การแนะนำผู้ร่วมงานวิจัยด้านวิทยาการคอมพิวเตอร์โดยใช้ความสัมพันธ์ตามช่วงเวลาและความคล้ายคลึงกันของงานวิจัย

นางสาวปวีณา ชัยวนารมย์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์ ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554
ลิขสิทธิ์ของจุฬาลงกรณ์มหาวิทยาลัย

COMPUTER SCIENCE RESEARCH COLLABORATOR RECOMMENDATION BASED ON TEMPORAL RELATION AND PUBLICATION SIMILARITY

Miss Paweena Chaiwanarom

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy Program in Computer Science
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2011

Thesis Title	COMPUTER SCIENCE RESEARCH COLLABORATOR RECOMMENDATION BASED ON TEMPORAL RELATION AND PUBLICATION SIMILARITY
By	Miss Paweena Chaiwanarom
Field of Study	Computer Science
Thesis Advisor	Professor Chidchanok Lursinsap, Ph.D.

Accepted by the Faculty of Science, Chulalongkorn University in Partial Fulfillment of the Requirements for the Doctoral Degree

..... Dean of the Faculty of Science
(Professor Supot Hannongbua, Dr.rer.nat.)

THESIS COMMITTEE

..... Chairman
(Associate Professor Peraphon Sophatsathit, Ph.D.)

..... Thesis Advisor
(Professor Chidchanok Lursinsap, Ph.D.)

..... Examiner
(Suphakant Phimoltares, Ph.D.)

..... Examiner
(Professor Boonserm Kijsirikul, Ph.D.)

..... External Examiner
(Associate Professor Kosin Chamnongthai, Ph.D.)

ปวีณา ชัยวนารมย์ : การแนะนำผู้ร่วมงานวิจัยด้านวิทยาการคอมพิวเตอร์โดยใช้ความสัมพันธ์ตามช่วงเวลาและความคล้ายคลึงกันของงานวิจัย. (COMPUTER SCIENCE RESEARCH COLLABORATOR RECOMMENDATION BASED ON TEMPORAL RELATION AND PUBLICATION SIMILARITY) อ. ที่ปรึกษาวิทยานิพนธ์หลัก: ศาสตราจารย์ ดร. ชิดชนก เหลือสินทรัพย์, 69 หน้า.

การเลือกผู้ร่วมงานวิจัยเป็นขั้นตอนที่สำคัญในการผลิตงานวิจัยที่มีคุณภาพ หากแต่งานวิจัยที่มีจุดมุ่งหมายเพื่อแนะนำผู้ร่วมงานยังมีจำนวนไม่มากนัก วิทยานิพนธ์ฉบับนี้เสนอขั้นตอนวิธีการใหม่ซึ่งประกอบไปด้วยสองแนวทางหลักสำหรับแนะนำผู้ร่วมงานวิจัยที่เหมาะสมที่สุดสำหรับนักวิจัยที่กำลังมองหาผู้ร่วมงาน แนวทางแรกเป็นการวิเคราะห์เชิงปริมาณโดยนำเสนอตัวแบบความเหนียวแน่นเชิงโครงสร้างบนพื้นฐานของความร่วมมือตามช่วงเวลา (SC-CoT) โดยมุ่งเน้นที่จำนวนผลงานตีพิมพ์บนเครือข่ายผู้แต่งร่วม แนวทางที่สองเป็นการวิเคราะห์เชิงความหมายโดยนำเสนอตัวแบบความคล้ายคลึงกันในเชิงความหมายจากพื้นฐานและแนวโน้มในงานวิจัย (SS-BaT) ในแนวทางนี้ เนื้อความในผลงานตีพิมพ์ถูกนำมาใช้ในการประเมินองค์ความรู้ในงานวิจัยด้านต่างๆ ให้แก่นักวิจัยแต่ละคนเพื่อใช้ในการกำหนดความคล้ายคลึงกันของพื้นฐานงานวิจัยให้แก่นักวิจัยสองคน นอกจากนี้ตัวแบบยังทำการกำหนดแนวโน้มของหัวข้องานวิจัยได้โดยอัตโนมัติ ผู้ร่วมงานที่นำมาแนะนำจะถูกเลือกตามทฤษฎีหกช่วงคนเพื่อลดเวลาการค้นหา จากการทดสอบขั้นตอนวิธีดังกล่าวที่สร้างจากผลงานตีพิมพ์ของหกสาขาย่อยในช่วงระยะเวลาหกปีพบว่า ตัวแบบที่นำเสนอให้ผลที่ดีกว่าตัวแบบที่มีอยู่ นอกจากนี้ยังพบว่า ตัวแบบ SC-CoT และ SS-BaT ควรถูกนำมาใช้ร่วมกันเพื่อแนะนำนักวิจัย ผลลัพธ์จากการทดลองในหลายลักษณะทำให้เชื่อมั่นได้ว่าขั้นตอนวิธีที่นำเสนอสามารถแนะนำผู้ร่วมงานวิจัยที่เหมาะสมในสาขาใดสาขาหนึ่งโดยเฉพาะให้แก่นักวิจัยที่กำลังมองหาผู้ร่วมงานได้เป็นอย่างดี

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต .....

สาขาวิชา วิทยาการคอมพิวเตอร์ ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

ปีการศึกษา 2554 .....

4973884423: MAJOR COMPUTER SCIENCE

KEYWORDS: SOCIAL NETWORK / CO-AUTHORSHIP NETWORK / GRAPH MINING / DATA MINING / DIGITAL LIBRARY / KNOWLEDGE DISCOVERY / RESEARCH COLLABORATION

PAWEENA CHAIWANAROM: COMPUTER SCIENCE RESEARCH COLLABORATOR RECOMMENDATION BASED ON TEMPORAL RELATION AND PUBLICATION SIMILARITY. THESIS ADVISOR: PROF. CHIDCHANOK LURSINSAP, Ph.D., 69 pp.

A collaborator selection is an important process of entering a new quality research. Unfortunately, research collaborator recommendation is still an open research topic which has not yet been sufficiently studied. In this dissertation, a new methodology was proposed to recommend the most suitable collaborators for an inquired researcher. There are two main models which are the structure approach and the semantic approach of analysis. First, a quantitative measure called *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model is introduced. It concentrates on the quantity of publication underlining the structure of co-authorship network over time. Second, *Semantic Similarity with Background and Trend of Research* (SS-BaT) model based on semantic approach has been proposed to identify a semantic similarity between researcher pairs. The content of published papers are used to assign the research knowledge in various topics for individuals before determining the research background similarity between researcher pairs. Also, a current research trend of each researcher is automatically detected. For reducing the search time, the potential collaborators are retrieved within six degrees of separation. With six years of published papers over six topic areas, the results of proposed methodology is outperformed the results of existing methods. Both SC-CoT and SS-BaT models must be comprehensively combined for recommending collaborators to the inquired researchers. The results from the various experiments convince that the methodology can recommend the suitable collaborators for a given researcher in a particular topic domain.

Department: Mathematics and Computer Science **Student's Signature** .....

Field of Study: Computer Science ..... **Advisor's Signature** .....

Academic Year: 2011 .....

Acknowledgments

This dissertation would not have been possible without the guidance and the help of several individuals who in one way or another contributed and extended their valuable assistance in the preparation and completion of this study.

Foremost, I would like to express my sincere gratitude to my advisor Professor Chidchanok Lursinsap for the continuous support of my Ph.D. study and research. His guidance helped me in all the time of research and writing of this thesis. Besides my advisor, I would like to thank the rest of my thesis committee: Associate Professor Peraphon Sophatsathit, Professor Boonserm Kijsirikul, Associate Professor Kosin Chamnongthai, and Dr. Suphakant Phimoltares, for their valuable comments and good-natured support. I am grateful to Associate Professor Suchada Siripant for providing me with an excellent work environment in the Advanced Virtual and Intelligent Computing Center (AVIC) through all my study.

I owe my deepest gratitude to the Rajamangala University of Technology Rattanakosin for funding source. This fund gave me a chance to pursue Ph.D. program. I greatly wish to thank the supporting foundations for six months of the internship program of the National Institute of Informatics (NII), Japan. I would also like to express my appreciation to Associate Professor Ryutaro Ichise for providing the valuable materials, comments, and discussion in NII. To the all members in Ichise laboratory, thank you for sharing their knowledge and helping me during the completion of the research in Japan.

My special thanks go to Dr. Koonlachat Meesublak and Dr. Kalika Suksomboon for providing the necessary materials and reviewing my papers. I have been impressed by their knowledge and by the speed of service. I cannot forget to thank my colleaques in AVIC who have helped me in accomplishing this dissertation and all other academic activities. Also, I would like to extend my thanks to all friends for their kindness and cheerful support during my Ph.D. student life.

Finally, I would like to dedicate my Ph.D. to my family who have given me the opportunity of an education from the best institutions and support throughout my life.

Contents

Chapter Page

List of Figures

List of Tables

CHAPTER I

INTRODUCTION

1.1 Problem and Motivation

When a researcher would like to start a work in a new research topic, a problem usually encountered by the researcher is who he should collaborate with. The potential collaborators in this case include researchers having written papers together and new collaborators who have never joined work. Because almost academic search engines focus on document search rather than people search, researcher searching is still an open research topic. Although some online digital libraries can retrieve the researchers whom relate to a given researcher, only the researchers having written papers together can be retrieved regardless the unknown persons. Thus, the current services cannot discover the new collaborators who never collaborated together. Lack of meeting the unknown persons makes the inquired researchers miss the satisfied partners.

Collaborator recommendation is an objective of academic collaboration analysis. The other objectives are the analysis for link prediction (given in Section 2.2.2), research community discovering (given in Section 2.2.3), and expert finding (given in Section 2.2.4). The techniques proposed to overcome link prediction problem can be used for collaborator recommendation problem in case of suggesting the new collaborators who never joined work. Unfortunately, such techniques cannot be used to calculate any relation weights between researcher pair having collaborated together. Research community discovering is used for clustering researchers to research communities. The researchers will be clustered based on their profiles and the contents of published papers. The output identifies who stays in the same communities. Nevertheless, the research community discovering analysis concentrates on the relation in community level rather than pairwise relation between two researchers. Moreover, no time evolution is concerned in the analysis. Thus, the technique proposed to discover communities may not be suitable to directly apply for recommendation problem which is sensitive to up-to-date data. Lastly, expert finding is one of applications to study the community discovering analysis. The output of expert finding is a set of expertises in a given research area compared with researchers in the same community. However, this is also the communities level analysis, not pairwise relation analysis. Since the appropriated collaborators for an inquired researcher need not the

expertises, the technique used in expert finding may not be suitable for collaborator recommendation problem.

In the viewpoint of inquired researchers, the summary properties of the potential researchers who dramatically expected to be the colleagues can be attributed to the following factors.

- *Productivity*. Researchers with more number of publications are likely to be selected.
- *Friendship*. Researchers with a shorter distance of friendship are more likely to be recognized than researchers with a further distance. For instance, we always meet a friend-of-friend before a friend-of-friend-of-friend.
- *Research background*. Senior researchers with the same research experiences are more likely to be selected.
- *Research trend*. Researchers who interest the same research topic are more likely to be selected.

According to the four requirements, a set of famous researchers in the given research topic seems to be the best choice. Unfortunately, relatively few researchers can reach that set. To increase the opportunity to meet the satisfied collaborators, the definition of appropriate collaborators have to be defined. In this dissertation, the appropriated researchers for a given researcher are the researchers who have the most research knowledge similarity and reach ability in a given topic. Research knowledge means research experiences in preceding years and currently interested research topics. According to the definition, recommended researchers are not necessary the famous or familiar researchers. Thus, the challenges in recommender prospective correspond to the requirements are as follows.

- *How to comprehensively analyze the four factors over time?* Based on the defined appropriate research, researchers with updated information of published papers along with year of publication have to be considered.
- *How many candidate researchers will be considered to recommend?* There is a trade-off between the number of candidates and searching time. Although the large number of candidates increase an opportunity to meet the satisfy collaborators, it also increases the searching time as well.

To overcome the both problems, the collaborators recommendation problem is formulated from the social network analysis and research semantic analysis between researchers. A method for visu-

alizing the relationship among the researchers from all published papers has to be designed. Subsequently, the significant features used to account the cohesion between a pair of researchers will be considered. Moreover, a method uses for indicating the semantic similarity between a pair of researcher utilized from the content of papers has to be proposed. The time evolution is also attended in the proposed solutions for taking into account to the up-to-date data. Consequently, all features will be dynamically considered based on the evolution over time.

1.2 Objectives

This dissertation aims at two objectives as follows.

1. Discovering a current research topic of researchers.
2. Proposing models used for recommending the appropriate collaborators for a given researcher in Computer Science research area.

1.3 Scope and Limitations

1. All data set will be obtained from SCOPUS which is the worlds largest database of abstract and citation of peer-reviewed literature and quality web sources. This dissertation considers to Computer Science researches in six topics, including Bio-informatics, Data Mining, Hardware, Neural Networks, Software Engineering, and Algorithm and Theory.
2. The new collaborators recommendation model will be proposed to investigate, analyze, and suggest the appropriate collaborators for an inquired researcher and an inquired research topic. The potential collaborators will be retrieved within six degrees of separation from an inquired researcher.

1.4 Contributions

Main contributions of this dissertation are as follows.

1. This dissertation proposed a model based on structure approach called *Structure Cohesion based on Collaboration Over Time* (`SC-CoT`). The two factors, i.e. productivity and friendship are used for determining the cohesion between researcher pairs. The `SC-CoT` carefully

analyzed the both factors over time. The model comprises of three studies: 1) power of researcher based on individual papers; 2) closeness between two researchers based on common papers; 3) mutual relation between two researchers based on common friends.

2. This dissertation proposed a model based on semantic approach called *Semantic Similarity with Background and Trend of Research* (SS-BaT). It is extended based on the existing unsupervised learning technique for clustering research topics from large documents called author-topic model (ATM) [1]. The proposed model uses the extracted words from various parts of papers, i.e. title, abstract, authors keyword, index keywords, venue, and references. The two factors, i.e. research background and research trend are combined to calculate the semantic similarity between two researchers. The output of research background analysis provides the similarity of research experience between two researchers. In order to measure the trend of research, the evolution of research interesting of individual are studied and the probability of the given topic in the near future are estimated. The potential researchers with the most background similarity and the most estimated probability in the same current topic should be assigned the highest priority to selected.

3. The both proposed models, namely SC-CoT and SS-BaT, are used to calculate the relevance scores between two researchers. These scores are used to rank the potential partners for collaborators recommendation task. The results are evaluated in various aspects.

1.5 Methodology

1. Reviewing and study the research papers related to the social network analysis, academic collaboration analysis, link prediction, research community discovering, expert finding, and information retrieval.

2. Preparing data sets by collecting the real-word bibliography data of papers from SCOPUS.

3. Constructing co-authorship networks based on the collected papers.

4. Retrieving a set of potential collaborators within six degrees of separation.

5. Developing *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model.

6. Developing *Semantic Similarity with Background and Trend of Research* (SS-BaT) model.

7. Using the outputs of SS-BaT model and SS-BaT model to calculate the relevance scores.

8. Ranking the potential collaborators based on the relevance scores and apply to recommendation task.

9. Evaluate the experimental results in the following various key aspects:

   (a) The best snapshot of time use for recommending.

   (b) The different results between structure approach and semantic approach in the two set of candidate researchers, namely researchers having collaborated and researchers not having collaborated.

   (c) The accuracy of semantic approach based on `SS-BaT` model.

   (d) The different results compared with other methods.

   (e) The accuracy of different inquired researchers, i.e. senior researchers who have more research experience, and junior researchers who are young researchers with a few experiences.

   (f) The accuracy of the particular topics within six degrees of separation.

1.6 Dissertation Organization

This dissertation is organized as follows. Chapter II provides necessary theoretical background and the related works. Chapter III describes the proposed methodology. Chapter IV describes the data sets preparation, experiment setting, evaluation method, and experimental results. Chapter V concludes the main findings in this dissertation and discusses about the future work.

CHAPTER II

BACKGROUND AND RELATED WORKS

2.1 Background

In this section, the social network concept and the theoretical background of techniques applied in this dissertation are reviewed.

2.1.1 Social Network

This section gives the valuable terms usually used in the social network area. Not only the concept of social network analysis but also the related networks and the important properties are explained.

2.1.1.1 Social Network Analysis

The social entities are called actors which may be discrete individual, corporate, or social units group. Actors are connected to one another by social ties. A social network consists of a group of a finite set of actors and connections among them [2]. The appealing of social network analysis are studying the relationships among social entities, and on the pattern and implications of these relationships. Network analysis assumes that the actors in a network can be connected to each other by using some important features of that network. The concepts of graph theory has been extended by researchers into social network analysis for the several decades. A social network can be represented by a graph based on graph theory in mathematical viewpoint where actors represented by nodes and ties represented by edges [3]. The results in graph structures are often very complex. There can be many kinds of edges between the node pairs. Social network analysis concentrates on the relationships between actors rather than the individual properties [4]. In recently, social network analysis has utilized as a key technique in modern sociology. It used as a tool for studying the structure of the interested social groups and communities. It is also applied in biology, anthropology, geography, communication studies, economics, organizational studies, and social psychology. For information science domain, social network analysis has been used in the field of graph mining [5] such as detection of abnormal subgraphs/edges/nodes, graph compression, web mining, link prediction, etc.

2.1.1.2 Small World Networks

Social networks have the small world property. The idea of a small world network was created in the 1960's by the American psychologist, Stanley Milgram [6]. The hypothesis of small world phenomenon is that a path which connects a person to another is generally short. A graph $G = (V, E)$ become to be a small world network if it is formulated from the following two conditions, i.e., local property, and global property. In local property, it has a high clustering coefficients. The clustering coefficient is a property of a node in a network. If the neighborhood is fully connected, the clustering coefficient is 1. But if value is close to 0, it means that the connections in the neighborhood are rarely appeared. For global property, it has a low average distance between nodes. The concept led to the famous phenomenon called *six degrees of separation* after Stanley Milgram tested the hypothesis in year 1967. In Milgram's experiment [6], a sample of individuals were asked to reach another target person by passing a message along a chain of acquaintances. The average length of achieved chains turned out to be about five intermediaries or six separation steps. This constituted a term of *six-degrees of separation* phenomenon.

Academic researchers have continued to examine this phenomenon as internet-based communication technology. A recent experiment at Columbia University found that about five to seven degrees of separation are sufficient for connecting any two people via e-mail [7]. The Internet does too, if we treat web pages as nodes and hypertext links as edges. Nowadays, the small world question is still a popular research topic, and many experiments are still being conducted.

2.1.1.3 Scale-Free Networks

Small world networks usually follow a *power law distribution* [8]. It means that the fraction of nodes with degree d (the number of links a node has) is proportional to $\frac{1}{d^c}$ where c is a constant. In other words, a large number of nodes have the average degree and very few have either very high or very low degrees. Such networks are said to be *scale-free* or power-law graphs. One important characteristic of scale-free networks is the *clustering coefficient distribution*. Since this distribution also follows a power law, we can say that scale-free network is a network whose degree distribution follows a power law. In scale-free networks, the path for reaching to a very high degree node is usually short, and the highest-connected nodes can quickly distribute the query to all nodes in the network. The highest-degree nodes are often called *hubs* which serve the specific purposes in their networks. Thus, the power law distribution highly influences in the network structure. The discovery of the power law distribution required the development of new modeling paradigms. A much used

assumption is that in scale-free networks a new link is likely to be connected to an existing node of higher degrees. This is a phenomenon labeled as *preferential attachment*.

Consider a social network in which nodes are people and links are acquaintance relationships between such people. It is easy to see that people tend to form communities, i.e., small groups in which everyone knows each other. Moreover, the people of a community have a few acquaintance relationships to people in other communities. However, a few are related to other people that they are linked to a large number of communities. Those people may be said as hubs responsible for the small-world phenomenon

2.1.1.4 Co-Authorship Networks

Co-authorship networks are a kind of social networks which are used for representing a network of researchers. An investigation of Nascimento et al. [9], Smeaton et al. [10], Liu et al. [11] and Huang et al., [12] showed their co-authorship network are a growing small word network. The networks have been used to determine the status of individual researchers and the structure of scientific collaborations. Similarly to the social networks, a co-authorship network can be represented by a graph $G = (V, E)$, where V is a set of researchers, and E is a set of joint publications between researcher pairs. In the academic communities, there are several types of collaboration networks, e.g. affiliation network, citation network, etc. The co-authorship network is one of the most important types of connections between academic networks. Several researches have been shown the co-authorship networks analysis in particular research communities for understanding the collaboration characteristics of the communities [13, 14]. In recent decades, the growing of scientific collaboration has been interested by a large number of researches. Studies into co-authorship have been focused on two different approaches. The first study concerns the analysis of reasons why authors collaborate and the consequences of such decision. The second approach is studying the structure of co-authorship based on a social network analysis [13, 15, 16].

2.1.2 Author-Topic Model

Author-topic model (ATM) is an unsupervised learning technique proposed by Steyvers et al., [1, 17–19]. The model comprehensively analyzed the relationship among authors, topics, and words. The authors classified to a likely topic by the model are not necessarily the expertise in that topic, but they are the authors who tend to generate the most words for such likely topic. The author-topic models can be applied in a variety of works which concentrate on mining the set of authors and

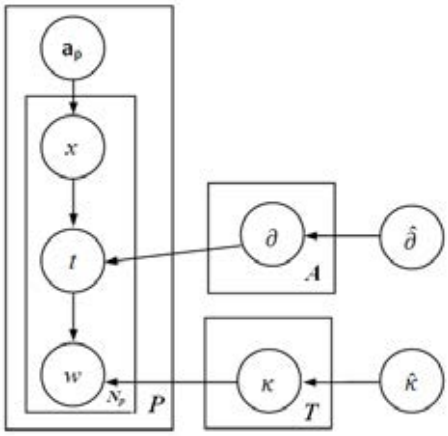


Figure 2.1 The generative process of the author-topic model with a graphical model.

words including, studying of topic trends over time, finding the authors who tend to write a new paper in a given topic, and detecting the interested papers for an author.

Author-topic model uses a probabilistic model representing each topic as a probability distributions over words and expressing each author as a probability distributions over topics. The model discovers not only what topics are represented in a document (topic-word), but also which authors are distributed to each topic (author-topic). The topic-word and author-topic distributions are learned from a set of documents, authors, and words based on an unsupervised model using a Markov chain Monte Carlo algorithm. Figure 2.1 illustrates the generative process of author-topic model with a graphical model. Nodes represent variables, a directed edge represents a conditional dependency between variables, and a box represents a repeated sampling process whose number of repetitions is given by the variable at the bottom of the box, i.e., A, T, P, N_p. The variables $\hat{\partial}$ and $\hat{\kappa}$ are the perimeters to decide probabilistic distribution. For learning, the model formulates a document as a bag of words by representing each document as a vector of word count. The frequency of a word appearing in the document is an element in a vector. Each author is associated with a multinomial distribution over topics. A document written by multiple authors is a mixture of the distributions corresponds to such authors. An author (from a set of A authors) gets a topic from his multinomial distribution over topics, denoted by ∂, when generating a document. Then, the author samples a word from the multinomial distribution over words corresponded to the given topic. All words appear in the document is repeated for this process. For each word in the document, an author x is uniformly sampled from a_p. Then, a topic t is sampled from the multinomial distribution ∂ corresponding to author x and a word w is sampled from a multinomial topic distribution κ corresponding to topic t. This sampling process is repeated N times.

There are two parameters to be estimated in author-topic model, i.e., the A author-topic distribution ∂, and the T topic distribution κ. The model used Gibbs sampling, a Markov chain Monte Carlo algorithm to sample from the posterior distribution over parameters. The posterior distribution on just x and t are evaluated and then use the results to infer ∂ and κ. Suppose P documents contains T topics expressed over W unique words. For each word, the topic and author assignment are sampled as follows:

$$P(t_i = j, x_i = a | w_i = m, \acute{t}, \acute{x}) \hat{\kappa} \frac{n_{-i,j}^{w_i} + \hat{\kappa}}{n_{-i,j} + W\hat{\kappa}} \frac{n_{-i,j}^{a_i} + \hat{\partial}}{n_{-i,j} + T\hat{\partial}} \,. \tag{2.1}$$

where $t_i = j$ and $x_i = a$ represent the assignments of the i^{th} word in a document to topic j and author a respectively, $w_i = m$ represents the observation that the i^{th} word is the m^{th} word in the lexicon, and $\acute{t}, \acute{x}$ represent all topic and author assignments not including the i^{th} word. Term $n_{-i,j}^{w_i}$ is the number of times word m is assigned to topic j, not include the current instance. Term $n_{-i,j}^{a_i}$ is the number of times author a is assigned to topic j, not include the current instance. Term $n_{-i,j}$ is a count that does not include the current assignment of t_i. Furthermore, the variables $\hat{\partial}$ and $\hat{\kappa}$ are the perimeters to decide probabilistic distribution as shown in Figure 2.1.

The algorithm is started by assigning words to random authors and topics. Each Gibbs sample takes equation (2.1) to all words in the set of documents. This sampling process is repeated for a defined iterations. Moreover, the estimate the topic-word distributions κ and author-topic distributions ∂ can be estimated by:

$$\kappa_{m,j} = \frac{n_{-i,j}^{w_i} + \hat{\kappa}}{n_{-i,j} + W\hat{\kappa}} \,. \tag{2.2}$$

$$\partial_{a,j} = \frac{n_{-i,j}^{a_i} + \hat{\partial}}{n_{-i,j} + T\hat{\partial}} \,. \tag{2.3}$$

2.1.3 Moving Average

Moving average is one of the most popular technique which is applied in various applications. It is commonly used to understand the stock price behavior and trend prediction for several decades [20]. Therefore, the moving average is usually taken by averaging the prices over a period of time producing a smoother line. There are different types of moving average and, therefore, different formulas to

calculate it. Two of the most common types of moving averages are the *simple moving average (SMA)* and the *exponential moving average (EMA)*. They are described in more detail below.

Simple moving average is the simplest form of a moving average which is calculated by taking the arithmetic mean of a given set of values. The mean average is computed by summing up a set of data over a specified period and dividing the summation by the number of such periods. A moving average moves because the newest period is added and the oldest period is dropped. This type of moving average is the most commonly used. Nevertheless, the results are lagging indicators because they are calculated based on data in a preceding period. They can only indicate a trend that is already in place. Thus, simple moving average method is not suitable for the shorter time-series data which more sensitive to the recent time periods. Let $\mathbf{X} = \{x_1, x_2, \ldots, x_n\}$ be an order set of n time-series data. Let $\mathbf{Y} = \{y_t, y_{t-1}, \ldots, y_{t-\hat{n}+1}\}$ be an order set of the last $\hat{n}$ elements over a specified time period t, where $\mathbf{Y} \in \mathbf{X}$ and $\hat{n} < n$. The simple moving average of $\hat{n}$ elements in a specified time period t denote by SMA_t can be computed by equation (2.4).

$$SMA_t = \frac{1}{\hat{n}} \sum_{y_i \in Y} y_i \ . \tag{2.4}$$

Exponential moving average was first suggested by Charles C. Holt [21]. It is considerably more complicated to reduce the lag in the simple moving average by applying more weight to the more recent data relative to older data. The weighting applied to the most recent data depends on the specified period of the moving average. The shorter the period is the more weight that will be applied to the most recent data. As a result, the exponential moving average will be more sensitive and will follow the up-to-date data closer than the simple moving average. Based on the above defined sets, exponential moving average of $\hat{n}$ elements in a specified time period t denote by EMA_t can be computed by equation (2.5).

$$EMA_t = \left[(y_t - EMA_{t-1}) \times \frac{2}{(\hat{n}+1)} \right] + EMA_{t-1} \ . \tag{2.5}$$

When using the equation (2.5) to calculate the first of result the exponential moving average, EMA_0, there is no value available to use as the previous exponential moving average. This small problem can be solved by starting the calculation with a simple moving average in equation (2.4) and continuing on with the equation (2.5) from there.

2.2 Related Works

This section reviews the researches in the academic collaboration analysis. These researches are grouped by the objective of works, namely link prediction, research community discovering, and expert finding. Lastly, the difference between such objectives and collaborator recommendation will be compared.

2.2.1 Academic Collaboration Analysis

In recent surveys, there are many papers studied in the academic collaboration [9,13,15,22–28]. The studies concentrated on analysis characterizing the pattern of network evolution using various network measurements. Such measurements are enable to understand the growth pattern based on the small world phenomenon. The examples of measures (metrics) in social network analysis are betweeness, bridge, centrality, closeness, clustering coefficient, cohesion, degree, path length, etc. These measures studied the general sense of collaboration and did not preserve historical research collaboration in particular researcher pairs. A study of Han et al. [29] concentrated on the closeness and distance between the researcher pairs. They proposed the supportiveness measurement in co-authorship network structure. The supportiveness from a researcher to another is used to measure how close the collaboration between them. However, this work used only the frequency of co-authorship for computing, regardless of time evaluation.

2.2.2 Link Prediction

Link prediction problem is predicting changes to a social network. The classical paper of link prediction [30] said that link prediction is seeking the accurately predict the edges that will be added to the future network time t' based on the structure of preceding network at time t. Consider a co-authorship network among researchers, there are many reasons why two researchers not having written papers together will do so in the next few years. Effective methods for link prediction could be used to analyze such a given co-authorship network and suggest the new collaborators with the most potential collaborators from a set of existing researchers and relations. All methods (predictors) assigned a connection weight to a researcher pair and then the weights are ranked in decreasing order. The summary of methods for link prediction summarized in [30] can be classified into three groups, namely basic graph-distance method, methods based upon node neighborhoods, and methods based upon the ensemble of all paths. Unfortunately, all methods concentrated on static network.

A simple software agents to observe the model social networks overtime was proposed in [31, 32]. An important contribution of these papers is that temporal metrics are an valuable new attention to link prediction, and should be used in future research and applications. Some further research to be performed include suggesting new temporal variations of static metrics and determining exactly the optimum number of time steps for computing temporal metrics. The experimental results of applying Bayesian analysis to link prediction problem were summarized in [32]. The latest link prediction approach was proposed in [33] which developed Supervised Random Walks, a new learning algorithms for link prediction and link recommendation. Their experiments on the Facebook social graph and large collaboration networks showed that the approach outperforms state-of-the-art unsupervised approaches as well as approaches that were based on feature extraction.

In summary, methods in link prediction problem can be applied for collaborator recommendation problem in case of finding the new researchers who have never written paper together. The most related work are Sachan's work [34]. They proposed a supervised learning method for building link predictors from co-authorship network structure and using some semantic attributes of the nodes like title and abstract information. Although the objective of the work is to overcome link prediction problem, the mining approach can be applied to finding the future collaborators. Nevertheless, time evolution and the citation of papers were not considered in [34].

2.2.3 Research Community Discovering

Community discovering has been a hot research in academic social network. A research community can be defined as a group of researchers that share similar attributes or connect to each other via certain interest. The objective of the researches is clustering researchers to research communities. The output of this research enables to know who stays in the same communities. The raw data usually are used for mining the communities are researcher's profiles, paper title, abstract, venue, citations. These data can be grouped into two mining approaches, i.e., profile-based approach, and document-based approach.

Su et al.'s work [35] is an example work in document-based approach. The work used author co-citation analysis method to identify researchers in related fields. The method can discover researchers with multiple communities. Productivity and author's names appear in the top of papers and reference of papers were used for mining. Other parts of papers, i.e, title, abstract, venue were not considered in the study. Zaïane et al. [36] proposed a method to discovering communities in document-based approach. They generated bipartite (author-conference) and tripartite (author-conference-topics) graph models, where topics are frequency of words extracted from paper title and

abstracts. The co-authorship relations were taken into account while designing the model. Nevertheless, the authors of this said that their model was still preliminary. ArnetMiner system [37, 38] is a research used both profile-based and document-based approach. It focuses on extracting researcher profiles from personal homepage. It first collects and identifies a set of homepage from the web, then uses a unified approach to extract the profile properties from such pages. It extracts publications from online digital libraries using rules. Then, it integrates the extracted profiles of researchers and the extracted publications by using the given researcher name.

In summary, all of the above researches only focused on the relation in community level rather than pairwise relation between researcher pairs. In addition, these researches considered in a snapshot of time, regardless dynamic evolution.

2.2.4 Expert Finding

Expert finding is one of applications extended study from community discovering analysis. The task of expert finding is identifying researchers with relevant expertise or experience for a given topic. The study of [39] focused on developing models for searching an organizations document repositories for experts on a given topic using generative probabilistic models. Tang et al. [40] presented a topic level expertise search framework for heterogeneous networks. Their proposed model was able to retrieve the related researchers based on several types of queries i.e., inquired researcher name, conference venue, topic. Zhang et al. developed a paper [41] extended from community discovering in [37, 38]. They used the extracted personal information of researchers to calculate an initial expert score for individuals. In additional, the relationships between researchers were involved to improve the accuracy of expert finding. The assumption is if a researcher knows many experts on a topic or if his name frequently co-occurs with another expert, then it is likely that he is an expert on that topic. Recently, Daud et al. [42] proposed Temporal-Expert-Topic (TET) model compared with Author-Conference-Topic (ACT) [37] which is non-generalized time topic modeling. They said that ACT used the conferences information without considering conferences influence. They proposed the generalized time topic modeling approach, TET, which could provide ranking of experts in different groups in an unsupervised way. The experimental results showed that the proposed generalized time topic modeling approach with outperformed the non-generalized time topic modeling approaches, due to simultaneously capturing conferences influence with time information.

In research collaborators recommendation task, the appropriated collaborators for a given researcher need not be the expertise. It is quite possible that the research backgrounds of two researchers might be similar although the they were classified in the different topics. Therefore, the expertise may

not be the most appropriated collaborators for a given researcher.

CHAPTER III

PROPOSED METHODOLOGY

The proposed methodology mainly consists of five parts, namely constructing co-authorship network (given in Section 3.1), retrieving the potential collaborators (given in Section 3.2), calculating cohesion weight based on *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model (given in Section 3.3), calculating semantic similarity based on *Semantic Similarity with Background and Trend of Research* (SS-BaT) model (given in Section 3.4), and calculating the relevance scores for ranking the potential collaborators (given in Section 3.5). The downloaded data set from SCOPUS[1] offers various attributes of papers. A set of valuable attributes used for both models are shown in Table 3.1. The last six attributes used to get words for discovering the topic domain of researchers in the semantic approach. Note that this dissertation use the terms node, author, and researcher interchangeably. The terms edge, link, and co-authorship are also used interchangeably. Additionally, subscripts of all weights in here can be represented into two types of relationship, i.e., undirected and directed relationship. Based on undirected relationship, weights which subscripted by (v_i, v_k) and (v_k, v_i) are the same meaning. On the other hand, subscript of weights denoted by $(v_i \rightarrow v_k)$ and $(v_k \rightarrow v_i)$ are not the same since it is used to represent directed relationship in the different view points.

3.1 Constructing Co-authorship Network

A collection of papers downloaded from SCOPUS are grouped by year of publication in two data sets, namely DB-Data set used for building the models and Test-Data set used for testing. Firstly, DB-Data set will be used to construct a co-authorship network in this section. According to the social network analysis concept, the network is constructed in this dissertation for two objectives, i.e., retrieving the potential collaborators within six degree of separation, and preserving the calculated statistical relationship changed over the steps of developing *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model.

For the first generation of co-authorship network, the network is represented by undirected multi-graph, whose nodes represent unique researchers and edges represent the published papers be-

[1] http://www.scopus.com

Table 3.1 Attributes of papers used in SC-CoT model and SS-BaT model.

Attribute Name	SC-CoT Model (Structure Approach)	SS-BaT Model (Semantic Approach)
Paper ID	√	√
Year	√	√
Authors	√	√
Title		√
Abstract		√
Authors Keywords		√
Index Keywords		√
Venue		√
References		√

Table 3.2 An example of published papers.

Paper ID	Year	Authors
p_1	2006	v_1, v_2, v_3
p_2	2007	v_1, v_2
p_3	2009	v_3, v_4

tween two researchers. The number of edges between two nodes represents the number of papers they co-authorship. Thus, the existing edges represent previous collaboration between researcher pairs. For example, Table 3.2 shows a set of collected papers represented by three attributes associated with the three attributes of SC-CoT model in Table 3.1. In the first row, paper p_1 was written by three authors, i.e., v_1, v_2, v_3, and published in year 2006. The associated co-authorship network $\mathbf{G} = (\mathbf{V},\mathbf{E},\mathbf{L})$ in Figure 3.1 shows the relationship among four authors in Table 3.2. Let $\mathbf{V} = \{v_1, v_2, \ldots, v_n\}$ be a set of n unique researchers in a co-authorship network. $\mathbf{E}$ is a set of edges which are the published papers written by the researcher pairs, v_i and v_k. $\mathbf{L}$ is the set of paper attributes associated with the edges connecting the researcher pairs. The labels on each edge (v_i, v_k) consist of (1) Paper ID, (2) the number of authors in the paper, and (3) year of publication. The frequency of co-authorship can be represented by the number of edges between node pairs, for instance, frequency of co-authorship between $(v_1, v_2) = 2$, and frequency of co-authorship between $(v_2, v_3) = 1$.

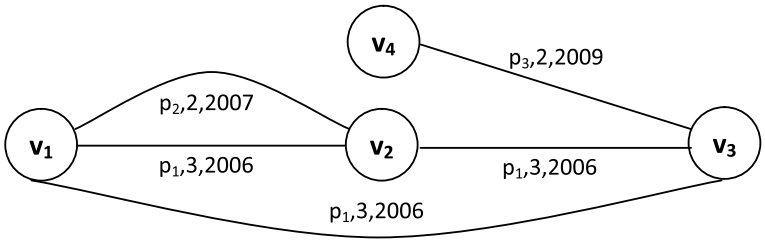


Figure 3.1 Co-authorship network of papers in Table 3.2 represented by undirected multigraph.

Table 3.3 Degree of separation among four researchers in Figure 3.1.

Researcher, (v_i, v_k)	v_1, v_2	v_1, v_3	v_1, v_4	v_2, v_3	v_2, v_4	v_3, v_4
Degree of separation, Δ_{v_i,v_k}	1	1	2	1	2	1

3.2 Retrieving the Potential Collaborators

Due to this dissertation limits to recommend the potential collaborators within six degree of separation from an inquired researcher, this section explains who is inquired researcher and who are his potential collaborators. According to the collected papers kept in two sets, i.e., DB-Data set and Test-Data set, the unique researchers are independently extracted from each set. Next, the common researchers who appeared in both sets are listed and called *core authors*.

Definition 1: Core authors. Core authors is a set of authors denoted as $\mathbf{A} = \{v_i \in \mathbf{V} | v_i$ is a researcher who appeared in both DB-Data set and Test-Data set$\}$. These are a set of inquired researchers who would like to find their collaborators.

The distance between two nodes in a graph is the number of edges in a shortest path connecting them. In this dissertation, the distance is used for representing degree of separation between arbitrary researcher pairs. The symbol Δ_{v_i,v_k} represents the degree of separation between a core author v_i and his potential collaborator v_k. Note that the relation between v_i and v_k is undirected relationship. Thus, Δ_{v_i,v_k} and Δ_{v_k,v_i} are always equal. For example in Figure 3.1, there are two paths between (v_1, v_3) that are $< v_1 - v_3 >$ and $< v_1 - v_2 - v_3 >$. The degree of separation between (v_1, v_3) is one, $\Delta_{v_1,v_3} = 1$, as shown in Table 3.3 because the shortest path between them comprises of one edge.

Selecting potential collaborators from all researchers may waste time in that they might not be met in the real world. The suitable degree of separation has to be defined for reducing search time and increase the number of potential collaborators. Fortunately, a study of Franceschet [27]

Table 3.4: Set of neighbors, friend researchers, and non-friend researchers of four researchers in Figure 3.1.

Core Author, v_i	Neighbors, $R(v_i)$	Friends, $Q(v_i)$	Non-Friends, $\overline{Q(v_i)}$
v_1	$\{v_2, v_3, v_4\}$	$\{v_2, v_3\}$	$\{v_4\}$
v_2	$\{v_1, v_3, v_4\}$	$\{v_1, v_3\}$	$\{v_4\}$
v_3	$\{v_1, v_2, v_4\}$	$\{v_1, v_2, v_4\}$	$\emptyset$
v_4	$\{v_1, v_2, v_3\}$	$\{v_3\}$	$\{v_1, v_2\}$

found the average distance of computer scientists is 6.41 which corresponds to the six degrees of separation in a popular small-world experiment [43]. Thus, this dissertation limits to recommend the potential collaborators within six degrees of separation. Starting from a core author v_i, the depth-first search (DFS) algorithm [44] is applied for traversing the co-authorship network and retrieving a set of nodes within six degrees of separation from v_i. In general, the potential collaborators are possibly appeared in three categories of friendship, i.e., friend researchers, non-friend researchers, and unknown researchers.

Definition 2: Neighbors. Neighbors of v_i is a set of authors with respect to author v_i denoted as $R(v_i) = \{v_k \in \mathbf{V} | v_k$ is a researchers who appear within six degrees of separation from $v_i\}$. $R(v_i)$ are the potential collaborators for core author v_i.

Definition 3: Friend researchers. Friend researchers is a set of authors with respect to author v_i denoted as $Q(v_i) = \{v_\partial \in R(v_i) | v_\partial$ is a researcher having collaborated with $v_i\}$. Thus, $Q(v_i) \subseteq R(v_i)$. In other words, these are the adjacent nodes which located in one degree of separation from core author v_i.

Definition 4: Non-friend researchers. Non-friend researchers is a set of authors with respect to author v_i denoted as $\overline{Q(v_i)} = \{v_\lambda \in R(v_i) | v_\lambda$ is a researcher not having collaborated with $v_i\}$. Thus, $\overline{Q(v_i)} = R(v_i) - Q(v_i)$. These appear between two and six degrees of separation from core author v_i.

Definition 5: Unknown researchers. Unknown researchers is a set of researchers who are not in $\mathbf{V}$. In other words, they are new researchers just first appearing in the Test-Data set.

Suppose all researchers in Figure 3.1 are core authors. The set of three categories of friendship

are shown in Table 3.4. Based on two assumptions measured according to the degree of separation, we have (a) the probability that non-friend pairs will meet each other in higher degree of separation is small; (b) the further distance between two researchers implies the less potential collaboration. Hence, a parameter called *distance parameter*, $d(v_i, v_k)$, is introduced and it is based on the concept of imitating process of discharging a capacitor. It is calculated by using the condition in equation (3.1). The parameter used for adjusting pairwise weights between v_i and v_k associated to the degree of separation.

$$d(v_i, v_k) = \begin{cases} 1.0, & \text{friends} \\ \frac{1.0}{2^{\Delta}}, & \text{non-friends.} \end{cases} \tag{3.1}$$

3.3 Structure Cohesion based on Collaboration Over Time (SC-CoT) Model

3.3.1 Overview of SC-CoT Model

To analyze the relationship between two researchers in co-authorship network, a new model called *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model is proposed. The model is a quantitative measure which concentrates on the quantity of publication underlining co-authorship network over time. Intuitively, productivity of papers of researchers over time and valuable information in friendship involved would suffice for measuring the correlation and cohesion between researcher pairs. Therefore, this model pays attention to how many papers, time of publication, and co-authors, regardless of what papers were written about.

The overall process of SC-CoT model is shown in Figure 3.2. The model comprised of three measurements associated with three analysis, namely power of researcher analysis (given in Section 3.3.2), common papers analysis (given in Section 3.3.3), and common friends analysis (given in Section 3.3.4). Outputs associated with the three measurements, i.e., power weights, closeness weights, and mutual weights, will be combined in the structure cohesion analysis process (given in Section 3.3.5). Finally, the output of the last process becomes the output of SC-CoT model called cohesion weight, $\mathcal{N}_{v_i \to v_k}$. Therefore, the $\mathcal{N}_{v_i \to v_k}$ is the cohesion weight of v_k with respect to v_i compared with all neighbors of v_i.

3.3.2 Power of Researchers Analysis

In this section, a weight for individual researcher called power weight is proposed for measuring the power of researcher. Researchers with more power weight is the researchers who has more

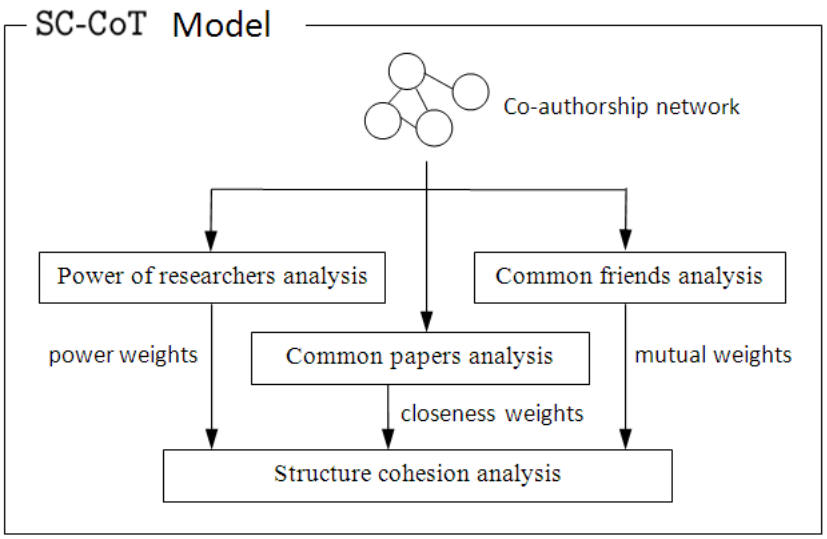
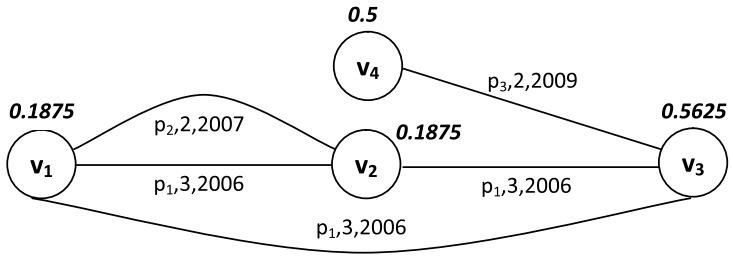


Figure 3.2 The overall process of SC-CoT model. Output is a set of cohesion weights $\mathcal{N}_{v_i \to v_k}$.

possibility to publish a new paper in the near future. The assumption is researchers who have a large number of recently published papers should have more power to connect with a new friend via a new paper. One reason why the recommending results for researcher pair are not symmetric is due to the unbalanced power of researchers. For example, v_i is an expertise having more power in data mining with a large number of recent papers and friends. He has been recommended as a collaborator to a lot of researchers, but a few of them are recommended as collaborators to him.

Since co-authorship network is an example of scale-free networks, it also has preferential attachment [45] property in the co-authorship network constructed in Section 3.1. This property is generally used to apply in various researches [46], [47]. The property defines most nodes are connected with a small number of friends whereas a few nodes are connected with a large number of friends, regardless of the number of papers. Thus, a new link is likely to be connected to an existing node with a large number of friends. Because the classical idea of preferential attachment is not concerned about the year of publication, the age of friendship is not captured. To overcome the recommendation problem which is more sensitive to up-to-date data, using only the number of friends without publication year may be rough for considering the power of researcher over time. The idea in this dissertation is researchers who published papers in the different years may not have the same power even though the number of friends is equal. Furthermore, if two researchers have the same number of friends, the researcher who has more recently published papers should has more power to publish a new one. Therefore, this dissertation adapts to use the number of papers for calculating the power researcher stead of directly use the number of friends like in the preferential attachment



p_e	Year	φ	$1/2^{\varphi}$
paper: p_1	2006	4	0.0625
paper: p_2	2007	3	0.125
paper: p_3	2009	1	0.5

$$\mathcal{P}_{v_1}: \quad a_{v_1,p_1} + a_{v_1,p_2} = \quad 0.0625 + 0.125 = \quad 0.1875$$
$$\mathcal{P}_{v_2}: \quad a_{v_2,p_1} + a_{v_2,p_2} = \quad 0.0625 + 0.125 = \quad 0.1875$$
$$\mathcal{P}_{v_3}: \quad a_{v_3,p_1} + a_{v_3,p_3} = \quad 0.0625 + 0.5 = \quad 0.5625$$
$$\mathcal{P}_{v_4}: \quad a_{v_4,p_3} = \quad 0.5 = \quad 0.5$$

Figure 3.3: Co-authorship network in Figure 3.1 extended with labeling nodes by power weights of researchers, $\mathcal{P}_{v_j}$, assume $\beta = 2011$.

property.

Firstly, *Spearman's rank correlation coefficient* is used to prove the correlation between the number of papers and the number of friends. The researchers are ranked with two approaches, namely ranking ordered by the number of papers and ranking ordered by the number of friends. After the ordered sets of researchers from the both approaches are compared, the result of correlation is 0.84. Then, the number of papers and the number of friends are perfectly monotonically related. The number of friend proportionally grows to the number of papers. Hence, for more precision over time, instead of using the number of friends, the number of papers are used for determining the power of researcher. Next, let $\mathbf{P} = \{p_1, p_2, \ldots, p_m\}$ be a set of published papers in DB-Data set. Let $\mathbf{P}_{v_j} = \{p_{j_1}, p_{j_2}, \ldots, p_{j_u}\}$ be a set of published papers of researcher v_j, where $\mathbf{P}_{v_j} \subseteq \mathbf{P}$. The power weight of researcher v_j denoted by $\mathcal{P}_{v_j}$ is calculated by using equation (3.2).

$$\mathcal{P}_{v_j} \quad = \quad \sum_{p_e \in P_{v_j}} a_{v_j,p_e} \ . \tag{3.2}$$

Suppose that v_j is a co-author in paper p_e which published in year α and the year of recommendation is β, where $p_e \in \mathbf{P}_{v_j}$. The power weight of v_j on paper p_e, denoted by a_{v_j,p_e}, is calculated by $\frac{1}{2^{\varphi}}$, where $\varphi = (\beta - \alpha) - 1$. The maximum value of a_{v_j,p_e} is one when the paper was published in one year ago prior to recommending. A calculated power weight, $\mathcal{P}_{v_j}$, is an attribute of individual researcher which is represented on the node label in co-authorship network. Figure 3.3 shows co-

authorship network extended from the network in Figure 3.1. The below of co-authorship network shows how to compute power weight of each researcher. For example, researcher v_1 wrote two papers, namely p_1 and p_2 in year 2006 and 2007, respectively. The power weight of v_1 on paper p_1 is $a_{v_1,p_1} = 2^4 = 0.0625$. The power weight of v_1 on paper p_2 is $a_{v_1,p_2} = 2^3 = 0.125$. In summary, the power weight of researcher v_1 associated with both papers is $\mathcal{P}_{v_1} = 0.1875$. Note that the power weight of v_1, 0.1875, based on two papers is less than the power weight of v_4, $\mathcal{P}_{v_4} = 0.5$, based on a paper since both papers of v_1 are older published than a paper of v_4.

3.3.3 Common Papers Analysis

In this section, a measurement used to determine pairwise weights between friend researchers utilized from co-authorship relation in their common papers is introduced. In addition, a method used to approximate pairwise weights between non-friend researchers is also proposed. The weights calculated from the both methods are called closeness weights. These weights are used to measure the pairwise weight based on common papers. Note that the idea of this measurement has been proposed in [48].

According to a list of authors in a paper, a set of closeness weights between them can be identified. Researchers who wrote the same papers are treated as friends of each other. The idea is two researchers who have recently published a larger number of papers with a small authors should be weighted more. A closeness weight between v_i and his friend in $Q(v_i)$ denoted by $\mathcal{S}_{v_i,v_\partial}$ where $v_\partial \in Q(v_i)$. Three features are used for computing as follows.

- *The number of authors in a paper*: author pair should be weighted more if their paper has few authors.
- *Year of publication*: author pair with recently co-authorship have a higher weight.
- *Frequency of co-authorship*: author pair with frequent co-authorship have a higher weight.

The number of authors in a paper and the year of publication are listed in the second and third order of an edge label as shown in the example in Figure 3.3 where an edge represents a paper. The last feature, frequency of co-authorship, is represented by the number of edges linked between node pairs. All features will be comprehensively analyzed for determining the closeness weight between a pair of authors. That is the first and the last features will be dynamically considered along with the evolution over time in the second feature.

Suppose the researchers v_i and v_∂ co-authored h papers in set $\mathbf{C}_{v_i,v_\partial} = \{p_{e_1}, p_{e_2}, .., p_{e_h}\}$, where $\mathbf{C}_{v_i,v_\partial} \subseteq \mathbf{P}$, $\acute{p}_e \in \mathbf{C}_{v_i,v_\partial}$. The closeness weight between v_i and v_∂, denoted by $\mathcal{S}_{v_i,v_\partial}$, based on their common h papers can be defined in equation (3.3).

$$\mathcal{S}_{v_i,v_\partial} = \sum_{\acute{p}_e \in C_{v_i,v_\partial}} w_{v_i,v_\partial,\acute{p}_e} \ . \tag{3.3}$$

$$w_{v_i,v_\partial,\acute{p}_e} = (z_{v_i,v_\partial,\acute{p}_e}) \times (t_{v_i,v_\partial,\acute{p}_e}) \ . \tag{3.4}$$

Term $z_{v_i,v_\partial,\acute{p}_e}$ is the result of the first feature, the number of authors in a paper. Suppose researchers v_i and v_∂ wrote paper $\acute{p}_e$, and $f(\acute{p}_e)$ is the number of authors in paper $\acute{p}_e$. The $z_{v_i,v_\partial,\acute{p}_e}$ between v_i and v_∂ on paper $\acute{p}_e$ is calculated by using equation (3.5). The maximum value of $z_{v_i,v_\partial,\acute{p}_e}$ is one when the paper has two authors. Note that all papers written by sole author are filtered out from this calculation.

$$z_{v_i,v_\partial,\acute{p}_e} = \frac{1}{f(\acute{p}_e) - 1} \ . \tag{3.5}$$

Term $t_{v_i,v_\partial,\acute{p}_e}$ is the result of the second feature, year of publication. Suppose researcher v_i and v_∂ co-author in paper $\acute{p}_e$ published in year α and the year of recommending is β. The weight between v_i and v_∂ on paper $\acute{p}_e$ is calculated by using equation (3.6), where $\varphi = (\beta - \alpha) - 1$. The maximum value of $t_{v_i,v_\partial,\acute{p}_e}$ is one when the paper is published in one year ago prior to recommending.

$$t_{v_i,v_\partial,\acute{p}_e} = \frac{1}{2^\varphi} \ . \tag{3.6}$$

The outputs of closeness weight between v_i and his neighbors in Table 3.4 are shown in Figure 3.4. The solid edges link between friend pairs and dash edges link between non-friend pairs. Each solid edge is labeled by closeness weight, $\mathcal{S}_{v_i,v_\partial}$, calculated in Table 3.5 (a).

On the other hand, two researchers not having co-authored any papers are called non-friend to each other. The closeness weight between them has to be approximated due to there are no edges between them. The transitive property on the paths between them is applied for determination. There could be several interaction paths between the two non-friend not directly collaborate but join work through a number of the other researchers in the network. The idea is two un-linked nodes tend to be connected in the near future if they connected with more short paths. The measure for identifying the approximated closeness weight between non-friend pairs is proposed as follows.

v_i, v_∂	$\sum_{\acute{p}_e \in C_{v_i,v_\partial}} w_{v_i,v_\partial,\acute{p}_e}$	$\mathcal{S}_{v_i,v_\partial}$
v_1, v_2:	$0.03125 + 0.125$	0.15625
v_1, v_3:	0.03125	0.03125
v_2, v_3:	0.03125	0.03125
v_3, v_4:	0.5	0.5

(a) The closeness weight between v_i and v_∂, $\mathcal{S}_{v_i,v_\partial}$, calculated by equation (3.3). There are the weights labeled on solid edges in Figure 3.4.

$v_i, v_\partial, \acute{p}_e$	$z_{v_i,v_\partial,\acute{p}_e}$	$t_{v_i,v_\partial,\acute{p}_e}$	$w_{v_i,v_\partial,\acute{p}_e}$
v_1, v_2, p_1	0.5	0.0625	0.03125
v_1, v_2, p_2	1.0	0.125	0.125
v_1, v_3, p_1	0.5	0.0625	0.03125
v_2, v_3, p_1	0.5	0.0625	0.03125
v_3, v_4, p_3	1.0	0.5	0.5

(b) The closeness weight between v_i and v_∂ on a paper $\acute{p}_e$, $w_{v_i,v_\partial,\acute{p}_e}$, calculated by equation (3.4).

$\acute{p}_e$	Year	φ	$1/2^\varphi$	$f(\acute{p}_e)$	$1/(f(\acute{p}_e)-1)$
paper: p_1	2006	4	0.0625	3	0.5
paper: p_2	2007	3	0.125	2	1.0
paper: p_3	2009	1	0.5	2	1.0

(c) The features of each paper, i.e., the number of authors in a paper and year of publication.

Table 3.5 Details to calculate the closeness weight between v_i and v_∂, $\mathcal{S}_{v_i,v_\partial}$, assume $\beta = 2011$.

Let v_i be a core author and v_λ be a non-friend researcher of v_i where $v_\lambda \in \overline{Q(v_i)}$. Let $\mathbf{X}_{v_i,v_\lambda} = \{x_{l_1}, .., x_{l_\eta}\}$ be a set of paths linked between v_i and v_λ. Approximated closeness weight between v_i and v_λ, denoted by $\tilde{\mathcal{S}}_{v_i,v_\lambda}$, based on η paths can be defined in equation (3.7). Prior to finding the value of $\tilde{\mathcal{S}}_{v_i,v_\lambda}$ from all paths, the average pairwise weight of each path x_l, denoted by $\hat{\mathcal{S}}_{x_l}$ must be computed by using equation (3.8). Due to the possible maximum number of weights in a path is six, the mean may be biased with an outlier weight if arithmetic mean is applied. Unlike the arithmetic mean, the harmonic mean gives less significance to outliers and provide a truer picture of the average. Thus, the harmonic mean is applied in equation (3.8).

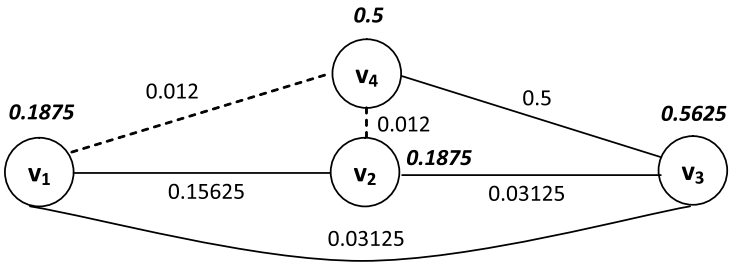


Figure 3.4: Co-authorship network represented by a simple graph derived from multi-graph in Figure 3.3. The solid edge link between friend pairs labeled by $\mathcal{S}_{v_i,v_\partial}$ calculated in Table 3.5 (a) whereas the dashed edge link between non-friend pairs labeled by $\tilde{\mathcal{S}}_{v_i,v_\lambda}$ calculated in equation (3.11) and equation (3.14).

$$\tilde{\mathcal{S}}_{v_i,v_\lambda} = \left(\sum_{x_l \in X_{v_i,v_\lambda}} \hat{\mathcal{S}}_{x_l} \right) \times d(v_i, v_k) \; . \tag{3.7}$$

$$\hat{\mathcal{S}}_{x_l} = \frac{\Delta}{\frac{1}{\mathcal{S}_{v_i,v_{l_1}} 2^{-o(v_i,v_{l_1})}} + \cdots + \frac{1}{\mathcal{S}_{v_{l_\Delta},v_\lambda} 2^{-o(v_{l_\Delta},v_\lambda)}}} \; . \tag{3.8}$$

Each path x_l, $x_l \in \mathbf{X}_{v_i,v_\lambda}$, consists of a set of $\Delta - 1$ intermediated nodes and a set of Δ edges, degree of separation, between v_i and v_λ. Let $\mathbf{H}_{v_i,v_\lambda} = \{v_{l_1}, v_{l_2}, \ldots, v_{l_{\Delta-1}}\}$ be the order set of nodes from v_i to v_λ on path x_l. The first edge connects v_i with v_{l_1}, but the last edge connects $v_{l_{\Delta-1}}$ with v_λ. For any path x_l, let $o_{(v_{l_b},v_{l_y})}$ be the order of edge (v_{l_b}, v_{l_y}) on this path x_l where $v_{l_b} \in \mathbf{H}_{v_i,v_\lambda}$, $v_{l_y} \in \mathbf{H}_{v_i,v_\lambda}$. The factor $2^{-o_{(v_{l_b},v_{l_y})}}$ attenuates the closeness weight, calculated from equation (3.3), of a further node v_{l_y}. This implies that the further distance v_{l_y} is from v_i, the less weight is assigned. Since the number of paths between two nodes proportionally grows to the degree of separation between them, term $d(v_i, v_k)$ is used to reduce the summation of $\hat{\mathcal{S}}_{x_l}$ from a large number of that paths.

Each dashed edge in Figure 3.4 labeled by the approximated closeness weight, $\tilde{\mathcal{S}}_{v_i,v_\lambda}$. From the example, there are two non-friend pairs in Table 3.4, i.e., (v_1, v_4) and (v_2, v_4). The approximated closeness weight of the first pairs, $\tilde{\mathcal{S}}_{v_1,v_4}$, can be computed on two paths in equation (3.11). For the other, $\tilde{\mathcal{S}}_{v_2,v_4}$, can be calculated on two paths in equation (3.14).

Paths between (v_1, v_4) : $x_1 = < v_1 - v_2 - v_3 - v_4 >$

$$x_2 = < v_1 - v_3 - v_4 >$$

$$\hat{\mathcal{S}}_{x_1} = \frac{3}{\frac{1}{0.15625(2^{-1})} + \frac{1}{0.03125(2^{-2})} + \frac{1}{0.5(2^{-3})}} = 0.019 \; . \tag{3.9}$$

$$\hat{\mathcal{S}}_{x_2} = \frac{2}{\frac{1}{0.03125(2^{-1})} + \frac{1}{0.5(2^{-2})}} = 0.028 \; . \tag{3.10}$$

$$\tilde{\mathcal{S}}_{v_1,v_4} = (0.019 + 0.028) \times \frac{1}{2^2} = 0.012 \; . \tag{3.11}$$

Paths between (v_2, v_4) : $x_1 = < v_2 - v_3 - v_4 >$

$$x_2 = < v_2 - v_1 - v_3 - v_4 >$$

$$\hat{\mathcal{S}}_{x_1} = \frac{2}{\frac{1}{0.03125(2^{-1})} + \frac{1}{0.5(2^{-2})}} = 0.028 \; . \tag{3.12}$$

$$\hat{\mathcal{S}}_{x_2} = \frac{3}{\frac{1}{0.15625(2^{-1})} + \frac{1}{0.03125(2^{-2})} + \frac{1}{0.5(2^{-3})}} = 0.019 \; . \tag{3.13}$$

$$\tilde{\mathcal{S}}_{v_2,v_4} = (0.028 + 0.019) \times \frac{1}{2^2} = 0.012 \; . \tag{3.14}$$

3.3.4 Common Friends Analysis

In this section, the last measurement underlining co-authorship analysis is proposed. The output of the measurement used for computing pairwise weight based on common friends is called mutual weight. The mutual weight uses basic idea of *Adamic/Adar* [49] which tries to predict friendship with something in common. The assumptions associated in this section are (a) two users shared more common items are weighted more than two users shared few common items; (b) items shared to a few user are weighted more similarity than items distributed in virtually all users. In addition, the concept also corresponds to *inverse document frequency (IDF)* [50] in information retrieval context. The *IDF* is a method used to measure importance of words. The *IDF* of word found in a few documents is high importance whereas the *IDF* of word occurs in almost entire documents is likely to be low. According to the both ideas, an algorithm for computing mutual weight based on common friends between core author v_i and his neighbor v_k denoted by $\mathcal{C}_{v_i,v_k}$ is proposed as follows.

1. Get a set of friend researchers (adjacent nodes) of v_i and keep in $Q(v_i)$ (see the second column in Table 3.6).

Table 3.6: The mutual weights for five researcher pairs associated with co-authorship network in Figure 3.5 .

v_i, v_k	v_i's Friends $Q(v_i)$	v_k's Friends $Q(v_k)$	Common Friends $Q(v_i, v_k)$	Distribution of s_{v_c}	Mutual Weight $\mathcal{C}_{v_i,v_k}$
v_1, v_2	$\{v_2, v_3\}$	$\{v_1, v_3\}$	$\{v_3\}$	log(4/3)	0.288
v_1, v_3	$\{v_2, v_3\}$	$\{v_1, v_2, v_4\}$	$\{v_2\}$	log(4/2)	0.693
v_1, v_4	$\{v_2, v_3\}$	$\{v_3\}$	$\{v_3\}$	log(4/3)	0.288
v_2, v_3	$\{v_1, v_3\}$	$\{v_1, v_2, v_4\}$	$\{v_1\}$	log(4/2)	0.693
v_2, v_4	$\{v_1, v_3\}$	$\{v_3\}$	$\{v_3\}$	log(4/3)	0.288

2. Get a set of friend researchers (adjacent nodes) of v_k and keep in $Q(v_k)$ (see the third column in Table 3.6).

3. Get a set of common friends between $Q(v_i)$ and $Q(v_k)$ keep in $Q(v_i, v_k)$. Thus, $Q(v_i, v_k) = \{v_c | v_c \in (Q(v_i) \cap Q(v_k))\}$, see the forth column in Table 3.6.

4. For each common friend v_c where $v_c \in Q(v_i, v_k)$ do

   (a) Count the number of friend researchers (adjacent nodes) of v_c denoted by $|Q(v_c)|$.

   (b) Calculate distribution of v_c, denoted by s_{v_c} with respect to nodes in co-authorship networks using equation (3.15). Term *n* is the number of nodes (unique researchers) in the co-authorship network. The distribution of v_c obtained by dividing the number of nodes by the number of v_c's friends and, then, taking the logarithm for reducing the value of that quotient. The high value of s_{v_c} occurred when v_c has a few friends. The minimum value of $|Q(v_c)|$ is two and the maximum is $(n-1)$. In other words, s_{v_c} represents the importance which v_c gives to his friends. The fifth column in Table 3.6 shows the distribution of v_c.

$$s_{v_c} = log\left(\frac{n}{|Q(v_c)|}\right) . \tag{3.15}$$

5. Calculated the total distribution of all common friends in $Q(v_i, v_k)$, where $\mathcal{C}_{v_i,v_k}$ is the mutual weight between them (see the last column in Table 3.6).

$$\mathcal{C}_{v_i,v_k} = \sum_{v_c \in Q(v_i,v_k)} s_{v_c} . \tag{3.16}$$

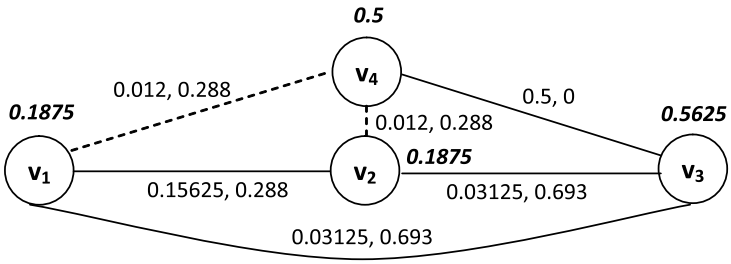


Figure 3.5: Co-authorship network extended from Figure 3.4. Edges are labeled by closeness weight, $\mathcal{S}_{v_i,v_\partial}$ or $\tilde{\mathcal{S}}_{v_i,v_\lambda}$ and mutual weight, $\mathcal{C}_{v_i,v_k}$, respectively.

The mutual weight $\mathcal{C}_{v_i,v_k}$ of five author pairs in Table 3.6 are represented in Figure 3.5. The weights are located at the second term of edge label. Note that the second term of edge label between node (v_3, v_4) is zero since they have no common friends.

3.3.5 Structure Cohesion Analysis

After computing the power weights, closeness weights, and mutual weights, all weighs are combined in the structure cohesion analysis. The output of this section is called cohesion weight which is also the final output of `SC-CoT` model. The three weights of the three measurement are treated as equal importance. Term $\mathcal{N}_{v_i \to v_k}$ in equation (3.17) represents the cohesion weight of neighbor v_k with respect to core author v_i. The neighbors with high for some weights, need not high for all weights, are accepted for recommending. Thus, weights are involved by addition instead of multiplication. Since the three weights are expressed bi-directed relationship calculated from the different scales, it have to be normalized prior to being combined together. Thus, the process of normalization is taken for two aims, namely converting from bi-directed relationship to directed relationship and taking the contribution from the different scales to the same scale.

$$\mathcal{N}_{v_i \to v_k} = \acute{\mathcal{P}}_{v_i \to v_k} + \acute{\mathcal{S}}_{v_i \to v_k} + \acute{\mathcal{C}}_{v_i \to v_k} \ . \tag{3.17}$$

Term $\acute{\mathcal{P}}_{v_i \to v_k}$ is the normalized power weight calculated by equation (3.18). The individual power of researchers are normalized to pairwise weight of neighbor v_k in the view point of core author v_i, where $v_k \in R(v_i)$. The $\acute{\mathcal{P}}_{v_i \to v_k}$ is a power weight which neighbor v_k will be selected to recommend to core author v_i compared with power weight of v_i's neighbors.

$$\acute{\mathcal{P}}_{v_i \to v_k} = \frac{\mathcal{P}_{v_k}}{\sum\limits_{v_k \in R(v_i)} \mathcal{P}_{v_k}} . \tag{3.18}$$

Term $\acute{\mathcal{S}}_{v_i \to v_k}$ is the normalized closeness weight computed by equation (3.19). The normalized weight shows the closeness of neighbor v_k in view point of v_i compared with v_i's neighbors. The closeness weights of friend pairs in $\mathcal{S}_{v_i,v_\partial}$ and approximated closeness weight of non-friend pairs in $\tilde{\mathcal{S}}_{v_i,v_\lambda}$, are combined together in equation (3.20) prior to being used in equation (3.19).

$$\acute{\mathcal{S}}_{v_i \to v_k} = \begin{cases} \frac{\mathcal{S}_{v_i,v_\partial}}{\mathcal{S}_{R(v_i)}}, & \text{friends} \\ \frac{\tilde{\mathcal{S}}_{v_i,v_\lambda}}{\mathcal{S}_{R(v_i)}}, & \text{non-friends} . \end{cases} \tag{3.19}$$

$$\mathcal{S}_{R(v_i)} = \sum_{v_\partial \in Q(v_i)} \mathcal{S}_{v_i,v_\partial} + \sum_{v_\lambda \in \overline{Q(v_i)}} \tilde{\mathcal{S}}_{v_i,v_\lambda} . \tag{3.20}$$

Lastly, term $\acute{\mathcal{C}}_{v_i \to v_k}$ is the normalized mutual weight of neighbor v_k in the view point of core author v_i compared with v_i's neighbors. The normalized weights are calculated by equation (3.21). If common friends does not exist then the value is assumed to be zero.

$$\acute{\mathcal{C}}_{v_i \to v_k} = \frac{\mathcal{C}_{v_i,v_k}}{\sum\limits_{v_k \in R(v_i)} \mathcal{C}_{v_i,v_k}} . \tag{3.21}$$

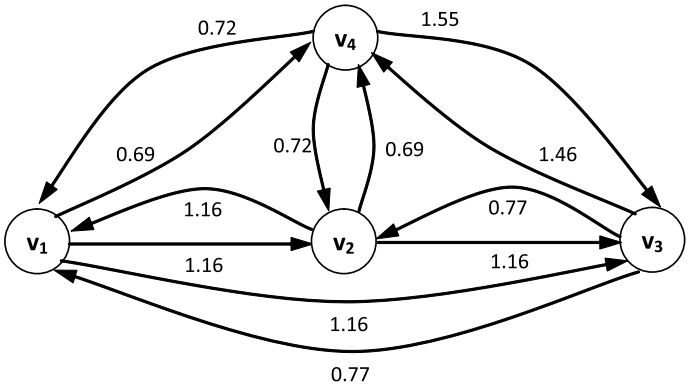


Figure 3.6: Directed co-authorship network which edges labeled by cohesion weight, $\mathcal{N}_{v_i \to v_k}$. The details of calculation are shown in Table 3.7.

After all cohesion weight, $\mathcal{N}_{v_i \to v_k}$, are computed, the directed co-authorship network is used for representing these directed relationship. From the same example, directed co-authorship network in Figure 3.6 is derived from undirected co-authorship network in Figure 3.5. The network shows

Table 3.7: The details show how to compute the cohesion weights, $\mathcal{N}_{v_i \to v_k}$, using equation (3.17). The calculated weights appeared on edge labels of co-authorship network in Figure 3.6.

$v_i \to v_k$	Normalized Weights			Cohesion Weights $\mathcal{N}_{v_i \to v_k}$
	Power Weights $\acute{\mathcal{P}}_{v_i \to v_k}$	Closeness Weights $\acute{\mathcal{S}}_{v_i \to v_k}$	Mutual Weights $\acute{\mathcal{C}}_{v_i \to v_k}$	
$v_1 \to v_2$	0.15	0.78	0.23	1.16
$v_1 \to v_3$	0.45	0.16	0.55	1.16
$v_1 \to v_4$	0.40	0.06	0.23	0.69
$v_2 \to v_1$	0.15	0.78	0.23	1.16
$v_2 \to v_3$	0.45	0.16	0.55	1.16
$v_2 \to v_4$	0.40	0.06	0.23	0.69
$v_3 \to v_1$	0.21	0.06	0.50	0.77
$v_3 \to v_2$	0.21	0.06	0.50	0.77
$v_3 \to v_4$	0.57	0.89	0.00	1.46
$v_4 \to v_1$	0.20	0.02	0.50	0.72
$v_4 \to v_2$	0.20	0.02	0.50	0.72
$v_4 \to v_3$	0.60	0.95	0.00	1.55

cohesion weight with directed relationships where edges labeled by cohesion weight, $\mathcal{N}_{v_i \to v_k}$. For example of $\mathcal{N}_{v_1 \to v_3}$ in Table 3.7, $\acute{\mathcal{P}}_{v_1 \to v_3} = 0.5625/(0.1875 + 0.5625 + 0.5) = 0.45$, $\acute{\mathcal{S}}_{v_1 \to v_3} = 0.03125/(0.15625 + 0.03125 + 0.012) = 0.16$, $\acute{\mathcal{C}}_{v_1 \to v_3} = 0.693/(0.288 + 0.693 + 0.288) = 0.55$, and $\mathcal{N}_{v_1 \to v_3} = 0.45 + 0.16 + 0.55 = 1.16$. Obviously, the weights between v_i and v_k are asymmetry, for instance, $\mathcal{N}_{v_1 \to v_3} = 1.16$ whereas $\mathcal{N}_{v_3 \to v_1} = 0.77$. Therefore, the weight in structural approach of neighbor v_3 in view point of inquired researcher v_1 compared with v_1's neighbors is 1.16.

3.4 Semantic Similarity with Background and Trend of Research (SS-BaT) Model

3.4.1 Overview of SS-BaT Model

In this section, the *Semantic Similarity with Background and Trend of Research* (SS-BaT) model is proposed. It is extended based on the author-topic model (ATM) [1] which is an existing unsupervised learning technique for extracting information from large documents. The output of SS-BaT model is the semantic similarity between two researchers based on the content of their papers, re-

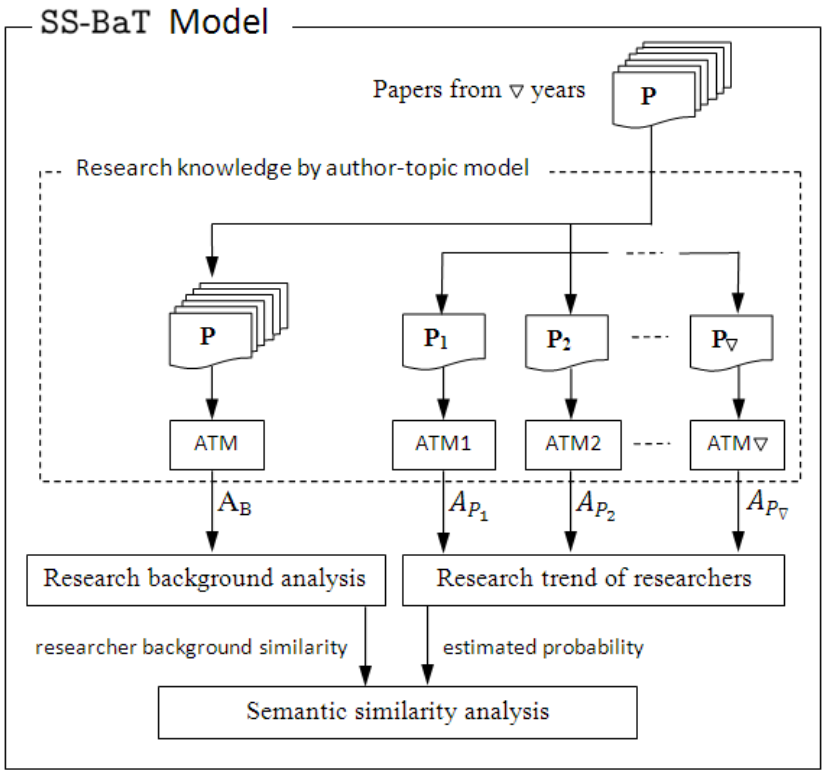


Figure 3.7: The proposed SS-BaT model for semantic approach. The "ATM" are abbreviated from author-topic model. Output of the model is a set of semantic similarity $\mathcal{M}_{v_i \to v_k}$.

gardless of the structure of co-authorship network. In order to get the most suitable neighbor v_k for inquired researcher v_i based on semantic approach, the following assumptions are set up as follows.

- *Background*: v_k is the researcher who has the most research experience similarity with v_i.
- *Trend*: v_k is the researcher who has the highest estimated probability in the current research topic of v_i.

The researchers v_k with the both assumptions should be selected first. To obtain the most suitable researcher, two levels of semantic similarity analysis are studied under different schemes, i.e., static scheme and dynamic scheme. The static scheme analyzes researcher background using whole papers over defined ∇ years. Output of this scheme is the similarity of research experience between two researchers during ∇ years. Since research interesting can change dynamically, researchers have to be investigated with the new trends and undertaken new research topics. In order to measure the trend of research, the dynamic scheme is also considered for solving the second assumption. Firstly, the probability of each topic in the recommended year β for individual researchers are estimated.

Next, the current research topic of inquired researcher v_i is automatically defined. Then, the probability of v_k which is the highest in v_i's current topic is determined. The overview of the semantic similarity calculation in SS-BaT model is shown in Figure 3.7 and details of the models are as follows.

3.4.2 Research Knowledge by Author-Topic Model

As mentioned in Chapter II, the model formulates a document as a bag of words by represents each document as a vector of word's counting. An element in the vector corresponds to the frequency of a word appearing in the document. Three input files, i.e., "Papers", "Authors", and "Words", have to be prepared for submitting to author-topic model running. The output of author-topic model treats each researcher as a probability distribution over topics and each topic is treated as a probability distribution over words. The process of how to prepare in input files and run by author-topic model are shown as follows.

3.4.2.1 Preparing Input Files for Author-Topic Model

For preparing the files, the content of attributes for semantic approach listed in Table 3.1 are considered. Some example is shown in Table 3.8. Firstly, content of two attributes, *Paper ID, Title*, are used to generate "Papers" file. Contents of two attributes, *Paper ID, Authors*, are used to generate "Authors" file. Lastly, contents of six attributes, *Paper ID, Title, Abstract, Author Keywords, Index Keywords, Venue, References*, are used to generate "Words" file which is the most difficultly created file. This file represents the frequency of various words appearing in the six attributes of papers. From the example in Table 3.8, some non-topic-related words have to be eliminated from the content of papers. The most awkwardly cleaned attribute is *References* due to it formulated from various BibTeX styles, i.e., article, book, proceeding, thesis, etc. Thus, the three steps for cleaning the contents of the six attributes are shown below.

1. Manually remove non-topic-related words such as pages, months, year of publication, address, abbreviated words, etc. from *Venue* and *References* attributes. An example of removed words in Table 3.8 are "IEEE", "pp. 177-205", "(2000)", "Trans.", "Aug".

2. Extract a set of nouns from the content of all attributes by using *SharpNLP library*[2]. The extracted nouns from Table 3.8 are shown in Table 3.9.

[2]www.codeplex.com/sharpnlp

Table 3.8 Content of an example paper with respect to attribute names in Table 3.1.

Attributes	Content
Paper ID	100
Year	2007
Authors	Jorgensen M., Shepperd M.
Title	A systematic review of software development cost estimation studies
Abstract	This paper aims to provide a basis for the improvement of software estimation research through a systematic review of previous work. The review identifies 304 software cost estimation papers in 76 journals and classifies the papers according to research topic, estimation ...
Author Keywords	Research methods; Software cost estimation; Software cost prediction; Software effort estimation; Software effort prediction; Systematic review
Index Keywords	Costs; Digital libraries; Engineering research; Estimation; Software cost estimation; Software cost prediction; Software effort estimation; Software effort prediction; Software engineering
Venue	IEEE Transactions on Software Engineering
References	Boehm, B., Abts, C., Chulani, S., Software Development Cost Estimation Approaches - A Survey (2000) Annals of Software Eng., 10, pp. 177-205; ... Jorgensen, M., Experience with the Accuracy of Software Maintenance Task Effort Prediction Models (1995) IEEE Trans. Software Eng., 21 (8), pp. 674-681., Aug; Jorgensen, M., Gruschke, T., Industrial Use of ...

3. Apply *Porter Stemming Algorithm*[3] to stem nouns. The algorithm will remove the commoner morphological and inflexional endings from nouns. For example, the stemming algorithm reduced the words "Computing", "computer", "Compute", and "computed" to the root word, "comput". The example of stemmed nouns are shown in the last column of Table 3.9.

After all nouns from all papers are completely stemmed, these nouns are used as words for generating "Words" file. The example of three input files are shown in Table 3.10 where Table 3.10 (c) shows the frequency of 30 words appeared in a paper, e.g. Paper ID $= 100$.

[3]http://www.tartarus.org/ martin/PorterStemmer

Table 3.9 List of nouns and stemmed nouns extracted from the content of six attributes in Table 3.8.

Attributes	Extracted Nouns	Stemmed Nouns
Title	review software development cost estimation studies	review softwar develop cost estim studi
Abstract	paper basis improvement software estimation research review work review software cost estimation papers journals papers research topic estimation	paper basi improv softwar estim research review work review softwar cost estim paper journal paper research topic estim
Author Keywords	research methods software cost estimation software cost prediction software effort estimation software effort prediction review	research method softwar cost estim softwar cost predict softwar effort estim softwar effort predict review
Index Keywords	costs digital libraries engineering research estimation software cost estimation software cost prediction software effort estimation software effort prediction software engineering	cost digit librari engin research estim softwar cost estim softwar cost predict softwar effort estim softwar effort predict softwar engin
Venue	software engineering	softwar engin
References	boehm abts chulani software development cost estimation approaches survey annals software experience accuracy software maintenance task effort prediction models software	boehm abt chulani softwar develop cost estim approach survei annal softwar experi accuraci softwar mainten task effort predict model softwar

3.4.2.2 Running Author-Topic-Model

After three files are completely created as an example shown in Table 3.10, these files and the number of required topics (clusters), τ, are submitted to author-topic model which applied on *PERL* programming. The output of model are probability distribution over τ topics for a particular researcher. Example of output with $\tau = 6$ are shown in Table 3.11. A topic number with the highest probability called likely topic will be assigned to each researcher. Since the designed `SS-BaT` model comprised of static scheme and dynamic scheme, each scheme is independently run on author-topic model as follows.

Table 3.10: An example of three input files generated from the paper in Table 3.8. The files used for submitting to author-topic model.

(a) Papers File

Paper ID	Title
100	A systematic review of software development cost estimation studies

(b) Authors File

Paper ID	Author ID	Author Name
100	1	Jorgensen M.
100	2	Shepperd M.

(c) Words File

Paper ID	Word ID	Word	Word Frequency
100	1	abt	1
100	2	accuraci	1
100	3	annal	1
100	4	approach	1
100	5	basi	1
100	6	boehm	1
100	7	chulani	1
100	8	cost	8
100	9	develop	2
100	10	digit	1
100	11	effort	5
100	12	engin	3
100	13	estim	10
⋮	⋮	⋮	⋮
100	25	softwar	17
100	26	studi	1
100	27	survei	1
100	28	task	1
100	29	topic	1
100	30	work	1

Running author-topic model for research background

A set of papers during ∇ years are used to generated the three input files and submitted to author-topic model for once running as shown in Figure 3.7. The structure of output is the same format as example in Table 3.11. Let $\mathbf{A_B}$ denote the output of author-topic model based on static scheme.

Running author-topic model for research trend

The yearly papers from ∇ years are used to analyze trends of researchers in topics over time. The papers are partitioned by year. Let $\mathbf{P} = \{\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_\nabla\}$ is an order set of papers grouped

Table 3.11: Example output from author-topic model. Assuming that the number of topics is six, $\tau = 6$.

Author	Topic 0	Topic 1	$\cdots$	Topic 5	Likely topic
Jorgensen M.	0.035	0.564	$\cdots$	0.110	1
Shepperd M.	0.020	0.320	$\cdots$	0.610	5
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$

by year where $\mathbf{P}_1$ are papers in the oldest year and $\mathbf{P}_\nabla$ are papers in the latest year. Let $\mathbf{P}_\alpha = \{p_{\alpha_1}, p_{\alpha_2}, \ldots, p_{\alpha_u}\}$ be a set of papers published in year α where $\mathbf{P}_\alpha \in \mathbf{P}$.

For each $\mathbf{P}_\alpha$, the papers $\mathbf{P}_\alpha$ are used to generate three input files for submitting to author-topic model, namely papers of the first year kept in $\mathbf{P}_1$ are submitted to the first running of author-topic model (ATM1) as shown in Figure 3.7. After all ∇ iterations of author-topic model are successfully process by ATM, yearly outputs are kept in $\mathbf{A_P}$. So, $\mathbf{A_P} = \{A_{P_1}, A_{P_2}, \ldots, A_{P_\nabla}\}$ be an order sets of output of author-topic model from the oldest year to the latest year associated with set $\mathbf{P}$. For instance, A_{P_1} is the output of papers in $\mathbf{P}_1$ run by ATM1 as shown in Figure 3.7. The structure of A_{P_α} where $A_{P_\alpha} \in \mathbf{A_P}$ is the same format as example in Table 3.11.

3.4.3 Research Background Analysis

After running the author-topic model in Figure 3.7, the output of author-topic model for research background is kept in $\mathbf{A_B}$. This output is utilized for calculating the background similarity between author pair. Although the likely topics treated by author-topic model between two researchers are different, it is quite possible that their research backgrounds might be similar. From Table 3.11, a researcher can be represented by a vector of probability distribution of τ topics. These vectors of probability distribution are treated as the input of computing the research background similarity between author pair. The research background similarity between core author v_i and each neighbor $v_k \in R(v_i)$ can be calculated by *cosine similarity* [51]. Given two vectors of probability distribution, Γ_{v_i} and Γ_{v_k}, the *cosine similarity*, θ, is represented using a dot product and magnitude as follows.

$$\mathcal{B}_{v_i,v_k} = cos(\theta) = \frac{\Gamma_{v_i} \cdot \Gamma_{v_k}}{\|\Gamma_{v_i}\| \, \|\Gamma_{v_k}\|} = \frac{\sum_{\phi=0}^{\tau-1} \gamma_{i_\phi} \times \gamma_{k_\phi}}{\sqrt{\sum_{\phi=0}^{\tau-1} \gamma_{i_\phi}^2} \times \sqrt{\sum_{\phi=0}^{\tau-1} \gamma_{k_\phi}^2}} \, . \tag{3.22}$$

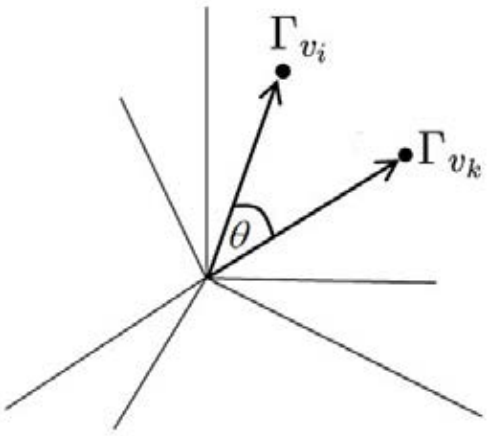


Figure 3.8 The probability distribution of v_i and v_k represented by vector Γ_{v_i} and Γ_{v_k}, respectively.

Let $\Gamma_{v_i} = \{\gamma_{i_0}, \gamma_{i_1}, \ldots, \gamma_{i_{\tau-1}}\}$ and $\Gamma_{v_k} = \{\gamma_{k_0}, \gamma_{k_1}, \ldots, \gamma_{k_{\tau-1}}\}$ are two points of v_i and v_k in τ-space as shown in Figure 3.8. The term γ_{i_ϕ} and γ_{k_ϕ} are the probability distribution of v_i and v_k in topic ϕ. The resulting similarity defines the cosine angle between the two vectors, with values between 0 and 1. As the angle between the vectors shortens, the cosine angle approaches 1, meaning that the two authors are getting the higher research background similarity.

3.4.4 Research Trend of Researchers

This section studies the evolution of research interest and determines research trend of individual researchers. The output of author-topic model kept in set $\mathbf{A_P}$ is used to estimate the probability of each topics of researchers. For a given researcher v_i, the neighbor with the highest probability in current research topic of v_i is the most appropriate collaborator. Suppose papers from six years are used to analyze trends of researchers, $\nabla = 6$. The three outputs from the set of six output of author-topic model, $\mathbf{A_P} = \{A_{P_1}, A_{P_2}, \ldots, A_{P_6}\}$, are shown in Table 3.12. The steps of determining research trend are as follows.

Firstly, the time-series tables are created for researchers i.e. one table for one researcher. The probability distribution of each topic of each researcher will be re-structured as time-series table by transposing data in the row major of Table 3.12 to column major of Table 3.13. Suppose that Table 3.12(c) is the output of author-topic model in year 2007, A_{P_6}, and Table 3.13 is the time-series table of *Shepperd M.* The probability distribution of *Shepperd M.* from Table 3.12(c) is listed in column 2007 of Table 3.13. For the year that researchers did not published any papers, i.e., year 2003, the values of all topics in the column 2003 are "null". These incomplete data have to be imputed by

Table 3.12: The three example outputs from the total six years of author-topic model for research trend. Suppose that A_{P_1}, A_{P_5}, and A_{P_6} are the outputs from years 2002, 2006, and 2007, respectively.

(a) Output from author-topic model in year 2002, A_{P_1}.

Author	Topic 0	Topic 1	$\cdots$	Topic 5	Likely topic
Jorgensen M.	0.125	0.398	$\cdots$	0.123	1
Shepperd M.	0.230	0.220	$\cdots$	0.330	5
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$

(b) Output from author-topic model in year 2006, A_{P_5}.

Author	Topic 0	Topic 1	$\cdots$	Topic 5	Likely topic
Jorgensen M.	0.026	0.459	$\cdots$	0.095	1
Shepperd M.	0.350	0.420	$\cdots$	0.430	5
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$

(c) Output from author-topic model in year 2007, A_{P_6}.

Author	Topic 0	Topic 1	$\cdots$	Topic 5	Likely topic
Jorgensen M.	0.035	0.564	$\cdots$	0.110	1
Shepperd M.	0.020	0.320	$\cdots$	0.610	5
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$

using the average based on its nearest columns. For example, Topic 5 in year 2003 will be imputed by $(0.330 + 0.620)/2$.

After the time-series tables of the researchers are completely created, the probability of each topic of researchers in the future can be approximated. Let $\Gamma_{v_{j_\phi}} = \{\gamma_{j_{\phi 1}}, \gamma_{j_{\phi 2}}, \ldots, \gamma_{j_{\phi \nabla}}\}$ be an order set of probability distribution of v_j on topic ϕ over ∇ years. For example, the completed probability distribution of *Shepperd M.* on topic 5 in Table 3.13 be $\Gamma_{v_{j_5}} = \{0.330, null, 0.620, 0.460, 0.430, 0.610\}$. The value 0.330 is the probability distribution of *Shepperd M.* on topic 5 in year 2002 as shown in Table 3.12(a). The value 0.430 is the probability distribution of *Shepperd M.* on topic 5 in year 2006 as shown in Table 3.12(b). The value 0.610 is the probability distribution of *Shepperd M.* on topic 5 in year 2007 as shown in Table 3.12(c). Note that the probability distribution of *Shepperd M.* on topic 5 in year 2004 and 2005 are assumed as 0.620 and 0.460, respectively. Since the example assumes *Shepperd M.* has no published papers in year 2003, the value of the second element of the set is null. For topic ϕ of researcher v_j, the estimated probability can be computed as follows.

Table 3.13: A time-series table over six years, $\nabla = 6$. Suppose this is the time-series table of *Shepperd M.* The probability distribution in years 2002, 2006, and 2007 are from Tables 3.12(a), 3.12(b), and 3.12(c), respectively.

Year-α	2002	2003	2004	2005	2006	2007
Topic 0	0.230	null	0.120	0.310	0.350	0.020
Topic 1	0.220	null	0.130	0.100	0.420	0.320
⋮	⋮	⋮	⋮	⋮	⋮	⋮
Topic 5	0.330	null	0.620	0.460	0.430	0.610

1. Create data curve of topic ϕ using data in set $\Gamma_{v_{j_\phi}}$. For example, $\Gamma_{v_{j_5}}$ can be plotted to curve *Prob.Dis.(Topic5)* in Figure 3.9.

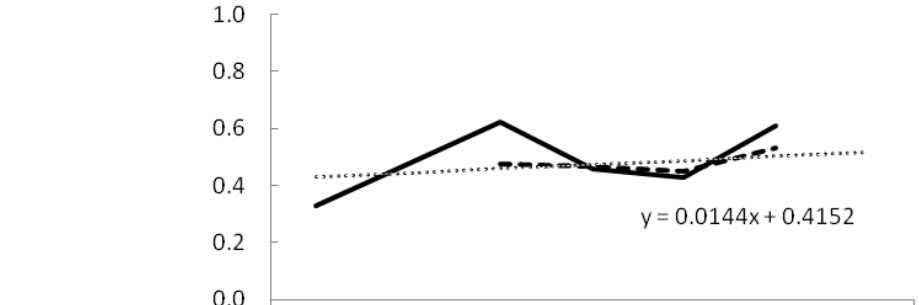
2. Create a average curve based on the plotted data curve. Since the recent data are more effect for prediction, *Exponential Moving Average (EMA)* is selected for this task. The *EMA* treats more weight to more recent data as described in Section 2.1.3. Suppose the time period used for averaging is defined as three years. Curve *EMA(Topic5)* in Figure 3.9 shows the moving average of Topic 5 based on curve *Prob.Dis.(Topic5)*. The first average, 0.48, appearing in year 2004 was calculated by simple moving average of the first three years as shown in equation (3.23). The other exponential moving averages were calculated by using equation (2.5). The example in equation (3.24) shows EMA in year 2005 is 0.47, where 3 is defined the time period, 2 and 1 are constant values in equation (2.5).

$$0.48 \quad = \quad (0.33 + 0.48 + 0.62)/3 \; . \tag{3.23}$$

$$0.47 \quad = \quad \left[(0.46 - 0.48) \times \frac{2}{3+1}\right] + 0.48 \; . \tag{3.24}$$

3. Create a trend line based on moving average curve by using simple linear regression analysis. Regression equation is also created in this step as well. For example, the equation of the trend line, *Trend(Topic5)*, in Figure 3.9 is $y = 0.0144x + 0.4152$.

4. Use regression equation to estimate a probability of topic ϕ of researcher v_j for recommended year, β. Let $\Theta_{v_j} = \{\theta_{v_{j_0}}, \theta_{v_{j_1}}, \ldots, \theta_{v_{j_{\tau-1}}}\}$ be a set of estimated probability of author v_j over



	2002	2003	2004	2005	2006	2007	2008
Prob.Dis.(Topic5)	0.33	0.48	0.62	0.46	0.43	0.61	
EMA(Topic5)			0.48	0.47	0.45	0.53	
Trend(Topic5)	0.4296	0.44	0.46	0.47	0.49	0.50	0.516

Figure 3.9: An example of data curve, average curve, and trend line for a researcher generated from time-series data Topic 5 in Table 3.13. Assume that the time period used for calculating the moving average is three years.

τ topics in year β. For example in Figure 3.9, the estimated probability of topic 5 of researcher v_j in year 2008 is $\theta_{v_{j_5}} = 0.516$ which can be calculated by equation (3.25).

$$0.516 \quad = \quad (0.0144 \times 7) + 0.4152 \; . \tag{3.25}$$

3.4.5 Semantic Similarity Analysis

Discovering a current research topic of researchers is an important ability according to the objectives of this dissertation. The discovered current research topic is assumed the inquired topic of the inquired researcher. In this section, an inquired topic, topic $\acute{\phi}$, will be automatically detected prior the semantic similarity between the inquired researcher v_i and his neighbors v_k will be determined. This section aims at two parts, namely inquired topic detection and semantic similarity calculation.

Inquired topic detection

Besides the inquired researcher's name, an inquired topic $\acute{\phi}$ is another requirement to be defined for selecting the collaborators. From the existing yearly outputs of author-topic model as in the example of Table 3.12, the likely topic of the latest year of v_i's publication is assumed to be an inquired topic of v_i. For making the convinced assumption, the papers in various years of publication are investigated. Suppose DB-Data set formulated from papers in year $2002 - 2007$ and Test-Data

set formulated from papers in year $2008 - 2009$. There are $1,681$ core authors (inquired researchers) who appeared in both sets. The latest year of publication of inquired researchers were classified in Table 3.14. The first row of table expresses 62% of inquired researchers publishing papers in the last year, 2007, 20% of inquired researchers publishing papers in two years ago, 2006, and so on.

Table 3.14 The proportion of core authors and their latest year of publication.

The Latest Year of Publication	#Inquired Researchers	Percent
2007	$1,038$	62%
2006	331	20%
2005	157	9%
2004	74	4%
2003	45	3%
2002	36	2%
	$1,681$	100%

For each inquired researcher, the likely topic of the latest year of v_i's publication was from set $\mathbf{A_P}$. An inquired topic of the inquired researcher denoted by topic $\acute{\phi}$. A set of published papers of each inquired researcher in Test-Data set were retrieved and checked the topic. Note that a particular paper belongs to a topic, namely paper p_1 belongs to topic 5, thus, an inquired researcher might wrote papers in various topics depending on his papers' topics. If the topic $\acute{\phi}$ is matched at least one topic of such inquired researcher, the defined inquired topic is correct. Table 3.15 shows 933 likely topics in year 2007 matched with the topic of $1,038$ inquired researchers. The last row shows the average correct inquired topics of 85%. Therefore, the likely topic of the latest year of v_i's publication was defined to be an inquired topic of v_i with confidence 85%.

Semantic similarity calculation

The outputs from two measurements have been already calculated, namely researcher background similarity between v_i and v_k kept in $\mathcal{B}_{v_i,v_k}$ and the estimated probability of individuals over τ topics in the recommended year kept in $\Theta_{v_j} = \{\theta_{v_{j_0}}, \theta_{v_{j_1}}, \ldots, \theta_{v_{j_{\tau-1}}}\}$. The semantic similarity of v_k with respects to v_i, denoted by $\mathcal{M}_{v_i \to v_k}$, can be calculated based on both outputs and the inquired topic $\acute{\phi}$. For each neighbor $v_k \in R(v_i)$ of an inquired researcher v_i, the steps are processed as follows.

1. Normalize research background similarity between v_i and his neighbors in $R(v_i)$. The simi-

Table 3.15 The correct defined inquired topic compared with the topics of papers in Test-Data set.

The Latest Year of Publication	#Correct Inquired Topics	#Inquired Researchers	Percent
2007	933	$1,038$	90%
2006	256	331	77%
2005	130	157	83%
2004	57	74	77%
2003	27	45	60%
2002	28	36	78%
	1,431	$1,681$	85%

larity will be converted from bi-directed relationship to directed relationship. The $\acute{\mathcal{B}}_{v_i \to v_k}$ is called background similarity of v_k with respect to v_i compared with v_i's neighbors.

$$\acute{\mathcal{B}}_{v_i \to v_k} = \frac{\mathcal{B}_{v_i,v_k}}{\sum\limits_{v_k \in R(v_i)} \mathcal{B}_{v_i,v_k}} . \tag{3.26}$$

2. Get the estimated probability distribution of v_k on topic $\acute{\phi}$ in the recommended year which is kept in $\Theta_{v_k} = \{\theta_{v_{k_0}}, \theta_{v_{k_1}}, \ldots, \theta_{v_{k_{\tau-1}}}\}$. The estimated probability denoted by $\theta_{v_{k_{\acute{\phi}}}}$ where $\theta_{v_{k_{\acute{\phi}}}} \in \Theta_{v_k}$.

3. Normalize the estimated probability distribution among neighbors in $R(v_i)$. The estimated probability will be converted from individual probability to pairwise probability with directed relationship. The $e_{v_i \overset{\acute{\phi}}{\to} v_k}$ is called estimated probability of inquired topic $\acute{\phi}$ of v_k with respects to v_i when being compared with v_i's neighbors. If $\theta_{v_{k_{\acute{\phi}}}}$ less than zero, then $\theta_{v_{k_{\acute{\phi}}}}$ will be treated as zero. Neighbor v_k with the highest $e_{v_i \overset{\acute{\phi}}{\to} v_k}$ is the neighbor who is forecast to produce papers in topic $\acute{\phi}$ with the highest probability compared with v_i 's neighbors.

$$e_{v_i \overset{\acute{\phi}}{\to} v_k} = \frac{\theta_{v_{k_{\acute{\phi}}}}}{\sum\limits_{v_k \in R(v_i)} \theta_{v_{k_{\acute{\phi}}}}} . \tag{3.27}$$

4. Compute the semantic similarity of v_k with respect to v_i, $\mathcal{M}_{v_i \to v_k}$, by combining research background similarity, $\acute{\mathcal{B}}_{v_i \to v_k}$, and the estimated probability, $e_{v_i \overset{\acute{\phi}}{\to} v_k}$. Since the best suitable recommended collaborator is the researcher having the most background similarity or the highest estimated probability, the both terms are combined by addition in equation (3.28). In case

of all neighbors having zero estimated probability, the neighbor with the most background similarity is still the best to be recommended.

$$\mathcal{M}_{v_i \to v_k} = \acute{\mathcal{B}}_{v_i \to v_k} + e_{\substack{\acute{\phi} \\ v_i \to v_k}} . \tag{3.28}$$

3.5 Ranking the Potential Collaborators Based on the Relevance Scores

After the collected papers are used to build the two proposed models, the outputs are set of the cohesion weights, $\mathcal{N}_{v_i \to v_k}$, from SC-CoT model and set of the semantic similarity, $\mathcal{M}_{v_i \to v_k}$, from SS-BaT model. The relevance score between core author v_i and neighbor v_k can be determined in four steps from Section 3.5.1 to Section 3.5.3.

3.5.1 Calculating the Raw Relevance Scores

Using the outputs from both models, the raw relevance score of v_k with respect to v_i in Δ degrees of separation, $\mathcal{W}^{\Delta}_{v_i \to v_k}$, is introduced for determining the relevance of v_k with respect to v_i and vice versa. In order to compare the efficiency of SC-CoT model and SS-BaT model in the Chapter IV, three methods for computing the raw relevance score are proposed. The first method uses only the cohesion weights, $\mathcal{N}_{v_i \to v_k}$, for calculating the raw relevance score as shown in equation (3.29).

$$\mathcal{W}^{\Delta}_{v_i \to v_k} = \frac{\mathcal{N}_{v_i \to v_k} \times d(v_i, v_k)}{\sum\limits_{v_k \in R(v_i)} \mathcal{N}_{v_i \to v_k} \times d(v_i, v_k)} . \tag{3.29}$$

The second method uses only semantic similarity, $\mathcal{M}_{v_i \to v_k}$, for calculating the raw relevance score as shown in equation (3.30).

$$\mathcal{W}^{\Delta}_{v_i \to v_k} = \frac{\mathcal{M}_{v_i \to v_k} \times d(v_i, v_k)}{\sum\limits_{v_k \in R(v_i)} \mathcal{M}_{v_i \to v_k} \times d(v_i, v_k)} . \tag{3.30}$$

The third method is a hybrid method considering both the cohesion weights and the semantic similarity as shown in equation (3.31). The both terms are combined by addition because either high cohesion weight or semantic similarity is accepted.

$$\mathcal{W}^{\Delta}_{v_i \to v_k} = \frac{\mathcal{N}_{v_i \to v_k} \times d(v_i, v_k)}{\sum\limits_{v_k \in R(v_i)} \mathcal{N}_{v_i \to v_k} \times d(v_i, v_k)} + \frac{\mathcal{M}_{v_i \to v_k} \times d(v_i, v_k)}{\sum\limits_{v_k \in R(v_i)} \mathcal{M}_{v_i \to v_k} \times d(v_i, v_k)} . \tag{3.31}$$

Since degree of separation are taken into account in this dissertation as described in Section 3.2, the distance parameter, $d(v_i, v_k)$, is weighted by equation (3.29), equation (3.30), and equation (3.31) for reducing the raw relevance score of neighbors in further degree of separation. Thus, $\mathcal{W}^1_{v_i \to v_k}$ means the raw relevance score of v_k with respect to v_i in one degree of separation (v_i and v_k is friend for each other). The $\mathcal{W}^2_{v_i \to v_k}$ means the raw relevance score of v_k with respect to v_i with two degrees of separation (v_i and v_k is friend-of-friend for each other).

3.5.2 Weighting the Raw Relevance Scores by Seniority

Seniority of researcher is a property of researcher which represents the research's experience via the number of published papers. This dissertation defined the seniority of researcher as follows.

Definition 6: Senior Researchers. Senior researchers are the researchers who published at least ten papers in a given topic.

Definition 7: Junior Researchers. Junior researchers are the researchers who published less than ten papers in a given topic. Thus, junior researchers are young researchers having fewer papers (less than ten papers).

This dissertation proposes a method to use the seniority of researcher for weighting the raw relevance scores of researcher pair. The idea is although the raw relevance score of two candidate neighbors are equal, the relevance score may not be equal if the neighbors have the different seniority. For example, an inquired researcher v_i has two neighbors, i.e. v_1 is a senior researcher and v_2 is a junior researcher. Although the raw relevance score of $\mathcal{W}^{\Delta}_{v_i \to v_1}$ and $\mathcal{W}^{\Delta}_{v_i \to v_2}$ are equal, both relevance scores may not be equal because of their different seniority. The important question is how to determine the parameter used to weight each seniority. After exploring the papers in DB-Data set between years $2000 - 2007$, the proportion of three seniority types of researchers over six topic areas are shown in Table 3.16. The table shows the potential seniority types of an unorder researcher pair consisting of three types, namely senior&senior, senior&junior, and junior&junior. For each seniority type of each research topic, the number of researcher pairs in such types is counting and divided by the number of researcher pairs in such research topic.

In case of bio-informatics area, 6%, 48%, and 46% of researcher pairs formulated from senior&senior, senior&junior, and junior&junior, respectively. In summary, (1) the possibility of collaboration among senior researchers is 6%; (2) the possibility of collaboration between senior researchers and junior researchers is 48%; (3) the possibility of collaboration between junior researchers and junior researcher is 46%. Note that "senior&junior" and "junior&senior" are the same. The proportions

Table 3.16: The proportion of three seniority types of researchers over six topic areas used as the seniority parameter, $\acute{\omega}_{v_i,v_k}$, for weighting the relevance scores.

Research Topics	Senior&Senior	Senior&Junior	Junior&Junior
Bio-informatics	6%	48%	46%
Data Mining	24%	42%	34%
Hardware	24%	46%	30%
Neural Network	19%	43%	38%
Software	12%	42%	46%
Algorithm&Theory	31%	50%	19%
Average	19%	45%	36%

are defined as the *seniority parameter*, $\acute{\omega}_{v_i,v_k}$, used to weight the raw relevance scored based on their seniority types. The calculated raw relevance score of v_k with respect to v_i in Δ degree of separation denoted by $\mathcal{W}^{\Delta}_{v_i \to v_k}$ will be weighted by the *seniority parameter*, $\acute{\omega}_{v_i,v_k}$ as shown in equation (3.32).

$$\acute{\mathcal{W}}^{\Delta}_{v_i \to v_k} = \mathcal{W}^{\Delta}_{v_i \to v_k} \times \acute{\omega}_{v_i,v_k} \ . \tag{3.32}$$

For example, if the inquired topic is neural network, the seniority of inquired researcher v_i is senior, and the seniority of neighbor v_k is junior, their *seniority parameter* will be defined as 0.43. The relevance score of v_k with respect to v_i, $\acute{\mathcal{W}}^{\Delta}_{v_i \to v_k}$, can be calculated by

$$\acute{\mathcal{W}}^{\Delta}_{v_i \to v_k} = \mathcal{W}^{\Delta}_{v_i \to v_k} \times 0.43 \ . \tag{3.33}$$

3.5.3 Adjusting the Relevance Scores from Directed Relationship to Bi-directed Relationship

Although the relevance score of two candidate neighbors are equal, the relevance score may not be equal if the relevance score in vice versa are highly different value. Suppose an inquired researcher v_i has two neighbors, i.e. v_1 and v_2. If the relevance score of $\acute{\mathcal{W}}^{\Delta}_{v_i \to v_1}$ and $\acute{\mathcal{W}}^{\Delta}_{v_i \to v_2}$ are equal, then the possibility to select v_1 to v_i should be assigned more if the value of $\acute{\mathcal{W}}^{\Delta}_{v_1 \to v_i}$ is very high and the value of $\acute{\mathcal{W}}^{\Delta}_{v_2 \to v_i}$ is low. Therefore, the (bi-directed) relevance score, $\acute{\mathcal{W}}^{\Delta}_{v_i \leftrightarrow v_k}$, can be determined by adding the two (directed) relevance scores, $\acute{\mathcal{W}}^{\Delta}_{v_i \to v_k}$ and $\acute{\mathcal{W}}^{\Delta}_{v_k \to v_i}$ by

$$\acute{\mathcal{W}}^{\Delta}_{v_i \leftrightarrow v_k} = \acute{\mathcal{W}}^{\Delta}_{v_i \rightarrow v_k} + \acute{\mathcal{W}}^{\Delta}_{v_k \rightarrow v_i} \ . \tag{3.34}$$

3.5.4 Ranking the Potential Collaborators

The neighbors of v_i kept in $R(v_i)$ will be ranked in descending order based on the (bi-directed) relevance scores $\acute{\mathcal{W}}^{\Delta}_{v_i \leftrightarrow v_k}$ in Section 3.5.3. Researcher v_k with the highest relevance score compared with all neighbors in $R(v_i)$ is the most suitable researcher for recommending to v_i. Such researcher will be ranked in the first order. Nevertheless, the results of ranking are asymmetric with respect to a viewpoint, thus, v_k may be ranked in the first order for v_i, but v_i might be ranked only in the third order for v_k.

CHAPTER IV

RESULTS AND DISCUSSION

This chapter describes the data sets preparation, experiment setting, method for ranking the potential collaborators. Besides, the results from the experiments are evaluated in various aspects.

4.1 Data Collection

The data sets are bibliographic data collected from SCOPUS [1], the largest abstract and citation database. In order to obtain the suitable papers for implementing in the proposed methods, steps of preparing data should be realized as follows.

1. Define the number of research topics. From the observation in Huang et al.'s work [12], the same topics were adapted in this dissertation.

2. Define the topics. The potential selected topics were explored from various journals and conferences in Computer Science. Finally, this dissertation decided to select publication papers in six topics, including Bio-informatics, Data Mining, Hardware, Neural Network, Software Engineering, and Algorithm and Theory.

3. Collect papers. Since the co-authorship network is a small world network which follows a *power law distribution*, it is extremely sparse. For obvious pattern study, the potential papers have to be carefully collected. To reduce the number of papers written by authors having published only one paper, this dissertation selected papers written by at least one senior author. After exploring the number of published papers of researchers in SCOPUS, about 100 senior researchers per a topic were chosen. A researcher might be the senior researcher in more than one topic. The selected the papers were published between years 2000 and 2009 written by selected senior researchers in each topic. The number of selected papers grouped by topic are shown in Table 4.1. Data set in the table was divided into two sets for experiments, i.e., DB-Data set for building model and Test-Data set for testing.

[1] http://www.scopus.com

Table 4.1 A collection of papers over six topics.

Topic Name	The Number of Papers
Bio-informatics	$1,672$
Data Mining	$1,848$
Hardware	$1,674$
Neural Network	$1,686$
Software	$1,728$
Algorithm&Theory	$1,692$

4.2 Experiment Setup

As mention in Section 3.1, the collected papers were separated in two data sets based on years of publication, i.e., DB-Data set used for generating the models, and Test-Data set used for testing. Since the papers were collected within ten years, $2000 - 2009$, the number of publication years for DB-Data set and Test-Data set must be defined. An important question is how many preceding years suitable for building the models. Fortunately, a study of Yoshikane [52] said that:

> "This study defines a newcomer in 1998 as one who published a paper (the "document type" of which is "article" in the SCI) as the first author in one of the core journals in 1998 and had not published a paper in the same domain for the preceding 7 years from 1991 to 1997".

According to the above study, it implies that an author may possibly re-publish papers during seven years. Therefore, this dissertation decided to use papers in preceding six years for building model called DB-Data set and two years for testing called Test-Data set. The reason why using two years for testing is these written papers may be delayed and published one year later. Note that setting preceding six years to create models and two years to analyze was also used in Sachan's work [34]. Consequently, the collected papers from years 2000 to 2007 were separated into three snapshots of time slice, each consisting of papers in six years. The first DB-Data set contains papers from year years 2000 to 2005. The second set contains papers from years 2001 to 2006, and the last one contains papers from years 2002 to 2007. Each DB-Data set was used to independently created the models underlining the proposed methodology. For Test-Data set which consists of papers from two years,

Table 4.2 Statistics of experimental data in particular snapshot for *Approach-A*.

	Snapshot-1		Snapshot-2		Snapshot-3	
	DB-Data	Test-Data	DB-Data	Test-Data	DB-Data	Test-Data
Year of publication	$2000-05$	$2006-07$	$2001-06$	$2007-08$	$2002-07$	$2008-09$
#Papers	$4,049$	$1,901$	$4,536$	$1,890$	$4,990$	$1,762$
#Researchers	$12,832$	$6,430$	$14,562$	$6,533$	$16,425$	$6,183$
#Unique researchers	$5,770$	$3,621$	$6,098$	$3,768$	$6,828$	$3,544$
#Core authors	$1,523$	$1,523$	$1,651$	$1,651$	$1,681$	$1,681$
#Linked researcher pairs	$16,962$	$9,068$	$18,617$	$11,308$	$21,986$	$10,299$
#Authors per paper	3.17	3.38	3.21	3.46	3.29	3.51

there are two approaches for testing. The first one is test with the same data set, *Approach-A*, and the second is test with the different data set, *Approach-B*.

Table 4.2 shows statistics of experimental data in particular snapshot based on *Approach-A*. A particular model created underline each DB-Data set was tested by the data set in next two years. For instance in Snapshot-1, the models were created by the papers in years 2000 to 2005 and tested with the papers in years 2006 and 2007. The number of unique researchers equal to the number of nodes in the associated co-authorship network. The unique researchers comprised of core authors and their co-authors in the selected papers. The number of core authors in DB-Data set and Test-Data set in the same snapshot is the same. Linked researcher pairs are arbitrary two authors having collaborated. It is equal to the number of edges in the associated co-authorship network. Based on the table, fifty-fifty researchers in Test-Data set are unknown researchers. For example in snapshot-3, $3,544$ unique researchers in Test-Data set consists of $1,681$ core authors who were existing authors in DB-Data set, so the remaining $1,863$ unique researchers were the new coming authors just appeared in Test-Data set. On the other hand, there are $1,681$ authors from all $6,828$ in DB-Data set who continued to produce their papers in Test-Data set. These can imply that the probability to select a researcher who continued to produce his papers from a set of unique researchers in the preceding six years was approximated as $1,681/6,828 = 0.25$.

Table 4.3 shows statistics of data set in a particular snapshot for *Approach-B*. The effect of time evolution to the accuracy of the collaborator recommendation experiments will be analyzed here. Thus, there are three DB-Data sets used to build the models and compare the results being tested with the same data in years 2008 and 2009. The number of core authors in Test-Data set was not shown since it was equal to the number of core authors in DB-Data snapshot which was used to build the model.

Table 4.3 Statistics of experimental data in a particular snapshot for *Approach-B*.

	DB-Data Set Snapshots			Test-Data Set
	1	2	3	
Year of publication	$2000-05$	$2001-06$	$2002-07$	$2008-09$
#Papers	$4,049$	$4,536$	$4,990$	$1,762$
#Researchers	$12,832$	$14,562$	$16,425$	$6,183$
#Unique researchers	$5,770$	$6,098$	$6,828$	$3,544$
#Core authors	$1,523$	$1,651$	$1,681$	-
#Linked researcher pairs	$16,962$	$18,617$	$21,896$	$10,299$
#Authors per paper	3.17	3.21	3.29	3.51

In semantic approach, the important setting is author-topic model running. In this dissertation, the number of topics, τ, was fixed at 6. The related setting was defined in the same way as in [1]. The variables $\hat{\partial}$ and $\hat{\kappa}$ (Figure 2.1) were set at 0.16 and 0.01, respectively. Author-topic model was run at $2,000$ iterations for learning.

Lastly for moving average, the number of the last $\hat{n}$ years was fixed at 3. It means that a $3-$year interval was used for computing the moving average in equation (2.5). The reason is that each DB-Data set snapshot consists of papers within six years, so a half of six is defined for each computing.

4.3 Evaluation Method

After neighbors in $R(v_i)$ are ranked by the relevance scores computed by a selected method from equation (3.29), equation (3.30), and equation (3.31), a set of neighbors with the highest relevance score will be recommended to v_i. The number of recommend researchers depend on the number of the real collaborators of v_i in Test-Data set. If the recommended researchers match in the real collaborator names, they are called *right recommended researchers*. For example, if v_i collaborates with two researchers in Test-Data set, the two researchers in $R(v_i)$ are recommended for v_i. If only one recommended researcher appears in the set of real collaborators, the accuracy is $1/2 = 0.5$. The method for calculating the accuracy is shown in equation (4.1) where R_{rec} is the number of *right recommended researchers* and R_{real} is the number of researchers collaborated with v_i in Test-Data set.

$$Accuracy(v_i) = \frac{R_{rec}}{R_{real}} \ . \tag{4.1}$$

4.4 Experimental Results

The accuracy of experiments was independently computed for a particular groups of friendship, i.e., friend researchers, non-friend researchers, and neighbors as described in Section 3.2. For each group, the accuracy of core authors v_i was summarized and divided by the number of all core authors in such group.

The column "Friends" represents the accuracy of recommending friend researchers to v_i. Column "Non-Friends" represents the accuracy in case of recommend non-friends researchers to v_i. Lastly, column "Neighbors", represents the average accuracy of recommending neighbors including friend researchers and non-friend researchers. For example, if v_i collaborates with two researchers in Test-Data set, one is friend in DB-Data set and the other is non-friend in DB-Data set. A friend with the highest relevance score in $R(v_i)$ will be recommended to v_i. If such recommended friend matches in Test-Data set, the accuracy for column "Friends" is $1/1 = 1.0$. On the other hand, a non-friend with the highest relevance score in $R(v_i)$ will be recommended to v_i. If he does not match in Test-Data set, the accuracy for column "Non-friend" is $0/1 = 0$. Finally, the average accuracy in column "Neighbors" is calculated by $1/2 = 0.5$.

The experimental results in this dissertation were evaluated in various aspects as follows.

4.4.1 Results of the Different Time Snapshots

As described in Section 4.2, there are three DB-Data sets tested with two approaches, namely *Approach-A* in Table 4.2 - test with the different data set and *Approach-B* in Table 4.3 - test with the same data set.

For each approach, the papers in a particular DB-Data set were used to build the models, i.e., `SC-CoT` model, `SS-BaT` model. For each model, the relevance scores were calculated by using the hybrid method based on equation (3.31). After checking the accuracy of recommending, the outputs of *Approach-A* and *Approach-B* were shown in Table 4.4(a) and Table 4.4(b), respectively. The outputs of each snapshot were separately analyzed into three columns of friendship, i.e., "Friends", "Non-friend", and "Neighbors". Column "Friends" shows the accuracy in case of selecting friend researchers for recommending. On the other hand, column "Non-friends" shows the accuracy of se-

Table 4.4: The accuracy of experiments compared with two approached of testing. The relevance scores are calculated based on the hybrid method.

(a) The accuracy of *Approach-A* in Table 4.2 - test with the different data set.

Snapshot	Year of publication	Friends	Non-friends	Neighbors
1	$2000-2005$	0.796	0.290	0.588
2	$2001-2006$	0.812	0.284	0.602
3	$2002-2007$	0.822	0.294	0.609

(b) The accuracy of *Approach-B* in Table 4.3 - test with the same data set.

Snapshot	Year of publication	Friends	Non-friends	Neighbors
1	$2000-2005$	0.627	0.195	0.424
2	$2001-2006$	0.712	0.234	0.503
3	$2002-2007$	0.822	0.294	0.609

lecting non-friend researchers for recommending. Lastly, column "Neighbors" shows the average of both friends and non-friends.

The results of *Approach-A* are shown in 4.4(a). The three results in the same group of friendship are a few difference. The three results of neighbors shows that snapshot-3 gave the highest accuracy. For *Approach-B*, using DB-Data set snapshot-1, snapshot-2, snapshot-3 were implied to use such data sets to predict the collaboration in the next 3, 2, and 1 years, respectively. From the results in Table 4.4(b), the third snapshot which contained papers published between years $2002-2007$ was the best for recommending the collaborators in years $2008-2009$. On the other hand, using data set in years $2000-2005$ provided the lowest accuracy because it was the oldest data set. From this obvious results, the accuracy of recommending strongly depended on up-to-date data. Thus, papers in DB-Data snapshot-3 were applied for all experiments in this dissertation.

Although it seem the accuracy of non-friends is small with 0.294, the average in neighbors with 0.609 is quite high compared with 0.25 which may due to the observation in Section 4.2. Thus, the accuracy of the hybrid model with 0.609 in Table 4.5 is dramatically higher than 0.25.

4.4.2 Results of the Different Relevance Score Methods

As proposed in Section 3.5, the raw relevance score $\mathcal{W}^{\Delta}_{v_i \to v_k}$ can be computed in the three alternative methods shown in equation (3.29), equation (3.30), and equation (3.31). In this experiment, the different accuracy compared among three methods were evaluated.

The results in Table 4.5 shows the hybrid method in equation (3.31) is the best for friends, non-friends, and neighbors. For all types of friendship, using only the structure approach in method-1 gave more accuracy than using only the semantic approach in method-2. The best approach used for recommending is the hybrid method based on both structure approach and semantic approach in method-3. For the remaining experiments in this dissertation, the relevance score method-3 was applied for recommending.

Table 4.5 The accuracy of experiments compared among three methods of the relevance scores.

Methods	Models	Equations	Friends	Non-friends	Neighbors
1	SC-CoT (structure)	(3.29)	0.781	0.246	0.568
2	SS-BaT (semantic)	(3.30)	0.735	0.227	0.539
3	hybrid (structure & semantic)	(3.31)	**0.822**	**0.294**	**0.609**

4.4.3 Results of Semantic Approach Based on SS-BaT Model

In this section, the efficiency of semantic approach based on SS-BaT model were evaluated. In Section 3.4.5, the semantic similarity of v_k with respects to v_i, denoted by $\mathcal{M}_{v_i \to v_k}$, can be calculated based on researcher background similarity, $\mathcal{B}_{v_i,v_k}$ and the estimated probability of research topics, Θ_{v_i}. The power research trends effect to the accuracy of SS-BaT model will be observed in this experiment. Instead of using equation (3.30) based on $\mathcal{M}_{v_i \to v_k}$, this experiment used only the researcher background similarity, $\mathcal{B}_{v_i,v_k}$, in equation (4.2) to calculate the relevance scores as follows.

$$\mathcal{W}^{\Delta}_{v_i \to v_k} = \frac{\mathcal{B}_{v_i \to v_k} \times d(v_i, v_k)}{\sum\limits_{v_k \in R(v_i)} \mathcal{B}_{v_i \to v_k} \times d(v_i, v_k)} \,. \tag{4.2}$$

The results in Table 4.6 show the relevance scores calculated based on semantic similarity without trend give the lower accuracy compared with the accuracy of SS-BaT model. So, both research background and research trend should be combined for collaborators recommendation based on semantic approach.

Table 4.6 The accuracy of semantic approach.

Models	Equations	Friends	Non-friends	Neighbors
Semantic similarity without trend	(4.2)	0.603	0.124	0.443
Semantic similarity with trend (SS-BaT)	(3.30)	0.735	0.227	0.539

4.4.4 Results Compared with Other Methods

Since no prior work has been proposed for directly overcome the problem of collaborators recommendation, the existing related works were selected for comparison. There are three related works used in here, i.e., Han et al.'s work [29], Liu et al.'s work [11], and Sachan's work [34].

Han et al.'s work proposed the supportiveness measure in co-authorship network structure. The supportiveness from researcher v_k to v_i is used to measure how close the collaboration from v_k to v_i. This approach used only the frequency of co-authorship for computing the closeness based on directed relationship. Since the closeness between un-linked researcher pairs cannot be determined, this work will be used for comparing in case of recommending friend researchers for v_i.

Liu et al.'s work proposed to determine the magnitude of the link between two researchers based on co-authorship network structure. They used the frequency of co-authorship and the number of authors in a paper for computing the weights of directed relationship. This work also used for comparing in friends researchers recommending .

The last one is Sachan's work. This work proposed a supervised model based on both co-authorship network structure and semantic of the content of papers. The main objective of work is to overcome link prediction problem which is a bi-directed relationship. Based on the definition of link prediction in Section 2.2.2, this work was used for comparing in non-friends recommending who never been collaborated together.

In case of recommending friend, the average accuracy (mean) of the proposed method was compared with the first two methods. In the same way, the proposed method was compared with the method of Sachan's work. The significance of differences between two means can be assessed using the *t*-test. The *t*-test is one of the most commonly used hypothesis tests. It is applied to compare whether the average difference between two groups is really significant. This dissertation defines two hypotheses, the null and alternative hypotheses as follows:

H_0: The average accuracy of the given two models are the same, $\mu_1 = \mu_2$,

Table 4.7 The accuracy compared with other methods.

Methods	Means ($\overline{x}$) and Variances (S^2)				Difference of Two Means		
	Friends		Non-friends		$t(\overline{x_a}, \overline{x_b})$	Alpha	Critical
	$\overline{x}$	S^2	$\overline{x}$	S^2		Value	Value
Proposed method	**0.822**	0.102	0.294	0.156			
Han et al.'s work [29]	0.591	0.178			$t(0.822, 0.591) = 1.920$	0.1	$\pm$ 1.645
Liu et al.'s work [11]	0.625	0.169			$t(0.822, 0.625) = 1.648$	0.1	$\pm$ 1.645
Sachan's work [34]			0.092	0.234	$t(0.294, 0.092) = 1.542$	0.2	$\pm$ 1.282

H_1: The average accuracy of the given two models are not the same, $\mu_1 \neq \mu_2$.

The test statistics between two means denoted by $t(\overline{x_a}, \overline{x_b})$ where $\overline{x_a}$ is the mean of the proposed method and $\overline{x_b}$ is the mean of another model are calculated by

$$t(\overline{x_a}, \overline{x_b}) \quad = \quad \frac{\overline{x_a} - \overline{x_b}}{\sqrt{\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}}} \, . \tag{4.3}$$

S_a^2 and S_b^2 are the variance of the two models and n_a and n_b are the number of test cases. Since the standard deviations of the models are not equal, the separate variance *t*-test is used to statistical test. The degrees of freedom that define the specific *t*-distribution for this is given by

$$df \quad = \quad \frac{\left(\frac{S_a^2}{n_a} + \frac{S_b^2}{n_b}\right)^2}{\frac{\left(\frac{S_a^2}{n_a}\right)^2}{n_a - 1} + \frac{\left(\frac{S_b^2}{n_b}\right)^2}{n_b - 1}} \, . \tag{4.4}$$

The results of friends in Table 4.7 show the accuracy of Liu et al.'s work which utilized from both the frequency of co-authorship and the number of authors in a paper is outperformed Han et al.'s work which used only the frequency of co-authorship. When the `SS-BaT` model including year of publications were involved in Liu et al.'s s work liked in the `SC-CoT` model, it gave the highest accuracy of 0.822. These showed the year of publication, power of researchers analysis, common friends analysis, and semantic meaning played the important role for determining the re-collaboration between friend pairs in the near future. To compare whether the mean between the proposed method and another work is really significant, the value of $t(\overline{x_a}, \overline{x_b})$ are shown in column difference of two means. Firstly, the *t*-value between the proposed method and Han et al.'s work is 1.920. The critical value of t is set with $df = 2,551$ and alpha $= 0.1$ is ± 1.645. Since the value 1.920 is greater than the

critical value 1.645, and falls into the critical region, the null hypothesis of equal means is rejected. The average accuracy of proposed method is not equal to the average accuracy of Han et al.'s work at a 90% confidence level. Secondly, the *t*-value between the proposed method and Liu et al.'s work is 1.648 The critical value of t is set with $df = 2,551$ and alpha $= 0.1$ is ± 1.645. Since the value 1.648 is greater than the critical value 1.645, and falls into the critical region, the null hypothesis of equal means is rejected. The average accuracy of proposed method is not equal to the average accuracy of Liu et al.'s work at a 90% confidence level.

For non-friends recommendation, the accuracy of the proposed hybrid method based on `SC-CoT` model and `SS-BaT` model is outperformed Sachan's work. The *t*-value between the proposed method and Sachan's work is 1.542. The critical value of t is set with $df = 2,551$ and alpha $= 0.2$ is ± 1.282. Since the value 1.542 is greater than the critical value 1.282, and it falls into the critical region, the null hypothesis of equal means is rejected. The average accuracy of proposed method is not equal to the average accuracy of Sachan's work at a 80% confidence level.

The accuracy of Sachan's work is very low, 0.092, when applied in recommendation problem. A reason of such low accuracy is the output the work represented in bi-directed relationship of link prediction. For more explanations, consider the four propositions as follows.

$$\begin{aligned} \text{proposition-}A &= \text{“}v_i \text{ selects } v_k\text{”} \\ \text{proposition-}B &= \text{“}v_k \text{ selects } v_i\text{”} \\ \text{proposition-}C &= A \wedge B \\ \text{proposition-}D &= \text{“}v_i \text{ and } v_k \text{ are linked”} \end{aligned}$$

Suppose a paper in test data is written by v_i and v_k. The test cases in recommendation task are separated in two cases, i.e., proposition-A, and proposition-B. Let proposition-C represent a symmetrical relationship, i.e., (v_i selects v_k) AND (v_k selects v_i). On the other hand, link prediction task concentrates on a link between v_i and v_k is present as shown in proposition-D, regardless any direction. In order to derive directed relationship in recommendation task to bi-directed relationship in prediction task, the proposition "(v_i selects v_k) AND (v_k selects v_i) equivalent to (v_i and v_k are linked)" is assumed. If the recommendation model recommends v_k to v_i, but not recommends v_i to v_k, the accuracy for recommendation task is $1/2 = 0.5$. In link prediction task, proposition-B is false because v_k selects v_i, but v_i is not recommended to v_k. So, proposition-D is also false and the predictor is false. The accuracy for link prediction task is $0/2 = 0$. Therefore, the chance of right recommendation is higher than the chance of right prediction. Moreover, one advantage of the proposed methods over Sachan's work is the work have to re-balance data set to prevent over fitting problem based on supervised learning model. Since the hybrid model is un-supervised method based on `SC-CoT` and `SS-BaT`,

Table 4.8 The accuracy of core authors which separated to senior authors and junior authors.

Inquired Researchers	Friends	Non-friends	Neighbors
Senior authors	0.762	0.198	0.563
Junior authors	0.831	0.315	0.622
Core authors	0.822	0.294	0.609

the model accepts imbalanced data set without any re-balancing. In summary, Sachan's work may be not suitable for overcoming the recommendation problem. It was designed only for link prediction task based on bi-directed relationship, not for directed relationship in recommendation task.

4.4.5 Results of the Different Inquired Researchers

As mention in Section 3.2, researchers who appeared in both DB-Data set and Test-Data set are called core authors. These are assumed that a set of inquired researchers would like to find their collaborators. The core authors can be divided into two types grouped by research experiences. The first type is senior authors which defined in the step of data collection as researchers who published at least ten papers in a particular topic. Another type is junior authors which is core authors who are not senior authors. The objective of experiments in this section is to observe the accuracy in two viewpoints of requirements: (a) recommend collaborators for senior authors and (b) recommend collaborators for junior authors.

For the first experiment, only senior authors are used as inquired researchers. Table 4.8 shows that the accuracy of recommending the collaborators to senior authors is lower than the accuracy of recommending collaborators to core authors. Note that results of core authors are the accuracy of both senior authors and junior authors which correspond to the results of hybrid model in Table 4.5. The reason behind these results is a particular senior researcher has a lot of published papers and he also has a lot of friends. Because the model has to select the suitable collaborators from a large number of his candidate friends, the possibility to select the right recommended researchers is small. In the same manner, the number of non-friends, i.e., friends of friends, proportionally grows in the number of friends, and the possibility to select the right recommended researchers is also small.

In case of recommending collaborators for junior researchers, the accuracy is higher than the accuracy of core authors. The reason is that a junior researcher has published a few papers, and the number of his friends and non-friends are small. It is not difficult to recommend a potential collaborators from a small set of candidate researchers.

Table 4.9 The accuracy of recommendation over six degrees of separation

Topics	Degree of Separation							Neighbors	Average Collaborators
	Friends	Non-friends							
	1	2	3	4	5	6	Average		
Bio-informatics	0.822	0.348	0.096	0.071	0.033	0.050	0.203	0.617	5.78
Data Mining	0.767	0.300	0.108	0.051	0.050	0.050	0.159	0.441	8.85
Hardware	0.783	0.304	0.123	0.097	0.067	0.042	0.180	0.501	5.73
Neural Network	0.928	0.842	1.000	0.500	0.500	0.167	0.732	0.895	4.11
Software	0.839	0.316	0.132	0.038	0.067	0.250	0.206	0.604	5.54
Theory	0.792	0.409	0.192	0.083	0.083	0.143	0.285	0.597	4.14
Average	0.822	0.420	0.275	0.140	0.133	0.117	0.294	0.609	5.70

4.4.6 Results of Six Degrees of Separation Over Six Topics

Now, the accuracy of recommendation is finely analyzed in a particular topic. Table 4.9 shows the accuracy of recommending friends based on the hybrid method. The accuracy are separately considered over the six topics. The highest accuracy values of friends, non-friends, and neighbors achieved by a neural network are 0.928, 0.732, and 0.895, respectively. On the other hand, the lowest accuracy is data mining. After the number of friends in DB-Data set is checked, the average number of friends is 4.11 which is lowest. Since the candidates researchers will be selected for recommending is the lowest, the possibility to select the right recommended researchers is highest when being compared with other topics. Therefore, the accuracy of recommendation is inversely proportional to the number of collaborators (friends) in the previous collaboration. In the same fashion, the hardest topic for recommending is data mining since the highest average collaborators is 8.85.

Moreover, Table 4.9 shows the further distance between two researchers with less accuracy. The last line of table shows the average accuracy of each degree of separation. The highest accuracy of recommending neighbors in the first degree (friend) is 0.822. The accuracy is decreased when the degree of separation is increased. So, the accuracy of recommending neighbor in the second degree is 0.420 whereas the sixth degree is 0.117. The reason is that the number of neighbors is increased when the degree of separation is increased. Therefore, the possibility to select the right recommended researchers is small, with respected to a large number of neighbors.

CHAPTER V

CONCLUSION AND FUTURE WORK

5.1 Conclusion

This dissertation presented a methodology for research collaborators recommendation i.e. a structure approach called *Structure Cohesion based on Collaboration Over Time* (SC-CoT) model and a semantic approach called *Semantic Similarity with Background and Trend of Research* (SS-BaT) model. The first model analyzed cohesion weights underlying the structure of co-authorship network evolution while the second model focused on semantic similarity analyzed from the content of papers. To evaluate the efficiency of recommendation, the models were examined and the results were summarized in various aspects as follows.

1. A likely topic of the latest year of inquired researcher's publication obtained from author-topic model can be used to assume as his inquired topic and his current research topic with confidence of 85%.

2. The accuracy of strong recommendation depends on the up-to-date data. Time evolution underlining the years of publication plays an important role for determining the cohesion in SC-CoT model and the semantic similarity in SS-BaT model. To overcome the collaborators recommendation problem, the recent data should be used as much as possible.

3. Recommending the collaborators considers the cohesion weights based on structure of co-authorship network proposed in SC-CoT model and the meaning of papers from SS-BaT model. In the same fashion, the hybrid method of both SC-CoT model and the meaning of papers from SS-BaT model should be comprehensively taken into consideration for recommending the researchers who do not join together.

4. The accuracy of recommending the collaborators having collaborated based on the hybrid method is 0.822 while the accuracy of other existing methods is about 0.625. The average accuracy of proposed method is not equal to the average accuracy of existing method at a 90% confidence level. Furthermore, the accuracy of recommending the collaborators not having collaborated is 0.294 whereas the accuracy of other existing method is 0.092. The average

accuray of proposed method is not equal to the average accuracy of existing method at a 80% confidence level.

5. Recommending collaborators for the senior researchers having more published papers is more difficult than recommending collaborators for the junior researchers with a few published papers.

6. The accuracy of recommendation is inversely proportional to the number researchers having collaborations in the past period. The hardest topic for recommending is data mining with the accuracy of 0.441 since the number of average collaborators in data mining is 8.85. In contrast, the neural network is the most easy topic with an accuracy 0.895 since the number of average collaborators is 4.11.

7. The higher degree of separation between two researchers gives the less accuracy. The accuracy of recommending neighbors in the first degree (friend) is the highest with 0.822. The accuracy is decreased when the degree of separation is increased. So the accuracy of recommending neighbor in the second degree is 0.420 until the sixth degree is 0.117.

From the above evidences, the proposed models are expected to be useful for finding the suitable collaborators for a given researcher. Moreover, the proposed the `SC-CoT` model and the `SS-BaT` model can be separately applied or incorporated to other applications. The `SC-CoT` can investigate the social network over time. The `SS-BaT` can help the works related to the meaning of research papers such as expertise searching and topic clustering.

5.2 Future Work

- Topic model processing. One limitation of the author-topic model is that it takes several days for learning the data set. Moreover, the classical author-topic model cannot capture the research knowledge over time. Hence, the year of publication should be involved as a parameter in the learning process. Nevertheless, other topic modeling may be used for discovering the research knowledge instead of author-topic model.

- Reducing the candidate researchers. Intuitively, the computational time is proportionally grown with the number of candidate researchers. Using higher degree of separation makes more the number of candidate potential researchers and more opportunity to meet the satisfied collaborators. At the same time, it also spends more computational time. Thus, a pruning algorithm should be pre-processed for getting rid of the irrelevance candidates.

REFERENCES

[1] Steyvers, M., Smyth, P., Rosen-Zvi, M., and Griffiths, T. L., Probabilistic Author-Topic Models for Information Discovery. KDD, (2004): 306-315.

[2] Wasserman, S. and Faust, K. Social Network Analysis: Methods and Applications. New York, USA : Cambridge University Press, 1994.

[3] Carré, B. Graphs and Networks. Oxford University Press, 1979.

[4] Otte, E. and Rousseau, R. Social Network Analysis: A Powerful Strategy, Also for the Information Sciences. Journal of Information Science 28 , 6 (2002): 441–453.

[5] Chakrabarti, D. and Faloutsos, C. Graph Mining: Laws, Generators, and Algorithms. ACM COMPUTING SURVEYS 38 , 1 (2006): 2.

[6] Travers, J. and Milgram, S. An Experimental Study of the Small World Problem. Sociometry 32 , 4 (1969): 425-443.

[7] Watts, D. J. Six Degrees: The Science of a Connected Age. W. W. Norton & Company, 2004.

[8] Aiello, W. and Chung, F., Random Evolution in Massive Graphs. in Proceedings of the 42nd IEEE symposium on Foundations of Computer Science (FOCS '01), (2001): 510.

[9] Nascimento, M. A., Sander, J., and Pound, J. Analysis of SIGMOD's Co-Authorship Graph. SIGMOD Rec. 32 (2003): 8–10.

[10] Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., and Sødring, T. Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of A Century?. SIGIR Forum 37 (2003): 49–53.

[11] Liu, X., Bollen, J., Nelson, M. L., and Sompel, H. Co-authorship Networks in The Digital Library Research Community. Inf. Process. Manage. 41 (2005): 1462–1480.

[12] Huang, J., Zhuang, Z., Li, J., and Giles, C. L., Collaboration Over Time: Characterizing and Modeling Network Evolution. in Proceedings of the International Conference on Web Search and Web Data Mining (WSDM '08), New York, NY, USA, ACM, (2008): 107–116.

[13] Newman, M. Scientific collaboration networks. I. Network Construction and Fundamental Results. Physical Review E 64 , 1 (2001): 016131.

[14] Acedo, F. J., Barroso, C., Casanueva, C., and Galán, J. L. Co-Authorship in Management and Organizational Studies: An Empirical and Network Analysis. Journal of Management Studies 43 , 5 (2006): 957-983.

[15] Barabási, A. L. *et al.* Evolution of the Social Network of Scientific Collaborations. Physica A: Statistical Mechanics and Its Applications 311 , 3-4 (2001): 590 – 614.

[16] Moody, J. The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. American Sociological Review 69 , 2 (2004): 213–238.

[17] Griffiths, T. L. and Steyvers, M. Finding Scientific Topics. in Proceedings of the National Academy of Sciences 101 , Suppl. 1 (2004): 5228-5235.

[18] Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P., The Author-Topic Model for Authors and Documents. in Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04, Arlington, Virginia, United States, AUAI Press, (2004): 487–494.

[19] Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P., and Steyvers, M. Learning Author-Topic Models from Text Corpora. ACM Trans. Inf. Syst. 28 (2010): 4:1–4:38.

[20] Hellstrm, T. and Holmstrm, K., Predicting the Stock Market. tech. rep.

[21] Holt, C. C. Forecasting Seasonals and Trends by Exponentially Weighted Moving Averages. International Journal of Forecasting 20 , 1 (2004): 5 – 10.

[22] Newman, M. E. Scientific collaboration networks. II. Shortest Paths, Weighted Networks, and Centrality.. Physical Review E - Statistical, Nonlinear and Soft Matter Physics 64 , 1 Pt 2 (2001): 016132.

[23] Vázquez, A. Growing Network with Local Rules: Preferential Attachment, Clustering Hierarchy, and Degree Correlations. Phys. Rev. E 67 , 5 (2003): 056104.

[24] Elmacioglu, E. and Lee, D. On Six Degrees of Separation in DBLP-DB and More. SIGMOD Rec. 34 (2005): 33–40.

[25] Sharma, M. and Urs, S. R., Small World Phenomenon and Author Collaboration: How Small and Connected is the Digital Library World?. in Proceedings of the 10th International Conference on Asian Digital Libraries: Looking Back 10 Years and Forging New Frontiers, ICADL'07, Berlin, Heidelberg, Springer-Verlag, (2007): 510–511.

[26] Bird, C. *et al.*, Structure and Dynamics of Research Collaboration in Computer Science. SDM, SIAM, (2009): 826–837.

[27] Franceschet, M. The Small World of Computer Science. CoRR abs/1010.4747.

[28] Franceschet, M. Collaboration in Computer Science: A Network Science Approach. Part II. CoRR abs/1104.4296.

[29] Han, Y., Zhou, B., Pei, J., and Jia, Y., Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach. in Proceedings of the SIAM International Conference on Data Mining, (2009): 1111-1122.

[30] Liben-Nowell, D. and Kleinberg, J. The Link-Prediction Problem for Social Networks. J. Am. Soc. Inf. Sci. Technol. 58 (2007): 1019–1031.

[31] Potgieter, A., April, K. A., Cooke, R. J. E., and Lockett, M. Adaptive Bayesian Agents: Enabling Distributed Social Networks. South African Journal of Business Management 37 (2006): 41–55.

[32] Potgieter, A., April, K. A., Cooke, R. J. E., and Osunmakinde, I. O. Temporality in Link Prediction: Understanding Social Complexity. Emergence: Complexity & Organization 11 , 1 (2009): 83–96.

[33] Backstrom, L. and Leskovec, J., Supervised Random Walks: Predicting and Recommending Links in Social Networks. in Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM '11), New York, NY, USA, ACM, (2011): 635–644.

[34] Sachan, M. and Ichise, R. Using Semantic Information to Improve Link Prediction Results in Network Datasets. International Journal of Engineering 2 , 4 (2010): 334–339.

[35] Su, Y.-M., Yang, S.-C., Hsu, P.-Y., and Shiau, W.-L. Extending Co-Citation Analysis to Discover Authors with Multiple Expertise. Expert Syst. Appl. 36 (2009): 4287–4295.

[36] Zaïane, O. R., Chen, J., and Goebel, R., Mining Research Communities in Bibliographical Data.. in Proceeding of WebKDD/SNA-KDD 5439 (2007): 59-76.

[37] Tang, J., Zhang, D., and Yao, L., Social Network Extraction of Academic Researchers. in Proceedings of the 7th IEEE International Conference on Data Mining, Washington, DC, USA, IEEE Computer Society, (2007): 292–301.

[38] Tang, J., Zhang, J., Yao, L., *et al.*, ArnetMiner: Extraction and Mining of Academic Social Networks. in Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '08, New York, NY, USA, ACM, (2008): 990–998.

[39] Balog, K., Azzopardi, L., and Rijke, M., Formal Models for Expert Finding in Enterprise Corpora. in Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '06, New York, NY, USA, ACM, (2006): 43–50.

[40] Tang, J., Zhang, J., Jin, R., *et al.* Topic Level Expertise Search Over Heterogeneous Networks. Mach. Learn. 82 (2011): 211–237.

[41] Zhang, J., Tang, J., and Li, J., Expert Finding in a Social Network. Advances in Databases: Concepts, Systems and Applications 4443 (2010): 1066-1069.

[42] Daud, A., Li, J., Zhou, L., and Muhammad, F. Temporal Expert Finding Through Generalized Time Topic Modeling. Knowledge-Based Systems 23 , 6 (2010): 615 – 625.

[43] Milgram, S. The Small World Problem. Physiology Today 2 (1967): 60–67.

[44] Cormen, T. H., Leiserson, C. E., Rivest, R. L., and Stein, C. Introduction to Algorithms. The MIT Press, 2nd ed., 2001.

[45] Newman, M. E. J. Clustering and Preferential Attachment in Growing Networks. Phys. Rev. E 64 (2001): 025102.

[46] Lü, L., Jin, C.-H., and Zhou, T. Similarity Index Based on Local Paths for Link Prediction of Complex Networks. Phys. Rev. E 80 , 4 (2009): 046122.

[47] Alexandridis, K., Coe, K., and Garnett, S. Semantic Analysis of Natural Language Processing in A Study of Nurse Mobility in the Northern Territory, Australia. Journal of Population Research 27 (2010): 15-42.

[48] Chaiwanarom, P., Ichise, R., and Lursinsap, C., Finding Potential Research Collaborators in Four Degrees of Separation. in Proceedings of the 6th International Conference on Advanced Data Mining and Applications, Berlin, Heidelberg, Springer-Verlag, (2010): 399–410.

[49] Adamic, L. and Adar, E. Friends and Neighbors on the Web. Social Networks 25 , 3 (2003): 211–230.

[50] Salton, G. and Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. Inf. Process. Manage. 24 (1988): 513–523.

[51] Manning, C. D. and Schuetze, H. Foundations of Statistical Natural Language Processing. The MIT Press, 1 ed., June 1999.

[52] Yoshikane, F., Nozawa, T., Shibui, S., and Suzuki, T. An Analysis of the Connection Between Researchers' Productivity and Their Co-authors' Past Attributions, Including the Importance in Collaboration Networks. Scientometrics 79 , 2 (2009): 435-449.

Appendix

Appendix

List of Publications

- Paweena Chaiwanarom, Ryutaro Ichise, and Chidchanok Lursinsap, "Finding Potential Research Collaborators in Four Degrees of Separation", in Proceedings of the 6^{th} International Conference on Advanced Data Mining and Applications (ADMA2010), Chongqing, China, (2010): 399-410.

- Paweena Chaiwanarom and Chidchanok Lursinsap, "Combining Missing Relative Distance and Information Retrieval Technique for Querying Missing Author in Directed Authorship Graph", in Proceedings of the International Conference on Computer and Automation Engineering (ICCAE 2009), Bangkok, Thailand, (2009): 36-40.

- Paweena Chaiwanarom and Chidchanok Lursinsap, "Comparison of Proposed PPMM with Other PPM Methods for Link Completion Problem", in Proceedings of the International Conference on Advanced Computer Theory and Engineering (ICACTE 2008), Phuket, Thailand, (2008): 341-345.

- Paweena Chaiwanarom and Chidchanok Lursinsap, "Link Completion using Prediction by Partial Matching", in Proceedings of the International Symposium on Communications and Information Technologies (ISCIT 2008), Vientiane, Laos, (2008): 675-680.

- Ohm Sornil and Paweena Chaiwanarom, "Combining Prediction by Partial Matching and Logistics Regression for Thai Word Segmentation", in Proceedings of the 20^{th} International Conference on Computational Linguistics (COLING 2004), Geneva, Switzerland, (2004).

- Ohm Sornil and Paweena Chaiwanarom, "Thai Syllable Segmentation Using Prediction by Partial Matching", in Proceedings of The Pacific Association for Computational Linguistics (PACLING'03), Halifax, Nova Scotia, Canada, (2003).

Biography

Personal Details:

Name: Miss Paweena Chaiwanarom

Date of Birth: March 28, 1978

Place of Birth: Bangkok, Thailand

Education:

- Ph.D. Program in Computer Science, Chulalongkorn University, Thailand (October 2006 - October 2011).
- Internship Student, National Institute of Informatics (NII), Japan (February 2010 - July 2010).
- M.Sc. Program in Computer Science, National Institute of Development Administration (NIDA), Thailand (June 2001 - March 2004).
- B.Sc. Program in Computer Science (2^{nd} Class Honors), King Mongkut's University of Technology North Bangkok, Thailand (June 1998 - September 2000).

Work Experiences:

- Lecturer, Rajamangala University of Technology Rattanakosin, Thailand (June 2004 - present).
- Computer Technical Officer, National Statistical Office, Thailand (March 2003 - May 2004).

Awards:

- Consolation Award (Information Technology Track), National Institute of Development Administration (NIDA) Research Award, Thailand (2005).

Scholarships:

- Teaching Assistant, Department of Computer Science, National Institute of Development Administration (NIDA), Thailand (2002).