KAEDAH DANGKAL BAGI PENGEKSTRAKAN RANGKAIAN SOSIAL AKADEMIK DARI WEB

MAHYUDDIN (MAHYUDDIN K. M. NASUTION)

UNIVERSITI KEBANGSAAN MALAYSIA

KAEDAH DANGKAL BAGI PENGEKSTRAKAN RANGKAIAN SOSIAL AKADEMIK DARI WEB

MAHYUDDIN (MAHYUDDIN K. M. NASUTION)

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

SUPERFICIAL METHOD FOR EXTRACTING ACADEMIC SOCIAL NETWORK FROM THE WEB

MAHYUDDIN (MAHYUDDIN K. M. NASUTION)

THESIS SUBMITTED IN FULFILMENT FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

PENGAKUAN

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

17 Januari 2013 MAHYUDDIN (MAHYUDDIN K. M. NASUTION)
P42116

PENGHARGAAN

Syukur Alhamdulillah kepada Allah S.W.T kerana dengan limpah kurniaNya telah memberikan saya kesihatan yang cukup, masa dan kematangan fikiran untuk menyiapkan kajian ini dalam bentuk sebegini rupa. Penghargaan ini ditujukan khas kepada semua insan yang terlibat secara langsung atau tidak langsung dalam membantu saya menyiapkan tesis ini.

Setinggi-tinggi penghargaan ditunjukan buat penyelia utama saya, Prof. Dr. Shahrul Azman Mohd Noah yang telah banyak meluangkan masa bagi menyelia dan memantau sepanjang penyelidikan ini dilaksanakan. Segala tunjuk ajar, bantuan, nasihat dan kritikan yang diberikan sangat berharga bagi memastikan penyelidikan ini dilaksanakan dengan sempurna. Terima kasih yang tidak terhingga diucapkan kerana tidak jemu memperbaiki setiap kesilapan saya dari segi pembangunan penyelidikan dan penulisan tesis ini.

Ribuan terima kasih juga ditujukan buat semua kakitangan Fakulti Teknologi Sains dan Maklumat (FTSM) di atas segala bantuan dan kemudahan yang diberikan. Penghargaan ini juga ditujukan kepada projek ScienFund 01-01-02-SF0097 dan skim bantuan UKM di atas sokongan dan sumbangan yang diberikan. Untuk rakan-rakan di makmal *Knowledge Technology* (KT) *Group Research – Center for Artificial Intelligence Technology* (CAIT) FTSM UKM yang tidak dapat disebutkan namanya, terima kasih atas segala ilmu pengetahuan, sokongan dan dorongan yang diberikan. Terima kasih juga ditujukan buat semua pihak pada Direktorat Pendidikan Tinggi Republik Indonesia atas sokongan dana pendidikan, terutama kepada Pihak Rektorat Universitas Sumatera Utara Medan Indonesia yang telah memberikan peluang dan waktu kepada saya untuk menyelesaikan program ini.

Buat istri (Maria Elfida) dan anak-anak (Maudy Maulina dan Raditya Macy Widyatamaka Nasution) serta seluruh keluarga tercinta, teristimewa buat ibunda Mahyon, doa restu, nasihat dan dorongan yang diberikan amat bermakna buat diri saya sehingga berjaya meneruskan pengajian di peringkat yang lebih tinggi. Namun demikian, dihujung pengajian ini, Ibunda telah kembali ke rahmatullah, Al-Fatihah. Sesungguhnya jasa semua insan yang terlibat, hanya Allah SWT sahaya yang mampu membalasnya. Al-Fatihah juga buat ayahanda (Khairuddin Matyuso Nasution) dan ibu mertua (Nursetia) yang telah kembali ke rahmatullah semasa penyelidikan ini dilaksanakan. Amin.

ABSTRAK

Penyelidikan telah menunjukkan kemungkinan pengekstrakan rangkaian sosial dari Web. Perkara yang paling penting dalam pengekstrakan rangkaian sosial adalah untuk mengenalpasti pelakon sosial yang sesuai dan hubungan yang mungkin wujud antara pasangan pelakon. Pengekstrakan ini adalah proses yang kompleks dan memakan masa. Terdapat dua aliran penyelidikan bagi pengekstrakan rangkaian sosial, iaitu pendekatan diselia dan tak diselia. Penyelidikan diselia melibatkan analisis korpus untuk mengenali entiti dan hubungan mereka dan juga label hubungan. Namun demikian, menamakan label hanya terhad kepada label yang telah ditetapkan dalam set latihan. Aliran tak diselia, sebaliknya, melibatkan hasil enjin carian untuk menjanakan rangkaian sosial tetapi dengan maklumat yang terhad. Sehubungan dengan itu, tujuan utama penyelidikan ini adalah untuk mempertingkatkan kaedah dangkal bagi pengekstrakan rangkaian sosial dengan melaksanakan teknik capaian maklumat, seperti nyahkekaburan nama. Kajian ini mencadangkan kaedah baru dengan mengeksploitasi *snippet* hasil carian Web yang dihasilkan oleh enjin carian. Pengekstrakan bermula dengan pengesanan hubungan antara sepasang pelakon yang dikemukakan sebagai kueri kepada enjin carian. Dalam kajian ini, diberikan tumpuan kepada empat hubungan akademik yang penting: ‘pengarang-bersama’, ’kumpulan-penyelidikan’, ’persidangan-saintifik’, dan ‘peranan-akademik’. Setiap hubungan kemudian ditakrifkan dengan senarai kata berkaitan. Maklumat dijana dari hasil carian seperti kiraan dan URL yang digunakan untuk menjanakan kekuatan hubungan, manakala petunjuk bagi hubungan didasarkan kepada frekuensi ternormalisasi dan nilai TF.IDF yang ditemukan dalam *snippet*. Pilihan label bagi setiap hubungan berasaskan nilai kekuatan dan nilai kesamaan perkataan. Kompleksiti kaedah dicadangkan adalah $O(mn)$ yang mana m dan n adalah bilangan pelakon, yang lebih baik daripada kompleksiti kaedah sebelumnya iaitu $O(n^2)$. Semasa proses penilaian, sebanyak 76 nama pensyarah Fakulti Teknologi dan Sains Maklumat UKM digunakan sebagai benih untuk menjanakan pelakon lain. Dalam usaha untuk mengesahkan hubungan yang dihasilkan, soal selidik telah diedarkan di kalangan penyelidik tersurat dan tersirat dengan menyebutkan hubungan antara pelakon yang mengambil bahagian. Keputusan menunjukkan potensi pendekatan dicadangkan meningkatkan prestasi pendekatan semasa untuk mengekstrak rangkaian sosial dari Web, iaitu dapatan semula daripada kaedah yang diajukan adalah tinggi (70-90%), akibatnya kejituan menurun (40-12%), tetapi masih lebih baik dibandingkan dengan kaedah sebelumnya.

SUPERFICIAL METHOD FOR EXTRACTING ACADEMIC SOCIAL NETWORK FROM THE WEB

ABSTRACT

Research has shown the possibility of extracting social network information from the Web. The most critical part in social network extraction is to identify appropriate social actors and their relations, and this is a very complex and time consuming process. There are two research streams for extracting social network, which are the supervised and unsupervised approach. The supervised research involves corpus analysis which results in the identification of entities and their relations as well as the label for the relations. However, naming of labels is only restricted to the predefined labels in the training set. The unsupervised stream, on the other hand makes use of search engines results to generate social network but information for labelling is limited. Hence, the main aim of this research is to enhance the superficial method for extracting social network by implementing the techniques of information retrieval, such as name disambiguation. This study proposed a new method by exploiting the web search results snippet produced by search engines. It starts with the detection of relation between pair of actors submitted as query to a search engine. In our study we focus on four important academic relations which are 'co-author', 'research group', 'scientific event', and 'academic role'. Each of these relations is further defined with list of associated terms. Information generated from the search results such as the hit count and URL are used to generate strength relation, whereas clue for relations is based on the normalised term frequencies and TF.IDF values of the associated terms found in the snippets. The choice of label for relation is then based on the product of strength value and relation values based on similarity of term. The complexity of proposed method is $O(mn)$ where m and n are the number of actors, which is better than the complexity of previous method, $O(n^2)$. During the evaluation, a total of 76 lecturer names from the Faculty of Information Science and Technology, UKM are used as seeds to generate other actors. In order to validate the generated relations, questionnaires are distributed among researchers that explicitly and implicitly mentioned the relation among the participating actors. Results show the potential of the proposed approach to improve the performance of the current approaches for extracting social network from the Web. The recall values are readily high (70-90%), consequently the precision decreases (40-12%), but it is better than the performance of previous method.

KANDUNGAN

Halaman

LAMPIRAN

SENARAI ILUSTRASI

SENARAI JADUAL

SENARAI SINGKATAN

ACM	*Association for Computing Machinery*
ACM DL	ACM *Digital Library*
ACT	*Author Conference Topic*
APT	*Author Persona Topic*
ARS	*Association Rule and Similarity*
ART	*Author Recipient Topic*
ASNQ	*Academic Social Network Questionnaire*
BSM	*Basic Superficial Method*
BSMp	*Basic Superficial Method based on pattern*
BSMV	*Basic Superficial Method based on variation*, contoh BSMV1 dan BSMV2
CEAS	*Conference on Email and Anti-Spam*
CEC-EEE	*Conference on E-Commerce Technology* (CEC) *and Enterprise Computing, E-Commerce and E-Services* (EEE)
CRF	*Conditional Random Field*
CSS	*Cascading Style Sheets*
DART	*Discriminative ART*
Das	Dapatan semula
DBLP	*Database System and Logic Programming, Digital Bibliography & Library Project*
DbSM	*Description based Superficial Method*
DDbSM	*Deep Description based Superficial Method*
DMO	*Data Mining and Optimization*
DNS	*Domain Name System*
DTIC	*Defense Technical Information Center*
EI	*Engineering Index* (*Compedex*)
EM	*Expectation Maximization*
ESWC	*Extended Semantic Web Conference*
FOAF	*Friend-of-a-Friend*
GPM	*Generative Probabilistic Model*
GOT	*Group Over Time*

GSI	*Group Switch Index*
GT	*Group Topic*
Har	*Harmonic mean*
HMRF	*Hidden Markov Random Field*
HITS	*Hyperlink-Induced Topic Search*
HTML	*HyperText Markup Language*
HTTP	*Hypertext Transfer Protocol*
IC	*Industrial Computing*
ICML	*International Conference on Machine Learning*
IDF	*Inverse Document Frequency*
IEEE	*Institute of Electrical and Electronics Engineers*
IJCAI	*International Joint Conferences on Artificial Intelligence*
INT	*Integration* (Perpaduan)
ISI	*The Institute for Scientific Information*, Thomson ISI
ISWC	*International Semantic Web Conference*
JAIR	*Journal of Artificial Intelligence Research*
KBK	Komunikasi Berperantaraan Komputer
KDD	*Knowledge Discovery and Data Mining*
Kej	Kejituan
KES	Peng(K)ESanan
KHL	Kekuatan Hubungan Langsung
KHtL	Kekuatan Hubungan tak Langsung
KT	*Knowledge Technology*
LinkKDD	*Link KDD*
MRR	*Mean Reciprocal Rank*
MSU	*Multimedia Software and Usability*
NIPS	*Neural Information Processing System*
NPD	Normalisasi pemusatan derajah
PM	Profesor Madya
PP	Pemusatan Perantaraan
PR	*Pattern Recognition*
Prof.	Profesor
RART	*Role-ART*

RbSM	*Rule based Superficial Method*
RSPB	Rangkaian Sosial Pengarang-Bersama
RSPS	Rangkaian Sosial Persidangan Saintifik
RSKP	Rangkaian Sosial Kumpulan Penyelidikan
SIAM	*Society for Industrial and Applied Mathematics*
SIGKDD	*ACM SIGKDD Conference on KDD*
SIS	*Strategic Information System*
SNA	*Social Network Analysis*
SS	*Service Science*
STP	*Software Technology and Programming*
TF	*Term Frequency*
TF.IDF	*Term Frequency – Inverse Document Frequency*
URL	*Uniform Resource Locators*
USM	*Underlying Superficial Method*
USR	*Underlying Strength Relation*
WWW	*World Wide Web* atau Merujuk kepada *WWW Consortium* (W3C)
XHTML	*eXtended* HTML
XML	*Extensible Markup Language*

SENARAI DEFINISI

BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Interaksi sosial dengan melibatkan teknologi telah memainkan peranan penting pada kehidupan seharian, yang pada amnya berguna untuk menubuhkan keselarasan sosial sama ada dalam satu komuniti ataupun pada satu negara. Setiap orang menjalankan komunikasi dengan orang lain sama ada secara konvensional ataupun membabitkan teknologi maklumat (Domingos 2005; Breslin & Decker 2007). Teknologi maklumat semasa penggunaannya terdiri daripada dua maksud yang berbeza: positif dan negatif. Kedua-dua perkara ini mempengaruhi ke arah mana perubahan sosial. Manusia moden telah bergerak jauh dari keadaan manusia nomad yang hidup dengan apa adanya. Manusia moden saat ini dalam hidup sehariannya memerlukan teknologi maklumat untuk membangun dan melakukan transformasi sebagai satu pakej perubahan untuk mencapai kesejahteraannya (positif), selain itu beberapa ahli sosial juga menggunakan teknologi maklumat untuk melakukan kejahatan atau sifatnya merosak (negatif) (Dietrich & Jones 2007). Komuniti seperti kumpulan penyelidik dan akademik, melakukan adaptasi dengan melibatkan perlindungan, kerjasama, dan persaingan semasa melakukan komunikasi atau interaksi dengan komuniti lain untuk merentasi perubahan dunia dan mempertahankan kewujudannya sebagai komuniti akademik (Wellman et al. 1996). Kumpulan orang dengan berbagai aktiviti demikian ialah rangkaian sosial pada rangkaian komputer global atau Internet. Rangkaian sosial, rangkaian komputer dan Internet ialah teknologi dan sebahagian daripada tamadun saat ini (Bruhn 2005).

Internet sebagai infrastruktur sosial moden bersama Web telah merekodkan banyak peristiwa dan aktiviti daripada interaksi manusia, yang memungkinkan untuk menghadirkan pelbagai konsep dan model hubungan antara manusia sebagai pelakon sosial. Malangnya, sebagai media sosial, Web adalah tidak terstruktur dan kurang maklumat semantik. Tambahan pula, Internet sebagai repositori mengandungi bilangan data atau laman Web yang terus bertambah dengan dinamik perubahan kandungan, bentuk dan pemaknaan. Oleh itu, diperlukan kaedah inovatif untuk mendapatkan pengetahuan berkenaan dengan struktur sosial daripada Web melalui proses pengekstrakan.

Perkara yang paling penting dalam pengekstrakan rangkaian sosial adalah untuk mengenalpasti pelakon sosial dan hubungan yang mungkin wujud antara pasangan pelakon itu (Matsuo et al. 2006b). Dalam bidang kecerdasan buatan, terdapat dua aliran penyelidikan untuk mengekstrak rangkaian sosial dari sumber maklumat heterogen (Tang et al. 2007). Aliran penyelidikan diselia melibatkan analisis korpus untuk mengenalpasti entiti sebagai pelakon sosial dan hubungan mereka, dan juga mendapatkan label hubungan (Cullota et al. 2004). Namun demikian, menamakan label hanya terhad kepada label yang telah ditetapkan dalam set latihan (McCallum et al. 2005a). Dalam hal ini, setiap pelakon sosial akan ditentukan dengan label yang diekstrak dari sumber maklumat (korpus). Label yang sesuai akan ditugaskan kepada hubungan dengan menggunakan teknik capaian maklumat, seperti model kebarangkalian generatif (McCallum et al. 2004). Parameter daripada model selalu digunakan sebagai modaliti untuk mendapatkan pengetahuan mengenai strukstur sosial dari korpus. Aliran penyelidikan tidak diselia pula hanya bertumpu kepada pengekstrakan rangkaian dengan melibatkan hasil enjin carian (Kautz et al. 1997b), tetapi mengabaikan isu penamaan hubungan atau pemberian petunjuk hubungan secara semula jadi. Rangkaian sosial yang dihasilkan oleh aliran tidak diselia secara amnya beroperasi dalam maklumat terhad (Mori et al. 2007), tetapi kaedah yang digunakan mudah untuk diadaptasi (Jin et al. 2008a). Dalam hal ini, kaedah dangkal ialah salah satu kaedah dalam aliran tidak diselia yang bersandarkan kepada analisis *co-occurrence* (Mori et al. 2006).

Banyak penyelidikan telah menunjukkan kemungkinan pengekstrakan rangkaian sosial dari Web (Mika 2005c; Matsuo et al. 2007b; Jin et al. 2007b). Berdasarkan teori graf, perkara ini melibatkan pengiraan dengan n^2 lelaran untuk bilangan n pelakon sosial. Dalam hal ini, pengekstrakan menggunakan perwakilan laman Web di dalam kueri untuk mendapatkan pelakon sosial dan hubungan antara pelakon. Dokumen Web dihasilkan oleh berjuta manusia. Pada sebarang masa, banyak dokumen baru diterbitkan dan mengandungi maklumat terbaru dari sebarang tempat di seluruh dunia. Sebahagian daripadanya mengandungi nama pelakon sosial dan hubungan antara mereka (Kudělka et al. 2009). Sebagai contoh, banyak kertas kerja diterbitkan dengan melibatkan Web sama ada dalam jurnal atau dalam prosiding. Setiap kertas kerja mengandungi sekurang-kurangnya satu nama pelakon sebagai pengarang, dan tidak sedikit kertas kerja ditulis oleh lebih daripada dua pengarang. Dalam hal ini, sebagai nod dalam rangkaian sosial, pelakon sosial baru terus muncul manakala pelakon lama tetap wujud. Terdapat pelakon sosial berbeza mempunyai nama yang sama, atau pelakon yang sama mempunyai nama yang berbeza. Dengan demikian, rangkaian sosial tertakluk kepada semua jenis perubahan dan perkembangan dinamik sumber maklumat (Snášel et al. 2009). Oleh itu, rangkaian sosial menjadi sangat kompleks karena bilangan nod dan hubungan yang terus berkembang.

1.2 PERNYATAAN MASALAH

Pengekstrakan rangkaian sosial daripada Web ialah proses yang sangat kompleks dan memakan masa. Banyak penggunaan rangkaian sosial bersumberkan data elektronik yang telah dijelaskan (Staab et al. 2005), tidak sedikit antaranya yang bersumberkan kepada Web (Jin et al. 2008b) dengan melibatkan pendekatan dan kaedah pengekstrakan yang berbeza. Walau bagaimanapun secara amnya pengekstrakan rangkaian sosial daripda Web bertumpu kepada satu komuniti sosial sahaja (Mika 2005c; Hamasaki et al. 2006b; Matsuo et al. 2007b; Jin et al. 2007b). Banyak kaedah pengekstrakan rangkaian sosial berasaskan pendekatan tidak diselia telah dibangunkan, yang pada amnya diperkenalkan sebagai kaedah dangkal, tetapi melibatkan multidimensi maklumat (*occurrence*, *co-occurrence*, *snippet*, frekuensi perkataan, kata kunci atau label) yang terhad (Mori et al. 2006).

Struktur formal rangkaian sosial melibatkan nod dan pinggir. Nod mewakili pelakon atau aktor, manakala pinggir mewakili hubungan antara pelakon. Untuk menjanakan pinggir antara setiap nod dalam rangkaian sosial, kaedah dangkal menggunakan produk Cartesian untuk menafsirkan pelakon sosial dan hubungan antara pelakon. Dalam hal ini, kaedah dangkal mengiktiraf sambungan tersurat dan sambungan tersirat yang mungkin wujud antara pelakon (Kirchhoff et al. 2008). Pada satu aspek, kaedah dangkal hanya tertumpu untuk menghasilkan kekuatan hubungan (*strength relation*) antara pelakon, iaitu hubungan yang ditaksir dalam [0,1], yang semula jadi bersifat kabur (*ambiguity*) dan timpang (*bias*). Untuk mempertingkatkan kaedah ini beberapa penyelidik menambahkan fitur ke dalam strategi yang digunakan, seperti menyediakan kata kunci tertentu kepada kueri (Mori et al. 2004; Jin et al. 2007a) atau mempertimbangkan bilangan k teratas daripada laman Web (Matsuo et al. 2006b). Dalam kes ini k merupakan bilangan *snippet* yang dikembalikan oleh enjin carian. Bagaimanapun, kaedah ini pada amnya tidak menamakan hubungan. Walau demikian, hanya segelintir penyelidik (Mori et al. 2007) memberikan kekuatan hubungan yang melibatkan sekumpulan kata kunci melalui penerokaan dimensi maklumat daripada nilai (contoh *hit count* dan frekuensi perkataan) *co-occurrence* dalam *snippet*. Pada aspek lain, kekuatan hubungan ialah ringkasan hubungan dari pelbagai hubungan yang mungkin wujud antara pelakon. Berasaskan dimensi maklumat yang ditakrifkan lebih dahulu, agregat (pengumpulan untuk menakrifkan semula) kekuatan hubungan perlu untuk memberikan makna yang berbeza mengikut struktur kepada rangkaian sosial yang diekstrak (Matuso et al. 2007b). Beberapa penyelidik melaksanakan agregat kekuatan hubungan berasaskan kepada label yang telah ditetapkan pada set latihan (McKelvey & Page 1986). Walau bagaimanapun, label hubungan dapat juga dijanakan dari *snippet*. Dalam hal ini, label diiktiraf sebagai konteks semasa kerana berasal dari dinamik perubahan daripada Web. Berasaskan kedua-dua aspek demikian, *snippet* mengandungi multidimensi maklumat, yang memungkinkan pelakon dan hubungan antara mereka dijelaskan. Namun demikian, rangkaian sosial demikian memerlukan pembangunan kaedah pengekstrakan yang sesuai. Dengan demikian, jika setiap *snippet* merupakan wakil daripada setiap laman Web, maka *snippet* akan dapat memberi makna kepada kekuatan hubungan dalam multidimensi maklumat secara semantik.

Secara praktikal tidak terdapat maklumat mengenai nod dan pinggir bagi menjelaskan dinamisme daripada rangkaian sosial di dalam Web (Yu & Wu 2010). Manakala nod, pinggir dan sumber maklumat bersama-sama menjadi bahagian terpenting dalam pengekstrakan rangkaian sosial. Secara asas, Web sebagai media sosial kurang maklumat semantik dan tidak terstruktur dan hanya boleh difahami oleh manusia. Namun demikian sebarang pengekstrakan maklumat dari Web, melibatkan repositori yang sangat besar yang hanya dapat diproses oleh mesin komputer. Ramai penyelidik (Kautz et al. 1997a; Mika 2004; Matsuo et al. 2006a; Tang et al. 2008a) oleh itu mempertimbangkan kelebihan daripada kaedah dangkal untuk mengesktrak rangkaian sosial dari Web berdasarkan kepada aspek mudah dan harmoni, yang mana enjin carian menjadi alat untuk mengakses maklumat. Namun demikian, secara kompleksiti untuk bilangan n pelakon sosial sebagai nod, kaedah dangkal yang bertumpu kepada hasil kueri akan mengekstrak $O(n^2)$ hubungan dari Web secara pengiraan mengikut prinsip daripada graf lengkap (Matsuo et al. 2007b). Dalam hal ini, kueri ialah suatu paradigma yang digunakan untuk memahami sumber maklumat. Walau bagaimanapun, mengikut skalabiliti setiap API (*Application Programming Interface*) enjin carian telah menghadkan kepada hanya m bilangan kueri per hari untuk setiap IP (*Internet Protocol*), sebagai contoh *Google API* mempunyai $m \leq 1000$ (Matsuo et al. 2007a). Had ini memberikan pemahaman kepada strategi pelaksanaan kaedah dangkal, yang juga akan mengambilkira multidimensi maklumat yang sedia ada dalam *snippet*.

Menurut prinsip dan struktur, enjin carian mengembalikan maklumat yang secara ontologi dapat dikategorikan dengan pemberat ringan dan pemberat berat (Pretschner & Gauch 1999). Secara leksikal, maklumat kiraan statistik (*hit count*) ialah pemberat ringan yang dikembalikan oleh enjin carian berdasarkan indeks daripada kata carian yang dirumuskan ke dalam kueri yang dikemukakan. Sehubungan dengan itu, *co-occurrence* merupakan maklumat semantik yang memberi makna kepada kekuatan hubungan (Hofmann 2001). Bagaimanapun, setiap enjin carian juga mengembalikan maklumat yang mewakili laman Web dalam bentuk ringkasan, dikenali sebagai *snippet* (Croft et al. 2010). *Snippet* mengandungi alamat URL, tajuk dan abstrak daripada laman Web. Pada ontologi pemberat berat, maklumat yang dikandung *snippet* cukup menyediakan kecerdasan semantik

(Bollegala et al. 2007) bagi mengenalpasti pelakon sosial dan menamai hubungan antara pelakon (Matsuo et al. 2007a). Oleh itu, *snippet* dicadangkan sebagai sumber pengetahuan untuk memberikan makna kepada rangkaian sosial yang diekstrak.

1.3 MATLAMAT DAN OBJEKTIF

Matlamat utama daripada penyelidikan ini adalah untuk mengkaji penggunaan kaedah dangkal dalam pengekstrakan rangkaian sosial daripada Web dengan melibatkan strategi capaian maklumat. Untuk mencapai matlamat yang dinyatakan, objektif berikut digariskan:

i. Untuk mencadangkan dan membangunkan kaedah automatik untuk pengekstrakan rangkaian sosial
   a. yang selaras dengan strategi daripada capaian maklumat, dan
   b. melibatkan multidimensi dari Web secara semantik.
ii. Untuk mengarang satu set garis panduan pelaksanaan dan mensahkan kaedah dangkal dalam satu metodologi yang mempertimbangkan kompleksiti dan skalabiliti, dengan
   a. menganalisis rangkaian sosial untuk mengambil keputusan, dan
   b. menilai dan kesahan kaedah dan metodologi.

1.4 KEPENTINGAN PENYELIDIKAN

Penyelidikan ini meneroka kemungkinan untuk membangun satu kaedah pengekstrakan rangkaian sosial yang tidak sahaja mengenalpasti pelakon sosial dan hubungan antara mereka, tetapi juga mempertimbangkan kompleksiti dan skalabiliti. Setiap rangkaian sosial menggambarkan struktur sosial, yang menunjukkan tingkah laku daripada ahli sosial (Wasserman & Faust 1994), dan arah perubahan yang terjadi pada satu masyarakat, yang sangat bermanfaat untuk melihat arah tuju pembangunan dan tamadun (Kudělka et al. 2010). Kaedah pengekstrakan rangkaian sosial berkaitan erat dengan sekumpulan dokumen atau laman Web sebagai sumber maklumat dan pengetahuan. Kumpulan dokumen yang besar atau repositori yang dinamik seperti Web, dari aspek pengurusan pengetahuan adalah sukar untuk

dikawal (Kobayashi & Takeda 2000), manakala keperluan pengetahuan dalam kehidupan seharian manusia atau pembuatan keputusan adalah berkaitan dengan seberapa tepat maklumat yang diberikan. Terdapat jurang pemisah antara keperluan pengurusan sumber maklumat dan pertumbuhan dinamik sumber maklumat. Pengekstrakan rangkaian sosial menghasilkan teknologi yang bertugas untuk manjawab keperluan maklumat dimaksudkan dengan mana dokumen atau laman Web yang sesuai terwakili (Alguliev et al. 2012). Dalam hal ini, rangkaian sosial dihasilkan (yang diekstrak) secara logik akan menjadi bahagian penting dari capaian maklumat.

1.5 SKOP PENYELIDIKAN

Pada perlombongan data, kaedah pengekstrakan maklumat dapat dibahagi ke dalam tiga aliran yang berbeza: aturan heuristik, kaedah diselia berasaskan pengelasan, dan kaedah tidak diselia berasaskan penggugusan (Tang et al. 2007). Pendekatan mengikut kaedah diselia berasaskan pengelasan dimulai dari penakrifan pembolehubah pendam yang mengandungi label ditetapkan yang seterusnya dijurutera untuk menghasilkan suatu model (Blei et al. 2003). Untuk mengekstrak rangkaian sosial, setiap model didasarkan kepada bentuk triplet, iaitu hubungan tiga serangkai antara pelakon, dokumen / korpus, dan perkataan, tetapi pemaknaan struktur sosial dapat ditingkatkan dengan mengubahsuai model (Wang et al. 2005). Beberapa penyelidik telah menguruskan korpus berbeza daripada beberapa sumber maklumat seperti emel, kertas kerja, atau dokumen elektronik lainnya, dan parameter berasaskan pembolehubah pendam digunakan sebagai modaliti untuk mendapatkan pengetahuan daripada korpus. Sebaliknya, pendekatan mengikut kaedah tidak diselia berasaskan penggugusan mencuba untuk mencari struktur tersembunyi di dalam data yang tidak berlabel seperti dokumen atau laman Web (Kautz et al. 1997a; Mika 2005c). Struktur dicadangkan dan dirumuskan ke dalam corak yang fleksibel dan mudah menangkap maklumat yang tersedia dalam sumber maklumat (Matsuo et al. 2006a). Dalam hal ini, penyelidik daripada pengekstrakan rangkaian sosial meneroka *occurrence* dan *co-occurrence* dengan menggunakan kueri dan mengukur kekuatan hubungan dengan melibatkan pengukuran kesamaan (Matsuo et al. 2006b). Selain kedua pendekatan itu, penyelidik juga menggunakan aturan heuristik bagi

mempercepat proses pengekstrakan penamanaan rangkaian sosial, iaitu petua yang menyediakan jalan pintas untuk menyelesaikan masalah yang sukar diurus oleh kedua-dua pendekatan terdahulu. Selaras dengan pendekatan tidak diselia, kajian ini melibatkan kaedah dangkal untuk meneroka pengekstrakan automatik rangkaian sosial daripada Web, tetapi mempertimbangkan prinsip pendekatan lainnya untuk memperkaya secara semantik rangkaian sosial.

Dalam konsep tradisional, penjanaan rangkaian sosial dimulai dari pengumpulan maklumat rangkaian sosial daripada kenyataan yang sedia ada dalam satu komuniti sosial atau masyarakat, dan menganalisis rangkaian sosial untuk mendapatkan pengetahuan tentang struktur sosial (Wasserman & Faust 1994). Penjanaan rangkaian sosial saat ini pula dapat dijalankan dengan mengekstraks sumber maklumat seperti Web (Kautz et al. 1997a, 1997b) atau dokumen elektronik lainnya (Mika 2007b). Namun demikian, pengekstrakan rangkaian sosial daripada Web, melibatkan beberapa tahap, iaitu selain pengekstrakan, penyelidik juga melakukan penggambaran, agregat, analisis, dan penilaian (Mika 2007b; Matsuo et al. 2007b). Penggambaran akan merinci struktur sosial secara jelas (Shen et al. 2006), agregat pula memberikan makna tertentu kepada rangkaian sosial (Alguliev et al. 2012), analisis ialah satu usaha untuk mengungkapkan tingkah laku pelakon dalam struktur sosial (Coscia et al. 2009), dan penilaian bertujuan untuk menguji keberkesanan kaedah pengekstrakan selain mengurus pengetahuan (Zhou et al. 2008). Set garis panduan daripada pengekstrakan rangkaian sosial dalam satu metodologi pengekstrakan rangkaian sosial mempertimbangkan ke semua tahapan ini. Pada tahap pengekstrakan, diperkenalkan satu kaedah pengekstrakan rangkaian sosial berasaskan benih yang telah ditakrif untuk menjanakan pelakon sosial lain, kemudian diperkenalkan satu kaedah pengekstrakan kata kunci untuk penyahkaburan nama dan untuk mereduksi kepincangan (Bekkerman & McCallum 2005).

Bagi tujuan kajian ini, skop kajian dihadkan untuk menyelesaikan pengenalpastian pelakon sosial dan pengenalpastian hubungan antara pelakon. 76 pelakon sebagai benih ditakrif daripada laman Web Fakulti Teknologi dan Sains Maklumat (FTSM) Universiti Kebangsaan Malaysia (UKM), yang digunakan untuk menjana pelakon lain dan hubungan antara pelakon daripada sumber maklumat

pangkalan data dalam talian DBLP (*Digital Bibliography & Library Project*). Terdapat beberapa pangkalan data dalam talian yang oleh penyelidik diambilkira sebagai sumber maklumat untuk menjanakan rangkaian sosial asas (Cox et al. 2003; Aleman-Meza et al. 2006; Hamasaki et al. 2006a). Pada amnya pangkalan data dimaksudkan ialah pangkalan data kertas kerja akademik dan penyelidikan, dan oleh itu rangkaian yang diekstrak ialah rangkaian sosial akademik (Tang et al. 2008a). Beberapa sistem pengekstrakan rangkaian sosial dihadkan kepada satu komuniti. Flink (Mika 2005c) ialah sistem untuk membina rangkaian sosial daripada komuniti Web Semantik dengan menggunakan profail semantik FOAF (*frient-of-a-frient*), POLYPHONET (Matsuo et al. 2007b) dibangun untuk mengenalpasti hubungan pada *Japan AI Conference*, dan ArnetMiner (Tang et al. 2008b) ialah sistem perlombongan rangkaian sosial akademik dan penyelidik. Senarai pelakon dalam kajian ini melibatkan pensyarah, pelajar dan penyelidik pada FTSM UKM yang digunakan sebagai pelaksanaan dan pengujian daripada kaedah pengekstrakan rangkaian sosial akademik. Walaupun begitu, kaedah pengekstrakan akan dapat dilaksanakan pada domain lain.

1.6 METODOLOGI PENYELIDIKAN

Metodologi penyelidikan untuk kajian ini terdiri dari tiga fasa. Fasa pertama ialah pemahaman yang berkaitan dengan kawasan penyelidikan yang sama dan perumusan kaedah yang dicadangkan untuk menyelesaikan masalah. Fasa kedua ialah pelaksanaan kaedah yang diperolehi daripada fasa pertama dengan membangun prosedur untuk mengesktrak rangkaian sosial daripada Web dan menjalankan eksperimen. Fasa ketiga adalah untuk menjalankan analisis dan kaedah penilaian untuk mensahkan hasil yang diperoleh dengan membanding hasil antara kaedah yang berbeza dan dengan kayu ukur yang telah ditakrif.

Fasa pertama dimulai dengan memahami secara menyeluruh tentang falsafah mengenai rangkaian sosial dan kajian yang berkaitan dengan penjanaan rangkaian sosial. Proses ini dilakukan dengan kajian kesusteraan dari beragam kertas kerja: buku, jurnal, prosiding dan laporan teknik yang berkaitan dengan kawasan berikut:

i. Penyelidikan mengenai fungsi dasar daripada pengekstrakan rangkaian sosial dan pemodelan pengekstrakan rangkaian sosial dari sumber maklumat. Penyelidikan yang menyediakan pemahaman tentang pengekstrakan rangkaian sosial dan bagaimana pengekstrakan disokong oleh sumber maklumat, dan bagaimana sumber maklumat diiktiraf dan ditafsirkan.

ii. Penyelidikan mengenai pelbagai kaedah pengekstrakan rangkaian sosial. Mengkaji pendekatan yang digunakan oleh penyelidik, model yang digunakan untuk menguraikan dan menafsirkan maklumat daripada sumbernya, jenis data sebagai masukan dan had yang diberikan.

iii. Penyelidikan mengenai capaian maklumat dan penggunaan capaian maklumat untuk menilai dan menguji kaedah yang digunakan.

Setelah menjalankan kajian semula kepada beberapa penyelidikan, kaedah dan hasil daripada pengekstrakan rangkaian sosial diamati untuk mengenali jurang yang wujud antara beberapa penyelidikan. Berasaskan kajian semula itu, dirumuskan pertanyaan penyelidikan yang perlu dijawab oleh penyelidikan ini.

Dari kajian semula kertas kerja yang sedia ada, dikenali bahawa ramai penyelidik telah secara automatis mengekstrak rangkaian sosial daripada Web (Mori et al. 2007) dengan mewujudkan sistem pengekstrakan, salah satunya adalah ArnetMiner (Tang et al. 2008b). Proses pengekstrakan dilakukan dengan menentukan rangkaian sosial asas melibatkan kaedah diselia, dan secara automatis boleh disahkan oleh pengguna atau pengarang (Tang et al. 2007). Untuk meningkatkan pendekatan ini, penyelidikan melibatkan sumber maklumat pangkalan data dalam talian untuk menjanakan pelakon sosial berdasarkan benih dan rangkaian sosial asas, dan disempurnakan secara automatis dengan menggunakan kaedah dangkal yang ditingkatkan. Dalam hal ini, proses pengayaan secara semantik daripada rangkaian sosial dijalankan kepada kekuatan hubungan melalui proses kecerdasan buatan berasaskan pengetahuan yang sedia ada dalam *snippet*. Untuk menyokong proses semantik ini, perlu perumusan berikut:

i. Penakrifan konsep *co-occurrence* langsung dan tidak langsung yang sedia ada dalam sumber maklumat. Konsep ini menyatakan wujudnya hubungan yang tersirat dan tersurat, yang memungkinkan kekuatan hubungan dapat diagregat.

ii. Penakrifan konsep pengayaan yang memodelkan dan mengelaskan atribut pelakon dan hubungan antara pelakon. Fungsi daripada konsep ini adalah untuk menafsirkan kandungan semantik daripada maklumat yang telah diekstrak mengikut lakaran beg perkataan sama ada berasaskan *occurrence* ataupun *co-occurrence*.

iii. Set prosedur garis pandu. Garis pandu digunakan untuk mengenali pelakon sosial dan hubungan antara pelakon sama ada melibatkan konsep *co-occurrence* untuk memaknakan kekuatan hubungan dalam rangkaian sosial ataupun untuk menapis konteks semasa mengikut lakaran beg perkataan.

Pada fasa kedua, eksperimen dijalankan untuk menunjukkan kebolehlaksanaan kaedah dan pendekatan yang dibuat. Eksperimen dilaksanakan melibatkan enjin carian Yahoo! dan beberapa pengukuran kesamaan (Deza & Deza 2006). Pengekstrakan rangkaian sosial ini akan menghasilkan rangkaian sosial dihasilkan berbeza dan akan diperoleh rangkaian sosial integrasi. Pengekstrakan rangkaian sosial integrasi dibahagikan ke dalam

i. Pengesanan hubungan – Objektif daripada peringkat ini adalah untuk mendapatkan *co-occurrence* yang mewakili hubungan secara binari daripada laman Web.

ii. Penyahkaburan nama – Objektif daripada peringkat ini adalah untuk mengurangkan kepincangan dan kekaburan nama dengan melibatkan kata kunci.

iii. Penjanaan kekuatan hubungan – Objektif daripada peringkat ini adalah untuk mengukur tingkat hubungan antara pelakon.

iv. Agregat kekuatan hubungan – Objektif daripada peringkat ini adalah untuk mewujudkan struktur sosial berbeza.

Fasa ketiga bertumpu kepada analisis dan penilaian. Analisis rangkaian sosial adalah pengekstrakan pengetahuan dari rangkaian sosial yang bertujuan untuk menyiasati struktur nod dalam rangkaian sosial dan digunakan untuk melihat tingkah laku sosial daripada pelakon (Coscia et al. 2009), manakala penilaian bertujuan untuk mensahihkan kaedah pengekstrakan dan menunjukkan kemungkinan pengurusan dokumen boleh dilakukan melalui rangkaian sosial yang diekstrak. Untuk menyokong penyiasatan dan pensahihan ini dijalankan aktiviti berikut:

i. Membandingkan pengiraan daripada keperluan mengemukakan kueri kepada enjin carian daripada setiap kaedah dan pendekatan.
ii. Pengukuran pemusatan berasaskan nod, pemusatan ketertutupan, dan pemusatan perantaraan untuk mendapatkan ciri-ciri nod dalam rangkaian sosial.
iii. Menakrifkan darjah nod untuk melihat tingkah laku pelakon sosial dari aspek kepemimpinan, ikatan, dan kepelbagaian.
iv. Menjalankan kaji selidik dengan mengagihkan borang soal selidik kepada sekumpulan pelajar siswazah pada FTSM UKM, untuk mendapatkan rangkaian sosial piawai yang digunakan sebagai kayu ukur.
v. Membandingkan secara kesamaan rangkaian sosial yang diekstrak dengan kaedah berbeza.
vi. Membandingkan dapatan semula dan kejituan setiap kaedah pengekstrakan rangkaian sosial.
vii. Melihat capaian maklumat dengan melibatkan pemangkatan berasaskan kekuatan hubungan.

1.7 ORGANISASI TESIS

Tesis ini terdiri daripada 10 bab. Bab I adalah pendahuluan tesis yang menjelaskan latar belakang pengekstrakan rangkaian sosial dan masalah yang berkaitan dengan pengekstrakan rangkaian sosial. Bab ini juga mendedahkan matlamat dan objektif daripada penyelidikan bersama dengan kepentingan dan skop penyelidikan. Ia juga menguraikan metodologi penyelidikan.

Bab II adalah kajian kesusasteraan yang meliputi dua subjek utama, kajian falsafah mengenai rangkaian sosial dan sumber maklumat daripada rangkaian sosial. Bab ini membincangkan fenomena dan paradigma rangkaian sosial, dan infrastruktur bagi kewujudan rangkaian sosial. Kawasan yang dibincangkan secara terperinci adalah tentang sejarah, takrif, dan data daripada rangkaian sosial. Kemudian, dibincangkan kelebihan dan kekurangan Web sebagai sumber maklumat, yang juga melibatkan enjin carian sebagai pintu gerbang akses kepada maklumat dalam Web.

Bab III adalah kajian kesusasteraan yang berkaitan dengan kaedah pengekstrakan dan perkara yang menyokong pengekstrakan rangkaian sosial. Bab ini membicangkan aliran penyelidikan yang melibatkan kaedah berbeza dan tujuan yang hendak dicapai. Pada bahagian akhir daripada bab ini dibincangkan konsep dan teori analisis rangkaian sosial, dan pentakrifan konsep dasar daripada capaian maklumat, dalam hal ini dengan menggunakan dapatan semula dan kejituan.

Bab IV adalah kaedah pengekstrakan rangkaian sosial yang bertujuan untuk membincangkan strategi pelaksanaan beberapa kaedah pengekstrakan rangkaian sosial. Beberapa aspek yang dipertimbangkan daripada pelaksanaan adalah seperti kompleksiti, skalabiliti, dan pemaknaan yang mungkin diberikan oleh kekuatan hubungan yang dihasilkan. Pendekatan terintegrasi didasarkan kepada pengekstrakan rangkaian sosial asas, kaedah pengesanan dan kaedah baru, tetapi juga melibatkan beberapa kaedah lain daripada kaedah dangkal, yang menjadi teras dalam satu pendekatan. Pendekatan ini juga menjelaskan kelebihan dan keburukan setiap kaedah yang digunakan untuk memperkaya rangkaian sosial yang boleh dipercayai. Oleh itu, pendekatan ini menjadi garis pandu pengekstrakan rangkaian sosial daripada Web.

Bab V menjelaskan kaedah dangkal dan petua sekutuan melalui eksperimen. Eksperimen ini menghasilkan pengekstrakan rangkaian sosial asas daripada laman Web DBLP. Bab ini juga menjelaskan keterkaitan antara unit analisis / properti daripada rangkaian sosial, seperti nod, pinggir, sumber maklumat dan benih, untuk meramalkan ketumpatan (*density*) rangkaian sosial yang dihasilkan.

Bab VI menjelaskan pengekstrakan kata kunci untuk mengekstrak rangkaian sosial, iaitu penyahkaburan nama melibatkan sumber maklumat dalam kontek semasa. Kata kunci setiap pelakon sosial dijanakan melalui kumpulan *snippet* yang dikembalikan oleh enjin carian. Bab ini juga menjelaskan pengujian dan penilaian kaedah yang diajukan untuk pengekstrakan kata kunci.

Bab VII menjelaskan pengekstrakan rangkaian sosial dengan melibatkan beberapa kaedah. Eksperimen yang dijalankan menghasilkan beberapa rangkaian sosial yang berbeza untuk senarai nama pelakon yang sama. Pengesanan hubungan memberikan dasar pengekstrakan rangkaian sosial terintegrasi, yang kemudian diagregat dan diberi label yang sesuai. Dalam eksperimen ini dihasilkan rangkaian sosial berasaskan pengarang-bersama, rangkaian sosial persidangan-saintifik, dan rangkaian sosial berasaskan kumpulan penyelidikan dan rangkaian sosial berasaskan peranan-akademik.

Bab VIII menjelaskan penyiasatan tingkah laku pelakon melalui struktur sosial yang diekstrak daripada Web. Beberapa kaedah analisis rangkaian sosial dibincangkan, yang utama adalah darjah nod, pemusatan ketertutupan, dan pemusatan keterantaraan. Pada awal bab ini dijelaskan kompleksiti dan skalabiliti setiap kaedah dan dibandingkan dengan pendekatan terintegrasi. Pada akhir bab dijelaskan analisis rangkaian sosial berasaskan darjah nod untuk kepentingan pembuatan keputusan.

Bab IX menguraikan penilaian dan pengujian kaedah yang digunakan untuk mengekstrak rangkaian sosial daripada Web. Pada bab ini juga dibincangkan kemungkinan menggunakan rangkaian sosial dalam pencapaian maklumat, dalam hal ini dilakukan eksperimen dengan memangkatkan dokumen berasaskan kekuatan hubungan.

Bab X atau bab terakhir membentangkan kesimpulan daripada tesis atas sumbangan penyelidikan bagi pengekstrakan rangkaian sosial. Bab ini juga menghuraikan perluasan kajian untuk masa hadapan yang dapat diterokai dan yang

mungkin akan mempertingkatkan kaedah pengekstrakan rangkaian sosial sama ada dari aspek kebolehpercayaan atau capaian maklumat.

BAB II

RANGKAIAN SOSIAL

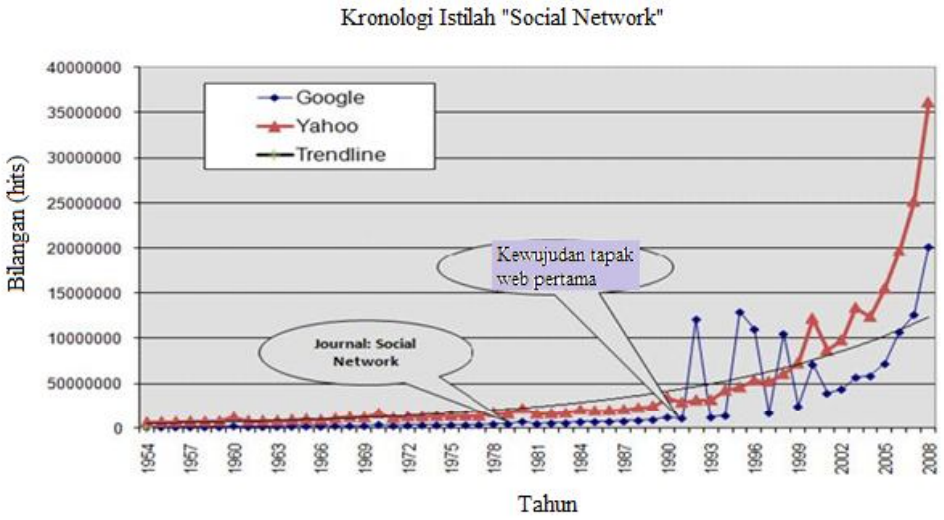
2.1 PENGENALAN

Istilah rangkaian mempunyai makna yang tidak sama dalam bidang yang berbeza. Dalam bidang sains sosial, rangkaian lazimnya mengandungi pelakon sosial yang membolehkan kewujudan sesuatu kehubungan. Dalam bidang sains komputer, terdapat mesin yang dihubungkan oleh media dengan mana pelakon sosial melakukan komunikasi antara mereka. Konsep rangkaian merupakan usaha untuk memahami fenomena yang ada dalam dunia dan menentukan paradigmanya. Rangkaian sosial menjadi paradigma dalam fenomenanya tersendiri bersama kewujudannya dengan rangkaian komputer, yang mana Web adalah sesuatu yang fenomenal dalam rangkaian komputer dan rangkaian sosial menjadi isu penting yang berasingan. Bab ini membincangkan rangkaian komputer dan Internet sebagai infrastruktur kepada kewujudan rangkaian sosial melalui persekitaran Web.

2.2 DEFINISI RANGKAIAN SOSIAL

Kajian rangkaian sosial telah lama bersinar, sejak John Barnes (Barnes 1954, 1969) memperkenal istilah ini pada tahun 1954, sebagai penegasan terhadap apa yang dipanggil oleh Moreno (Moreno 1946) sebagai *sociogram* (Churchill & Halverson 2005). Kajian dalam bidang ini telah secara meluas diterokai, sebahagian daripadanya adalah kajian berkenaan dengan struktur sosial dan sebahagian yang lain sebagai aplikasi yang menyokong kehidupan seharian tamadun moden seperti menganalisis kestabilan dan pertumbuhan industri (Bengtson & Knock 1999; Battiston 2004; Jin et al. 2008a), analisis organisasi sosial (Papakyriazis &

Boundourides 2001; Kilduff & Tsai 2003; Yeung 2005), strategi koperatif di kalangan organisasi (Chung 1996; Lazega & Pattison 1999), penentuan gaji atau ganjaran bagi prestasi kerja (Soo-Hoon & Keng-Howe 2000; Fontaine 2008), dan aplikasi lain yang berkaitan dengan rangkaian sosial (Staab et al. 2005; Jin et al. 2008b). Pada masa ini, penyelidikan dalam bidang ini menjadi lebih menarik dengan kemunculan sumber elektronik, seperti pangkalan data dan dokumen elektronik. Rajah 2.1 jelas menunjukkan peningkatan bilangan dokumen berindeks berkaitan dengan rangkaian sosial dalam enjin carian Google dan Yahoo!. Secara langsung menunjukkan populasi istilah '*social network*' dalam dunia Web.



Rajah 2.1 Peningkatan istilah "social network" dalam enjin carian Yahoo! dan Google

Kajian rangkaian sosial yang melibatkan dokumen elektronik sama ada dalam korpus atau dalam talian telah membuka pintu kepada topik baru kajian rangkaian sosial yang antaranya adalah seperti penganalisisan arkib mel (Kautz et al. 1996), penapisan emel (Boykin & Roychowdhury 2005), pengesanan spam dalam emel (Lam & Yeung 2007), analisis kandungan weblog untuk mencari kemungkinan pelan keganasan dan jenayah (Yang & Ng 2007), menganalisis sokongan sosial melalui laman web (Coromina & Coenders 2006; Vehovar et al. 2008) seperti bagi kepentingan pilihan raya atau hubungan antara ahli parlimen dengan isu sosial (Wang et al. 2005), dan menentukan persahabatan melalui foto peribadi dalam

layanan rangkaian sosial (Kim et al. 2012). Mungkin, penerokaan dokumen elektronik, khususnya dengan melibatkan semantik, telah membawa keterujaan baru pada penyelidikan rangkaian sosial, dan ini dirumuskan sebagai anjakan dalam pelbagai bidang berkaitan, terutamanya dalam bidang sains sosial dan sains maklumat.

2.2.1 Apakah Rangkaian Sosial?

Rangkaian sosial merupakan satu pendekatan bagi mewakili hubungan antara individu, kumpulan atau organisasi. Rangkaian sosial secara tradisional tertumpu kepada persepsi antara peribadi dan ditakrif sebagai koleksi kehubungan sosial atau antara peribadi individu dalam kumpulan sosial (Pattison 1993). Setiap manusia ialah makhluk sosial yang memerlukan sokongan dan persahabatan orang lain sepanjang hidup mereka. Kerjasama sosial telah memainkan peranan penting dalam kelangsungan hidup manusia sebagai spesies. Manusia primitif mesti belajar bekerjasama dalam usaha untuk melindungi dirinya daripada spesies yang lebih kuat dan untuk berjaya dalam memburu haiwan besar, dan oleh itu kelangsungan hidup mereka bergantung kepadanya. Manusia moden telah melangkah jauh dari keadaan sosial berburu dan berkumpul, ke suatu keadaan yang memerlukan interaksi sosial (Lorenz 1965), iaitu suatu keadaan dengan sentiasa mengambil dan memberi sokongan sosial antara mereka sahaja. Sokongan sosial adalah pengalaman peribadi yang berasaskan persepsi, dan merupakan satu set keadaan objektif, atau satu set proses interaksi (Shinn et al. 1984). Apa yang dinilai dan apa yang tidak dihormati? Pada konteks ini, komunikasi diguna untuk meningkatkan perubahan interaksi untuk terus hidup, iaitu komuniti sosial tumbuh dan di sini peranan telah diambil oleh pelakon sosial sendiri.

Terdapat banyak takrifan rangkaian sosial berdasarkan bidang berbeza. Bidang sains sosial menakrifkan rangkaian sosial sebagai satu set pelakon sosial yang boleh memiliki kehubungan antara satu dengan yang lain. Rangkaian sosial boleh mempunyai sedikit atau ramai pelakon, dan satu atau lebih jenis hubungan antara pelakon. Rangkaian sosial yang mewakili jenis hubungan tunggal di kalangan

pelakon dinamakan simpleks (*simplex*), manakala rangkaian sosial diwakili lebih daripada satu jenis hubungan diberi nama multipleks (*multiplex*).

Jadual 2.1 *TwoBugis People*

Nama Pelakon	Ciri-ciri
Dato' Sri Mohammad Najib Abdul Razak http://www.facebook.com/najibrazak (Facebook). http://twitter.com/NajibRazak (Twitter)	6th Prime Minister of Malaysia (2009-…), 9th Deputy Prime Miniister of Malaysia (2004-2009), Deputy Minister of Energy, Telecommunications and Post (1978), Minister of Culture, Youth and Sports (1986), Minister of Defense (1991), Minister of Education (1995), Minister of Finance (2008), United Malays National Organization (UMNO), University of Nottingham, B.Sc. in Industrial Economics (1974), Born 23 July 1953, Kuala Lipis, Businessman, Islam (religion), Bugis Ethnic.
Drs. H. Muhammad Jusuf Kalla	10th Vice President of Indonesia (2004-2009), Minister of Industry of Trade (2000), Minister of State owned Enterprises (2000), Coordinating Minister of People's Welfare (2001), Golkar Party, Hasanuddin University (1967), The European Institute of Business Administration (1977), Born 15 May 1942, Watampone South Sulawesi, Businessman, Islam (religion), Bugis Ethnic.

Rangkaian sosial juga diwakilkan sebagai kajian model matematik ke atas interaksi sosial. Satu rangkaian sosial boleh dimodelkan oleh teori graf $G(V,E)$, iaitu satu bentuk formal yang mempunyai set nod $V \neq \emptyset$ dan satu set tepi atau pinggir E. V terdiri daripada individu dalam sesebuah organisasi atau komuniti. Satu nod yang mewakili pelakon merupakan properti utama rangkaian sosial, seorang pelakon yang kewujudannya mungkin dianggap mustahil dalam satu rangkaian sosial – sebagaimana seorang petani di Sulawesi Indonesia, dan Perdana Menteri Keenam Malaysia Dato' Sri Mohamad Najib bin Abdul Razak – sebenarnya mungkin lebih berkait rapat daripada sebaliknya yang pernah diutarakan. Teori ini berdasarkan idea enam darjah (Travers & Milgram 1969), yang diutarakan dalam permainan '*Six*

Degrees of Separation' oleh John Guare (Guare 2007). Hubungan antara mereka sebagai properti daripada rangkaian sosial dipanggil ikatan (*tie*) (Cook 2001). Satu ikatan boleh terdiri daripada satu atau lebih hubungan antara setiap pasangan pelakon. Ikatan sosial antara dua pelakon dikatakan sebagai pasangan (*dyad*). Satu konsep untuk menentukan hubungan antara pelakon sosial (Casciaro 1998). Oleh itu model rangkaian bersandarkan kepada teori graf adalah bagi menyederhanakan konsep hubungan sosial yang mungkin wujud antara dua pelakon sosial. Sifat rangkaian ini dikaji sebagai subset graf (Garton et al. 1997), yang mewakili rangkaian sosial sebagai struktur sosial, iaitu nod mewakili sebarang pelakon sosial, dan nod ini dihubungkan oleh satu atau lebih hubungan tertentu yang mungkin melibatkan persahabatan, perkawinan dan persaudaraan.

Dalam sains sosial, individu, kumpulan, atau organisasi amnya dikenali sebagai pelakon sosial atau ejen sosial, manakala dalam bidang sains komputer dikenali sebagai entiti atau objek. Entiti ialah satu objek yang unik dan dapat dikenali dalam persekitaran tertentu, dan setiap entiti memiliki satu atau lebih atribut. Dalam rangkaian sosial berdasarkan semantik, nod ialah contoh konsep yang mewarisi maklumat tersembunyi daripada pelakon di sebalik pengetahuan mengenai latar belakang pelakon itu. Setiap pelakon sosial mempunyai peranan dalam sosial. Setiap pelakon pula dicirikan oleh fiturnya secara bersendirian, dan fitur ialah label bagi pelakon. Contoh label bagi 2 (dua) orang Bugis dipaparkan pada Jadual *TwoBugis People*s (Jadual 2.1).

Fitur (ciri-ciri) umumnya dikenali sebagai atribut yang diperoleh daripada maklumat untuk memahami situasi atau masalah (latar belakang pelakon). Atribut ialah ciri-ciri yang menguraikan tentang sesuatu entiti (seorang pelakon). Oleh itu, nod rangkaian tidak lagi homogen. Pelakon, atribut dan hubungan boleh diwakilkan secara formal seperti berikut.

Pelakon

$A \neq \emptyset$ ialah satu set pelakon yang terdiri daripada a_i, $i = 1,\ldots,n$, iaitu a_i sebagai unsur daripada A adalah entiti, atau $A = \{a_1,\ldots,a_n\}$. Simbol $|.|$

mewakili saiz daripada set atau bilangan unsur dalam satu set. $|A| = n$ ialah bilangan unsur dalam A.

Atribut

$Z \neq \emptyset$ ialah satu set atribut yang terdiri daripada z_j, $j = 1,\ldots,m$, iaitu unsur z_i ialah atribut yang boleh wujud untuk pelakon sosial, atau $Z = \{z_1,\ldots,z_n\}$. Pasangan (A,Z) adalah sebagai kejadian dan entiti, Z_i ialah subset daripada Z. Z_i ialah set atribut daripada setiap entiti, iaitu (a_i,Z_i), $i = 1,\ldots,n$.

Hubungan

R ialah set hubungan yang mungkin antara pelakon sosial $R = \{r_j | j = 1,\ldots,k\}$ dengan mana $r_j : A \times A \rightarrow r_j(a,b)$ dalam R, atau $A \times A = \{(a,b) : a,b \text{ dalam } A\}$.

Ramai penyelidik rangkaian sosial tidak menghiraukan ciri-ciri khusus individu, seperti jantina, umur, pendidikan, pekerjaan (Wagner & Sternberg 2004), atau seperti kumpulan perkataan dalam *snippet* yang mewakili pelakon sosial untuk mempertimbangkan hubungan dan pertukaran di kalangan pelakon sosial (Memon & Alhajj 2010). Para penyelidik hanya mempertimbangkan pertukaran modal sosial dan interaksi yang mencipta dan mengekalkan hubungan sosial. Apakah hubungan itu berbeza, berasaskan semua atribut sosial seperti struktur keluarga atau seperti persaudaraan atau persahabatan di dalam masyarakat, atau mempunyai sifat transaksi seperti kehubungan perdagangan antara negara, ialah elemen yang memberi erti berbeza dalam rangkaian sosial. Lagipun, jenis sumber boleh menjadi banyak dan pelbagai, sumber ini boleh menjadi ketara seperti barangan dan perkhidmatan atau tidak ketara seperti pengaruh atau sokongan sosial (Wellman 1992). Oleh itu, pinggir dalam rangkaian juga tidak homogen, dan digunakan untuk menerangkan hubungan antara nod. Hubungan mempunyai ciri seperti berikut:

i. Kandungan daripada satu hubungan merujuk kepada sumber yang dipertukarkan. Sumber ialah barangan bernilai dalam masyarakat, sesuatu yang menggalakkan dan menyokong diri seseorang untuk terus hidup dan mengekalkannya (Lai et al. 1998; Molm et al 2006). Dalam konteks rangkaian komputer, sepasang pelakon bertukar-tukar pelbagai jenis

maklumat, seperti komunikasi mengenai pantadbiran, perkara peribadi, perkara yang berkaitan dengan kerja atau sosial. Hubungan yang berdasarkan rangkaian komputer melibatkan penghantaran fail data atau aturcara komputer serta menyediakan sokongan emosi atau mengatur mesyuarat seperti rangkaian perbincangan elektronik dengan menggunakan emel korporat (Wu et al. 2004; Tyler et al. 2005), penyenaraian emel (Smith 1997; Gloor et al. 2003), membina sokongan bagi peserta dalam pilihan raya melalui penggunaan blog, ruang komuniti dalam talian dan perkhidmatan rangkaian sosial seperti Facebook, Twitter, dan sebagainya (Kayaalp et al. 2009; Uesugi 2011). Dengan peningkatan perdagangan secara elektronik (sebagai contoh, sistem catat pesanan berasaskan Web, perbankan elektronik), maklumat yang dipertukarkan melalui Komunikasi Berperantaraan Komputer (KBK) juga dipertimbangkan mungkin sesuai dengan pertukaran wang, barang atau perkhidmatan dalam dunia sebenar.

ii. Hubungan boleh mempunyai arah (berarah) atau tidak (tidak berarah) (Brink & Gilles 2000; Robins et al. 2009). Sebagai contoh, seseorang boleh memberi sokongan sosial kepada orang lain. Terdapat dua kehubungan yang boleh wujud di sini: memberi sokongan dan menerima sokongan. Struktur yang menghubungkan Web dianggap sebagai proksi untuk hubungan dunia sebenar sebagai pautan yang dipilih oleh pengarang laman Web untuk menyokong maklumat yang ada di dalam laman web itu (ini dipanggil juga sebagai *rangkaian sosial Web* (Mika 2005a)). Selain itu, pelakon boleh berkongsi kehubungan persahabatan yang tidak mempunyai arah, iaitu, kedua-dua sekali mengekalkan hubungan dan tidak ada arah khas dijanakan. Dalam kes khusus, konsep ontologi juga berkaitan dengan hubungan (Mika 2005a, 2005b, 2007a), sebagai contoh jika pelakon berkongsi konsep yang sama, konsep ini mungkin mewujudkan hubungan (Matsuo et al. 2006a). Walau bagaimanapun, sementara kedua-duanya berkongsi persahabatan, hubungan boleh tidak seimbang: satu pelakon boleh menuntut persahabatan yang rapat, tapi pelakon lain hanya boleh menjana persahabatan yang lemah, atau komunikasi boleh dimulakan lebih kerap

oleh seorang pelakon daripada yang lain. Oleh itu, ketika hubungan dikongsi, ungkapannya mungkin tidak simetri (asimetri).

iii. Hubungan juga berbeza daripada segi kekuatan (Mollenhorst et al. 2008). Kekuatan hubungan (*strength relation*) boleh wujud dalam beberapa cara (Marsden & Campbell 1984; Wellman & Worley 1990; Matsuo et al. 2007b; Jin et al. 2007b). Dalam era komunikasi, setiap pasangan pelakon boleh (kerap, jarang-jarang, atau sekali-sekala) melakukan komunikasi sepanjang hari, seminggu, atau setahun. Mereka boleh bertukar-tukar banyak atau sedikit modal sosial: wang, barang, atau perkhidmatan. Mereka boleh membekalkan maklumat yang penting atau remeh. Aspek kehubungan ini menyebabkan kekuatan hubungan boleh menjadi berbeza. Jenis hubungan penting dalam penyelidikan rangkaian komputer, KBK dan rangkaian sosial, oleh kerana melibatkan pertukaran maklumat yang kompleks atau sukar (Fish et al. 1993). Pada setiap keadaan, boleh ada sokongan emosi, komunikasi yang tidak menentu atau samar-samar, dan ada komunikasi untuk menjana idea, mewujudkan konsensus, menyokong kerja, sesuatu yang mengeratkan kehubungan, untuk menyokong komuniti maya, atau perkongsian pengetahuan.

Konsep rangkaian ialah salah satu paradigma yang boleh menakrifkan pelbagai perkara pada era moden atau untuk memahami fenomena yang wujud dalam masyarakat. Pendekatan rangkaian membolehkan untuk mencam interaksi antara unit analisis seperti pelakon, hubungan, label, dan lainnya bersandar kepada pembolehubah yang dijana bagi sesuatu konsep. Pembolehubah semantik pelakon sosial saling berkait melalui unit hubungan antara pasangan pelakon yang dikenali sebagai ikatan. Dalam hal ini, rangkaian sosial berkongsi struktur yang sama. Sepasang pelakon boleh mengekalkan kehubungan berasaskan satu ikatan sahaja, contohnya sebagai ahli dalam organisasi yang sama. Walau bagaimanapun, mereka boleh mengekalkan hubungan multipleks berasaskan kepada kehubungan seperti berkongsi maklumat, memberi sokongan kewangan dan menghadiri persidangan bersama-sama. Contoh kehubungan pengarang bersama kertas kerja dan kehubungan hadir dalam persidangan saintifik ialah hubungan multipleks (Fienberg et al. 1985).

Ikatan sedemikian juga boleh berbeza kandungan, arah dan kekuatan. Setiap ikatan boleh mempunyai arah, iaitu berasal dari pelakon sumber dan mencapai sasaran seorang pelakon yang lain, atau terikat seri antara pasangan pelakon (tidak berarah). Contohnya hubungan antara Simba, Mustafa, Sarabi, dan Scar dalam cerita "*The Lion King*" mempunyai ikatan yang berbeza. Dalam hal ini, "Sarabi dan Mustafa menjadi ibu bapa daripada Simba", tapi tidak sebaliknya (ikatan yang terikat arah), manakala "Mustafa ialah saudara daripada Scar" dan "Scar ialah saudara daripada Mustafa" (ikatan terikat seri). Dalam contoh lain, pengarang bersama kertas kerja merupakan satu bentuk dari *co-occurrence* yang menghasilkan ikatan seri seperti pengarang "Mahyuddin K. M. Nasution" dan "Shahrul Azman Mohd Noah" yang menulis kertas kerja "*Superficial method for extracting social network for academic from Web snippet*", manakala ikatan dengan arah dari "Mahyuddin K. M. Nasution" atau "Shahrul Azman Mohd Noah" terhadap "Yutaka Matsuo" kerana kedua-dua pengarang pertama melakukan petikan dari kertas kerja yang ditulis oleh "Yutaka Matsuo". Oleh itu, hubungan juga disokong oleh nama kehubungan sebagai properti rangkaian sosial yang dipanggil sebagai label seperti '*citation*'.

2.2.2 Data untuk rangkaian sosial

Jenis data berlainan boleh digunakan untuk membina rangkaian sosial. Pada tahap awal pengumpulan data untuk rangkaian sosial memerlukan strategi, agar data rangkaian boleh diorganisasi dan diuji untuk memberikan proses yang signifikan apabila melibatkan analisis yang berbeza (Tichy et al. 1979). Proses pengumpulan data terdiri daripada penggunaan kaedah manual atau penggunaan semula rekod elektronik yang sedia ada. Secara klasik kajian rangkaian sosial dilakukan melalui pemerhatian secara manual, dan kaedah ini melibatkan ramai pekerja secara intensif dan mengambil lebih banyak masa dan sumber. Sebagai contoh pemerhatian dengan menggunakan soal selidik tentang kehubungan pada sekumpulan manusia (Vehovar et al. 2008), atau menyiasat sampel populasi melalui temuduga seorang demi seorang pelakon sosial mengenai hubungan mereka.

Rekod elektronik seperti dokumen atau pangkalan data digital adalah antara penyelesaian kreatif bagi masalah pengumpulan data. Internet menjadi pilihan

sebagai prasarana asas bagi mendapatkan maklumat digital pelbagai topik dari seluruh dunia. Perkara ini dapat dilakukan dengan penganotasian maklumat semantik daripada dokumen Web.

2.3 *WORLD WIDE WEB* DALAM RANGKAIAN KOMPUTER

World Wide Web (WWW) atau Web, mengandungi banyak maklumat, tidak hanya menjadi pangkalan data teks terbesar dalam rangkaian komputer yang pernah wujud dalam sejarah, tetapi menjadi lebih kompleks selaras dengan saiz yang terus tumbuh dengan kadar yang luar biasa (Nikravesh et al. 2002). Web dibina diatas rangkaian global yang menghubungkan semua orang diseluruh dunia. Sumber dalam Web dicipta oleh berjuta-juta orang. Pada setiap masa, banyak dokumen baru diterbitkan yang mengandungi maklumat terkini dari pelbagai tempat di seluruh dunia (Jan et al. 2006).

Untuk mendapatkan faedah daripada sumber Web dengan baik, memerlukan kejuruteraan data Web untuk membawa maklumat kepada nilai yang tinggi bagi pengguna akhir. Enjin carian memainkan peranan sebagai pintu gerbang untuk mengakses Web (Arrue et al. 2008). Pengguna boleh mendapatkan maklumat dengan mengemukakan kueri kepada enjin carian. Enjin carian memproses lebih dahulu sumber di dalam Web dan menyimpannya ke dalam pangkalan data indeks. Kemudian apabila menerima permintaan daripada pengguna, enjin carian dengan pantas boleh mencari dalam pangkalan data indeks dan mengembalikan dokumen hasil kepada pengguna (Bruno et al. 2011).

Bahagian-bahagian berikut menjelaskan apa dan kenapa Web sebagai sumber maklumat bagi pengekstrakan rangkaian sosial, yang menguraikan alasan bahawa bilamana rangkaian komputer menghubungkan orang, itu bermakna rangkaian sosial.

2.3.1 Apakah Web?

Pada bulan Ogos 1991, Berners-Lee mengeluarkan perisian Web dan mencipta sebuah laman Web pertama, yang boleh bekerja secara optimum dengan protokol

HTTP pada Internet (Berners-Lee 1996). Dalam hal ini, Internet (*Interconnected-networking*) menjadi infrastruktur penting bagi komunikasi di dunia. Internet ialah rangkaian komputer yang disambungkan secara global dengan melibatkan penggunaan TCP/IP (*Transmission Control Protocol/Internet Protocol*) sebagai protokol pertukaran pakej (Baggio & van Steen 2005).

Sejarah Internet yang bermula pada akhir 1960-an boleh dikata sebagai titik mula rangkaian sosial berasaskan elektronik, kerana telah menghubungkan entitas tertentu. Apabila *Defence Advance Research Projects Agency* (DARPA) Amerika Syarikat membina satu rangkaian intranet komputer untuk tujuan penyelidikan tentera, dan kemudian penggunaannya diperluaskan untuk menghubungkan institusi dan universiti pada tahun 1970-an. Perkara ini secara sosial telah membangun satu rangkaian antara institusi rasmi pemerintah. Seterusnya, usaha sentiasa dibuat untuk menyambung rangkaian komputer yang jauh bersama-sama untuk mewujudkan rangkaian tersambung secara global, yang menghubungkan benua yang dipisah oleh lautan, hinggalah ke awal 1990-an, Internet boleh sampai kepada organisasi, syarikat, isi rumah, dan kini hampir di semua tempat yang didiami oleh manusia, yang menyebabkan kehadiran rangkaian sosial lebih luas kerana melibatkan semua lapisan sosial masyarakat. Perkara ini disokong oleh usaha melibatkan hasil penyelidikan daripada perkakasan dan perisian. Internet global dibina atas infrastruktur rangkaian gergasi bertenaga (Tanenbaum 2003). Internet telah mengeskploitasi kemajuan terkini dalam teknologi semikonduktor, teknologi perkakasan dan teknologi nano (Li et al. 2007) yang mempermudah penyampaian maklumat dan komunikasi antara ahli sosial, memunculkan satu bentuk interaksi sosial maya, yang memungkinkan pendekatan ekstraksi rangkaian sosial boleh dibuat. Lagi pula, dengan kemunculan Web telah banyak mengubah kaedah pertukaran maklumat. Web telah mengurangkan kos penerbitan maklumat, telah memendekkan masa akses kepada maklumat, dan telah menurunkan tembok antara penerbit maklumat dan penerima maklumat (Berberich et al. 2007). Akibatnya juga akan mengurangkan kos dan bilangan orang yang diperlukan untuk mendapatkan suatu rangkaian sosial.

2.3.2 Mengapa Web?

Pada abad ini, maklumat telah menjadi keperluan asas bagi semua orang. Maklumat tidak hanya perlu pantas tetapi juga boleh diperbaharui (Liu et al. 2002). Mengikut tradisi, sesuatu maklumat melalui banyak peringkat untuk boleh sampai kepada penerima dari pengarang: pengkoleksian, rekod, penyuntingan, ratifikasi, percetakan, penghantaran, pengiklanan, dan kemudian transaksi. Peringkat ini merupakan kaedah utama bagi menerbitkan maklumat seperti buku, surat khabar, berita radio, program televisyen, dan sebagainya. Satu-satunya cara aliran maklumat ialah dari penerbit kepada penerima. Penghebahan maklumat secara klasik melibatkan ramai orang dan pelbagai peralatan, dengan itu kos yang dikehendaki meningkat. Kos yang tinggi dalam kaedah ini telah menghadkan bilangan penerbit dan pengarang yang boleh mencipta dan mengagihkan maklumat. Untuk mencetak buku atau untuk menghasilkan program televisyen mengambil masa berbulan sampai tahunan persiapan sebelum maklumat itu boleh sampai kepada penerima. Oleh itu, antara ramai orang yang mahu untuk menghebahkan maklumat, hanya sebahagian daripada mereka yang boleh menerbitkan. Dengan demikian, laju interaksi sosial terhad kepada maklumat yang diakui oleh segelintir ahli komunitas sosial tertentu sahaja, yang menyebabkan maklumat tentang rangkaian sosial terhad kepada dokumen klasik sebagai sumber maklumat, dan tidak menyokong pengekstrakan rangkaian sosial yang melibatkan penggunaan teknologi informasi.

Pada era Web, segala-galanya telah berubah. Kos untuk menggunakan Web sangat murah sehingga semua orang boleh menyiarkan maklumat dengan hampir percuma (Sommerville et al. 1998). Organisasi seperti syarikat atau universiti mengendalikan pelayanan mereka sendiri untuk mengiklankan maklumat mereka. Akhbar, syarikat, dan penerbit menghasilkan berita, rencana, dan buku dalam Web selain buku cetakan dan surat khabar konvensional. Selain organisasi, mana-mana individu juga boleh meminta ruang cakera keras dan menubuhkan *homepage* sendiri atau blog, seperti blog http://scholarlyoa.com/publishers/ yang menyenaraikan penerbit yang meragukan (sebagai satu penyelidikan tentang literatur) dan beberapa pandangan dari berbagai belahan dunia (sebagai interaksi sosial akademik), dan kajian ini diterbitkan pada jurnal *Nature* pada tahun 2012 (Beall 2012).

Perkembangan semasa Web, yang dipanggil sebagai "Web 2.0" (Chang et al. 2008) membolehkan semakin ramai orang untuk menerbitkan maklumat tanpa pengetahuan teknikal yang mahir. Selain itu, dokumen Web semakin pintar: maklumat tertentu tidak sahaja ditandai oleh tag tetapi juga metadata khas yang membolehkan maklumat itu dikenali secara unik melalui teknologi tertentu (Lai & Turban 2008), yang memungkinkan sebarang pengekstrakan maklumat untuk memperoleh pengetahuan tertentu boleh dilakukan, seperti pengekstrakan rangkaian sosial akademik.

Bilangan maklumat di Web telah meningkat setiap hari dan Web telah menjadi satu persekitaran heterogen (Zhao & Ram 2007). Sejak itu, bilangan penerbit maklumat di Web tumbuh dengan pantas, pangkalan data Web bercampur-campur dari dan dengan pelbagai dokumen. Dokumen Web boleh berasal dari pelbagai jenis sumber, dengan topik pelbagai dan gaya penulisan mereka yang tidak seragam. Format dokumen berbeza-beza. Format rasmi juga berbeza dalam dokumen sebab diwujudkan oleh pertubuhan yang berbeza sebagaimana bentuk kertas kerja daripada Springer, IEEE dan lainnya. Format tidak rasmi dalam dokumen juga beraneka-ragam oleh individu seperti yang terdapat dalam laman blog. Dokumen organisasi lazimnya mengandungi maklumat lebih dipercaya manakala dokumen daripada individu mungkin mengandungi maklumat sedikit dipercayai. Oleh sebab pelbagai alasan demikian, secara keseluruhan dokumen Web dipertimbangkan sebagai tidak berstruktur, dan memerlukan teknologi untuk mendapatkan sebarang pengetahuan, iaitu yang dikenali sebagai pengekstrakan.

Secara semantik, penjanaan sebarang maklumat yang sesuai adalah sangat penting untuk merealisasikan Web sebagai sesuatu yang dipercaya atau sebagai gambaran sosial sebenar, yang memungkinkan penaksiran kebolehpercayaan dan dapat dipercayai (Golbeck & Hendler 2004). Lagi pula, teknik perwakilan dokumen adalah juga pelbagai. Sebelum ini, kebanyakan laman Web statik dalam kandungan, tetapi sekarang ini laman Web baru kerap muncul dengan kandungan yang dinamik bertukar. Laman Web yang boleh berinteraksi dengan pengguna dan memberikan maklumat masa nyata (konteks terkini) semakin meningkat, yang secara tidak langsung menggambarkan dinamik perubahan sosial, yang menyebabkan perlu

meneroka Web untuk mendapatkan sebarang pengetahuan atau konteks terkini tentang sesuatu perkara termasuk rangkaian sosial dan membangun kaedah pengekstrakan yang sesuai.

Web boleh dianggap sebagai cermin dunia sebenar. Perubahan dalam dunia sebenar digambarkan dalam Web maya (Gruber 2008). Akhbar senantiasa mengemaskini berita dari seluruh dunia pada setiap saat. Orang menerangkan komen mereka, pendapat mereka, dan ulasan mereka dalam laman Web dan blog mereka. Orang yang mempunyai minat atau hobi yang sama datang bersama-sama untuk membentuk komuniti dan rangkaian sosial. Pengguna Internet kini hidup dalam dua dunia pada masa yang sama: satu dunia sebenar, satu lagi dunia Web maya. Mereka boleh hidup, bekerja dan menikmati bersama dengan dunia maya. Perkhidmatan permainan di Internet dan e-dagang ialah jenis pemakaian Web yang memberi pengguna nilai baru dan manfaat yang baru, yang mengukuhkan interaksi sosial dan hubungan antara mereka melalui modal sosial yang mengalir bersama Web, dan jaringan sosial semakin penting untuk menggambarkan struktur sosial yang berbeza-beza berdasarkan Web sebagai sumber maklumat yang boleh dipercaya (Wang et al. 2004). Dalam perkataan lain, rangkaian sosial sebenar akan sepadan dengan rangkaian sosial hasil pengekstrakan.

2.3.3 Struktur Laman Web

Web ialah sumber data yang kaya dan sentiasa tersirat dengan pelbagai jenis objek sebenar dan maklumat kehubungan yang sepadan. Apa sahaja objek (sama ada entiti atau atribut itu sendiri) boleh ditakrifkan secara literal, seperti teks "*Social Network*". Oleh itu semua maksud objek yang berasaskan perkataan boleh diwakili oleh objek literal itu sendiri. Secara formal, perkataan w ialah unit asas data diskret, yang ditakrifkan sebagai item dari perbendaharaan kata yang diindeks oleh $\{1,\ldots,K\}$, yang mana w_k =1 jika k ada dalam K, dan w_k = 0 jika sebaliknya. Entri asas dalam web ialah laman (atau dokumen) yang menakrifkan dan menerangkan satu entiti atau konsep.

Dokumen

Dokumen ialah satu urutan n perkataan yang ditandakan oleh $d = \{w_i | i=1,\ldots,n\}$, iaitu w_i ialah perkataan ke-i dalam urutan. Saiz dokumen $|d| = n$.

Laman Web sebagai dokumen Web mempunyai beberapa tag yang memberi struktur kepada teks.

Laman Web

Laman Web diwakilkan dengan ω ialah pepohon dokumen (*document tree*) yang terdiri daripada

- Set tag, yakni $TAG = \{tag_i | i=1,\ldots,m\}$, $TAG \neq \emptyset$, dan
- Set perkataan, yakni $W = \{w_j | j=1,\ldots,n\}$.

dan terdiri daripada sebarang bilangan cawangan yang memenuhi aturan mengenai item apa yang boleh ada pada setiap cawangan.

Unsur pepohon dokumen terdiri daripada akar, cabang dan daun, iaitu

i. Unsur akar pepohon dokumen ialah *doctype html*. Sebagai contoh

```
<!DOCTYPE html PUBLIC "-//W3C/DTT HTML 1.0 Transitional//EN"
    http://www.w3.org/TR/xhtml1/DTD/xthml1-transitional.dtd>
```

Jika unsur akar wujud dalam konteks dokumen yang dikenalpasti oleh *doctype* sebagai XHTML, maka unsur HTML juga memerlukan atribut XMLNS (XML *namespace*).

ii. Unsur *head* mengandungi maklumat metadata yang menerangkan dokumen itu sendiri, atau bersekutu dengan sumber yang berkaitan, seperti skrip dan gaya kunci. Sebagai contoh

```
<head>
   <title> Tajuk Laman </title>
</head>
```

Head laman web mengandungi:

- Unsur tajuk atau dikenali sebagai *title* ialah nama yang diberikan kepada laman web, pada asasnya sebagai bahagian penting daripada *head* laman web, bertujuan untuk mengenalpasti kandungan dokumen web. Tajuk, sekeping maklumat yang sangat penting dari segi menyediakan ringkasan bermakna untuk mewakili halaman itu untuk enjin carian, yang memaparkan kandungan tajuk semasa hasil carian dikembalikan.
- *Base* ialah unsur tambahan kepada unsur tajuk, sebagai tapak *uniform resource locators* (URLs) bagi pautan atau sumber pada laman itu, dan tetingkap sasaran untuk membuka kandungan yang berkaitan.
- *Link* (pautan) adalah sebagai unsur tambahan kepada unsur tajuk, untuk merujuk kepada sumber daripada berbagai jenis, kerap ini berkenaan dengan kepingan gaya yang menyediakan arahan mengenai cara menyediakan pelbagai unsur pada laman Web.
- *Meta* ialah unsur tambahan kepada unsur tajuk untuk menyediakan maklumat tambahan mengenai laman itu. Sebagai contoh, pengkodan aksara laman yang digunakan, ringkasan kandungan laman, arahan untuk enjin carian sama ada melalui kandungan indeks ataupun tidak, dan lain-lain.
- Objek juga merupakan unsur tambahan yang mewakili bekas serbaguna generik yang diperuntukkan bagi objek media.
- Skrip adalah sebagai unsur tambahan sama ada bagi membenam atau merujuk kepada sebuah skrip luaran.
- Gaya menyediakan kawasan yang menambah unsur tajuk untuk menakrif gaya CSS yang dibenam.

iii. Unsur tubuh ialah unsur yang mengandungi sebahagian besar laman Web, iaitu elemen yang berkaitan dengan perenggan, senarai, pautan, imej, jadual, dan banyak lagi.

Oleh itu, laman web secara unik dikenalpasti melalui tajuk dan alamat URL (*uniform resource locator*) (Barners-Lee et al. 2004; Lee et al. 2005). Organisasi logik web boleh dilihat sebagai hierarki, dan URL laman web merupakan petunjuk

mengenai kedudukan logik dalam struktur hierarki. Secara sintaksis, URL mewakili sumber di Internet yang boleh ditakrif secara diskret seperti berikut.

Komposisi URL

Komposisi URL mengandungi set token $U = \{s, d_1,\ldots,d_m, p_1,\ldots,p_{n-1}\}$ yang memenuhi struktur: $s://d_m.\cdots.d_2.d_1/p_1/p_2/\cdots/p_{n-1}$, rentetan terdiri daripada skema, autoriti dan laluan.

Pada amnya, URL dipisahkan oleh garis condong (*slash* atau "/") kepada beberapa lapisan dan setiap lapisan boleh dianggap sebagai nama domain, direktori atau fail. Secara terperinci, setiap token mewakili komponen URL. Token s mewakili skema dan mengandungi protokol yang diguna untuk berkomunikasi dalam Internet, seperti http dan https. Rentetan seperti $d_m.\cdots.d_2.d_1$ ialah autoriti, suatu komponen mempunyai 3 subkomponen maklumat iaitu pengguna, *host*, dan *port*.

i. Maklumat pengguna boleh terdiri daripada nama pengguna, dan maklumat pilihan-skema khas mengenai bagaimana mendapatkan kebenaran untuk mengakses sumber. Lazimnya, perkara ini diikuti oleh *commercial at-sign* ("@") yang memberi sempadan kepada *host* jika ada, seperti dalam alamat emel.
ii. *Host* mengandungi lokasi pelayan Web, pada lokasi ini boleh diterangkan sama ada sebagai sistem nama domain (*domain name system*, DNS) mahupun sebagai protokol internet (*internet protocol*, IP).
iii. *Port* ialah nombor tertentu. Sebagai contoh, nombor pelabuhan lalai (*default port*) adalah 80 untuk HTTP iaitu $s://d_m.\cdots.d_2.d_1:80/$. Simbol noktah bertindih (":") merupakan awalan sebelum nombor pelabuhan.

Rentetan terakir token, $/p_1/p_2/\cdots/p_{n-1}$ merupakan laluan, iaitu satu komponen mengandungi direktori yang melibatkan laman web dan nama fail daripada laman itu, direktori dan fail dipisah oleh garis condong. Token terakhir daripada laluan kadang-kadang datang dengan dua komponen lain: *query* and *fragment*. Kueri (*query*) ialah satu komponen yang mengandungi nama parameter dan nilai yang boleh dibekalkan kepada penggunaan Web. Token laluan dan kueri dipisah oleh simbol tanda soal

(“?”). Bentuk kueri ialah `nama=[value]`, iaitu terdapat simbol sama dengan (“=”) antara nama parameter dan nilai parameter. Sepasang `nama=[value]` dipisah antara satu sama lain oleh simbol *ampersand* (“&”). *Fragment* ialah komponen bagi menunjuk kepada bahagian parameter dokumen. Komponen terakhir ini dan bahagian sebelum ini dimediasi oleh simbol *sharp* (“#”). Sebagai contoh http://search.yahoo.com/search;_ylt=AkMEIMkaXLZz0ZciMgGgURmbvZx4?p=Mahyuddin+K.+M.+Nasution&toggle=1&cop=mss&ei=UTF-8&fr=yfp-t-701.

Bentuk kanonik URL (Lee et al. 2005)

Bentuk kanonik URL ialah komposisi komponen-komponen dalam U = {s,d,p,q} = {*scheme,authority,path,query*}, iaitu bentuk rentetan $s://d_m.\cdots.d_2.d_1/p_1/p_2/\cdots/p_{n-2}/x$, $x = p_{n-1}$ atau $x = p_{n-1}?q$, dengan mana URL mempunyai n lapisan dan setiap bahagian dipisahkan oleh garis condong.

Oleh itu, semua kandungan daripada Web seperti teks literal, alamat URL, sama ada tajuk atau ringkasan daripada Web boleh diekstrak untuk mendapatkan rangkaian sosial berasaskan asumsi bahawa

i. Laman Web selalu mengandungi sebarang entiti sebagai agen atau pelakon sosial, terutama nama orang secara literal.
ii. Laman Web ditulis oleh, dicipta untuk atau mewakili satu atau lebih entiti.
iii. Laman Web cenderung mempunyai konsep yang sama secara literal teks atau format.
iv. Laman Web yang hampir sama cenderung ditempatkan oleh editor pada direktori yang berdekatan atau mempunyai kesamaan alamat URL.

2.3.4 Enjin carian

Bagaimanapun untuk mendapatkan maklumat daripada Web seperti rangkaian sosial, tidak hanya memerlukan teknologi seperti pengekstrakan, tetapi juga memerlukan alat seperti enjin carian. Oleh sebab, Internet telah tumbuh menjadi koleksi berbillion

laman Web (Hirate et al. 2008). Dengan kata lain, Web telah membesar kepada lebih daripada satu billion dokumen Web unik dan terus tumbuh lebih kurang pada kadar satu juta dokumen setiap hari (Bharat & Broder 1998). Kuantiti besar data dalam Web ialah satu kelebihan yang boleh diambil kira bagi pengguna bilamana mereka mahukan maklumat, seperti maklumat rangkaian sosial akademik. Pengguna boleh mendapatkan maklumat pada bila-bila masa. Pengguna boleh memanipulasikan data dengan bebas. Pengguna mempunyai banyak pilihan untuk mendapat pengetahuan baru dan mewujudkan nilai baru bagi data. Walau bagaimanapun, untuk menggunakan Web dengan cara yang berkesan merupakan satu masalah besar. Bagaimana mendapatkan maklumat yang berguna? Sebab akses kepada data merupakan perkara yang sama bagi sesiapa sahaja, setiap orang yang mengeksploitasi Web dengan pantas dan berkesan ialah pemenang dalam era Web. Enjin carian ialah antarmuka penting bagi maklumat yang luas dalam Internet. Enjin carian kini mempunyai peran sebagai alat yang penting untuk membantu pengguna untuk memperoleh maklumat (Broder 2002). Oleh itu, dalam pengekstrakan rangkaian sosial secara automatik dari Web, enjin carian akan berperan sangat penting.

Sebelum kemunculan enjin carian, orang ramai perlu tahu alamat URL laman Web untuk mendapatkan maklumat. Walau bagaimanapun, pengguna hanya boleh ingat dan boleh memasukkan bilangan alamat URL yang sangat sedikit. Sebelum enjin carian wujud, sesetengah perkhidmatan direktori web cuba untuk mengkategorikan laman-laman web mengikut tajuk. Tajuk akan diorganisasikan dalam struktur hierarki untuk membantu pengguna menavigasi melalui direktori dengan mudah. Walaupun direktori Web boleh membantu pengguna untuk mencari tajuk penting melalui struktur topik hierarkinya, pengorganisasian direktori Web memerlukan lebih kerja. Ini disebabkan oleh pangkalan data Web berubah dengan pesat, dan menjadi perkara yang mustahil bagi direktori laman Web kekal bersama perubahan itu. Oleh itu, untuk mengorganisasi direktori Web selari dengan perkembangan Web adalah dengan membangunkan struktur yang sesuai dan boleh mencerminkan kandungan laman Web (Horvat et al. 2009).

Boleh dikatakan bahawa enjin carian yang menyebabkan revolusi dalam kaedah mendapatkan maklumat. Enjin carian cuba untuk menyimpan semua kandungan yang wujud dalam Web. Pertama enjin carian merayapi Web dengan mengikuti hiperpautan antara laman Web. Kemudian, teks yang ada dalam satu laman dirayapi dan kemudian diindeks bagi mewakilkan kandungan laman, dan menyimpannya dalam pangkalan data pengindeksan. Enjin carian juga menganalisis hiperpautan antara laman Web (Brin & Page 1998; Qiao et al. 2010) untuk menentukan kedudukan laman Web bagi kepentingan, kualiti dan kebolehpercayaan. Keperluan maklumat pengguna diwakili dalam kueri yang dikemukakan kepada enjin carian. Satu kueri carian Web ialah kueri yang pengguna kemukakan kepada enjin carian Web untuk memenuhi keperluan maklumat mereka. Kueri carian Web adalah tidak berstruktur dan mengandungi kata carian (*search term*) yang secara tidak langsung berperan menyediakan multidimensi maklumat dalam pengekstrakan rangkaian sosial.

Kata carian

Kata carian t_k terdiri daripada sekurang-kurangnya satu perkataan atau $t_k = (w_1,\ldots,w_l)$, $l \leq k$, k ialah bilangan parameter yang mewakili perkataan. $|t_k| = k$ ialah bilangan perkataan dalam t_k, dan l ialah bilangan perbendaharaan perkataan dalam t_k.

Bagi setiap kata carian, enjin carian melihat kepada kandungan pangkalan data pengindeksan dan mendapatkan semula dokumen yang dekat dengan kueri pengguna dalam masa kurang daripada sedetik. Enjin carian Web menyekutukan kueri yang tidak berstruktur ke pangkalan data berstruktur. Sebagai contoh, kueri berkenaan dengan pelakon mungkin disekutukan dengan pangkalan data yang mengandungi huraian dan spesifikasi pelakon.

Enjin carian

Satu set laman Web diindeks oleh enjin carian menjadi Ω, iaitu set yang mengandungi pasangan tersusun kata carian t_{ki} dan laman Web ω_{kj}, atau (t_{ki},ω_{kj}), $i = 1,\ldots,n$, $j = 1,\ldots,m$. Jadual hubungan yang terdiri daripada dua kolum t_k dan ω_k ialah perwakilan daripada (t_{ki},ω_{kj}), $\Omega_k = \{(t_k,\omega_k)_{ij}\}$ ialah

subset daripada Ω atau $\Omega_k = \{w_{k1},\ldots,w_{kj}\}$. Kardinaliti daripada Ω ditandakan dengan $|\Omega|$, dan fungsi kebarangkalian massa ialah $P : \Omega \rightarrow [0,1]$.

Enjin carian kini merupakan satu perkhidmatan yang menjadi pintu laluan penting untuk akses ke Web. Walau bagaimanapun, manakala pengguna menikmati bilangan besar maklumat yang terdapat di Web, ini menimbulkan masalah kepada pengguna dan enjin carian: Kerelevanan dokumen yang dipulangkan dalam set hasil enjin carian masih kurang. Ruang Web wujud pada persekitaran berbilang pengguna akhir dan pelbagai perkhidmatan penggunaan. Penerbit maklumat menjadi pelbagai jenis, dokumen Web juga adalah sangat pelbagai. Laman web boleh berasal dari dokumen organisasi, syarikat atau individu. Laman web ialah dokumen rasmi dan tidak rasmi, mempunyai tajuk, gaya penulisan, dan format berbeza. Oleh itu, data dalam Web adalah sangat hingar dan sukar untuk membina algoritma yang boleh memproses dengan baik sebarang data Web. Akibatnya, setiap kali pengekstrakan sebarang maklumat dari Web dilakukan, penilaian dan pengujian dilaksanakan untuk mendapatkan pemahaman tentang maklumat yang dihasilkan (Mika 2007b).

2.3.5 Maklumat Rangkaian Sosial Dalam Web

Maklumat mengenai pelakon terkandung dalam sebahagian besar daripada seluruh laman Web. Oleh kerana Web ialah cermin dunia sebenar, Web ialah rekod mengenai pelakon sosial yang boleh terdiri daripada pelakon itu sendiri atau catatan daripada pelakon lain (Yang et al. 2008). Pelakon sosial muncul dalam dokumen Web sebagai entiti dalam entiti seperti negara, organisasi, syarikat, universiti, produk, dan lain-lain. Entiti ini berinteraksi antara satu sama lain dalam dunia sebenar dan mewujudkan kisah atau sejarah dalam Web maya. Pelakon sosial muncul dalam sumber yang berbeza seperti akhbar, laman e-dagang, laman web syarikat dan laman web pengguna (Han & Zhao 2009a, 2009b). Oleh itu, keperluan untuk mengesan individu yang sama yang terdapat pada laman berbeza adalah penting apabila pengguna mahu mengumpul maklumat yang berguna tentang seseorang pelakon sosial, termasuk dalam pengekstrakan rangkaian sosial dari Web (Matsuo et al 2007b; Bollegala et al 2006b). Terdapat banyak pelakon mempunyai nama yang sama dan tidak sedikit pelakon mempunyai lebih daripada satu nama. Ini dikenali

sebagai masalah nyahkekaburan nama (*name disambiguation*). Ada dua sebab asas kepada keperluan proses nyahkekaburan nama (Han et al. 2003, 2004, 2005a, 2005b):

i. Masalah hiponimi: Pelakon yang berbeza boleh berkongsi nama yang sama (kekaburan leksikal)
ii. Masalah sinonimi: Entiti tunggal boleh diwakilkan oleh berbilang nama (kekaburan rujukan).

Untuk mencari maklumat yang berguna tentang pelakon sosial secara berkesan atau untuk memastikan nod dalam rangkaian mewakili satu pelakon sosial, kedua-dua masalah diatas perlu diselesaikan. Masalah hiponimi boleh diselesaikan dengan menganalisis dokumen dan mengukur kesamaan konteks yang berkaitan dengan nama pelakon dalam dokumen itu. Manakala, masalah kedua adalah lebih kompleks kerana adalah sukar untuk menghitung banyak calon dengan nama yang sinonim (Llyod et al. 2005; Song et al. 2007b).

Maklumat pelakon dalam dokumen web adalah sangat berbeza daripada maklumat untuk pelakon yang sama dalam dokumen seperti artikel berita dan penerbitan akademik. Dalam dokumen artikel berita atau penerbitan akademik, keseluruhan dokumen adalah berkaitan dengan pelakon berkenaan, manakala dalam dokumen Web tidak semua kandungan dokumen adalah berguna (Vu et al. 2009). Terdapat juga data hingar yang perlu disaring.

Jika dianalisis, terdapat pelbagai jenis sambungan sosial antara pelakon dalam persekitaran Web. Struktur pertautan (*linking*) Web menghubungkan kepada sumber maklumat lain yang dianggap berwibawa dan cukup relevan untuk disebut. Manakala *co-occurrence* boleh ditafsirkan sebagai petunjuk berdekatan atau secara semantik didefinisikan sebagai suatu kewujudan bersama-sama atau berkaitan antara satu sama lain. Oleh itu, sambungan sosial boleh dikategorikan seperti berikut (Kirchoff et al. 2008).

i. Sambungan sosial secara langsung dan tersurat: Struktur pertautan secara tersurat bertujuan untuk menggambarkan kehubungan sosial. Contoh, adalah seperti laman rangkaian sosial seperti Facebook, LinkedIn, MySpace, dan Orkut. Berjuta-juta pengguna yang mengekalkan profil peribadi dengan senarai kawan untuk berinteraksi dan berkomunikasi dengan mereka (Boyd & Ellison 2008), lihat Facebook Dato' Sri Mohammad Najib Abdul Razak dengan alamat URL: http://www.facebook.com/najibrazak. Paparan awam daripada sambungan dilihat sebagai isyarat pengenalan penting dan sedang diguna untuk mengekalkan pengurusan teraan (Boyd 2004; Donath & Boyd 2004), yang secara langsung memenuhi syarat nyahkekaburan nama (Bekkerman & McCallum 2005; Ono et al. 2008), dengan mana secara tersurat profil pengguna menjadi perwakilan terstruktur tatah bawah daripada pengguna (Fong et al. 2009). Contoh lain ialah protokol *Friend-of-a-friend* (FOAF) yang secara tersurat menjelaskan nyahkekaburan (Paolillo & Wright 2005; Mika 2005b) dan kehubungan di antara pelakon sosial (Finin et al. 2005; Mika & Gangemi 2004; Mika 2005b).

ii. Sambungan sosial secara tidak langsung dan tersurat: Struktur pertautan daripada sebarang laman web yang tersambung ke tapak web lain. Contoh, laman Web "*Category:Bugis people*" (http://en.wikipedia.org/wiki/category: Bugis_people) tertaut dengan laman Web "Category:Malaysian Bugis people" (http://en.wikipedia.org/wiki/Category:Malaysian_Bugis_people) yang menyambungkan secara tidak langsung Wakil Presiden ke-10 Indonesia dengan Perdana Menteri ke-6 Malaysia (Jadual 2.1). Sambungan ini berkaitan dengan rangkaian hiperpautan (Hamasaki et al. 2006a; Kretschmer et al. 2007), tetapi berbeza apabila perkara ini wujud dalam blog, yang memberi makna lebih peribadi dalam pautan sosial (Abel et al. 2010; Missen et al. 2010). Secara tidak langsung, sambungan Web seperti "*Category: Bugis people*" menuas URL untuk nyahkekaburan nama (Lu et al. 2007).

iii. Sambungan sosial secara langsung dan tersirat: Sambungan yang diekstrak daripada maklumat teks yang terdapat pada laman web, yang

dengan jelas menunjukkan kehubungan sosial antara pelakon yang berbeza (Kautz et al. 1997b; Cullota et al. 2004; McCallum et al. 2004; Mika 2005c; Wang et al. 2005; Matsuo et al. 2007b; Mori et al. 2007; Jin et al. 2008a; Tang et al. 2008b). Contoh kehubungan seperti itu ialah pengarang-bersama (*co-authorship*) dalam penerbitan sainstifik. Pengarang-bersama yang terdapat dalam dokumen ini boleh diguna sebagai sumber bagi pengekstrakan rangkaian sosial pengarang (Sun et al. 2011; Tu & Seng 2011). Pada amnya, tajuk laman Web, URL, dan teks dalam Web dapat digunakan sebagai label bagi setiap pelakon sosial, kerana Web mewakili aktiviti setiap pelakon sosial (Wei et al. 2006; Zhu et al. 2010). Contoh: kata kunci bagi setiap pelakon akademik dapat dijanakan melalui kumpulan kertas kerja yang telah diterbitkan dalam Web (Bollegala et al. 2006a, 2006b; Pereira et al. 2009).

iv. Sambungan sosial secara tidak langsung dan tersirat: Sambungan diekstrak pada kandungan laman web, yang mungkin menunjukkan kehubungan sosial yang samar atau lemah antara pelakon. Contoh paling signifikan ialah jaringan kolaboratif yang berlaku dalam Amazon.Com. Objek tumpuan seperti buku dan mengandaikan buku sebagai mewakili individu yang membelinya. Maka hasilnya akan membentuk rangkaian sosial. (Krebs 2000), data pemesanan buku (Yanagimoto et al. 2010; Yang & Lee 2011), atau petikan dalam penerbitan karya saintifik (Hou & Chen 2011).

Selain itu, rangkaian sosial boleh dibentuk antara laman Web dengan hiperpautan kepada laman Web yang lain, dan ini dikenali sebagai rangkaian sosial Web, yang memodelkan laman Web sebagai nod dan hipertautan sebagai pinggir daripada rangkaian itu (Mika 2007b). Dengan pendekatan yang sama terhadap kandungan laman Web, secara semantik bilamana dalam laman Web terdapat hubungan semantik atau memiliki konsep yang sama, maka dengan mempertimbangkan dokumen Web atau konsep Web sebagai nod, rangkaian sosial antara mereka boleh dibangunkan sekurang-kurangnya melalui penggunaan kesamaan atau hubungan semantik tertentu (dalam hal ini dipanggil sebagai rangkaian sosial Web semantik).

2.4 PENUTUP

Kaedah tradisional dalam pembinaan rangkaian sosial melibatkan masa, tenaga dan dana yang banyak. Data elektronik dan Web membuka ruang dan penyelidikan baru dalam pembinaan rangkaian sosial. Sifat semulajadi Web khususnya memberikan ruang pengaplikasian teknik pengekstrakan maklumat dalam membina rangkaian sosial. Walau bagaimanapun dokumen Web lazimnya tidak berstruktur apabila berhadapan dengan pemenuhan struktur yang kuat seperti rangkaian sosial, dan ini secara langsung mengundang kekompleksan dalam proses pengekstrakan, meskipun Web mengandungi tag atau metada sebagai petunjuk bagi kandungannya. Bab berikutnya membincangkan secara terperinci konsep dan teori pengekstrakan rangkaian sosial daripada sumber digital ini.