bahawa andaian kesempurnaan maklumat tidak akan pernah dapat dipenuhi, kerana koleksi teks tidak akan pernah cukup besar untuk mencari kesemua ciri-ciri konteks yang diperlukan.

Pelbagai percubaan untuk mengatasi isu kejarangan dan hingar data telah diterokai. Cimiano dan Staab (2005) misalnya memperkenalkan GAHC yang meksploitasi WordNet dan membina HO sebagai panduan kepada pengekstrakan hypernim dan hyponim yang lebih jitu daripada teks. GAHC menghasilkan keputusan yang lebih baik berbanding teknik lain tanpa pengawasan seperti HAC atau FCA. Zakree (2011) trut dipengaruhi GAHC ketika merekabentuk CLONALG dan CLOSAT. Walau bagaimanapun, pendekatan yang berdasarkan yang hampir sama pola mengalami penarikan balik yang sangat rendah yang disebabkan oleh hakikat bahawa pola lexico-sintaktik yang sangat jarang ditemui.

Lebih teknikal dan khusus suatu teks, maka lebih kurang atau jarang pengetahuan asas dalam bentuk taksonomi yang boleh ditemui Cimiano (2006). Selain itu, Tze & Hussien (2006) menyatakan bahawa terlalu kerap istilah multi makna wujud pada WordNet yang menyebabkan makna untuk domain yang spesifik sukar dipilih. Zakree (2011) juga menyatakan WordNet tidak sesuai diterjemahkan sebagai perwakilan makna Melayu kerana penterjemahan bukan sekadar suatu perbuatan pemindahan bahasa, tetapi ia juga melibatkan interaksi budaya dan agama. Pemindahan makna (terjemahan) yang melibatkan budaya dan agama adalah suatu isu yang mungkin melibatkan isu sensitif. Geffet dan Dagan (2005) membuktikan bahawa penggunaan hipotesis pengagihan persamaan Harris boleh digunakan untuk 'menangkap' makna semantik.

Secara ringkasnya, jurutera ontologi bahasa Melayu memerlukan garis panduan mengenai keberkesanan, kecekapan dan keseimbangan kaedah yang berbeza untuk membuat keputusan yang teknik-teknik untuk memohon di mana tetapan. Selain Zakree (2011), setelah 6 tahun berlalu tidak ada kerja perbandingan yang sistematik untuk menganalisa teknik-teknik dan algoritma yang berbeza dalam bidang pembelajaran taksonomi dari teks Melayu. Senario ini adalah penghalang utama untuk kajian ini kerana ia membawa kepada masalah kekurangan garis panduan (Cimiano 2006). Kekurangan garis panduan merujuk kepada kekurangan penyelidikan di bidang berkaitan yang menyebabkan tiada garis panduan tentang keberkesanan, kecekapan dan keseimbangan kaedah yang berlainan untuk menyokong penciptaan automatik ontologi daripada teks Melayu.

Jelas sekali cadangan penyelesaian kepada masalah-masalah ini mesti beradaptasi dan mantap. Ia mesti menyesuaikan diri untuk dua sebab, (i) masalah kejarangan data dan (ii) 'hingar' dalam istilah ciri-ciri dan diekstrak. Kerja-kerja sebelum ini telah menunjukkan bahawa AKK mempunyai ciri-ciri yang memenuhi kriteria penyelesaian masalah ini. Terdapat beberapa motivasi untuk menggunakan AKK sebagai inspirasi. AKK telah berevolusi dan berkembang dengan pesat semenjak ia diperkenalkan pada 2008. Fister et al. (2012) telah mengkaji sifat dan kebolehan AKK dan menyifatkan bahawa AAK ini:

1) Mempunyai ciri kepelbagaian mod.
2) Boleh mengendalikan masalah kepelbagaian dengan efisyen
3) Mempunyai kadar penumpuan (*convergen*) yang pantas
4) Boleh digunakan untuk menyelesai masalah gelintaran umum mahu pun global, malah boleh digunakan sebagai gelintaran heuristik tempatan.
5) Boleh diaplikasi dalam kebanyakan domain.

Fister et al. (2012) menyimpulkan bahawa AKK menyelesaikan beberapa isu dan permasalah berikut: pengoptimuman selanjar, pengoptimuman kombinatorial, pengoptimuman kekangan, pengoptimuman multi-objektif dan juga persekitaran yang dinamik dan hingar. Namun sehingga kini, AKK masih belum ditemui dalam mana-mana kajian dan liputan kesusasteraan yang ia telah diaplikasi untuk pembelajaran taksonomi. Justeru, keberkesanan AKK dalam pembelajaran taksonomi merupakan suatu persoalan kajian yang perlu dijawab sebagai suatu sumbangan kepada dunia pembelajaran ontologi dan web semantic. Ketiadaan kajian berkenaan AKK sebagai algoritma pengelompokan berhierarki memberikan justifikasi bahawa kajian ini sangat penting dijalankan kerana ia belum terbukti berkesan dalam domain pembelejaran mesin. Oleh yang demikian, objektif kajian kesusateraan ini telah tercapai dalam mencari jurang yang terdapat pada AKK.

Kekurangan pendekatan sedia ada dalam menangani isu kejarangan dan hingar data khususnya untuk perlombongan teks Melayu telah diperbincangkan untuk mencari ruang sumbangan tesis ini. Tesis ini telah menemui pernyataan masalah yang

jelas yang jika dikaji akan memberi sumbangan signifikan kepada badan ilmu. Tesis ini memilih untuk meneroka penggunaan AKK untuk pembelajaran taksonomi.