



DISSERTATION – EE186601

**Modeling Scholar Profile in
Expert Recommendation based on
Multi-Layered Bibliographic Graph**

**Diana Purwitasari
07111660010201**

**Supervisors:
Prof. Dr. Ir. Mauridhi Hery Purnomo, M.Eng.
Dr. Surya Sumpeno, ST., M.Sc.
Dr.Eng. Chastine Fatichah, S.Kom., M.Kom.**

**Doctoral Program
Department of Electrical Engineering
Faculty of Intelligent Electrical and Informatics Technology
Institut Teknologi Sepuluh Nopember
Surabaya 2020**

STATEMENT OF APPROVAL

DISSERTATION

Modeling Scholar Profile in Expert Recommendation based on Multi-Layered Bibliographic Graph

This dissertation was prepared to fulfil one of the requirements for obtaining
a Doctoral Degree (Dr.)
at Institut Teknologi Sepuluh Nopember

Diana Purwitasari
NRP. 07111660010201

07111660012001

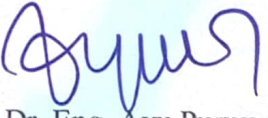

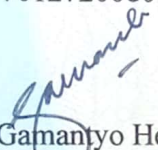
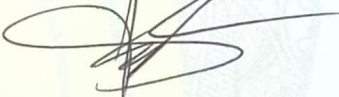

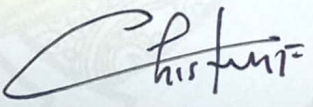
Examination date: January 31st, 2020

Graduation period: March, 2020

Approved/ Accepted by:

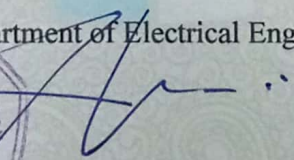
Examiners

Supervisors

- | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 
1. Dr. Eng. Ayu Purwarianti, ST.,
MT.
NIP. 197701272008012011 | 
1. Prof. Dr. Ir. Mauridhi Hery Purnomo,
M.Eng.
NIP. 195809161986011001 |
| 
2. Prof. Ir. Gamantyo Hendratoro,
M.Eng, Ph.D.
NIP. 197011111993031002 | 
2. Dr. Surya Sumpeno, ST., M.Sc.
NIP. 196906131997021003 |
| 
3. Dr. I Ketut Eddy Purnama, ST.,
MT.
NIP. 196907301995121001 | 
3. Dr.Eng. Chastine Fatichah, S.Kom.,
M.Kom.
NIP. 197512202001122002 |

Head of Department of Electrical Engineering




Dedet Candra Riawan, S.T., M.Eng., Ph.D.
NIP. 197311192000031001

STATEMENT OF ORIGINALITY


I hereby certify that any and all parts of my dissertation with the title

**"MODELING SCHOLAR PROFILE IN EXPERT RECOMMENDATION
BASED ON MULTI-LAYERED BIBLIOGRAPHIC GRAPH"**

is actually an independent intellectual work, is accomplished without the use of materials that are not allowed, and is not the work of others whom I consider as my own work.

All references cited or referenced have been written completely in the bibliography. If it turns out that I violate this statement, I am willing to accept any penalty in accordance with applicable regulations.

Surabaya, January 27th 2020

A handwritten signature in black ink, appearing to read 'Diana Purwitasari', written over a horizontal line.

Diana Purwitasari

ABSTRACT
**Modeling Scholar Profile in Expert Recommendation based on
Multi-Layered Bibliographic Graph**

By : Diana Purwitasari
Student Identity Number : 07111660010201
Supervisors : Prof. Dr. Ir. Mauridhi Hery Purnomo, M. Eng.
Dr. Surya Sumpeno, ST., M.Sc.
Dr.Eng. Chastine Fatichah, S.Kom., M.Kom.

A recommendation system requires the profile of researchers which called here as Scholar Profile for suggestions based on expertise. This dissertation contributes on modeling unbiased scholar profile for more objective expertise evidence that consider interest changes and less focused on citations. Interest changes lead to diverse topics and make the expertise levels on topics differ. Scholar profile is expected to capture expertise in terms of productivity aspect which often signified from the volume of publications and citations. We include researcher behavior in publishing articles to avoid misleading citation. Therefore, the expertise levels of researchers on topics is influenced by interest evolution, productivity, dynamicity, and behavior extracted from bibliographic data of published scholarly articles. As this dissertation output, the scholar profile model employed within a recommendation system for recommending productive researchers who provide academic guidance.

The scholar profile is generated from multi layers of bibliographic data, such as layers of author, topic, and relations between those layers to represent academic social network. There is no predefined information of topics in a cold-start situation, such that procedures of topic mapping are necessary. Then, features of productivity, dynamicity and behavior of researchers within those layers are taken from some observed years to accommodate the behavior aspect. We experimented with AMiner dataset often used in the following bibliographic data related studies to empirically investigate: (a) topic mapping strategies to obtain interest of researchers, (b) feature extraction model for productivity, dynamicity, and behavior aspects based on the mapped topics, and (c) expertise rank that considers interest changes and less focused on citations from the scholar profile. Ensuring the validity results, our experiments worked on standard expert list of AMiner researchers. We selected Natural Language Processing and Information Extraction (NLP-IE) domains because of their familiarity and interrelated context to make it easier for introducing cases of interest changes. Using the mapped topics, we also made minor contributions on transformation procedures for visualizing researchers on maps of Scopus subjects and investigating the possibilities of conflict of interest.

Keywords: modeling scholar profile, bibliographic data for academic social network, productivity and dynamicity features, behavior-based features, expertise rank

ACKNOWLEDGEMENTS

Al-hamdu lillahi rabbil 'alamin, praise belongs to Allah, the Lord of the worlds, and His blessing for giving me the strength, the chance, and endurance to complete this dissertation for the Doctoral Program in the Department of Electrical Engineering, Faculty of Intelligent Electrical and Informatics Technology, Sepuluh Nopember Institute of Technology (ELECTICS – ITS), as a part of a grand design research at Multimedia Computing Laboratory.

I would like to express my appreciation to numerous people who have helped me in invaluable ways to complete my research works over the last three years.

First and foremost, I would like to sincerely thank my supervisors Prof. Mauridhi Hery, Dr. Chastine Fatichah, and Dr. Surya Sumpeno. I cannot thank them enough for all their guidance personally and professionally, understanding, patience and most importantly, positive encouragement along with a warm spirit that they have shared me to finish this dissertation. It has been a great pleasure and honor to have them as my supervisors.

Appreciation and gratitude I also extend to a number of institutions for supporting my doctoral program: the Indonesia Endowment Fund for Education (LPDP) that provides an opportunity through the Indonesian education scholarships and Ministry of Research and Higher Education Indonesia through the sandwich-like scholarship (*Peningkatan Kualitas Publikasi Ilmiah*, PKPI-2018) in Groningen, the Netherlands; the management of the Department of Informatics for giving study leave permission and moral supports; the management and staff at the Department of Electrical Engineering who have provided administrative support and facilities during my research works.

I thank the examiners, Dr. Ayu Purwarianti, Prof. Gamantyo Hendrantoro, and Dr. I Ketut Eddy Purnama, who gave constructive comments and challenged my thinking with questions and assumptions, and viewed issues from multiple perspectives for the

perfection of this dissertation. I would also like to extend enormous gratitude to Prof. (Bart) Verkerke and Dr. Christian Steglich for their supports and patience during PKPI program.

To my colleagues, students of Doctoral and Master's programs under the guidance of Prof. Mauridhi Hery in B204 Lab and colleagues of Department of Informatics that I could not mention one by one, great kudos for your friendship, collaboration, fun, and supportive environment for insightful discussions, aside of all advices and non-academic chit-chat during interesting lunch plus coffee-break talks. They truly made me enjoy the last three years as a student again.

Last but not least, my deepest gratitude goes to all of my beloved family members for their never-ending support all along the way throughout this challenging work. Thank you for your understanding, care, and patience in each time I need rushing to meet deadlines during this dissertation program, which is quite a bit due to some inevitable procrastination moments.

To those who indirectly contributed in this research, your kindness means a lot to me. As a closing, although this is still far from being perfect, therefore, I loved any suggestions given about feedback to enhance this research. May Allah gives blessing upon knowledge for whoever read this dissertation. I sincerely hope that you enjoy this dissertation and find that it meets your needs. The same expectation is brought, so this dissertation could contribute in promoting the progress of science, advancing the national research productivity, and eventually to yield another solutions to real-life problems

Aamiin.

Surabaya, January 27th 2020

Diana Purwitasari

TABLE OF CONTENTS

STATEMENT OF APPROVAL	i
STATEMENT OF ORIGINALITY.....	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS	v
TABLE OF CONTENTS	vii
LIST OF FIGURES	ix
LIST OF TABLES.....	xi
LIST OF ABBREVIATIONS, TERMS AND SYMBOLS.....	xiii
Chapter 1. INTRODUCTION	1
1.1. Research Background.....	1
1.1.1. Interest changes influence research expertise	2
1.1.2. Rank expertise based on research domain	3
1.2. Problem Statement	4
1.3. Research Objectives and Benefits	5
1.4. Research Roadmap.....	6
1.5. Original Contributions.....	6
1.6. List of Publications	8
1.7. Research Scope and Limitation	9
1.8. Book Structure	10
Chapter 2. BIBLIOGRAPHIC DATA FOR ACADEMIC SOCIAL NETWORK.....	11
2.1. Academic Search Systems.....	11
2.2. Modeling Bibliographic Data	14
2.3. Expertise Rank.....	18
2.4. Recommendations with Considering Interest Changes	20
2.5. Visualizing Bibliographic Data.....	24
2.6. Summary	26
Chapter 3. MULTI-LAYERED BIBLIOGRAPHIC GRAPH FOR MODELING	
SCHOLAR PROFILE.....	27
3.1. Abstracting Multi-Layered Bibliographic Graph.....	27
3.2. Research Framework.....	28
3.3. Data Acquisition and Preparation	31
Chapter 4. CLUSTERING FOR IDENTIFYING TOPICS OF RESEARCHERS.....	33
4.1. Clustering with Various Word Embedding	33
4.2. Mapping Topics to Articles and Researchers in AMiner dataset.....	38
4.3. Evaluating Topics for Recommendations.....	40
4.3.1. Cross-Domain Collaborating for Researchers	42
4.3.2. Visualizing Researchers based on Topics.....	50

4.3.3. Extracting Conflict-of-Interest-based Features	55
4.4. Summary	60
Chapter 5. EXTRACTING PRODUCTIVITY-DYNAMICITY FEATURES	63
5.1. Data preparation.....	63
5.2. Extracting productivity features.....	64
5.3. Extracting dynamicity features	65
5.4. Selecting Productivity and Dynamicity Features	67
5.5.1. Feature selection with correlation test.....	67
5.5.2. Create validation dataset for expertise on topics.....	69
5.5. Summary	70
Chapter 6. EXTRACTING BEHAVIOR FEATURES	73
6.1. Graph Theories related to Researcher Representation	74
6.2. Extracting Exploration and Consistency features	76
6.3. Experiments Exploration Feature with Stochastic Actor-oriented Model (SAOM).....	79
6.4.1. Preparation for RSIENA.....	79
6.4.2. RSIENA Scripts	82
6.4.3. RSIENA Results	83
6.4.4. RSIENA Evaluation Functions	86
6.4. Summary	92
Chapter 7. EXPERTISE RANK USING SCHOLAR PROFILE	93
7.1. Features in Behavior-Based Scholar Profile.....	94
7.1.1. Extracting networks of author-topic related features	95
7.1.2. Extracting citation related features.....	96
7.2. Expertise Rank with Weighted-Sum Method	98
7.3. Expertise Rank with Linear Regression	99
7.4. Summary	105
Chapter 8. CONCLUSIONS AND FUTURE WORKS.....	107
REFERENCES	109
Appendix 1. RSIENA scripts for examining exploring feature.....	113
Appendix 2. Sample data of AMiner experts	117
Appendix 3. Weights for expertise rank with R package DecisionAnalysis	119
Appendix 4. Sample data for expertise rank for Query T2	121
Appendix 5. Sample results of expertise rank	123
Appendix 6. Mathematical functions for evaluating network evolution.....	125

LIST OF FIGURES

Figure 1-1	Group of studies in Scholar Profile for Expert Recommendation.....	6
Figure 1-2	Fishbone diagram of research contributions	7
Figure 1-3	Dataset used in this dissertation	9
Figure 2-1	Sample of SINTA scores for Institut Teknologi Sepuluh Nopember.....	12
Figure 2-2	Sample of SINTA networks for a researcher	12
Figure 2-3	Sample academic research data used in Microsoft Academic, Google Scholar, Scopus, and Aminer.....	13
Figure 3-1	Networks of one-mode (co-author) and two-mode (bipartite) abstracted from article metadata.....	28
Figure 3-2	Scholar Profile based on Multi-layered Bibliographic Graph	28
Figure 4-1	Research topics with keywords extracted using Latent Semantic Indexing from student thesis of Informatics Engineering	35
Figure 4-2	Clusters in AMiner NLP-IE domain transformed with LSI.....	36
Figure 4-3	Pseudo code for function MapArticleTopic()	39
Figure 4-4	Pseudo code for function MapResearcherTopic ().....	39
Figure 4-5	Matrix of researchers, topics, and article numbers as sources for extracting features	40
Figure 4-6	Approaches for intra- departmental recommendation system	41
Figure 4-7	Pseudo code for recommendation using model M1-M2.....	43
Figure 4-8	Pseudo code for recommendation using model M3-M4.....	44
Figure 4-9	Possible cross-domain collaborative studies using TF-IDF (M2).....	45
Figure 4-10	Possible cross-domain collaborative studies using Latent Semantic Indexing (M3)	45
Figure 4-11	Possible cross-domain collaborative studies using Word Vector (M4).....	45
Figure 4-12	Precision comparison of cross-domain collaboration for five faculties	48
Figure 4-13	K-Means Clustering result without Word2Vec.....	49
Figure 4-14	K-Means Clustering result with Word2Ve	49
Figure 4-15	System architecture for visualizing academic experts on a subject domain map of cartographic-alike.....	51
Figure 4-16	Visualization experts with subject domains.....	53
Figure 4-17	Pseudo code for coloring grids on the base map of Scopus subject areas	54
Figure 4-18	Conflict of interest indication based on Scopus trends.....	56
Figure 4-19	Illustration for conflict of interest indication	56
Figure 4-20	Pseudo code for obtain CoI1	57
Figure 4-21	Pseudo code for obtain CoI2	58
Figure 4-22	Pseudo code for obtain CoI3	58
Figure 5-1	Illustration of matrices from raw data to productivity and dynamicity features.....	67
Figure 5-2	t-SNE visualization of scaled data with labels from FCM approach.....	70

Figure 6-1	Transitive triad relation on a one-mode (co-author) network (left) and cycle relation on a two-mode (author-topic) network (right)	74
Figure 6-2	Process for extracting Exploration feature.....	77
Figure 6-3	Process for extracting Consistency feature	77
Figure 6-4	Topics with increasing popularities over time from AMiner NLP-IE dataset to illustrate research trends.....	78
Figure 6-5	Snippet of RSIENA script for assigning input	82
Figure 6-6	Snippet of RSIENA script for assigning networks.....	82
Figure 6-7	Snippet of RSIENA script for assigning effects.....	83
Figure 6-8	Sample of RSIENA output file from the observed model	85
Figure 6-9	Log-odds plot for co-author selection based on career age	87
Figure 6-10	Hindsight on co-authoring collaborations from AMiner NLP.IE experts.....	91
Figure 7-1	Features for expertise rank	94
Figure 7-2	Features for expertise rank with networks of author topics and citation related information	94
Figure 7-3	Procedures to extract author-topic related features from networks of topics.....	95
Figure 7-4	Data descriptive on citation related feature.....	98
Figure 7-5	Weight values for computing weighted-sum of expertise rank	99
Figure 7-6	Scatter plots for observing relations between h-index and some expertise rank features	100
Figure 7-7	Trends of R-squared values for models with various training data on T4.....	103
Figure 7-8	Trends of correlation values for models with various training data on T4.....	103
Figure 7-9	A possible implementation for the proposed scholar profile	106

LIST OF TABLES

Table 2-1	Feature comparison in some academic search systems	14
Table 2-2	Comparisons in modeling bibliographic data for expert finder	15
Table 2-3	Literature studies about modeling heterogeneous bibliographic information network	15
Table 2-4	Literature studies about expertise rank	18
Table 2-5	Literature studies about time factor in research interest finding	21
Table 2-6	Literature studies about visualizing bibliographic data	25
Table 3-1	Our approaches for generating scholar profile	30
Table 3-2	JSON schema for AMiner dataset	31
Table 4-1	Clustering results with Silhouette indicators for goodness of measurement	34
Table 4-2	Clusters in Aminer NLP-IE Domain with transformed positions by LSI	37
Table 4-3	Topic words identified from probabilistic model and K-Means clustering	38
Table 4-4	Some topics with their words within and the possible domains in NLP.IE	40
Table 4-5	Dunn Index of Clustering Results	46
Table 4-6	Student questionnaires	47
Table 4-7	Major cooperation departments extracted from clustering results (graph-based k-Means with MST, $k=\{7\}$)	50
Table 4-8	Expert classification accuracies using CoI features with various similarity methods, classifiers and interest threshold values	59
Table 5-1	Collections in <i>Dbehavior</i> dataset.....	64
Table 5-2	Productivity features for each researcher in particular topic	65
Table 5-3	Dynamicity features for each researcher in particular topic	66
Table 5-4	Correlation values between productivity-dynamicity based features.....	68
Table 5-5	Combinations of correlation values and feature sets	68
Table 5-6	Scaling criteria for productivity-dynamicity features.....	69
Table 5-7	Classification accuracies with combinations of productivity- dynamicity features.....	71
Table 5-8	Fuzzy rules generated with FCM labels on scaled data of selected productivity-dynamicity features	71
Table 6-1	Levels for publishing related behavior based feature values	77
Table 6-2	RSIENA data input for experiments to observe exploring feature	80
Table 6-3	RSIENA Effects for observing changes in networks and attributes...	81
Table 6-4	RSIENA results for AMiner NLP-IE dataset.....	84
Table 6-5	RSIENA results for descriptive values	85
Table 6-6	RSIENA evaluation functions to observe co-author selection.....	87
Table 6-7	RSIENA evaluation results based on publishing behavior	89
Table 6-8	RSIENA evaluation results based on exploring behavior	89
Table 6-9	RSIENA evaluations related to the collaboration dynamics of co-authors.....	90

Table 6-10	Number of articles and citations in a sample of AMiner NLP.IE experts.....	92
Table 7-1	Sample of topic MST for authors	97
Table 7-2	Expertise rank with selected features and correlation values as performance indicator.....	99
Table 7-3	Regression results with varied models for expertise rank features on T4.....	101
Table 7-4	Regression results with R-squared values for various training data on T4.....	103
Table 7-5	Regression results with correlation values for various training data on T4.....	104

LIST OF ABBREVIATIONS, TERMS AND SYMBOLS

AMiner	ArnetMiner, Academic Research Network Miner
CoI (features)	Conflict-of-Interest (related features)
DI	Dunn Index as indicators for evaluating the clustering results
FCM	Fuzzy C-Means
frbs (package)	Fuzzy Rule-Based Systems (R package)
Gensim (package)	Generate Similar (Python package)
KMeans++	K-Mean clustering with additional steps to select better initial cluster centers to speed up convergence
LDA	Latent Dirichlet Allocation
LSI	Latent Semantic Indexing
LSTM	Long Short Term Memory
mlogit	Multinomial Logit Model
MST	Minimum Spanning Tree
NLP-IE	Natural Language Processing - Information Extraction
NLTK	Natural Language Toolkit
PCA	Principal Component Analysis
RSIENA (package)	R package in Simulation Investigation for Empirical Network Analysis
SAOM	Stochastic Actor-Oriented Model
SVM	Support Vector Machine
TF-IDF	Term Frequency–Inverse Document Frequency
t-SNE	t-Distributed Stochastic Neighbor Embedding
WSM	Weighted-Sum method
a_i	Author-i or researcher-i
$beh_{const}(a_i, wt)$	Behavior matrix of exploiting topics or consistency for Author-i in periode-t
$beh_{exp}(a_i, wt)$	Behavior matrix of exploring topics for Author-i in periode-t

bip_{wt}	Bipartite (two-mode) networks of author-topic relations for periode-t
CA	Collection of co-authors based on article metadata
c_k	Cluster-k or topic-k
d_j	Article-j or document-j
$D_{small-title}$	\pm 4800 articles from 70 AMiner experts on NLP-IE domain
L	Collection of labels for each article; the labels are cluster ID
	c_k
S	Collection of citations for each article d_j that has been cited n_j times on year y

Chapter 1.

INTRODUCTION

1.1. Research Background

Researchers develop or manage their academic networks to foster knowledge sharing in collaboration along with career development [1]. Academic searches such as Google Scholar, Scopus, or AMiner [2] help researchers in finding potential collaborators in expert finders [3]. A recommendation system of expert finder usually evaluates researcher expertise indicated from published articles as the output of research activities [4]. Information of articles or bibliographic data is abstracted as academic research networks [5]. Representing researcher expertise typically applies (1) statistical language modeling for content analysis of bibliographic texts [6] [7] [8], (2) graph modeling for structural analysis on bibliographic networks [9] [10] [11] [12], and (3) both models [13] [14] [15]. Those approaches consider expertise evidence [16] [17] from any combinations of textual information used in statistical language modeling, social interaction data used in graph modeling, and scientometric features such as citations. The system evaluates those expertise evidence and returns beneficial researchers in terms of productivity and relevancy.

Although researchers who have published and been cited more are generally considered as productive, there is a possibility of biased citation issue leading to questionable expertise status. Citation is often exploited to measure the performance of researchers through scores such as h-index [18] such that higher values refer to expertise status. However, the expertise of researchers should not be solely measured in quantitative manner. The productivity and consistency on topics in which the researchers claim their expertise should be consider as well. Another issue is the relevancy of research domain in which researchers could have interest changes, and makes the expertise status is varied from time to time. Following the time period, some studies presented h-index in annual term to accommodate expertise caused by career length [19].

The works in this dissertation contribute on modeling unbiased scholar profile, which consider interest changes and less focused on citations since more

objective expertise evidence are required. Then, an expert recommendation system appraises the scholar profile to rank expertise without bias.

1.1.1. Interest changes influence research expertise

Researcher who have different expertise could be influenced from others when they work as co-authors in publishing articles. The expertise of researchers on specified domain increases or decreases after some periods of time because of the influence [12]. The term “research domain” in this dissertation is shortened into the term “topic”, while the term “research interest” refers to some topics that become the interest of researchers, which leads to their expertise. The challenge of varying research interests is to recommend researchers who are focusing on certain topics for a defined period, such as statistical language modeling with Temporal-Expert-Topic (TET) [20]. Other works showed features extracted from analyzing structural (graph modelling) and time [9] [10] or combinations of content (statistical language modeling), structural, and time [21] [22]. In general, statistical approach obtained topic distribution that became the interest of researchers in a period of time [21], while for tracking interest, context similarities to previously published articles were semantically evaluated [22]. Some methods focused on time without content analysis [12], or reversely considering topics and ignoring time factor [23].

Researchers generally prefer others who are productive in recent times and those previous approaches do not immediately capture researcher productivity. Other works have implemented the term of researcher productivity on detecting the potency of rising stars [24]. Those features of rising stars explored the dynamicity on the researcher performance for productivity, impact and sociability from bibliographic data and the represented graphs. However, those indicators of expertise evidence for each researcher are not related to topics yet. Thus, this study attempts to derive topic information on productivity-dynamicity features to generate more objective expertise evidence and responsive on periods of time for handling interest changes. The extraction process should produce a number of features. Because some of the features might indicate similar evidence, feature selection is necessary.

Interest changes or called as topic drift [25] has effected the productivity of researchers and transformed their relations to others called as network churn [26].

The possibilities of exploration or exploitation (consistency) to topics for a researcher has been discussed [26] [27], but the change level has not been measured yet. However, researchers rarely take a leap on their interests, and thus their topics are likely connected which is termed as inter-related topics. Graph modeling enables a measure on relations between nodes, such that we denote topics as connected nodes and the context distance between topics as their relations. The distance between topics could complement expertise evidence of researchers to represent how far their exploration is.

Therefore, as parts of the contribution considering to interest changes, this study works on mechanisms to extract expertise evidence related to productivity and dynamicity of researchers based on topics, and also distance values to know the range extent of topic spread. Values of productivity and dynamicity features based on topics in different periods might be varied. For a researcher who has interest changes but still on inter-related topics and being supported with high performance of the research productivity would have evidences to signify his or her expertise on the topics.

1.1.2. Rank expertise based on research domain

Researchers are recommended according to expertise scores on specified research domain or topics as formerly substituted. Topic information is required in querying an expert recommendation system. For example, statistical language model formulates the probability of researchers as experts from a text collection of title-abstract according to a query topic [6]. For graph modelling, each query topic invokes to generate a network of researchers whose articles related to the topic, and then score them with a random walk model [9] [10] [11] [12].

Researchers have several sources of evidence for rank expertise [16] [17], in which citation as a scientometric features [8] [12] is every so often excessively exploited [28] [29] [30] [31]. Some studies evaluated the relation between citation and article content to ensure its fairness usage [32] [33]. Those studies could not be applied in this dissertation problem because they do not consider interest changes. Researchers could have strategies for exploration and or exploitation topics as their interest to become productive and confirm their expertise. Each strategy has different tradeoffs to make it difficult for selecting only one strategy. Thus, the works in this

dissertation do not decide the best strategy, but evaluate the levels of both strategies related to the expertise on topics. Therefore, the proposed mechanisms to extract expertise evidence should not only accommodate productivity-dynamicity of researchers on topics. The mechanisms incorporate less excessive usage on citations for rank expertise, and to compensate it with behavior of researchers in exploration and exploitation.

Studies related to social relations of researchers through academic networks validated that the research performance correlates to ego network of each researcher [34]. Efficiently collaborated researchers tend to become productive. This finding could be followed with the influence from other researchers while co-authoring articles. Since influence effect from others is not constant, collaborations of researchers also have productivity and dynamicity aspects. Referring to time factor in interest changes, we attempt to elicit the behaviors of researchers related to productivity and dynamicity in producing the output of research activities. There are two possible behaviors of researchers with regards to topics, exploration and exploitation (consistency), which require to be quantified into level values. Then, varying behavior levels for each topic in different periods will be gauged as evidence of research expertise, which is more objective than biased citations.

Proposed mechanisms to extract expertise evidence for both considerations of interest changes and less focused on citations as a scholar profile require topic information. This becomes a problem in a cold-start situation in which bibliographic data might be the only available metadata. Before implementing the proposed mechanisms, topics should be obtained through processing texts of title-abstract in the metadata. Then, the next process is using the topics to obtain a scholar profile that yields to numerous evidence of research expertise. Experiments of rank expertise using different combinations of evidence is necessary to ensure that the scholar profile could compensate citation based features.

1.2. Problem Statement

Previous studies about procedures to acquire evidence for capturing expertise of researchers on specified topics still left questions regarding to interest changes and less focused on citations. Before extracting any features as expertise evidence from

bibliographic data, information about topics are required. We identify topics and mapping them as the interest of researchers by processing texts of title-abstract which followed by relating the identified topics to researchers. Then, we validate the mapped topics as the interests of researchers by applying them in some recommendation situations. Thus, the problems related to extracting any features as expertise evidence are formulated into several points as follows.

- a. How to extract productivity and dynamicity features for scholar profile of researchers on each topic?
- b. How to extract behavioral features based on researcher behavior in publishing articles for scholar profile?
- c. How good the scholar profile is in representing the expertise of researchers on specified topics by considering some feature combinations?

1.3. Research Objectives and Benefits

The main objective is to analyze productivity, dynamicity, and behavior of researchers from their published articles, which provide evidence that considering on interest changes and less focused on citations to indicate unbiased expertise on specified topics as a scholar profile. Then, the objectives of this research are:

- a. To obtain topics from processing unstructured article texts of title-abstract by considering weighting schemes to get representative words for each topic
- b. To identify topics for articles followed by knowing the interests of researchers from their published articles which is called as topic mapping
- c. To evaluate the identified topics as the interests of researchers in empirical experiments that provide some recommendations
- d. To extract and select features of researchers from their productivity, dynamicity, and behavior with the published articles to represent scholar profile that considers interest changes and less focused on citations
- e. To evaluate the extracted features as evidence for the expertise of researchers on specified topics as a scholar profile in empirical experiments to provide some recommendations

The scholar profile model has an applicable benefit in a recommendation system, such as expert finder, for recommending productive researchers who could become collaborators and provide academic guidance.

1.4. Research Roadmap

We have conducted some studies in Figure 1-1 that show the preparation process for scholar profile using bibliographic data. After investigating previous works, we have implemented empirical experiments and published the results into a number of published articles in journals or proceedings. Works in details within Figure 1-2 described researches on (a) content analysis in bibliographic data, (b) modeling the data with graph, then (c) expertise rank. Additional studies also include (d) fairness aspects on expertise as well as (e) visualizing researchers according to their expertise.

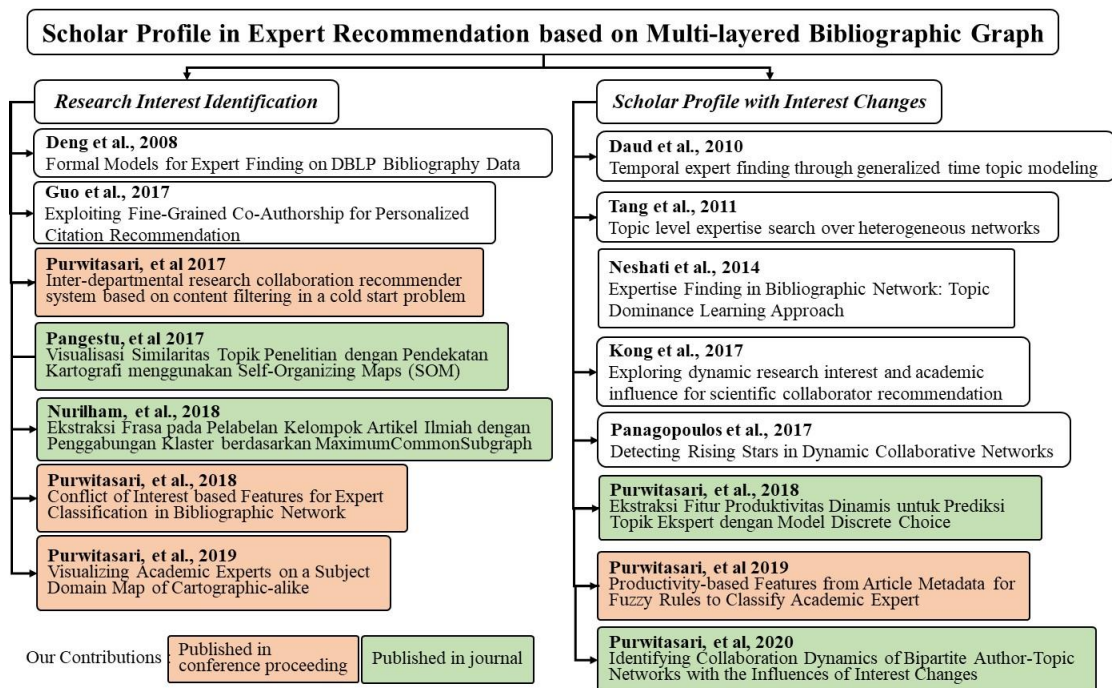


Figure 1-1 Group of studies in Scholar Profile for Expert Recommendation

1.5. Original Contributions

Within the last 15 years, expert recommendation system for researchers generally applies content and or structural analysis using articles as research activity output for identifying expertise evidence of specified topic. Evidence for researchers

in the form of textual, social interaction and scientometric have been extensively investigated.

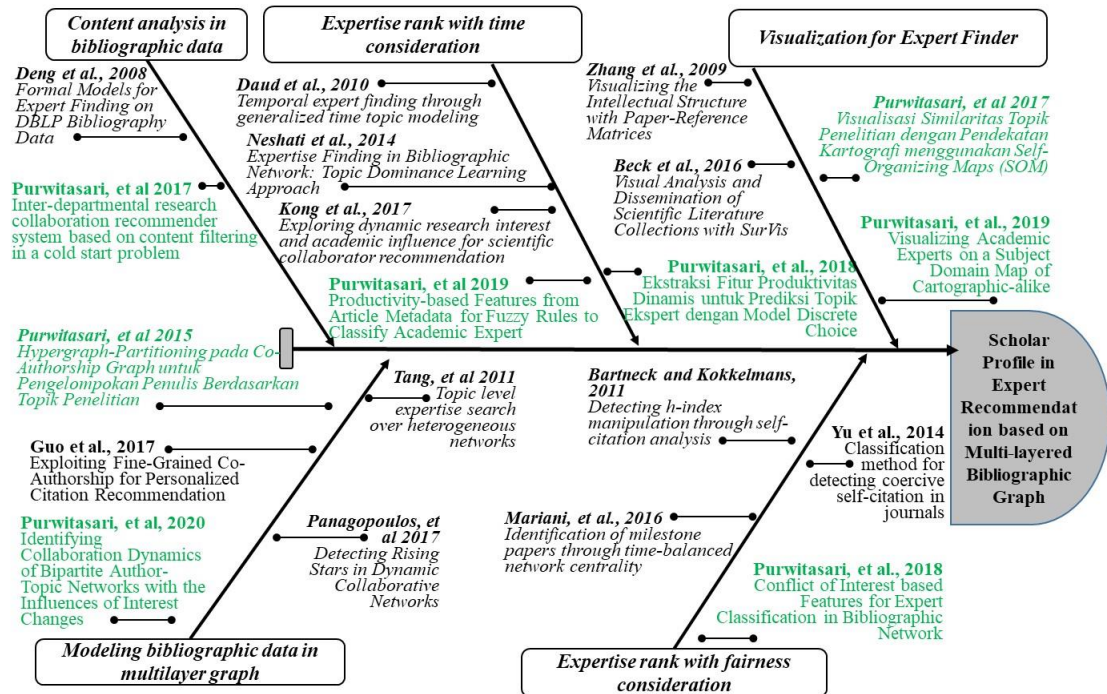


Figure 1-2 Fishbone diagram of research contributions

With opportunities for enhancement in identifying expertise because of the problems about interest changes and less focused on citations, in this dissertation we work on mechanisms to model unbiased scholar profile. The proposed mechanisms utilize both analysis for all three types of evidence with the main contributions that become the originalities and novelties of this research, along with publications to disseminate the contributions are listed below.

1. Expertise evidence on specified topics for productivity-dynamicity features of researchers have been extracted as a scholar profile [35].
2. Expertise evidence on specified topics for behavioral features related to researcher behavior in publishing articles also have been extracted as a scholar profile [40].
3. Combination of those features conditioned in a situation without topic information, or cold-start, to rank expertise on topics has not been investigated before. From empirical experiments using the scholar profile to obtain scores for

expertise of researchers, it reveals that the scores are comparable to the ones with citation related features.

Then, other supporting contributions along with publications to disseminate the contributions are also listed.

4. Productivity-dynamicity features without topics, which originally applied to identify rising stars in other works [24], was adapted for extracting mechanisms of expertise evidences on topics, and evaluated to reduce the similar evidences, then applied to observe the performance of evidences in giving recommendations.
5. Mechanisms to identifying, mapping and validating topics for researchers have been designed [36], with some previously investigated approaches were implemented in empirical experiments to provide some recommendations [37] [38] [39].
6. The performance of behavior features of researchers have been validated through empirical experiments that observe the influence of interest changes from other researchers [40].

1.6. List of Publications

- a. “Inter-departmental research collaboration recommender system based on content filtering in a cold start problem”, IEEE 10th Intl. Workshop on Computational Intelligence and App., 11-12 Nov. 2017, Hiroshima, Japan [37]
- b. “Conflict of Interest based Features for Expert Classification in Bibliographic Network”, Intl. Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM), Surabaya-Indonesia, 26-27 Nov. 2018 [38]
- c. “Ekstraksi Fitur Produktivitas Dinamis untuk Prediksi Topik Ekspert dengan Model Discrete Choice”, Jurnal Nasional Teknik Elektro dan Teknologi Informasi (JNTETI) Vol 7, No. 4, November 2018 pp. 418-426 [35]
- d. “Productivity-based Features from Article Metadata for Fuzzy Rules to Classify Academic Expert”. The 10th International Conference on Awareness Science and Technology (iCAST), Morioka-Japan, 23-25 Oct 2019 [36]

- e. “Visualizing Academic Experts on a Subject Domain Map of Cartographic-alike”. 4th International Conference on Computer, Communication and Computational Sciences (IC4S), Bangkok-Thailand, 11-12 Oct. 2019 [39]
- f. “Identifying Collaboration Dynamics of Bipartite Author-Topic Networks with the Influences of Interest Changes”, Springer International Journal Scientometrics, 2020 (Scopus Q1). (First Online: 14 January 2020)

1.7. Research Scope and Limitation

Our experiments used a well-known dataset of experts from AMiner. The dataset contains metadata of scientific articles in the “computer science” domain. For observing the performances of our scholar profile model, we selected some AMiner experts especially in domains of Natural Language Processing and Information Extraction (NLP-IE). The selection reasons are familiarity issue and interrelated context between those two domains to illustrate real conditions for differentiating experts. However, our approaches are not limited to certain domains, so the procedures are applicable in any cold-start situations.

Regarding cold-start situations, we also performed some empirical experiments with article metadata in our university which come from undergraduate theses, called as ITS (Institut Teknologi Sepuluh Nopember) dataset. In that case, the researchers are lecturers and the theses are their publication output. Some procedures related to topic mapping applied on AMiner dataset are modified and implemented in ITS dataset (Figure 1-3).

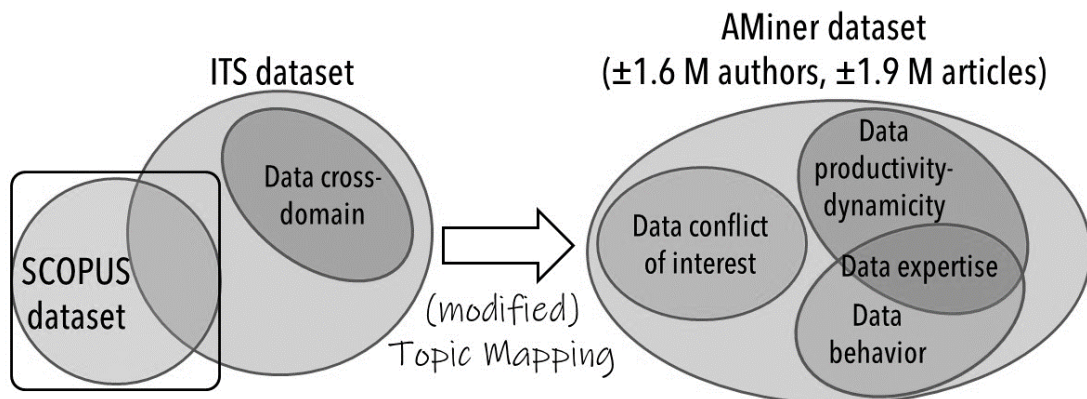


Figure 1-3 Dataset used in this dissertation

Although this dissertation starting the works from problems of interest changes and biased citations, but for evaluating the performance of the scholar profile we compared expertise score with common indicator H-index of researchers, which basically derived from citations. This approach for evaluation scenario is motivated from other studies with similar situations of unavailable ground truth [41].

1.8. Book Structure

This dissertation report starts with an introduction to a problem in academic recommendation systems for generating scholar profile that considers interest changes and less focused on citations (Chapter 1). Then, our introduction continues to common representation of academic social network for scholar profile extracted from bibliographic data of researchers (Chapter 2). Some important issues are discussed such as modelling bibliographic data with topic mapping and using the mapped results for rank expertise as well as for visualizing the researchers for evaluation purpose.

Our focus on this dissertation is about unbiased scholar profile considering interest changes and less focused on citations. Besides the research framework (Chapter 3), we also report other approaches to identify the topics of researchers based on the contents of research outputs of scientific article metadata in Chapter 4. Then we also investigate some works to employ the topics to visualize researchers. The following chapters discusses the works to prepare the scholar profile with behavior-based features as our main contributions (Chapter 5, Chapter 6). We also put some of those features into several situations of predicting expertise.

Finally, the behavior-based features are observed for ranking expertise (Chapter 7), then we conclude our reports with discussions for future works (Chapter 8). Our empirical experiments for ranking expertise using three groups of features: productivity of researchers based on topics, their behaviors on exploring and exploiting topics, and then the last states of researchers based on published articles and received citations. Some findings and essential settings are reported in the last chapter, including further work scenarios for generalizing the findings since this dissertations still worked on the controlled experiments.

Chapter 2.

BIBLIOGRAPHIC DATA FOR ACADEMIC SOCIAL NETWORK

Starting with well-known academic search systems, this chapter discusses the studies of academic social network created from bibliographic database which is explored further for any mining tasks. Next sections are about modeling bibliographic data, and followed by studies about ranking expertise for researchers. The last sections are studies on issues that are going to be solved in this dissertation.

2.1. Academic Search Systems

Some academic search systems mentioned in this dissertation (Google Scholar, Scopus, Microsoft Academic Search, AMiner) began as a university research project like AMiner [2] for academic researcher social network building, search, and mining in China. AMiner has collected more than 130,000,000 researcher profiles and 100,000,000 papers from multiple publication databases since 2006 until 2016. Unlike any search model with keyword matching, AMiner offers topical analysis of the academic data [7] to help users know the experts, give recommendation of scientific articles, publication venues, topics-subtopics and their evolving in the past years, along with any relations or influences between research works.

Similar to Scopus, Science and Technology Index (SINTA) as a citation and expertise center supported by Indonesian government also shows relations of Indonesian researchers. There are >70.000 verified researchers in SINTA with articles of ± 40.000 journals and ± 17.000 conferences. For each researcher profile, information of articles per year according to Scopus, citations per year according to Google Scholar, Scopus score and Google Scholar score among other information are available, as shown in Figure 2-1, while their academic network is also shown in Figure 2-2. Users are expected to explore SINTA nodes themselves to know the expertise of a researcher in Figure 2-3. Feature differences for comparing some academic search systems are shown in Table 2-1.

Information related to topics such as discovery and evolution are interesting. With recommender system in a bibliographic network, a user may be interested in the most similar article or researcher for a given query. The input query could be combination of researchers, articles, and keywords. A recommender system such as expert finder is expected to return a list of other researchers who have similar situations, i.e. young researchers who require to find any possible collaborators.

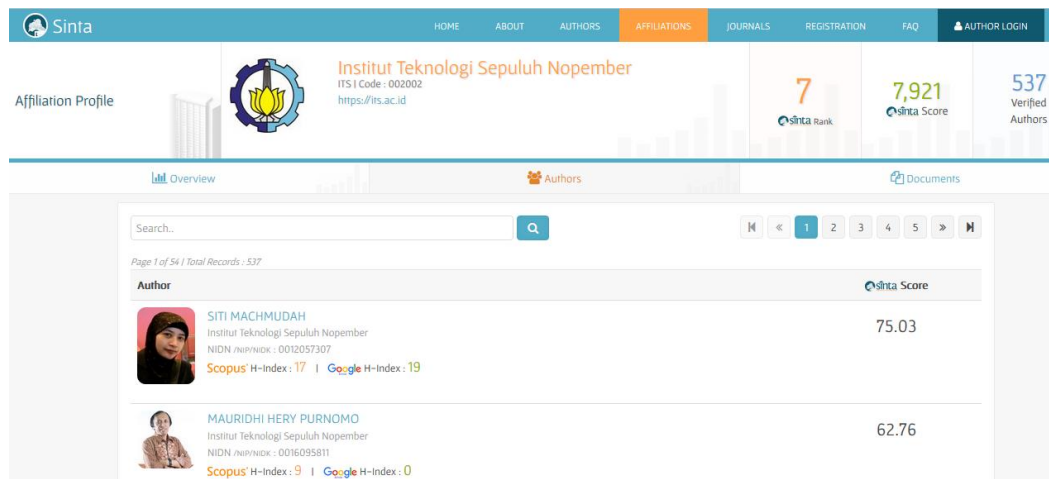


Figure 2-1 Sample of SINTA scores for Institut Teknologi Sepuluh Nopember

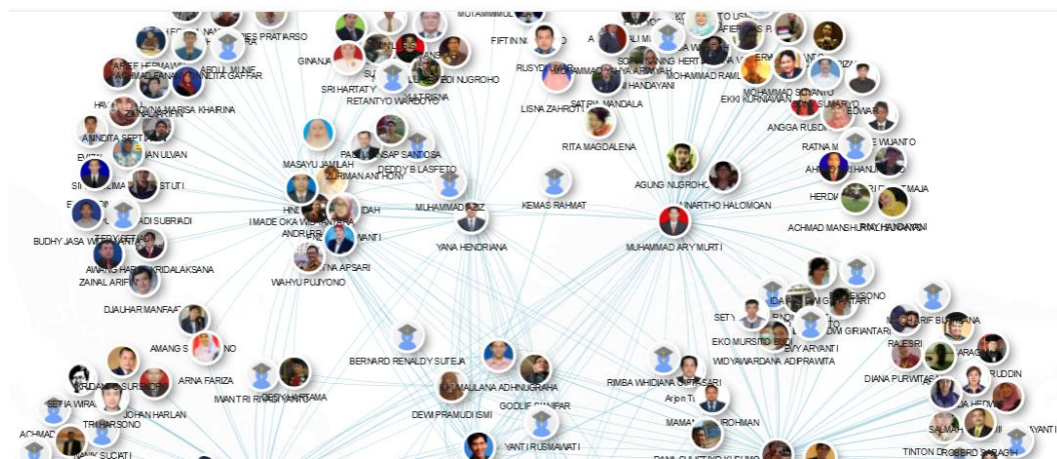


Figure 2-2 Sample of SINTA networks for a researcher

Microsoft Academic Mauridhi Hery Purnomo Sign in or Sign up

Mauridhi Hery Purnomo
Sepuluh Nopember Institute of Technology

238 PAPERS 307 CITATIONS*

Top Publications

Malaria parasite identification on thick blood film using genetic programming
2013, *International Conference on Instrumentation, Communications, Information Technology, and Biomedical Engineering*, pp 194-198
I. Ketut Eddy Purnama (Sepuluh Nopember Institute of Technology), Farah Zakiah Rahmanti, Mauridhi Hery Purnomo (Sepuluh Nopember Institute of Technology)
This blood film is used to know type and phase of the malaria parasite, but which is widely used in Indonesia is the thick blood film. Therefore we need a method that can identify parasites in thick blood film image with a high percentage of accuracy. This research aims to ...
Citations (9) * Source Share Cite

Co-authors
Ardyono Priyadi
Mochamad Hariadi
I Ketut Eddy Purnama
Mochamad Ashari
Margo Pujiantara
Show More

Co-author Affiliations
Sepuluh Nopember Institute of Technology
Kumamoto University
Udayana University

Temporary short circuit detection in induction motor winding using combination of wavelet transform and neural network

Google Scholar

Mauridhi Hery Purnomo
Institut Teknologi Sepuluh Nopember
Verified email at if.its.ac.id
Artificial Intelligence

Cited by

	All	Since 2012
Citations	1042	931
h-index	15	13
i10-index	23	22

VIEW ALL

Keynote Speaker II: Biomedical Engineering Research in the Social Network Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media
MH Purnomo, S Sumpeno, EI Setiawan, D Purwitasari
Procedia Computer Science 116, 3-9

Investigation of Symmetrical Optimum PI Controller based on Plant and Feedback Linearization in Grid-Tie Inverter Systems
M Facta, A Priyadi, MH Purnomo
International Journal of Renewable Energy Research (IJRER) 7 (3), 1228-1234

Scopus Search Sources Alerts Lists Help SciVal Institut Teknologi Sepuluh Nopember Surabaya

Author details About Scopus Author Identifier

The Scopus Author Identifier assigns a unique number to groups of documents written by the same author via an algorithm that matches authorship based on a certain criteria. If a document cannot be confidently matched with an author identifier, it is grouped separately. In this case, you may see more than one entry for the same author.

Return to search results 1 of 3 Next

Hery Purnomo, Mauridhi Follow this Author

Institut Teknologi Sepuluh Nopember, Department of Electrical Engineering, Surabaya, Indonesia
Author ID: 6602604153

Other name formats: Purnomo, Mauridhi Hery Purnomo, Maundhi Hery Hery Purnomo, M. Purnomo, Mauridhi Hery Purnomo, Mauridhi H. Purnomo, Mauridhi Hery Purnomo, M. H.

Subject area: Computer Science Engineering Mathematics Energy Physics and Astronomy Social Sciences Business, Management and Accounting
Chemical Engineering Medicine Agricultural and Biological Sciences Multidisciplinary Materials Science Psychology View all

Document and citation

h-index: 9 View h-graph

Documents by author: 189 Analyze author output

Total citations: 413 by 351 documents View citation overview

KEEST Miner Whatever comes to your mind EN Login

Purnomo Mauridhi Hery Follow

Login to view email, homepage, and external links

Update

Research Interests

- Biomedical Research
- Gray Tone Spatial Dependency Matrix (gicm)
- Classification
- Contrast Limited Adaptive Histogram Equalization (clahe)
- Knee Osteoarthritis

Ego Network

Similar Authors: Purnomo, Maundhi Hery, Hery Purnomo, M., Purnomo, Mauridhi Hery, Purnomo, M. H., Purnomo, Mauridhi Hery, Purnomo, M. H.

D-Core: Purnomo, Maundhi Hery, Hery Purnomo, M., Purnomo, Mauridhi Hery, Purnomo, M. H., Purnomo, Mauridhi Hery, Purnomo, M. H.

Ego Network: Purnomo, Maundhi Hery, Hery Purnomo, M., Purnomo, Mauridhi Hery, Purnomo, M. H., Purnomo, Mauridhi Hery, Purnomo, M. H.

Figure 2-3 Sample academic research data used in Microsoft Academic, Google Scholar, Scopus, and Aminer

Table 2-1 Feature comparison in some academic search systems

Feature List	Scopus	AMiner	SINTA	Google Scholar	MAS
Most active researchers	V	V	V	V	V
Recommendation of articles	V	V	V	V	V
Recommendation of publication venues		V			
Topics-subtopics and their evolving in the past years		V			
Relations or influences between research works		V			

2.2. Modeling Bibliographic Data

Various approaches in recommendation systems using bibliographic data are mainly classified into two types: based on expertise information (content analysis) and based on social relations of experts (structural analysis). As content analysis, scientific articles published by researchers could become useful indicators for research expertise. Content analysis focused on topical terms in titles, abstracts or keywords which semantically interpreted in the expertise extraction using language modeling of generative probabilistic [6] [7], clustering of word vector representation [42], or concept domain with ontology [13]. Different topics could have context relations, such that extracting the expertise of researchers needs to consider semantic similarity of texts in their published articles (i.e. ontology, Word2Vec [43], latent topics).

Other than content analysis, the structural analysis usually incorporates graph modeling [13] [42]. Other key aspects in modeling bibliographic data (Table 2-2) are the query object such as keywords that representing expertise area or the changes of topics compared to some different years. The changes of topics is defined as interest changes.

Table 2-2 Comparisons in modeling bibliographic data for expert finder

Refs	Content analysis	Structural analysis	Query object	Interest changes
[6]	Generative probabilistic	-	Topic keyword	-
[7]	Generative probabilistic	-	Topic keyword	Same topic model for each given year
[13]	Consider semantic similarity (concept domain with ontology)	concept layer and researcher layer	Researcher name to generate scholar profile	-
[42]	Consider semantic similarity (clustering of vector representation from word occurrences with Word2Vec)	paper-paper citation, author-paper, paper-word, co-authorship	Topic keyword Researcher name to generate scholar profile	-

Table 2-3 Literature studies about modeling heterogeneous bibliographic information network

Study Focus	Method	Summary	Advantages or Drawbacks
1. Formal Models for Expert Finding on DBLP Bibliography Data [6]			
Expert-finding in academic field uses entities of researchers, articles (title, abstract, keyword), and citations	Statistical language modeling <ul style="list-style-type: none"> Weighted language model Topic-based model Hybrid model (language and topic) 	Using Bayes theorem to calculate the probability $p(ca q)$ of a candidate ca being an expert given the query topic q . [44] Dataset: DBLP and supplement data <ul style="list-style-type: none"> Weighted language model considers that documents have different importance therefore the document priors need weight score. The weight factor is estimated using the citation number of document. 	Advantages: <ul style="list-style-type: none"> Considering document rank based on citation number in each document Not using graph modeling but utilizing varied entities of bibliographic data Drawbacks: <ul style="list-style-type: none"> Incomplete data need procedure for expertise resource selection (fetch abstract and index terms, collect predefined topics)

... Table 2-3 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
		<ul style="list-style-type: none"> • Topic-based model associates the query topic with pre-defined latent topic. Therefore, the model needs topic selection algorithm to calculate the similarity score between query topic and pre-defined topics. • Hybrid model aggregates the advantage of the language model and the topic-based model with some weight factors in a linear form. 	<ul style="list-style-type: none"> • Query object is a topic, not a scholar profile • Not handling time factor • Not capturing relations between entities of bibliographic data (usually exists in graph modeling)
2. Topic Level Expertise Search over Heterogeneous Networks [7]			
Research-paper recommender and expert finder that use entities of researchers, articles (title, abstract, keyword), citations, and publication venues	Combination of statistical modeling (generative probabilistic) and graph modeling. <ul style="list-style-type: none"> • Author-Conference-Topic (ACT) • Citation-Tracing-Topic (CTT) 	<ul style="list-style-type: none"> • ACT uses a latent topic layer to connect the objects and ignores the link information. ACT simulates writing process of a scientific paper using a series of probabilistic steps. • CTT captures topic distributions and topic relations between papers using two correlated generative processes. • Proposing a topical random walk algorithm that integrates the topic modeling results • Searching objects by combining the topic model and the word-based language model (generative probabilistic) • Query object is a topic or a scholar 	Advantages: <ul style="list-style-type: none"> • Discovering latent topics (“semantic” aspects) associated with each object in the academic network even though not using graph modeling • Estimating the relative importance of bibliographic object that considers the topic information Drawbacks: <ul style="list-style-type: none"> • Handling time factor, but topic model for each given year cannot be the same because experts can change their research interest
3. Combining Social Network and Semantic Concept Analysis for Personalized Academic Researcher Recommendation [13]			

... Table 2-3 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
Personalized expert finder uses entities of researchers, articles (title, abstract, keyword), and publication venues	<p>A two-layer network model (graph modeling): concept layer and researcher layer.</p> <p>The principle is that researchers are interested in others who have similar research areas and social relations.</p>	<ul style="list-style-type: none"> • Concept layer represents semantic relationships between research expertise areas • Researcher layer represents social relationships occurring in academic activities • The links between both layers represent that researchers may have more expertise in some particular research areas • Considered as a graph search problem starts from a particular researcher node and ends with a set of researchers' nodes • Hopfield net algorithm starts from one or some of the target nodes and walking through the two-layer network and links between 	<p>Advantages:</p> <ul style="list-style-type: none"> • Capturing relations between researchers and their expertise along with their socials (graph modeling) • Query object is a scholar profile • Discovering latent topics (“semantic” aspects) as research area domains <p>Drawbacks:</p> <ul style="list-style-type: none"> • Not handling time factor
4. Exploiting Fine-Grained Co-Authorship for Personalized Citation Recommendation [42]			
Research-paper recommender and expert finder that use entities of researchers, articles (title, abstract, keyword), and citations	<p>Fine-grained co-authorship modeling combines co-author network and publication topics.</p> <p>Multi-layered graph of paper-paper citation relation, author-paper relation,</p>	<ul style="list-style-type: none"> • Publication topics support content based analysis • Co-author network support collaborative based analysis <p>Procedures for fine-grained co-authorship modeling are:</p> <ul style="list-style-type: none"> • Using word2vec to generate word vector representations • K-means clustering on word vectors to identify topics • Mapping authors to particular research topics 	<p>Advantages:</p> <ul style="list-style-type: none"> • Capturing relations between researchers that show collaboration influence distributions (graph modeling) • Query object can be researchers, papers, and keywords

... Table 2-3 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
	paper-word relation, and co-authorship relation with fine-grained co-authorship modeling.	<ul style="list-style-type: none"> • Random walking with restart on co-authorship graph of a specific topic to generate the ranking score of each researcher in particular topic • Using ranking results to measure researcher similarity of collaboration influence • To generate the recommended papers or researchers: using graph-based paper ranking from random walk with restart on multi-layered graph 	<ul style="list-style-type: none"> • Discovering latent topics (“semantic” aspects) as research area domains Drawbacks: <ul style="list-style-type: none"> • Not handling time factor

2.3. Expertise Rank

Recommendation systems give the results of relevant researchers based on expertise scores. Generally, a co-occurrence of a researcher with the topic terms in the same context is assumed to be evidence to the suggested expertise. Features of content and structural modals from published articles are extracted through sensors [16]. Text sensor alone can extract more than one features from texts, so an approach is necessary to resolve the conflict when the sensors have different results. Structural features by itself can be used to rank researchers by utilizing PageRank approach. Content as well as structural perspective of bibliographic data are combined to have better performance in the recommendation system, which often needs a particular method in obtaining expertise scores.

Table 2-4 Literature studies about expertise rank

Study Focus	Method	Summary	Advantages or Drawbacks
1.	Finding Academic Experts on a Multisensor Approach using Shannon's Entropy [16]		

... Table 2-4 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
Expert finder in academic field uses entities of researchers, articles (title, abstract, keyword), and citations	A multi-sensor fusion to find researchers: <ul style="list-style-type: none"> • text sensor, • profile sensor and • citation sensor 	<ul style="list-style-type: none"> • Each sensor detects various sets of events • Text sensor measures term co-occurrences between query topics and articles: term frequency, aggregated/ averaged/ maximum Jaccard coefficient or Okapi BM25 of documents • Profile sensor measures total publication: number of publications or years since first publication/journal with (out) the query topics <p>Citation sensor measures researcher authority from citation graphs: number of citations for papers with (out) the query topics</p> <ul style="list-style-type: none"> • Combination of Dempster–Shafer theory with Shannon’s entropy resolves conflict from incompatible sensor 	<p>Advantages:</p> <ul style="list-style-type: none"> • A combination of multiple sources of evidence • Capturing relations between researchers and their expertise along with their socials • No dependency on hand-labeled data based on personal relevance judgments <p>Drawbacks:</p> <ul style="list-style-type: none"> • Not discovering latent topics (“semantic” aspects) • Not handling time factor
2. ExpertRank: A topic-aware expert finding algorithm for online knowledge communities [23]			
Expert finder in online knowledge communities	A heuristic combination of expertise relevance and social importance within community.	<ul style="list-style-type: none"> • Calculating similarity score between researcher profile and the query • Representing user–thread relationships on posting discussions using a graph • Modifying PageRank algorithm to handle participation in different discussion threads for calculating authority scores as participant frequentness level. 	<p>Advantages:</p> <ul style="list-style-type: none"> • applied to situations that do not have a knowledge ontology, have low information quality, and are rich in social media • Expertise score: weighted linear, progressive sequence, and scaling multiplication

... Table 2-4 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
		<ul style="list-style-type: none"> The modification is for weighted reference relationship caused by different interests of a community user. 	<ul style="list-style-type: none"> Extracting topic phrases Accommodating varied interests Drawbacks: Not handling time factor
3. TimeRank: A dynamic approach to rate scholars using citations [12]			
Expert rank with time factor. Uses entities of researchers and citations	A temporal citation network among researchers. The network starts with all researchers have the same rating, then updates ratings with citation rewards computed sequentially.	<ul style="list-style-type: none"> Nodes are researchers and an edge is a citation between researchers at certain time where initially all edges have the same rating value The edge value has a reward updated when the source node has been cited. Different with PageRank that uses the ratings at the same time for all citations, TimeRank incorporates the timing of citations (different times for different citations) 	Advantages: <ul style="list-style-type: none"> Considering the relative position of two authors at the time of the citation among them Drawbacks: <ul style="list-style-type: none"> Not handling time factor for weighting expertise score Not discovering latent topics (“semantic” aspects)

2.4. Recommendations with Considering Interest Changes

The challenge of varying research interests is not only to recommend researchers but also to answer question of who are the experts on certain topic for a defined year such as Temporal-Expert-Topic (TET) approach [20]. TET considers a researcher is responsible for generating latent topics of publication venues, while other works showed features extracted from analyzing structural and time [9] [10] or combinations of content, structural, and time [22] [21].

Users prefer researchers who works on similar topics in recent times. Using networks of co-authors and or citation, structural and time approaches give penalized values in computing expertise scores to measure the activity impact in different times. Detecting researchers who have rising star potential [24] is not an expert finder, but

the evolution features can represent research longevity. The evolution features capture the performance dynamics of a researcher through time in terms of productivity, impact and sociability. Other approaches using content-structural-time make the analysis with [21] or without [22] a topic model. In a period of time, the first method obtained topic distribution of researchers, and the second method relied on semantic relatedness within articles. With interest changes, recommendation system still has problem that leaves much room for improvement. Therefore, this dissertation offers a framework for accommodating it.

Table 2-5 Literature studies about time factor in research interest finding

Study Focus	Method	Summary	Advantages or Drawbacks
1. Time-aware PageRank for bibliographic networks [9]			
Expert rank with time factor. Using entities of researchers and citations	Modifying PageRank by adding or removing more weights to citations nodes.	Citation between two researchers: <ul style="list-style-type: none"> • who often collaborate with each other is considered less valuable • who have never co-authored a single publication is considered more valuable Those values are changing because of penalized citations by colleagues.	Advantages: <ul style="list-style-type: none"> • Combining time information from citation and collaboration graphs to rank • Avoiding too much citations Drawbacks: Ignoring content analysis
2. Temporal Expert Finding through Generalized Time Topic Modeling [20]			
Expert finder in academic field	Answering question of who are the experts on topic Z for year Y	<ul style="list-style-type: none"> • Semantics and Temporal Information based on Maven Search (STMS) calculates count matrices based on time factor • A researcher generates latent topics of the conferences on the basis of <ol style="list-style-type: none"> a) semantics-based text information b) researcher correlations with consideration of time information 	Advantages: <ul style="list-style-type: none"> • Capturing any relation types of word by taking time factor into account • Considering time factor and semantic aspects plus researchers and conferences influence

... Table 2-5 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
		<ul style="list-style-type: none"> Deriving a Bayes Theorem to determine topically related experts for different years 	
3. Expertise Finding in Bibliographic Network: Topic Dominance Learning Approach [22]			
Expert finder in academic field	<p>Topic dominance (supervised) learning assigns more scores to researchers who are more dominant.</p> <p>The used features are:</p> <ol style="list-style-type: none"> structural, temporal, activity-based semantic relatedness 	<ul style="list-style-type: none"> Assumed as classification problem, with relevant experts as positive data. For each pair of researchers of a document, determine which one should be ranked higher Structural features are based on researcher position in co-authors network Temporal features represent the research longevity of an expert Activity-based features indicate diversity and quality of researchers Semantic relatedness feature is similarity score of previous articles 	<p>Advantages:</p> <ul style="list-style-type: none"> Returning more experienced researchers in an article as relevant experts Considering time factor and semantic aspects for expertise score <p>Drawbacks:</p> <p>Because of different numbers of associated articles of each researcher, the variance range are too wide. The model cannot be generalized to find a ranking function.</p>
4. How to Choose Appropriate Experts for Peer Review: An Intelligent Recommendation Method in a Big Data Context [45]			
Expert-finding in academic field for peer review	<p>The model has</p> <ul style="list-style-type: none"> relevance analysis uses keyword-document matrix quality analysis uses article-journal matrix and project-type matrix 	<ul style="list-style-type: none"> Relevance analysis calculates similarities between researchers and applicants Subjective relevance uses self-identified keywords Objective relevance uses article keywords Quality analysis evaluates the expertise level of researchers 	<p>Advantages:</p> <ul style="list-style-type: none"> Considering personalities for recommendation Taking time factor into consideration when calculating the weight of keywords

... Table 2-5 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
	<ul style="list-style-type: none"> connectivity analysis uses researcher-applicant matrix 	<ul style="list-style-type: none"> Connectivity analysis excludes researchers with conflicts of interest to ensure review fairness 	Drawbacks: Requiring experiments for empirical evidences
5. Detecting Rising Stars in Dynamic Collaborative Networks [24]			
Clustering researchers with time consideration.	Analyzing researchers based on scientific performance, collaboration features, and their evolution over time	<ul style="list-style-type: none"> analyzing citations defining collaboration graphs computing metrics for each researcher based on graphs using the evolution of these metrics over time as the input to clustering finding researcher types and their main features to summarize their profile 	Advantages: Clustering researchers according to their performance indexes not ranks them based on scores of rising-star, rising, non-rising. Drawbacks: do not discover latent topics
6. Exploring Dynamic Research Interest and Academic Influence for Scientific Collaborator Recommendation [21]			
Expert-finding in academic field Using entities of researchers, articles (title, abstract, keyword), and citations	Beneficial Collaborator Recommendation (BCR) model learns on <ul style="list-style-type: none"> topic distribution, interest changes over time researchers' impact in collaborators network 	<ul style="list-style-type: none"> Dividing articles by year considering interest changes Making topic clustering process on researchers' publications Obtaining topic distribution of research interest in each year Highlighting topics by an increasing time function to fit the interest changes Combining the academic impact with the similarity results to fix the rank score Conducting top-N MBC recommendation according to fixed rank score 	Advantages <ul style="list-style-type: none"> finds researchers with high academic level and relevant research topics Handling time factor Discovering latent topics
7. MVCWalker: Random Walk-Based Most Valuable Collaborators Recommendation Exploiting Academic Factors [10]			

... Table 2-5 continues			
Study Focus	Method	Summary	Advantages or Drawbacks
Expert-finding in academic field uses entities of researchers, articles, and citations without texts	Defining link importance in academic social networks with <ul style="list-style-type: none"> • coauthor order, • latest collaboration time, and • times of collaboration 	<ul style="list-style-type: none"> • inspired by productive researchers tend to be more collaborative • relationship between first two researchers is the closest, while to the rest is relatively weak (coauthor order) • A monotonically increasing function over time reflects the dynamic feature of co-authorship (latest collaboration time) • Measuring the impact of different times of coauthoring (times of collaboration) 	Advantages: <ul style="list-style-type: none"> • does some guidance when skipping to next node according to link importance • Capturing relations between researchers (link importance with graph modeling) • Handling time factor on link dynamics Drawbacks <ul style="list-style-type: none"> • Not discovering latent topics

2.5. Visualizing Bibliographic Data

Visualization could become one way to evaluate recommendations. Although the works related to visualization in this dissertation do not become the main contributions, but we attempt to explore some methods to display researchers and their research expertise.

Visualization in the field of information science can be from content perspective (i.e. mapping research domains) or structural perspective (i.e. network based). Map-like knowledge domain visualization uses cartographic approach to mapping nongeographic information of research domains [46]. However 2D map alone cannot convey more structural information of bibliographic data such as citation, co-citation, or co-authorship. For research domain visualization, the view of cross-domain and their relations is another interface for aiding users in understanding trends or new information [47]. Node-link network to visualize co-citation relationship has become a routine for research domain analysis [48]. However, from an expert finder perspective, the substantial visualization is about researchers, expertise and their relations. Thus, node-link network visualization is not preferable.

Table 2-6 Literature studies about visualizing bibliographic data

Study Focus	Method	Summary	Advantages or Drawbacks
1. A Cartographic Approach to Visualizing Conference Abstracts [46]			
Visualizing content (research domains) of bibliographic information	Cartographic approach to map nongeographic information of research domains	<ul style="list-style-type: none"> • Using vector-space modeling and Self-Organizing Map (SOM) on publication texts • Computing a hierarchical cluster solution from SOM resulted neurons to support a multi-scale zoom able visualization • Applying geometric and topological transformations 	<p>Advantages: Providing a rich and interactive 2D map of research domains</p> <p>Drawbacks: Visualizing content but not structural aspect of bibliographic information</p>
2. Visualizing the Intellectual Structure with Paper-Reference Matrices [48]			
Visualizing content (research domains) and structural (citations) of bibliographic information	<p>Visualizing co-citation relationships from paper-reference matrix.</p> <p>Using FP-tree for data transformation</p>	<ul style="list-style-type: none"> • Co-citation analysis transforms article-reference list, builds up header tables and sorts article-reference lists • Creating FP-tree from sorted article -reference list 	<p>Advantages: Visualizing structural aspect of bibliographic information</p> <p>Drawbacks: Not visualizing expertise of researchers</p>
3. A Text Visualization Method for Cross-Domain Research Topic Mining [47]			
Visualizing content (research domains and their correlations) of bibliographic information	Using hierarchical topic model to construct a hierarchical and network structure of the cross research topics	<ul style="list-style-type: none"> • Using term co-occurrence network for recursively constructing topic hierarchy • Getting the evolutionary relationships: co-occurrence network in defined time, topical frequency and topical term ranking • Topic mapping to obtain the relative space information 	<p>Advantages: Providing entry points to a domain for non-experts and trends/ new information for experts</p> <p>Drawbacks: Not visualizing expertise of researchers</p>

2.6. Summary

Section 2.1 has described some implementations of recommendations system in cases of academic search systems, i.e. returning relevant researchers as experts. For an expert recommendation system, identifying expertise of researchers requires some procedures of modeling bibliographic data illustrated in Section 2.2 as sources of expertise evidence, which is followed by computations for knowing expertise rank in Section 2.3. The problem in this dissertation is originated from obtaining unbiased expertise evidence with considering interest changes and less focused on citations. Rank procedures with citations has been mentioned in Section 2.3, while Section 2.4 studied about interest changes in procedures for recommendation researchers. As one way for evaluating the recommendations, visualization approach in Section 2.5 also has been explored.

Chapter 3.

MULTI-LAYERED BIBLIOGRAPHIC GRAPH FOR MODELING SCHOLAR PROFILE

This chapter aims to outline initial methods used in mechanisms to model unbiased scholar profile with evidence of research expertise, which is identifying topics, mapping them as the interest of researchers, and followed by some validations through recommendation cases. The next chapters on other following methods in the proposed mechanisms are about utilizing the topics in extracting the evidences for scholar profile, and using the profile to represent the expertise of researchers. The evidences for unbiased scholar profile of productivity, dynamicity, and behavior of researchers on specified topics require structural analysis of represented graphs from bibliographic data.

3.1. Abstracting Multi-Layered Bibliographic Graph

Before discussing summarized works in this dissertation, we show graph assumptions for abstracting bibliographic data. Previous studies combined content and structural analysis of bibliography data, and generated multi-layered graphs of content and relation of researchers [13]. The content was derived from article texts published by the researchers, while the relation from co-authors within the articles. Motivated by those works, we restructure bibliographic data into two types of networks: one-mode and two-mode.

One-mode network refers to a graph with homogeneous nodes, which is co-author network, while two-mode or bipartite network refers to a graph with heterogeneous nodes, which is author-topic network. The term author means to accentuate the role of researchers in their published articles as authors, and they could influence other researchers during the process of co-authoring articles as co-authors. Abstraction of those networks is illustrated in Figure 3-1 with original bibliographic data of articles that consist of author information, followed by identifying topics and mapping them to articles and researchers, then apply information of authors and topics to generate bipartite network.

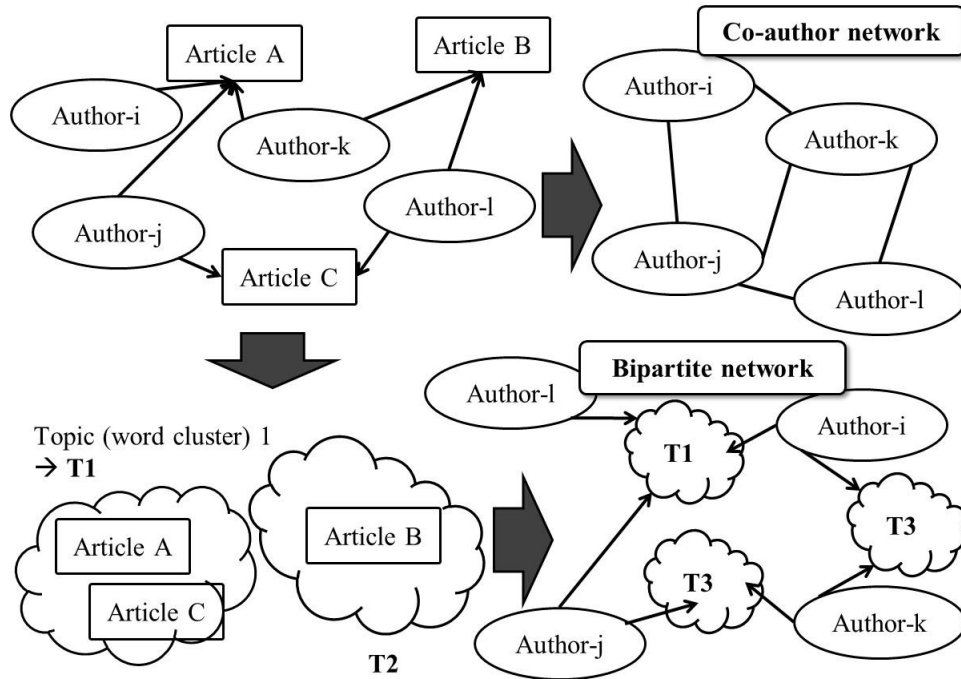


Figure 3-1 Networks of one-mode (co-author) and two-mode (bipartite) abstracted from article metadata

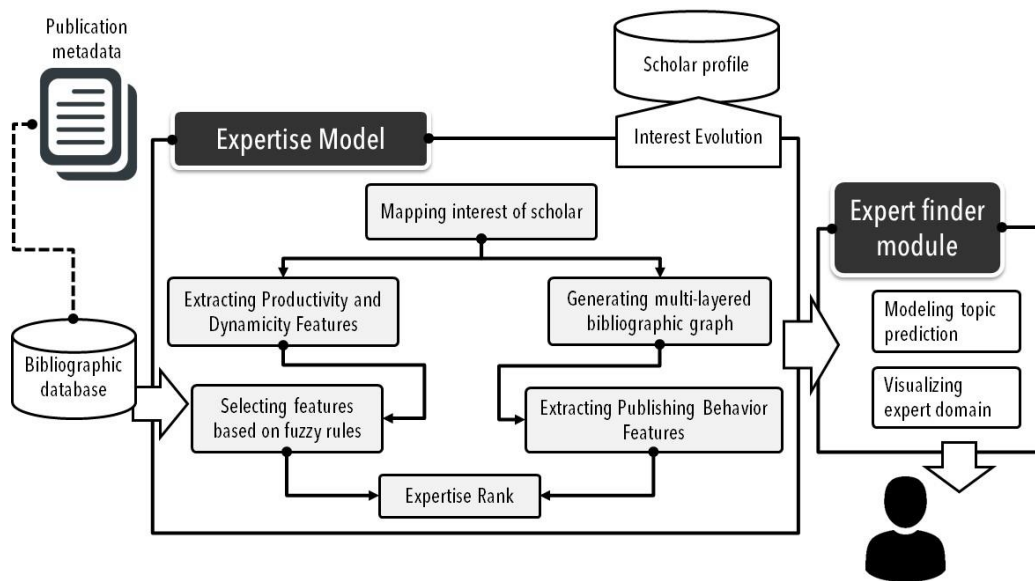


Figure 3-2 Scholar Profile based on Multi-layered Bibliographic Graph

3.2. Research Framework

This dissertation models on multi-layered graphs from bibliographic data as described in Figure 3-2 to present a scholar profile that considers interest changes and less focused on citations. The benefit from our model is to have a better

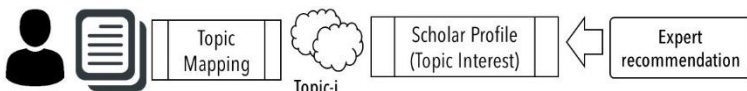
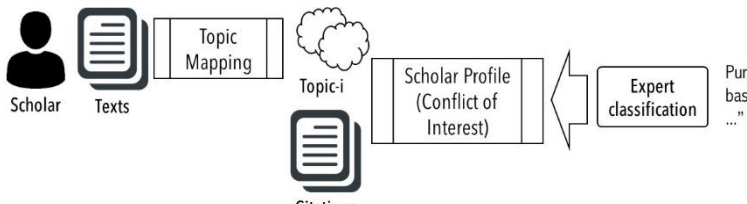
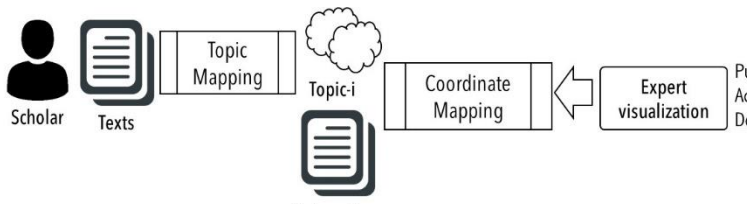
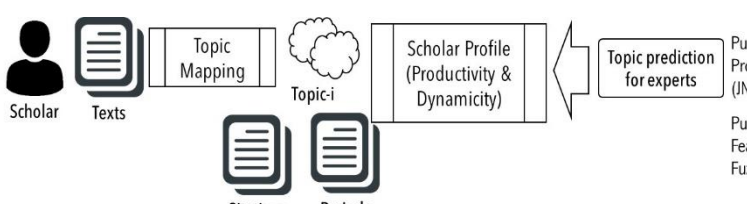
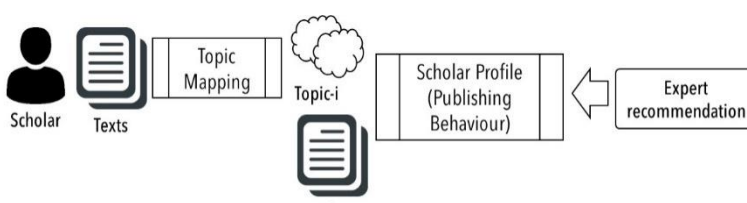
understanding scholar performance holistically from bibliographic data to find productive researchers as role models. The framework has modules to process bibliographic data into recommendations: Expertise Module and Expert Finder Module. Multi-layered in the framework refers to multi perspectives of bibliographic data derived from articles as the output of research activities. Starting from preprocesses metadata, then Expertise Module generates the scholar profile which is useful for Expert Finder Module. Those modules contained several stages that have been evaluated and became our research output as mentioned in Section 1.6.

We identified topics from texts of title-abstracts, which are taken from published articles of researchers. Those texts consist of words, such that identifying topics is equal with clustering the words. The results are groups or clusters of words with similar context, i.e. a cluster contains words of 'routing', 'experimental', 'error', 'evaluation', 'integrating', 'rules', 'inference', 'representation', and 'domains'. In cold-start situations, articles and researchers do not have information of topics. We prepared dataset, performed clustering with various settings to acquire topics, then topics or clusters of words should be mapped onto articles and researchers.

The mapping results were utilized for other processes in Figure 3-2. Those processes that include evaluations for recommendations are basically illustrated in Table 3-1 (a). For evaluating topics identified with clustering approach, we visualized the researchers based on their topics (c), or applied the topics to investigate conflict of interest that might boost citations and lead to overstated expertise (b).

Other processes include evidence related to researchers or termed as scholar to connect with a scholar profile. The evidence is the behavior of researchers that related to productivity in publishing articles as their research output and the possibility of interest changes, called as dynamicity, because of continuous interaction with other researchers as co-authors (e). For evaluating the productivity and dynamicity based features, we predict the topics of researchers (d), as well as rank expertise without dependency on citations. Unusual increasing value of citation often occurred due to some conflicts of interest. We compared the results of expertise rank by using citation related features and our scholar profile. We observed their correlations to the actual h-index values of researchers, which often used as the performance indicator for researchers.

Table 3-1 Our approaches for generating scholar profile

<p>a</p>	 <p>Purwitasari, et.al. "Inter-departmental Research Collaboration Recommender ..." (IWCAIA2017)</p> <p>Determining on how to identify topics of researchers for the scholar profile</p>
<p>b</p>	 <p>Purwitasari, et.al. "Conflict of Interest based Features for Expert Classification ..." (CENIM2018)</p> <p>Determining on how to identify researchers with the possibility of conflict of interest based on their mapped topics in the scholar profile</p>
<p>c</p>	 <p>Purwitasari, et.al. "Visualizing Academic Experts on a Subject Domain Map..." (IC4S2019)</p> <p>Determining on how visualize researchers based on their topics in the scholar profile</p>
<p>d</p>	 <p>Purwitasari, et.al. "Ekstraksi Fitur Produktivitas Dinamis untuk ..." (INTEI2018)</p> <p>Purwitasari, et.al. "Productivity-based Features from Article Metadata for Fuzzy Rules ..." (ICAST2019)</p> <p>Determining on how to generate scholar profile with productivity and dynamicity features from multi-layered graph</p>
<p>e</p>	 <p>Purwitasari, et.al. "Identifying Collaboration Dynamics ..." (SCIENTOMETRICS2020)</p> <p>Determining on how to generate scholar profile with behavior based features from multi-layered graph that considering interest changes</p>

3.3. Data Acquisition and Preparation

In this dissertation, bibliographic data were acquired from AMiner in the form of metadata texts for titles, author names, venues of conference and journals, abstracts, and citations as illustrated in Table 3-2. AMiner is a large dataset with more than 130,000,000 researchers and 100,000,000 articles [2]. The venues of AMiner articles can be journals or conferences, and same conferences held in different years are considered as different venues. We only employed researchers in AMiner corpus who have articles with at least receiving five citations. After observing the numbers of article for each venue, the availability of abstract texts, we utilized ± 500.000 articles with ± 360.000 authors and $\pm 1.100.000$ relations of co-authors.

Table 3-2 JSON schema for AMiner dataset

Field Name	Field Type	Description	Example
id	string	paper ID	013ea675-bb58-42f8-a423-f5534546b2b1
title	string	paper title	Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors
authors	list of strings	paper authors	["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"]
venue	string	paper venue	Journal of Computational Chemistry
year	int	published year	2017
references	list of strings	citing papers' ID	["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"]
abstract	string	abstract	This paper studies ...

Data preparation includes text preprocessing steps, i.e. changing into lowercase, case folding (removing delimiters), stemming (returning words to basic words), and removing stop words (frequent words that considered have no meaning). The results are words and occurrence numbers of words articles generated as vectors of articles. The vectors become data input for clustering to obtain topics, which is discussed in the next chapter.

As mentioned in Section 1.7, the works in this dissertation are about expertise of researchers on specified topics. Therefore, we investigated our mechanisms using AMiner standard dataset of researchers who have expertise on domains of NLP.IE

[49]. Thus, text preprocessing was applied on texts of title-abstract from articles published by 70 researchers of AMiner NLP.IE, in addition to other researchers who have published at least seven publications in total with any of them. There are 212 researchers and ± 4800 articles as AMiner dataset for identifying topics, extracting expertise evidence as features in the scholar profile, and rank expertise with the profile.

Chapter 4.

CLUSTERING FOR IDENTIFYING TOPICS OF RESEARCHERS

Topics are often extracted with a generative model of word distributions from texts of title-abstract. Identifying researchers and their topics is derived from that word distributions over texts into topics distributions [6] [7] [8]. With the issues of interest changes and less focused on citations, the distributions should consider any combinations of topic, citation, and or period for the mixture models. However, unsupervised approach of clustering is preferable to identify coherence words within topics instead of the generative model [21][50]. To support main contributions on extracting expertise evidences as a scholar profile, this chapter discussed designed mechanisms on identifying, mapping and evaluating topics in empirical experiments. The process of identifying and mapping for topics to be used in the later process of extracting evidences have made used of AMiner NLP.IE dataset.

4.1. Clustering with Various Word Embedding

Word embedding process converts texts into numbers by mapping words using a dictionary to vector representations. In general, word embedding is classified into frequency based embedding (i.e. Count Vector, TF-IDF Vector, Co-Occurrence Vector) and prediction based embedding (i.e. Word2Vec [43]). As mentioned before, research topics are groups or clusters of related words. We performed clustering experiments with Python Library NLTK for preprocessing, Gensim [51] for word-embedding (TFIDF, Word2Vec), and scikit-learn for clustering (KMeans).

We explored some scenarios for clustering to get better representations for identifying topics as displayed on Table 4-1 with three datasets. The first dataset D_{big} had all texts of title-abstract from AMiner. The second dataset D_{small} only focused on texts from AMiner experts (<https://aminer.org/data#Expert-Finding>), i.e. Natural Language Processing (NLP) and Information Extraction (IE). The third dataset $D_{small-title}$ was similar to the second one with only titles. $D_{small-title}$ contains \pm 4800 articles from 70 AMiner experts on NLP-IE domain.

Table 4-1 Clustering results with Silhouette indicators for goodness of measurement

Dataset	Min DF	#clusts.	No	Avg.Silh	Algorithm	Matrix Size	Features
<i>D_{small}</i> (±14.000 data x 200 dimension)	10	100	1	0,028	KMeans++	± 3.500 x 200	DF
	3	100	2	0,005	KMeans++	± 7.000 x 200	
	3	50	3	0,068	KMeans++	± 7.000 x 200	
	10	100	4	-0,022	KMeans++	± 3.500 x 200	
	10	100	5	0,264	GaussMix	± 3.500 x 2	FeatAgglo(2)
	10	100	6	0,128	GaussMix	± 3.500 x 2	PCA(2)
	10	50	7	-0,028	KMeans++	± 3.500 x 200	
<i>D_{big}</i> (±62.500 data x 200 dimension)	10	100	8	-0,046	KMeans++	± 12.000 x 200	
<i>D_{small}</i> (±14.000 data x 100 dimension)			9	0,135	KMeans++	± 3.500 x 100	
			10	0,364	GaussMix	± 3.500 x 2	FeatAgglo(2)
		100	11	0,003	GaussMix	± 3.500 x 10	FeatAgglo(10)
			12	0,256	GaussMix	± 3.500 x 2	PCA(2)
		10	13	0,104	GaussMix	± 3.500 x 10	PCA(10)
		50	14	0,115	KMeans++	± 3.500 x 100	
			15	0,179	KMeans++	± 3.500 x 100	
<i>D_{small-title}</i> (±4.200 data x 100 dimension)			16	0,177	GaussMix	± 3.500 x 2	FeatAgglo(2)
			17	0,158	GaussMix	± 3.500 x 10	FeatAgglo(10)
			18	0,215	KMeans++	± 600 x 100	
		10	19	0,527	GaussMix	± 600 x 2	FeatAgglo(2)
			20	0,651	GaussMix	± 600 x 2	FeatAgglo(2)
		21	0,206	GaussMix	± 600 x 2	PCA(2)	

From those articles, there are ± 4200 indexed words and each word has a word vector of 100 dimensions. After using Word2Vec, we defined several combinations of algorithms, such as KMeans++ and Gaussian Mixture, along with features extraction approaches of Document Frequency (DF), Feature Agglomeration, and Principal Component Analysis (PCA).

Clustering results showed that *D_{small-title}* and Word2Vec-KMeans++ were employed as main procedures in topic mapping for this dissertation. There were 30 clusters as shown in Figure 4-2 with their manually labeled keywords in Table 4-2.

Silhouette indicator was used to measure the goodness of clustering results with higher values means that words within the clusters have closer semantic relations [52]. Word2Vec transformation has resulted in a matrix of important words and their word vectors. Word vectors represent their weights in a semantic-kind-of feature space in which words with closer positions should be semantically related



Figure 4-2 Clusters in AMiner NLP-IE domain transformed with LSI

The points and the labels indicated that those words were semantically close. However, it should be noted that the topics were obtained from K-Means clustering with Word2Vec as word embedding. LSI usage was applied after clustering process and aimed for visualizing the topics to help manual evaluation. Another manual evaluation was performed by randomly checking the coherence between words in a topic. Table 4-3 illustrated a cluster that contains the term “ranking” from K-Means clustering and Latent Dirichlet Allocation (LDA) [54], which still used in recent studies for topic modeling.

Table 4-2 Clusters in Aminer NLP-IE Domain with transformed positions by LSI

C	Keyword Label	PosX	PosY
1	['approximate inference']	2.85	-0.48
2	['systems design', 'documents retrieval']	3.20	0.01
3	['answer finding', 'annotation framework', 'hierarchy topics']	2.92	0.27
4	['task support', 'algorithms natural']	3.16	-0.18
5	['ontology learning', 'cognitive science']	2.74	-0.26
6	['user modeling']	2.90	0.06
7	['abstract', 'results', 'domain', 'indexing']	2.87	-0.32
8	['expression content', 'joint bilingual', 'parser rule']	3.23	0.45
9	['document management', 'interaction interactive']	3.17	0.38
10	['classifier features', 'method automatic']	3.09	0.90
11	['online multimedia', 'words probabilistic']	2.78	-0.15
12	['discourse model', 'search databases', 'performance information']	3.17	0.49
13	['improve plans', 'analysis agent']	2.82	-0.46
14	['parsers corpus', 'fields extracting']	2.85	-0.36
15	['measures entailment']	2.95	-0.51
16	['annotations platform', 'paraphrases textual', 'extract link']	3.23	-0.26
17	['internet wrapper', 'searching browsing']	2.27	-0.04
18	['argument relation', 'dictionary tagging', 'predicting story']	2.85	-0.53
19	['language interpretation', 'intelligent query', 'data acquisition']	3.11	-0.43
20	['identifying noun', 'extraction question', 'summaries scientific']	3.31	1.18
21	['semantic parsing', 'linguistic models', 'computational lexicon']	2.88	-0.45
22	['collaborative filtering', 'sentence classifiers']	2.88	-0.09
23	['trees formal', 'software agents', 'agents software']	2.90	-0.57
24	['annotating', 'sentiment']	3.20	-0.17
25	['structures efficient', 'networks distributed']	3.04	0.62
26	['induction techniques', 'grammar rules']	3.05	0.52
27	['speech recognition', 'structure knowledge']	3.02	-0.26
28	['algorithm system', 'state methods']	2.92	0.74
29	['theory', 'inferring', 'role']	3.12	-0.12
30	['building collections', 'electronic dictionaries']	2.82	-0.36

Some words were in the phrase forms which was extracted using graph-based analysis [55]. We identified topics from $D_{small-title}$ after embedding using Word2Vec, then set parameters of document frequency (DF) and dimensions of word vectors. The results demonstrated that K-Means gave more semantically related words within the topic, in which the results using parameters of DF:10 - dims:100 gave more inter-related words than other K-Means results.

Table 4-3 Topic words identified from probabilistic model and K-Means clustering

LDA DF:3, dims: 100	KMeans (filter) DF:3, dims: 100	KMeans DF:10, dims: 100	KMeans DF:10, dims: 200
accomplished augmentative beliefs clarify companion competence continue distant enabled encoding modal portals precisely preprocessing ranking relied robustness roles scoring separately transparency	effectiveness feedback ranking relevance result term topic trec	approaches constraints design documents explanations future generation patterns ranking retrieval structured systems time	approaches constraints documents experiment explanations future generating patterns ranking retrieval university

4.2. Mapping Topics to Articles and Researchers in AMiner dataset

Mapping process includes mapping identified topics for each article, and then mapping them to researchers as their interest which also become their expertise. Each article could be a mixture of topics. We used domains of NLP.IE and identified 30 topics that makes a higher possibility of inter-related topics. Pseudo code for mapping topics to articles is listed on Figure 4-3 with typical similarity method between vectors called as Cosine Similarity. After manual analysis on articles in the dataset, each article customarily is a mixture of 2-3 topics, since the domains of NLP and IE are inter-related. Researcher could be recognized to have interest on a topic from published articles in that particular topic. Therefore, mapping topics of a researcher depends on listed topics from his or her articles. We obtained a list of topics as the interest for researchers by using pseudo code in Figure 4-4. With data collection of

researchers, topics, mapped articles, and mapped interests, we generated matrix input as sources for extracting features as illustrated in Figure 4-5. Maximum number of columns in the matrix for representing topics were 30, so the column numbers were 32. Then, maximum number of rows in the matrix were depended on observed years and number of researchers.

```

MapArticleTopic() # labeling 2-3 topics to each article
1.  Input:
2.   D: set of articles {dj}, dj is a word vector, an array structure
3.   C: set of clusters as topics {ck}, ck is a word vector
4.  Output:
5.   L: set of labels for articles {(dj, {ck})}, a dictionary structure

6.  For each article dj in set D
7.   Lj = array of similarity values of article dj
8.   For x = 1 ... |C|
9.     Lj[x] ← cosim(dj, cx)cx ∈ C # Cosine Similarity between two vectors
10.  Sort Lj in descending order
11.  L ← (dj, {Lj[1], Lj[2]})
12.  If |Lj[2] - Lj[3]| ≤ 0.001: L[dj] ← Lj[3]

```

Figure 4-3 Pseudo code for function MapArticleTopic()

```

MapResearcherTopic() # obtaining topics as the interest of researchers

Input:
A: Collection of authors {ai}, an array structure
CA: Collection of co-authors {(dj, {ai})}
L: Collection of labels for articles {(dj, {ck})}, a dictionary structure
Output:
RI: Collection of interest of researchers based on published articles {(ai, {ck})}

1.  For each author {ai}
2.   Set Dai as collection of articles authored by ai obtained from CA
3.   For article dj in Dai
4.     Set Ldj as list of mapped topics from dj obtained from L
5.     Iterate Ldj and add each topic as topic of ai into RI if the topic has not
       been added yet

```

Figure 4-4 Pseudo code for function MapResearcherTopic ()

AuthorID	Period	T1	T2	...	T30
A1	Year1	2	3	...	0
A1	Year2	0	2	...	3
...

Figure 4-5 Matrix of researchers, topics, and article numbers as sources for extracting features

Table 4-4 Some topics with their words within and the possible domains in NLP.IE

AMiner domain	NLP (89%)	NLP (71%)	NLP.IE	IE (67%)
Topic ID	T2	T10	T13	T29
Words within a topic	constraint	automatic	agent	artificial
	design	classifier	analysis	environments
	document	features	improve	inferring
	explanation	inductive	library	role
	future	message	plans	selection
	generation	refinement	program	strategies
	pattern	relevance	student	theory
	ranking	sets	work	
	retrieval	topic		
	structured	tree		
	system			
	time			

The words as labels of topics in Table 4-2 indicated that they are in NLP domain. After mapping topics to articles, listing topics of researchers and selecting three topics as their main interest, then retrieving only the mostly mapped topics to obtain T2, T10, T13, and T29 in Table 4-4. We selected some researchers for rank their expertise later. Then, we analyzed their mapped topics and their domains of NLP.IE. Those four topics were mapped to both domains in some degrees, T2 for 89% NLP, T10 for 71% NLP, T29 67% IE, and T13 accommodated researchers in both. This findings indicated the words related to IE domain were not distinctive.

4.3. Evaluating Topics for Recommendations

We have identified topics using clustering with Word2Vec word embedding beforehand on title texts of published articles by AMiner NLP-IE researchers. This section described our implementations after identifying topics with clustering on

some situations related to recommendations of researchers. The investigated focus is on clustering approach and not on identified topics from AMiner NLP-IE researchers. Section 4.3.1 discussed identified topics on ITS dataset as mentioned in Section 1.7 to investigate whether the approach of straightforward clustering is applicable on recommendations in finding researchers as collaborators. Section 4.3.2 also described the same dataset in a visualization based on Scopus research area. Then, Section 4.3.3 used identified topics from AMiner in a case of contextually inconsistency between the contents of cited and citing articles.

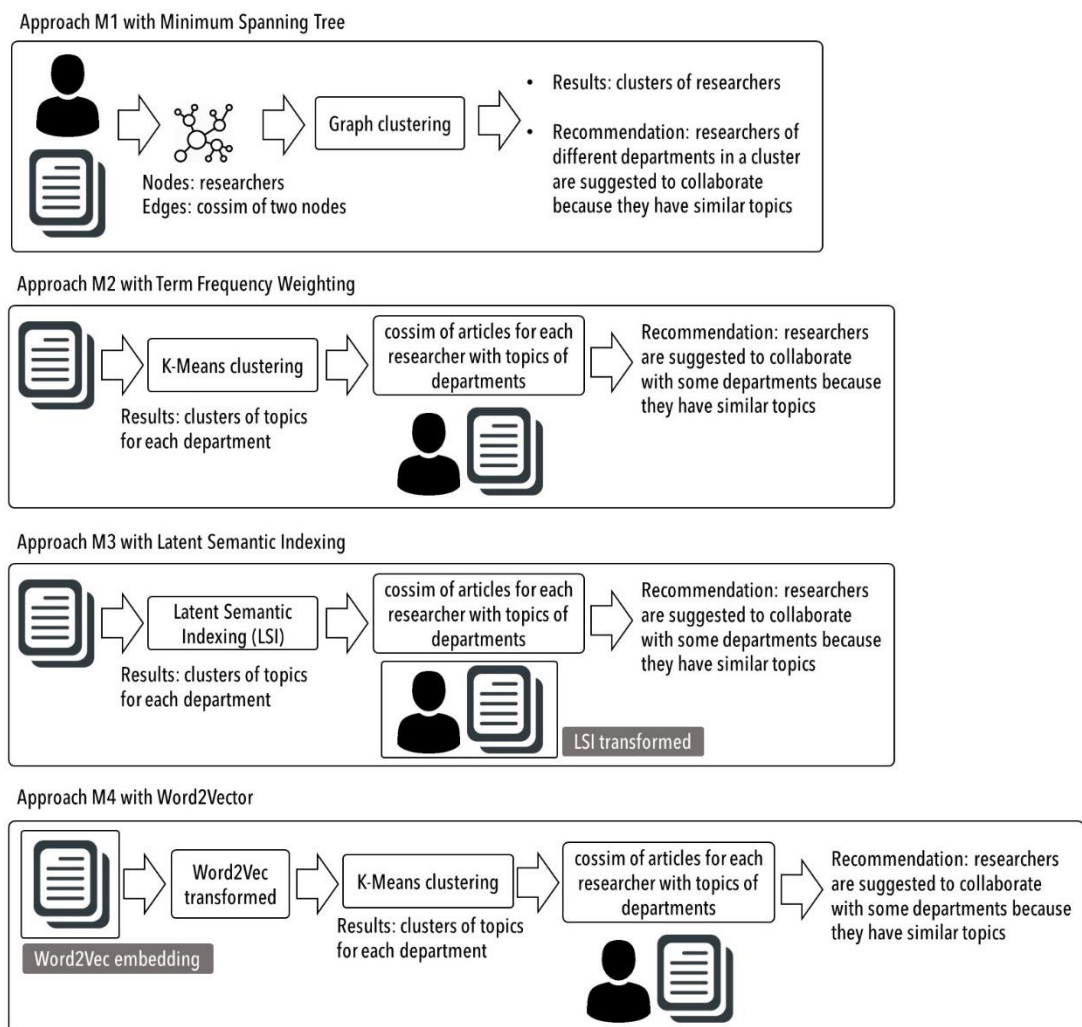


Figure 4-6 Approaches for intra- departmental recommendation system

4.3.1. Cross-Domain Collaborating for Researchers

We have investigated topics from clustering for identifying the topics of researchers (Figure 4-6) with dataset from our university (ITS dataset) because of the familiarity aspect. Topics were not clustered from the previous dataset $D_{small-title}$ that came from AMiner NLP.IE dataset, but from title texts of undergraduate and graduate thesis.

The identified topics were expected to become the references when supporting inter-departmental collaborations. We compared the topics using word-embedding (Word2Vec) with other representations like term frequency matrix as well as its projected forms (Figure 4-6): from term frequency, graph-based, Latent Semantic Indexing (LSI) until state-of-the-art word vector Word2Vec.

All approaches (term frequency weighting, Latent Semantic Indexing (LSI), Word2Vector) had text preprocessing and only focused on verbs-nouns after checking parts of speech to each word with the Indonesian Thesaurus, Kateglo. Beginning with preprocessing and then identifying research topics (topic modeling and competence mapping for researchers in departments), before recognizing collaboration across different departments. Pseudo codes for those approaches are listed in Figure 4-7 and Figure 4-8. Our experiments utilized Python, SQL Server, Gephi environments and other libraries such as Sastrawi for Indonesian Stemmer text preprocessing, Vis.js and NodeXL for visualizing graph, and Gensim for topic modeling.

A graph for Model M2 was created with 983 researcher nodes and 460,361 edge relations, and then simplified using minimum spanning tree (MST) into 958 edges. Visualization results of our models were in Figure 4-9, Figure 4-10, and Figure 4-11. Our data experiments came from the following faculties with ± 14.000 texts in total from 2005-2015, which have different compositions compared to the faculties and departments of ITS in 2020:

- F1 for Mathematics and Science (red),
- F2 for Industrial Technology (green),
- F3 for Civil Engineering and Planning (blue),
- F4 for Marine Technology (yellow) and
- F5 for Information Technology (black).

Preprocessing Steps

1. for each researcher l_j in dataset L do
2. select thesis titles from the collection D with the corresponding supervisors into researcher's collection D_j
3. for each student thesis title text in the selected collection D_j do
4. index terms (of nouns and verbs) and calculate inverted document frequency weight for each indexed term, the result is term matrix M_j with rows are records of student thesis and columns are tokens of thesis titles
5. generate an indexed term vector of current researcher $vl_j = \{w_m\}$ saved in vector set V
6. for each faculty-x department-y dep_{xy} do
7. select thesis titles from the collection D with the corresponding department into collection D_{xy}
8. for each student thesis title text in the selected collection D_{xy} do
9. index terms (of nouns and verbs) and calculate inverted document frequency weight for each indexed term, the result is term matrix M_{xy} with rows are documents of student thesis and columns are tokens of thesis titles (called as term weighting)
10. do K-Means clustering on term matrix of thesis titles M_{xy} from the corresponding department collection D_{xy} , saved in list of topic clusters $T_{xy} = \{tp_{xy1}, \dots, tp_{xya}\}$
11. convert each resulted topic cluster from set of student thesis titles $tp_{xy.i} = \{t_o\}$, into set of tokens $tp_{xyi} = \{w_m\}$ in which the tokens exist in the student thesis title

Model M1 graph-based Minimum Spanning Tree

1. for each combination of researchers l_i, l_j in L do
2. select texts from the collection D with the corresponding supervisors into researchers' collection D_{ij}
3. calculate edge value e_{ij} from the selected collection D_{ij} and researcher vectors of vl_i, vl_j saved in edge set E
4. do graph-based clustering algorithm [15] on generated co-network with information from sets of L and E , the result is clusters of researchers in which each cluster can contain researchers from different departments. The cluster can be the reduced number of edges cross-domain collaborative recommendation.

Model M2 matrix-based with term frequency weighting

1. for each researcher l_j in dataset L do
2. for each faculty-x department-y dep_{xy} do with condition that $l_j \notin dep_{xy}$
3. for each topic of dep_{xy} in T_{xy} (use K-Means clustering results in the preprocessing steps)
4. calculate cosine similarity of $vl_j = \{w_m\}$ and $tp_{xy.i} = \{w_n\}$, $cossim(vl_j, tp_{xyi})$
5. if the similarity value $> thresh_2$ then $tp_{xy.i}$ is recommended as cross-domain topic for l_j

Figure 4-7 Pseudo code for recommendation using model M1-M2

<p><u>Model M3 Latent Semantic Indexing</u></p> <ol style="list-style-type: none"> 1. for each faculty-x department-y dep_{xy} do 2. transform Latent Semantic Indexing of term matrix M_{xy} into LSI_{xy} (use term weighting results in the preprocessing steps) 3. for each topic cluster of dep_{xy} in T_{xy} (use K-Means clustering results in the preprocessing steps) 4. do LSI projection for $tp_{xyi} = \{w_m\}$ based on LSI_{xy} 5. for each researcher l_j in dataset L do 6. for each faculty-x department-y dep_{xy} do with condition that $l_j \notin dep_{xy}$ 7. for each topic of dep_{xy} in T_{xy} 8. do LSI projection for vector $vl_j = \{w_m\}$ based on LSI_{xy} 9. do step 4 and step 5 from Model M2 ($tp_{xy.i}$ use step 4) <p><u>Model M4 word vector based model</u></p> <ol style="list-style-type: none"> 1. for each faculty-x department-y dep_{xy} do 2. transform Word2Vec of D_{xy} into WV_{xy} (use D_{xy}, convert $D_{xy} = \{t_o\}$ into $D_{xy} = \{w_m\}$ from the preprocessing steps) 3. for each topic of dep_{xy} in T_{xy} (use K-Means clustering results in the preprocessing steps) 4. do Word2Vec projection for $tp_{xyi} = \{w_m\}$ based on WV_{xy} 5. for each researcher l_j in dataset L do 6. for each faculty-x department-y dep_{xy} do with condition that $l_j \notin dep_{xy}$ 7. for each topic of dep_{xy} in T_{xy} 8. do Word2Vec projection for vector $vl_j = \{w_m\}$ based on WV_{xy} 9. do step 4 and step 5 from Model M2 ($tp_{xy.i}$ use step 4) <p><u>Create validation set for the cross-domain recommended topics of M2</u></p> <ol style="list-style-type: none"> 1. for each researcher l_j in dataset L do 2. for each faculty-x department-y dep_{xy} do with condition that $l_j \in dep_{xy}$ 3. calculate cosine similarity of $vl_j = \{w_m\}$ and $tp_{xyi} = \{w_n\}$, $cosim(vl_j, tp_{xyi})$ 4. if the similarity value $> thresh_1$ then tp_{xyi} is topic competences for l_j 5. for each researcher l_j in dataset L do 6. for each faculty-x department-y dep_{xy} do with condition that $l_j \notin dep_{xy}$ 7. calculate cosine similarity of all topic competences of researcher l_j and all topics in list T_{xy} (use K-Means clustering results) 8. if the similarity value $> thresh_2$ then tp_{xyi} is recommended as cross-domain validation set for l_j

Figure 4-8 Pseudo code for recommendation using model M3-M4

Sphere nodes with varying size showed the faculties while triangle nodes showed the researchers with their representative faculty colors. Bigger size of a sphere node indicates there are more researchers in the particular faculty who have collaborated with other researchers from different departments.

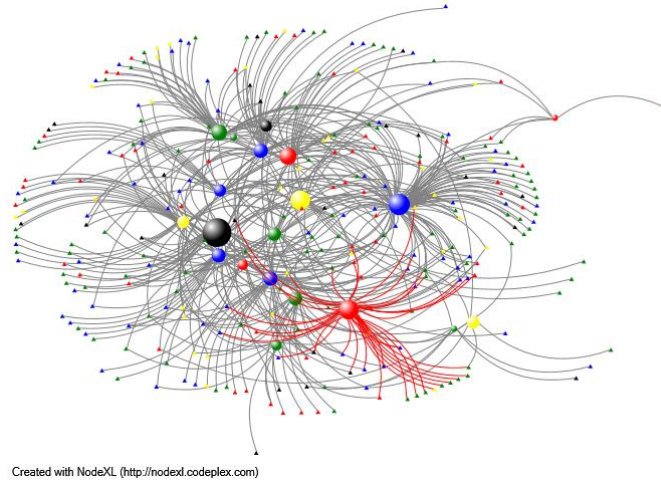


Figure 4-9 Possible cross-domain collaborative studies using TF-IDF (M2)

Social media network connections

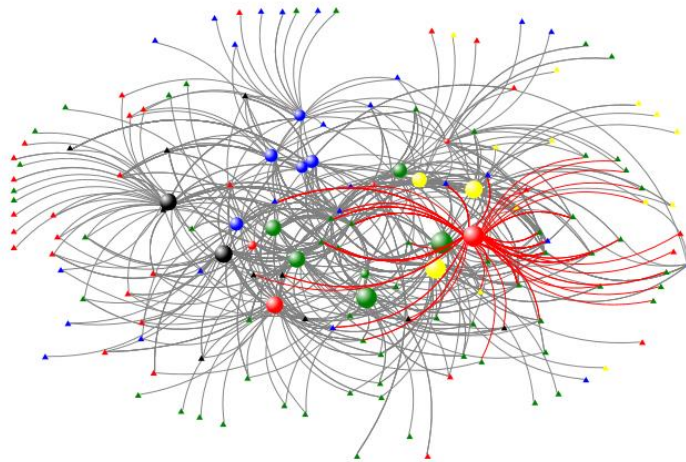


Figure 4-10 Possible cross-domain collaborative studies using Latent Semantic Indexing (M3)

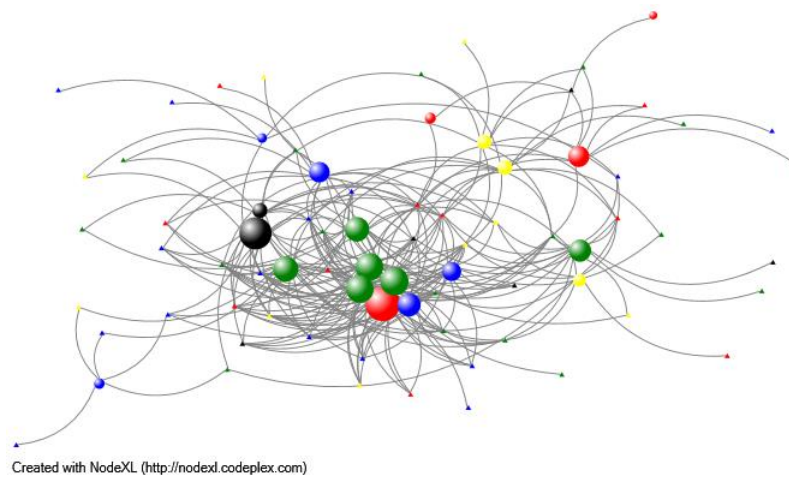


Figure 4-11 Possible cross-domain collaborative studies using Word Vector (M4)

Even with the same dataset, the visualization results emphasized on different faculty nodes caused by the approaches of M1-M4. The usage of Model M2 in Figure 4-9 suggested excessive recommendations with entangled visuals. To simplify the models, we transformed term matrix into a latent-topic matrix using LSI in model M3 as shown in Figure 4-10.

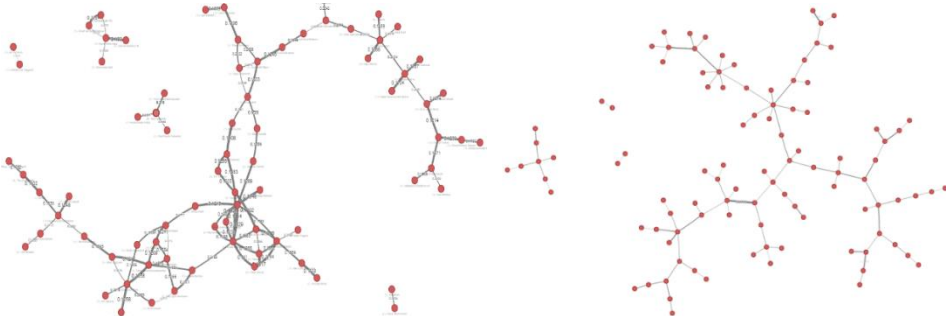
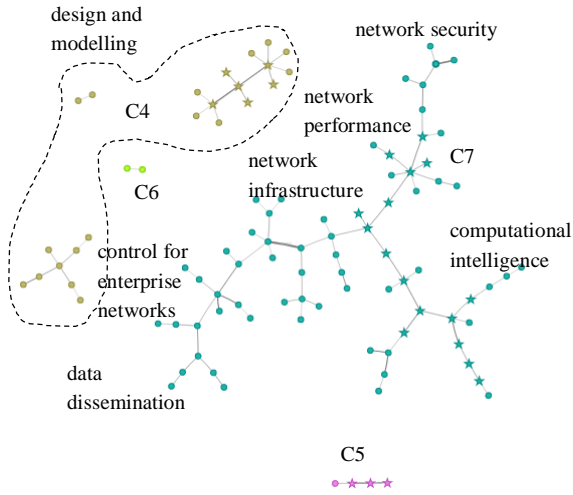
LSI extracted principal features of latent-topics from term matrix. Although LSI took time but it reduced matrix dimension and showed a less entangled visual. For example, faculty node F5 (black color, Informatics Engineering and Information System) have relations with lecturers from departments of F1 (red color, Mathematics, Statistics) and F2 (green color, Electrical Engineering). Then, the illustration of model M4 demonstrated a much less entangled visual compared to other models of M2 and M3, although the correctness of cross-domain topics was still questionable.

In our experiments for model M1, we compared the clustering results using K-Means with and without MST, K values = 5, 7, 10, and 16 as shown in Table 4-5. Dunn Index (DI) evaluation with higher index scores refer to better clustering results. Since words of topics can have different lexical words but convey the same meaning in contexts, implementation of synonym expansion was also explored. Compared to clustering results of graph-based K-Means with MST + synonym expansion, K-Means without MST gave better DI score. Setting C (minimal threshold weight value of edge) as a replacement setting of K (or N, number of clusters) in graph-based K-Means with MST showed that the value of C reduces the number of remaining edges. Due to synonym expansion, some nodes were connected to lecturers from other faculties and made less DI scores.

Table 4-5 Dunn Index of Clustering Results

K Value	K-Means without MST	Graph-based K-Means with MST	Graph-based K-Means with MST + synonym
5	0.701	0.847; C=0.81; N=7	0.697 ; C=0.4; N=7
7	0.701	0.858 ; C=0.84; N=15	0.660; C=0.4; N=5
10	0.695	0.844; C=0.84; N=10	0.660; C=0.4; N=4
16	0.566	0.810; C=0.84; N=7	0.652; C=0.4; N=2

Table 4-6 Student questionnaires

No	Question Descriptions
1	<p>Q: Which co-authorship network representation that reflects more on researchers' specialties at FTIF (with or without MST)?</p> <p>A: There are 28 students chose right figure because the representation of co-authorship network cannot be shown due to too many network edges.</p>  <p>Visualization of co-authorship networks at faculty level (Faculty of Information Technology, FTIF-ITS) before clustering without MST and with MST</p>
2	<p>Q: Are FTIF researchers with higher centrality scores consistent with their positions as centroids of MST?</p>  <p>The cluster anomaly of design and modelling in Informatics department is still related to research fields of Industrial Engineering department but it focuses on designing and modelling for industrial purposes. The research field of computer networking security is similar and can be clustered to network and radio communication which belongs to another department (Electrical Engineering in Faculty of Industrial Technology).</p>

We distributed a questionnaire about model M1 as shown in Table 4-6 to 32 students in the 7th semester of Informatics Department, who actively participate in laboratories for assisting other students or becoming administrators. Questionnaire contents were about evaluation on co-authorship network of FTIF researchers.

Students were requested to observe the networks of FTIF researchers from different experiment scenarios. Those students attended elective courses and were going to complete their courses in 7th or 8th semester. They have already made some preliminary studies, and made them familiar with expertise of the researchers.

The absence of reliable cross-domain recommendations to compare our models of M2-M4 created an evaluation obstacle. We asked domain experts to manually check recommended topics to investigate whether the topics were mapped correctly. The average precision of suggested topics for lecturers in all faculties compared to the results from domain experts is shown in Figure 4-12. It shows that mapping topics for cross-domain recommendations are better when words of topics have been transformed like LSI and Word2Vec, although Word2Vec showed better precisions in most faculties.

For manual evaluation, Figure 4-13 and Figure 4-14 illustrated a topic from KMeans with and without Word2Vec transformation on a cluster result in which the term “retrieval” exist. LSI result was omitted because of listing words of one topic requires to analyze word positions in the particular latent topic before deciding insertion of the words. Each line in a sample topic in those figures represented a title. Those titles were grouped in sub clusters manually for evaluation purpose. More titles were existed in Figure 4-13, but have been disregarded because of context disparity. Sub clusters in Figure 4-14 were more coherence especially in Sub-cluster 2 about game related works.

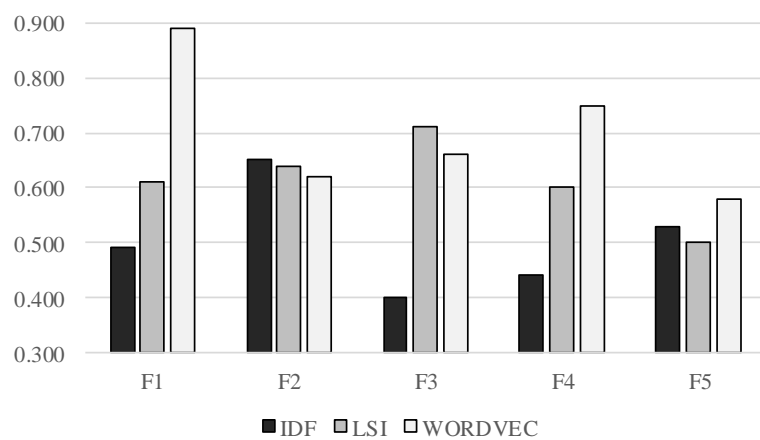


Figure 4-12 Precision comparison of cross-domain collaboration for five faculties

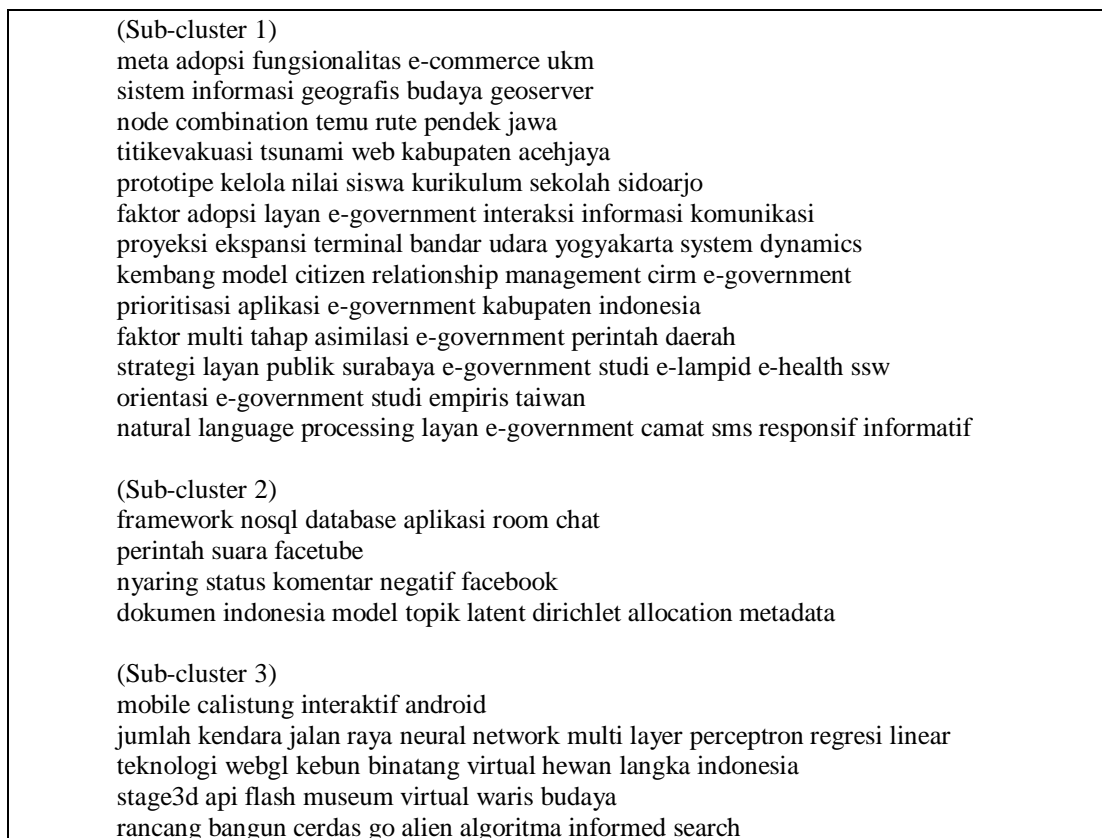


Figure 4-13 K-Means Clustering result without Word2Vec

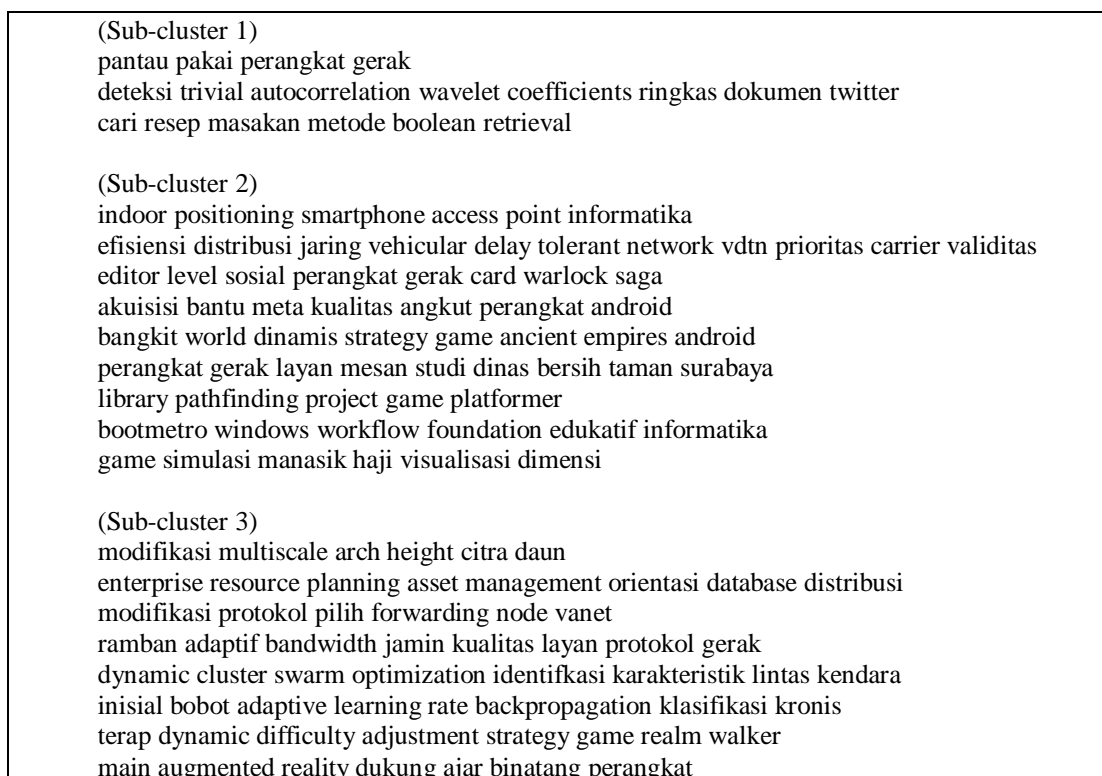


Figure 4-14 K-Means Clustering result with Word2Ve

A closer inspection of mapping topics to find preferences on other researchers' topics, leads to a finding related to cross-domain research works. Researchers in the departments of basic sciences often do the same topics with the applied departments, such as Mathematics and Statistics, which have closed connections with applied departments of Informatics Engineering, Information System, and Electrical Engineering. The university management through its research policy can nurture this phenomenon of cold-start recommendation further. Some of those similar departments are listed on Table 4-7.

Table 4-7 Major cooperation departments extracted from clustering results (graph-based k-Means with MST, $k=7$)

Cluster	Faculty	Consisted Departments
C1	F1, F2	Chemistry, Physics, Engineering Physics, Material & Metallurgical Engineering
C2	F1, F2, F3	Business Management, Industrial Engineering, Civil Engineering, Environmental Engineering, Chemistry, Biology, Mathematics, Geophysics Engineering
C3	F2, F3, F4	Electrical Eng., Marine Engineering, Urban & Regional Planning, Architecture, Geomatics Engineering
C4	F2, F4, F5	Naval Architecture, Informatics Engineering, Electrical Engineering, Statistics, Marine Engineering, Marine Transportation, Mechanical Engineering, Material & Metallurgical Engineering
C5	F2, F3, F4, F5	Industrial Product Design, Interior Design, Electrical Engineering, Multimedia & Network Engineering, Information System, Mechanical Engineering, Marine Engineering
C6	F2, F5	Informatics Engineering, Information System, Electrical Engineering
C7	F1, F2, F3, F4	Ocean Engineering, Geomatics Engineering, Biology, Chemical Engineering

4.3.2. Visualizing Researchers based on Topics

Previous section has demonstrated a straightforward clustering to identify topics of researchers with other matching process was applicable for cross-domain recommendation. A visualization could complement recommendations to confirm the topics of researchers. Studies introducing visual approaches to gain insights into science mapping are roughly categorized into topic content, topic relationship, and topic evolution [56]. For topic content, a cartographic approach displayed scientific literatures from keywords and their semantic relations on the map [57]. For topic

relationship, tree-based structures represented co-citations [48], but for topic evolution there was a visualization of emerging topic statuses with semantic relatedness among keywords [58].

By focusing on topic content, we defined four main processes to visualize researchers on standardized map of topics as shown in Figure 4-15 [39]:

- a. collecting metadata of researchers and domain related to Scopus subject area
- b. transforming metadata according to Scopus subject area
- c. scaling for the transformed articles to display the experts on the base map

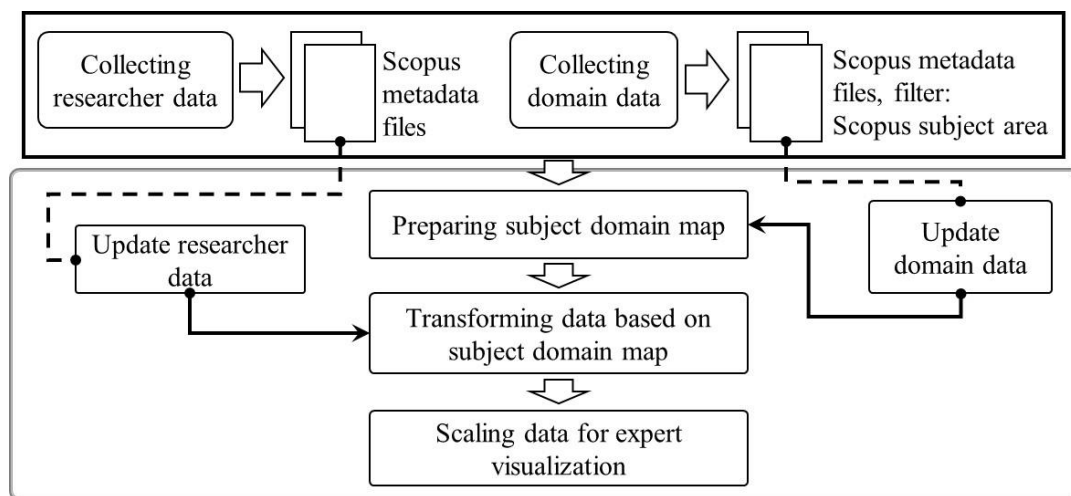


Figure 4-15 System architecture for visualizing academic experts on a subject domain map of cartographic-alike

Data Collection

We had two datasets for visualization called as Researcher Data and Domain Data which basically were Scopus article metadata of title-abstract texts. For Research Data, we listed top 200 researchers in our university based on Scopus h-index and manually downloaded their metadata of 3182 articles with “computer science” keyword existing in Scopus descriptions. Although the dataset is ITS data, but we did not employ all researchers. Only those of 200 researchers from eight faculties taken in the year of 2019, and they should have at least ten Scopus published articles. The following list of researchers demonstrated that “computer science” domain was applied on several fields with mostly on FTI, FTE, FTIK, and FMKD.

1. FTI (industrial technology)	46 researchers,
2. FTE (electrical technology)	44 researchers,
3. FTIK (information and communication technology)	39 researchers,
4. FMKD (mathematics, computation, and data science)	24 researchers,
5. FTSLK (civil, environmental, and geo engineering)	14 researchers,
6. FS (basic science)	9 researchers,
7. FTK (marine technology)	8 researchers, and
8. FBMT (business and technology management)	1 researchers.

For Domain Data, there are two levels of categories in Scopus subject area which resulted in 26 Scopus subject areas. We collected at least ± 2000 article metadata of title-abstract published from 2017-2018 for each subject area, and resulted into 51,939 bib-items of articles. However, for the next processes we only used information of 1st tier subjects (Health Sciences-HS, Life Sciences-LS, Physical Sciences-PS, and Social Sciences-SS). For all those works, we used Python packages, i.e. BibtexParser for parsing raw Scopus metadata and Natural Language Toolkit (NLTK) for text processing.

Data Transformation

In this phase, we processed metadata with Word2Vec approach to get semantic relations between keywords. Word embedding with Word2Vec of 200 dimensions was required to set some parameters, such that after experimenting on a number of combinations, we selected Skip Gram and set distance window to five terms for checking semantic relations with nearest neighbor terms on Gensim package. We performed word embedding with those parameters on Domain Data after tagging title-abstract texts based on parts of speech and only processed the noun keywords. The result is an embedded matrix of $\pm 75K$ keywords x 200 dimensions called as Domain Dictionaries, which is updatable for recent metadata of Scopus subject areas, i.e. articles after 2018.

After embedding data, we transformed Researcher Data and Domain Data using Domain Dictionaries to make the articles into 200 dimensional vectors called as Domain-based Article Data ($\pm 55K$ articles). To visualize the articles on 2D map

of Scopus subject areas, all vectors in Domain-based Article Data had 2D transformation of t-SNE (t-Distributed Stochastic Neighbor Embedding) [59] into x-y coordinates. Two closer article-points means that both articles have similar context and there is a high likelihood of articles contain similar keywords. For all those works, we used Natural Language Toolkit (NLTK) for text processing and Gensim for word embedding. If the Domain Dictionaries is updated with recent metadata, we should also perform 2D transformation of t-SNE on the new Domain-based Article Data.

We already had 2D positions of articles in Researcher Data and Domain Data. Then, the position of a researcher a_i was obtained from the average coordinate value of all article points belonging to a_i .

Data Visualization

From article-points of Researcher Data and Domain Data, we obtained minimum and maximum values of x-y coordinates and made a base map. Then, the map was divided into square grids of 3x3 units and all researcher-points were placed on the map as shown in Figure 4-16.

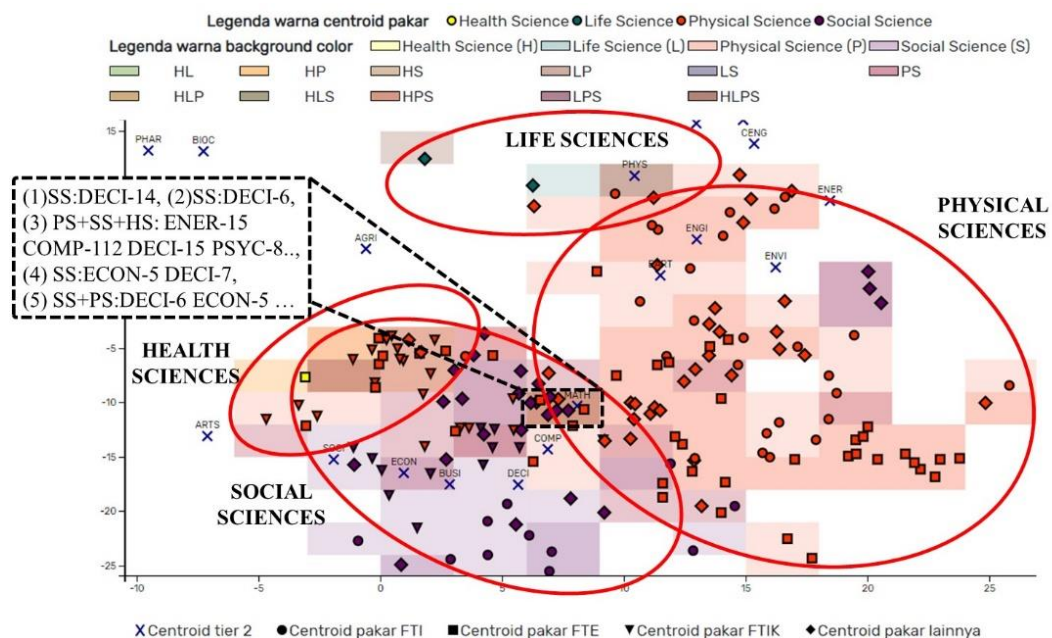


Figure 4-16 Visualization experts with subject domains

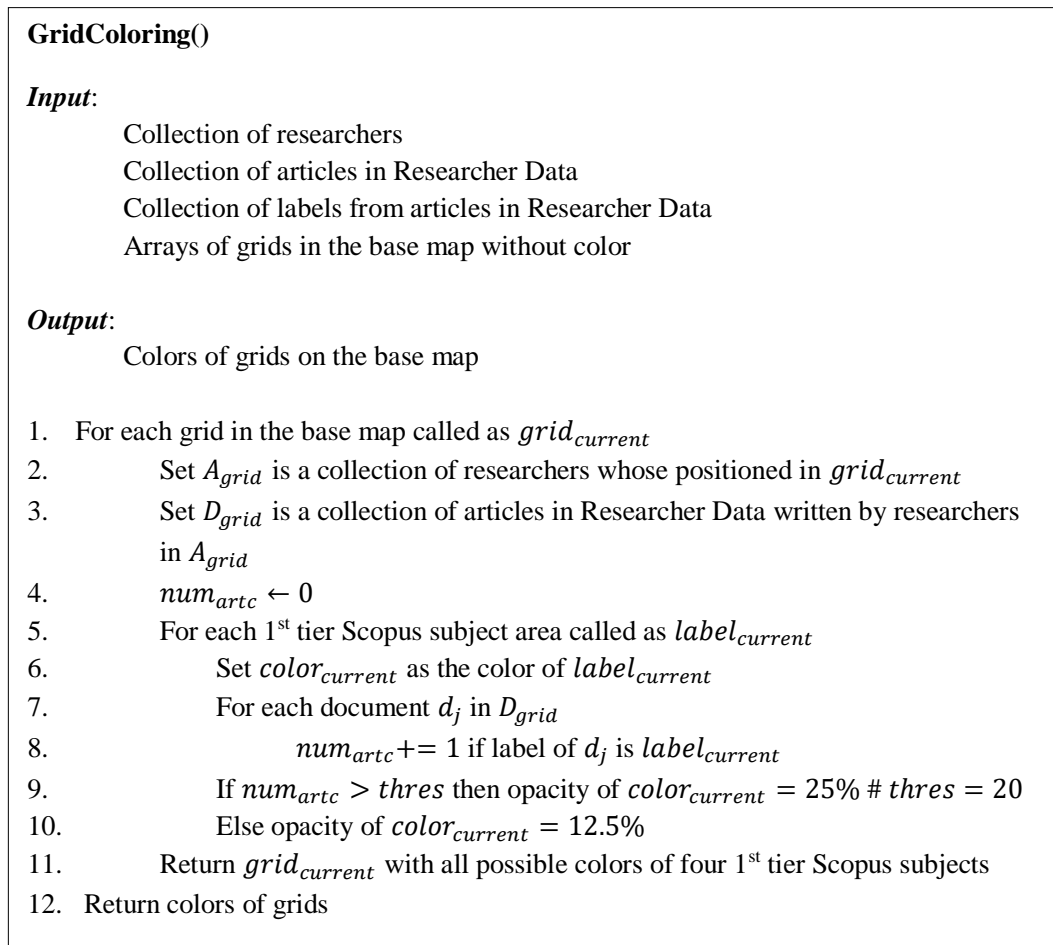


Figure 4-17 Pseudo code for coloring grids on the base map of Scopus subject areas

The color of each grid in the map was depended on subject areas of all article-points belong to researchers in the current grid as shown in Figure 4-17. We labeled the articles of Researcher Data according to the closest distance to article-points of Domain Data with kNN (k=100) (K Nearest Neighbour).

There were two Scopus labels for articles of Domain Data, 1st tier of four subjects and 2nd tier of 26 subjects. However, after experimenting on labeling with two types of labels, the results showed that 1st tier subjects were giving better representation. Before coloring the grids in the base map, all articles of Researcher Data had labels of Scopus 1st tier subject areas. For all those works, we used Scikit-learn for labeling, Seaborn and Matplotlib for visualization, in addition to Mpld3 for bringing the visual into web browser. As shown in Figure 4-16, grid colors were

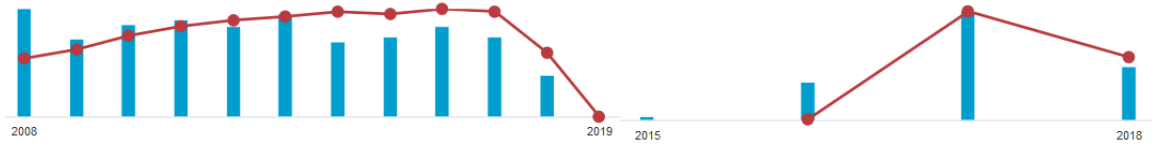
varied according to the opacity colors, i.e. sample grid with combinations of HS-25% + PS-25% + SS-25%.

The visualization in Figure 4-16 could be in commonplace level compared to the existing tools for visualization of sciences such as ScienceScape, Tableau Maps, VOSviewer, or CiteSpace for displaying topics of researchers, topic relations, etc. However, there are some limitations to use those tools such as not easily modified. This dissertation is motivated by the need to profile experts which does not necessarily researchers. The experts could be in organizational context which makes applying those tools is infeasible. Other difficulties occurred when the institution requires the visualization feature is directly connected to other internal systems. Therefore, further works for visualizing researchers according to specified topics to represent the expertise is still promising.

4.3.3. Extracting Conflict-of-Interest-based Features

Although research collaborations are encouraged, there are some disagreeable effects such as the rise of citation number for inflating h-index value which is conducted by co-authors or co-authors of co-authors. Using features of citation quantity could be insufficient to describe the expertise of researchers because of the citation misconduct possibility as shown in Figure 4-18. A researcher who is expertise on specified topics supposed to be productive as shown by citation quantity and be recognized as shown by citation quality. The anonymous researcher A in Figure 4-18 has demonstrated his or her expertise through normal trends of documents and citations. However, the anonymous researcher B who was awarded because of his or her productivity, which was corrected then because of some complaints, has demonstrated anomaly in the trends. Thus, condition for researcher B should be avoided.

Those anomaly could be caused by inflating citations which is a misconduct behavior for a researcher. As an illustration in Figure 4-19, there are two researchers a_x and a_y who became co-authors and a_x in article p_i cited by article p_j . When there is a researcher a_z who never becomes co-authors a_x and a_y and citing both articles p_i and p_j , it validates content relations between articles.



Document and citation Scopus trends of undisclosed authors without (Author A in the left, 300 citation difference between 2016-2017) and with (Author B in the right, >700 citation difference between 2016-2017) conflict of interest indication

Figure 4-18 Conflict of interest indication based on Scopus trends

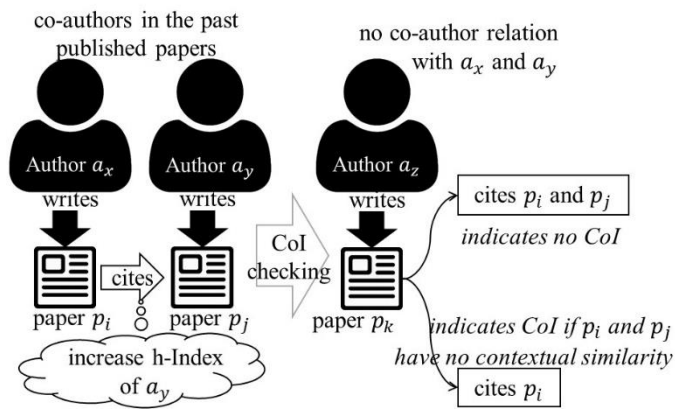


Figure 4-19 Illustration for conflict of interest indication

However, if a_z only cites article p_i , it shows a possibility of Conflict of Interest (CoI) situation between a_x and a_y in articles p_i and p_j that require further assessment. In order to avoid CoI anomaly-like bias caused by self-citation influence [31] or citations to unrelated works that interfere with the purpose of research track records [32], we defined the following three CoI-based features [38] :

- a. Self-citation feature, CoI_1 indicates researcher misconduct to increase the expertise quality by inflating h-index [60] from biased citations as shown in Figure 4-20.

With a range value of 0.0 – 1.0, unbiased researchers on citing behavior are expected to have lower values of CoI_1 . To calculate CoI_1 , additional information aside from

title and abstract texts is necessary, such as list of citations or references for each article.

Calculate_CoI1(a_i) # parameter author identifier
Input:
 Collection of articles
 Collection of citations
Output:
 A value for self-citation feature, higher value means that the author is likely to have misconduct on inflating h-index
Needed Functions:
 self.cite $_{a_i}(d_j)$ returns number of self-citations by author a_i in article d_j
 num.cite(d_j) returns number of references in article d_j

1. Set D_{a_i} to be a collection of articles authored by a_i
2. $temp_{d_j} = 0.0$
3. For each article d_j in D_{a_i}
4. $temp_{d_j} += \frac{\text{self.cite}_{a_i}(d_j)}{\text{num.cite}(d_j)}$
5. Return $temp_{d_j}/|D_{a_i}|$

Figure 4-20 Pseudo code for obtain CoI_1

b. Researcher-interest feature, CoI_2 indicates relatedness between a researcher to others whose articles have been cited through similarities on research interests as shown in Figure 4-21.

Calculating this feature requires known list of research interests for each researcher. We have experimented CoI features on selected data of AMiner dataset. We used Word2Vec model of pre-trained word and phrase vectors from Google News dataset (code.google.com/archive/p/word2vec) to transform and then calculate semantic similarities between texts. For our experiments, we mapped all texts into 100-dimensional vectors with the pre-trained Word2Vec model. Similar to using a range value of CoI_1 , researchers who cited articles of others without conflict of interest are expected to have higher values of CoI_2 . To calculate CoI_2 , the necessary information is a list of interest for each researcher.

c. Contextual similarity feature, CoI_3 ensures subjects between an article and its citations to have connected concepts as shown in Figure 4-22.

We used deep learning approach with Siamese architecture based on Long Short Term Memory (LSTM) to check subject relations [61] from texts of title-abstract. Researchers who cited articles of others without conflict of interest are expected to have higher values of CoI_3 similar to CoI_2 .

Calculate_CoI2(a_i) # parameter author identifier
Input:
 Collection of articles
 Collection of research interests
Output:
 A value for author-interest feature

1. Set D_{a_i} to be a collection of articles authored by a_i
2. Set RI_{a_i} to be a collection of research interest for author a_i
3. $temp_{d_j} = 0.0$
4. For each article d_j in D_{a_i}
5. $RI_{co.ai.d_j} \leftarrow$ Get a list of research interest from authors of d_j
6. $temp_{d_j} +=$ get semantic similarities between RI_{a_i} and $RI_{co.ai.d_j}$ using cosine similarities after transforming the texts with Word2Vec pre-trained model
7. Return $temp_{d_j}/|D_{a_i}|$

Figure 4-21 Pseudo code for obtain CoI_2

Calculate_CoI3(a_i) # parameter author identifier
Input:
 Collection of articles
 Collection of citations
Output:
 A value for contextual similarity feature

1. Set D_{a_i} to be a collection of articles authored by a_i
2. $temp_{d_j} = 0.0$
3. For each article d_j in D_{a_i}
4. $temp.cite_{d_j} = 0.0$
5. For each citation of d_j , $temp.cite_{d_j} +=$ get semantic similarities between title-abstract texts of d_j with the citation article
6. $temp_{d_j} +=$ average $temp.cite_{d_j}$ with total citation number of d_j
7. Return $temp_{d_j}/|D_{a_i}|$

Figure 4-22 Pseudo code for obtain CoI_3

For our empirical experiments [38], we prepared two datasets by selecting AMiner dataset for Siamese model based on LSTM, called as Siamese dataset, and

for observing the performances of CoI features, called as CoI dataset. AMiner dataset consists of $\pm 2M$ articles (2,092,356 articles) published between 1980 and 2014 and $\pm 1.6M$ researchers in Computer Science topics (<https://aminer.org/data>). We used K-Means for clustering the original AMiner data and selected $K=10$ after some observations based on Silhouette Index. We selected articles in each cluster with distances < 0.1 (closer distances means articles with similar subjects) or > 0.7 to the cluster centroid (farther distances make the articles have more varied subjects). Therefore, our Siamese dataset consisted of $\pm 2K$ articles.

Since there is no public dataset for conflict of interest case, we performed an initial validation on AMiner dataset during 12 years, 2001–2012 because more articles were published on that period. Then, we selected 80 researchers who wrote ± 150 -200 scientific articles during that period as CoI dataset with $\pm 15K$ articles and $\pm 430K$ citations related to the researchers. Three CoI-based features were extracted based on three window-times, 2001-2004, 2005-2008 and 2009-2012, to generate an input matrix of 80 researchers x 9 feature dimensions.

Table 4-8 Expert classification accuracies using CoI features with various similarity methods, classifiers and interest threshold values

Contextual Similarities	Classifier	Interest Threshold <i>intr. thres</i>				
		0.30	0.35	0.40	0.45	0.50
with Siamese architecture + Cosine	KNN	0.525	0.533	0.525	0.542	0.642
	Decision Tree	0.525	0.442	0.467	0.492	0.558
	Random Forest	0.533	0.550	0.450	0.592	0.667
Average Accuracy		0.528	0.508	0.481	0.542	0.622

with Cosine	KNN	0.558	0.558	0.517	0.525	0.575
	Decision Tree	0.592	0.592	0.542	0.575	0.617
	Random Forest	0.567	0.567	0.600	0.617	0.667
Average Accuracy		0.572	0.572	0.553	0.572	0.619

with Jaccard Coefficient	KNN	0.550	0.617	0.542	0.475	0.592
	Decision Tree	0.492	0.575	0.533	0.500	0.550
	Random Forest	0.625	0.542	0.642	0.625	0.650
Average Accuracy		0.556	0.578	0.572	0.533	0.597

The researchers were manually labeled with classes of 48 positive (1, no CoI indication) and 32 negative (0, any CoI indication) when the feature values were larger than standard deviation for each feature. Then, we experimented the input matrix of 80 researchers x 9 CoI features with classifiers of K-Nearest Neighbor (KNN) as voting-based classifier, Decision Tree as rule-based classifier, and Random Forest as the combinations in Table 4-8. CoI-based features rely heavily on additional information such as detail citations and the interest of researchers. The experiments were applied empirically with heavy assumptions on creating the validation dataset. The interest threshold displayed topic similarities of researchers, so 0.5 means at least half of the researchers in citing paper had similar interest with the ones in cited paper. Interest threshold was needed as cut-off value whether to include or ignore a research interest in computing similarities in CoI2. Higher values threshold was inclined to have better classification accuracies. As expected, CoI3 with deep learning approach showed better performance compared to related text similarity methods of Cosine and Jaccard Coefficient. Moreover, the Random Forest classifier with combined approach of voting and rule-based had better performance as well.

The results confirms that the proposed feature extraction methods could help to recognize the possibilities of misconduct behaviors. However, CoI1 feature has indecisive reason because most researchers in the selected dataset utilize around 30-35% of their previous works to show research track records. Then, CoI2 also requires phrase list of research interests, although the feature performed better than CoI1. The findings suggest that the implementation of CoI features needs more preliminary data, which might not be available. Despite of the limitations, the clustering approach could accommodate the need of topic identification.

4.4. Summary

In this chapter we have discussed on topic identification using clustering. This rather straightforward approach requires some validations, which is applied on the following cases:

- recommending cross-domain collaboration with ITS dataset and visualizing them on a standardized map from Scopus subject area, then
- identifying conflict of interest possibilities with AMiner dataset.

Through those cases we have shown topic identification with clustering is sufficient to be implemented in situations without existing subjects. However, aside of clustering to obtain topics, additional modifications were contextualized with the case situations.

We applied word embedding to convert words into vectors and utilizing the vectors in clustering process. The values represent word correlations between topics since word embedding identifies the context between words. Our approaches made word vectors become the representation for articles, topics, or researchers. Better clusters are obtained when word weight values are the embedding results. Although embedding with title texts produced more coherence clusters than title-abstract texts, because of the widespread of scientific fields within “computer science” domain in AMiner dataset, clustering still resulted into a low Silhouette score. However, the results presented a reasonable score when the vector space of words was transformed into two principle components. Nevertheless, finding relations between different entities, i.e. mapping topic interest, could be performed through cosine similarities.

The potency for recommending collaborations between departments in universities has been explored, although the findings still required more refining processes for recommendations. One recommendation example is a collaboration on the hot subject of smart home revolution, which suggested cooperation between departments in ITS, such as

- Industrial Product and Interior for designing,
- Electrical and Mechanical for connecting devices and appliances, in addition to
- Multimedia & Network along with Information System to provide application controllers for supporting handheld devices in a smart home.

For conflict of interest, more systematic procedures for creating validation dataset are necessary, i.e. manual checking on the researcher web profile and discussion with some researchers who become domain experts to evaluate the odds of misconducts. The possibility of misconduct behavior related to publishing articles in the form of biased citation could be categorized as conflict of interest. However, there is no reason to distrust researchers for performing biased citation when the

topics of citing articles are related to the topics of cited articles, since researchers with high citations do not necessarily bad scientist. Research is a continuing process. Thus, it is typical doings for making self-citations to previous works when publishes an article, as long as the number of self-citations is proportional to all citations within the article.

After applying topics of clustering results to articles and researchers, in the following chapters, we discuss our main contributions for extracting researcher features to generate unbiased scholar profile without much focusing on citations: productivity-dynamicity and behavior.

Chapter 5.

EXTRACTING PRODUCTIVITY-DYNAMICITY FEATURES

Productivity of researchers could be indicated from numbers of published articles and received citations on certain observed years [62], and could be quantified differently according to the time information [24] which often stated as dynamicity. In short, features related to productivity-dynamicity should represent quantifiable values of expertise evidence influenced by time periods. In this chapter we performed how to extract those features motivated by productivity-dynamicity of researchers who become rising stars [24]. We adapted the approaches in the studies of rising stars to accommodate topic information, since the issues in this dissertation are about expertise of researchers on specified topics. Then, we also performed some selection procedures to reduce the number of extracted features, such as the standard approach to remove highly correlated features. Since the quality of features could be assessed by applying them to solve a problem, we appointed the features in a topic prediction for researchers.

5.1. Data preparation

We have used AMiner dataset called as $D_{small-title}$ which contains four collection data: list of 70 NLP-IE researchers who at least have published 20 articles, list of articles from those researchers, and list of citations from those articles. The original expert list of NLP-IE contains 54 NLP researchers and 91 IE researchers. After manually validating the researchers with AMiner data, the list has reduced to 70 researchers (37 NLP researchers and 33 IE researchers).

Then, we selected other researchers from $\pm 1,600,000$ original AMiner authors who at least have seven publications with the initial 70 NLP-IE researchers in $D_{small-title}$. Therefore, there were 212 researchers including the initial ones for our empirical experiments in this chapter which called as $D_{behavior}$ dataset.

Table 5-1 Collections in $D_{behavior}$ dataset

No	Description	Notation
1.	Collection of researchers	$A = \{a_i\} = \{a_1 \dots a_{212}\}$, $ A = 212$, for validation purpose we manually collected h-index of each researcher from Scopus
2.	Collection of articles	$D = \{d_1 \dots d_j\}$, for each article d_j there are metadata of published year, its researchers, and its citation number. Each article d_j has been processed, so it only contains important words (more than three characters, not stop words, has >10 document occurrences)
3.	Collection of topics	$C = \{c_1 \dots c_{30}\}$, $ C = 30$, each c_k contains a number of semantically related words
4.	Collection of citations	$S = \{(d_j, n_j, y)\}$ means that article d_j has been cited n_j times on year y
5.	Collection of co-authors	$CA = \{(d_j, \{a_i\})\}$ means that article d_j has several researchers as authors. For our empirical experiments, we only listed a_i who exists in set A .
6.	Collection of article topics	$L = \{(d_j, \{c_k\})\}$, means that article d_j has 2-3 topics by function MapArticleTopic() in Figure 4-3

The current dataset contained a list of 212 NLP-IE researchers, a list of \pm 4800 articles from those researchers, a list of co-authors, and a list of citations from those articles. We assumed the additional 142 researchers have an interest on NLP-IE domain. Since AMiner dataset does not provide research topics, there is no mapping between researchers and articles to the topics. We clustered only title texts to obtain topics such that clustered words are often used in the particular topic. We used 30 topics from sub section 0 in $D_{behavior}$ as listed in Table 5-1.

5.2. Extracting productivity features

Performance of researchers in terms of productivity and collaboration are often influenced by a period of time [24]. The productivity was about publishing articles and getting citations over some observed years without consideration on topics. We modified the functions for extracting productivity as listed in Table 5-2. Features F_1, F_2, F_3 are about publishing articles, and features F_4, F_5, F_6 are about getting citations from the published articles. Four features describe continuing productivity of the researchers in the matter of cumulative values (F_2, F_3, F_5, F_6).

Table 5-2 Productivity features for each researcher in particular topic

No	Description
$F_1(a_x, c_i, t_n)$ certain period	Number of articles published by a_x which are labeled as topic c_i in year t_n .
$F_2(a_x, c_i, t_m, t_n)$ cumulative	Cumulative of number of articles published by a_x which are labeled as topic c_i during years of t_m and t_n . $F_2(a_x, c_i, t_m, t_n) = \sum_{t_y \in t_m \dots t_n} F_1(a_x, c_i, t_y)$ (1)
$F_3(a_x, c_i, t_m, t_n)$ cumulative, time penalty	Cumulative of number of articles after being weighted by time periods which are published by a_x and labeled as topic c_i during years of t_m and t_n . $F_3(a_x, c_i, t_m, t_n) = \sum_{t_y \in t_m \dots t_n} \frac{F_1(a_x, c_i, t_y)}{t_y - t_m + 1}$ (2)
$F_4(a_x, c_i, t_n)$ certain period	Total citation number of articles published by a_x which are labeled as topic c_i in year t_n . Function $ncite(d_k, t_n)$ returns total citation number of an article d_k which is published by a_x and labeled as topic c_i . $F_4(a_x, c_i, t_n) = \sum_{d_a \in c_i} ncite(d_k, t_n)$ (3)
$F_5(a_x, c_i, t_m, t_n)$ cumulative	Cumulative of total citation number of articles published by a_x which are labeled as topic c_i during years of t_m and t_n . $F_5(a_x, c_i, t_m, t_n) = \sum_{t_y \in t_m \dots t_n} F_4(a_x, c_i, t_y)$ (4)
$F_6(a_x, c_i, t_m, t_n)$ cumulative, time penalty	Cumulative of total citation number of articles after being weighted by time periods which are published by a_x and labeled as topic c_i during years of t_m and t_n . $F_6(a_x, c_i, t_m, t_n) = \sum_{t_y \in t_m \dots t_n} \frac{F_4(a_x, c_i, t_y)}{t_y - t_m + 1}$ (5)

Then, two features of those also consider the penalty impact caused by time periods (F_3, F_6). Researchers are seldom to have published articles in all topics or continuously publishing in the observed years for particular topic. Thus, the productivity feature matrix of researchers is often sparse. Next is processing productivity features for each topic to evaluate dynamic performance of the researchers during the observed years to describe tenacity behavior in their interest.

5.3. Extracting dynamicity features

Researcher performance in a set of time periods is about changes in minimum, maximum, last, total and overall representation from productivity features called as

dynamic based features [24]. By considering topics, those changes are shown from Table 5-2 into Table 5-3.

Table 5-3 Dynamicity features for each researcher in particular topic

No	Dynamic function for productivity features with certain period (F_1, F_4)	Dynamic function for productivity features with cumulative (and time penalty) (F_2, F_3, F_5, F_6)
chg (6)	$F_{1.chg}(a_x, c_i, t_{y-1}, t_y)$ $= F_1(a_x, c_i, t_y)$ $- F_1(a_x, c_i, t_{y-1})$	$F_{2.chg}(a_x, c_i, t_a, t_{y-1}, t_y)$ $= F_2(a_x, c_i, t_a, t_y)$ $- F_2(a_x, c_i, t_a, t_{y-1})$
min (7)	$F_{1.min}(a_x, c_i, t_a, t_z)$ $= \min_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{1.chg}(a_x, c_i, t_{y-1}, t_y)$	$F_{2.min}(a_x, c_i, t_a, t_z)$ $= \min_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{2.chg}(a_x, c_i, t_a, t_{y-1}, t_y)$
max (8)	$F_{1.max}(a_x, c_i, t_a, t_z)$ $= \max_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{1.chg}(a_x, c_i, t_{y-1}, t_y)$	$F_{2.max}(a_x, c_i, t_a, t_z)$ $= \max_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{2.chg}(a_x, c_i, t_a, t_{y-1}, t_y)$
end (9)	$F_{1.end}(a_x, c_i, t_{z-1}, t_z)$ $= F_1(a_x, c_i, t_z)$ $- F_1(a_x, c_i, t_{z-1})$	$F_{2.end}(a_x, c_i, t_{z-1}, t_z)$ $= F_2(a_x, c_i, t_a, t_z)$ $- F_2(a_x, c_i, t_a, t_{z-1})$
sum (10)	$F_{1.sum}(a_x, c_i, t_a, t_z)$ $= \sum_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{1.chg}(a_x, c_i, t_{y-1}, t_y)$	$F_{2.sum}(a_x, c_i, t_a, t_z)$ $= \sum_{\substack{t_{y-1} < t_y; \\ t_{y-1}, t_y \in t_a \dots t_z}} F_{2.chg}(a_x, c_i, t_a, t_{y-1}, t_y)$
rep (11)	$F_{1.rep}(a_x, c_i, t_a, t_z)$ $= \frac{1}{t_z - t_a + 1} \sum_{t_y \in t_a \dots t_z} F_1(a_x, c_i, t_y)$	a productivity feature with cumulative (F_2, F_5) $F_{2.rep}(a_x, c_i, t_a, t_z) = \frac{F_2(a_x, c_i, t_{z-1}, t_z)}{t_a - t_z + 1}$ a productivity feature with cumulative and time penalty (F_3, F_6) $F_{3.rep}(a_x, c_i, t_a, t_z) = F_3(a_x, c_i, t_{z-1}, t_z)$

For example, values of feature F_1 of a_x in one particular topic c_i during some observed years ($t_a \dots t_z$) are evaluated into five values of dynamicity features. The features are $F_{1.min}$ for describing minimum change (7), $F_{1.max}$ for maximum change (8), $F_{1.end}$ for last change (9), $F_{1.sum}$ for total change (10), and $F_{1.rep}$ for overall change (11). The functions (7)-(11) for extracting dynamicity features on particular topics need a change function in (6). Table 5-3 only lists the features for F_1 and F_2 , but others features should follow the same rules. Matrix changes for those extraction procedures are illustrated in Figure 5-1 with matrices for productivity and then dynamicity features are displayed for a_x .

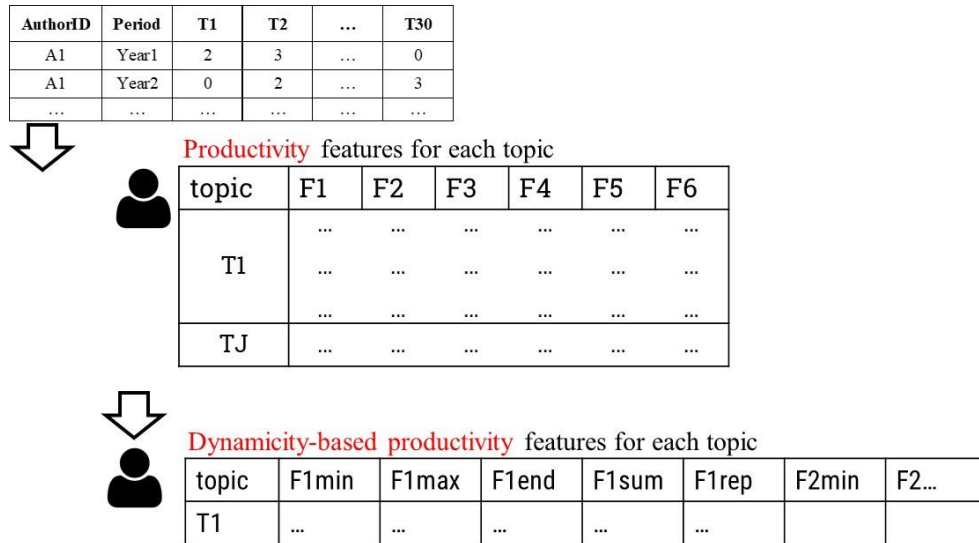


Figure 5-1 Illustration of matrices from raw data to productivity and dynamicity features

We have applied those features to an empirical experiment for predicting topics with an approach of discrete choice model [35] using R-package Multinomial Logit Models. Since there was no ground truth, we defined the status true and false of research interest by thresholding the summation value from a linear combination of those features. Log-likelihood values to compare model fitness showed that the combination of productivity-dynamicity features for article-based and citation-based gave better results. However, there were some inconsistencies to predict research interest in the testing data based on the fitted model from the training data. Therefore, in the next section we used different approaches to set ground truth data, select some features and define model for topic prediction.

5.4. Selecting Productivity and Dynamicity Features

5.5.1. Feature selection with correlation test

For selecting features from 30 productivity-dynamicity based features, we used AMiner dataset of 70 researchers from NLP-IE domain during 10 years of observation (2000-2009). We obtained correlation values as shown in Table 5-4. For example, $corr(F1_{min}, F3_{min}) = 0.96$ is quite high such that F_1 related to number of articles has the same meaning with F_3 related to cumulative number of articles after

being penalized. After iterations to eliminate pairs of strongly correlated features as shown in Table 5-5, there were six selected features as mentioned below.

Table 5-4 Correlation values between productivity-dynamicity based features

	f1_min	f1_max	f1_end	f1_sum	f1_rep	f2_max	f2_end	f2_sum	f2_rep	f3_min	f3_max	f3_end	f3_sum	f3_rep	f4_min	f4_max	f4_end	f4_sum	f4_rep	f5_min	f5_max	f5_end	f5_sum	f5_rep	f6_min	f6_max	f6_end	f6_sum	f6_rep
f1_min	1.000	-0.802	0.107	0.041	-0.668	-0.816	-0.216	-0.653	-0.480	0.958	-0.817	0.117	-0.334	-0.498	0.323	-0.234	-0.023	-0.142	-0.348	-0.224	-0.281	-0.211	-0.352	-0.311	0.298	-0.248	-0.071	-0.299	-0.307
f1_max	-0.802	1.000	0.061	0.304	0.697	0.933	0.442	0.716	0.525	-0.782	0.984	0.105	0.567	0.632	-0.248	0.245	0.091	0.217	0.346	0.252	0.292	0.263	0.355	0.330	-0.228	0.253	0.137	0.332	0.337
f1_end	0.107	0.061	1.000	0.539	0.133	0.084	0.638	0.131	0.127	0.092	0.054	0.929	0.272	0.275	-0.033	0.070	0.210	0.152	0.115	0.092	0.107	0.180	0.112	0.114	-0.045	0.082	0.197	0.135	0.146
f1_sum	0.041	0.304	0.539	1.000	0.290	0.360	0.766	0.378	0.214	0.056	0.333	0.685	0.822	0.539	0.049	0.090	0.153	0.284	0.129	0.089	0.136	0.221	0.148	0.135	0.062	0.100	0.188	0.252	0.203
f1_rep	-0.668	0.697	0.133	0.290	1.000	0.845	0.548	0.991	0.833	-0.722	0.707	0.188	0.656	0.922	-0.489	0.323	0.124	0.320	0.664	0.557	0.483	0.461	0.670	0.639	-0.416	0.364	0.214	0.564	0.621
f2_max	-0.816	0.933	0.084	0.360	0.845	1.000	0.547	0.864	0.647	-0.868	0.963	0.137	0.682	0.776	-0.329	0.291	0.105	0.289	0.492	0.369	0.389	0.361	0.509	0.468	-0.301	0.316	0.176	0.461	0.474
f2_end	-0.216	0.442	0.638	0.766	0.648	0.547	1.000	0.654	0.549	-0.235	0.449	0.806	0.761	0.847	-0.216	0.235	0.217	0.338	0.428	0.356	0.348	0.412	0.432	0.423	-0.204	0.261	0.283	0.433	0.462
f2_sum	-0.653	0.716	0.131	0.378	0.991	0.864	0.654	1.000	0.816	-0.702	0.732	0.188	0.726	0.924	-0.408	0.314	0.122	0.334	0.644	0.536	0.470	0.453	0.653	0.619	-0.384	0.354	0.211	0.565	0.608
f2_rep	-0.480	0.525	0.127	0.214	0.833	0.647	0.549	0.816	1.000	-0.574	0.507	0.161	0.396	0.843	-0.316	0.229	0.099	0.262	0.606	0.587	0.398	0.405	0.608	0.654	-0.307	0.261	0.175	0.461	0.571
f3_min	0.958	-0.782	0.092	0.056	-0.722	-0.808	-0.235	-0.702	-0.574	1.000	-0.789	0.114	-0.307	-0.549	0.330	-0.237	-0.021	-0.145	-0.379	-0.264	-0.293	-0.223	-0.383	-0.349	0.308	-0.250	-0.070	-0.313	-0.332
f3_max	-0.817	0.984	0.054	0.333	0.707	0.963	0.449	0.732	0.507	-0.789	1.000	0.099	0.614	0.640	-0.249	0.250	0.087	0.231	0.351	0.239	0.299	0.268	0.362	0.328	-0.224	0.261	0.136	0.350	0.343
f3_end	0.117	0.105	0.929	0.685	0.188	0.137	0.906	0.188	0.161	0.114	0.099	1.000	0.417	0.406	-0.023	0.091	0.219	0.200	0.129	0.102	0.131	0.220	0.127	0.130	-0.030	0.100	0.228	0.173	0.180
f3_sum	-0.334	0.567	0.272	0.822	0.656	0.682	0.761	0.725	0.396	-0.307	0.614	0.417	1.000	0.761	-0.191	0.230	0.139	0.346	0.367	0.248	0.319	0.347	0.385	0.333	-0.161	0.259	0.209	0.444	0.400
f3_rep	-0.498	0.632	0.275	0.539	0.922	0.776	0.847	0.924	0.843	-0.549	0.640	0.406	0.761	1.000	-0.354	0.299	0.159	0.358	0.623	0.540	0.460	0.480	0.629	0.622	-0.334	0.336	0.252	0.554	0.615
f4_min	0.323	-0.248	-0.033	0.049	-0.439	-0.329	-0.216	-0.408	-0.316	0.330	-0.249	-0.023	-0.191	-0.354	1.000	-0.638	-0.120	-0.231	-0.775	-0.528	-0.727	-0.467	-0.766	-0.674	0.989	-0.686	-0.205	-0.589	-0.649
f4_max	-0.234	0.245	0.070	0.090	0.323	0.291	0.235	0.314	0.229	-0.237	0.250	0.091	0.230	0.299	-0.638	1.000	0.748	0.846	0.647	0.326	0.949	0.865	0.669	0.563	-0.613	0.993	0.804	0.881	0.777
f4_end	-0.023	0.091	0.210	0.153	0.124	0.105	0.217	0.122	0.099	-0.021	0.087	0.219	0.139	0.159	-0.120	0.748	1.000	0.881	0.347	0.155	0.689	0.847	0.358	0.306	-0.135	0.734	0.985	0.687	0.586
f4_sum	-0.142	0.217	0.152	0.284	0.320	0.289	0.338	0.334	0.262	-0.145	0.231	0.200	0.346	0.358	-0.231	0.846	0.881	1.000	0.514	0.275	0.819	0.914	0.549	0.479	-0.213	0.835	0.926	0.881	0.743
f4_rep	-0.348	0.346	0.115	0.129	0.664	0.492	0.428	0.644	0.606	-0.379	0.351	0.129	0.367	0.623	-0.775	0.647	0.347	0.514	1.000	0.867	0.843	0.762	0.997	0.972	-0.767	0.706	0.453	0.839	0.949
f5_min	-0.224	0.252	0.092	0.089	0.557	0.369	0.356	0.536	0.587	-0.264	0.239	0.102	0.248	0.540	-0.528	0.326	0.155	0.275	0.867	1.000	0.568	0.545	0.855	0.912	-0.537	0.387	0.244	0.585	0.775
f5_max	-0.281	0.292	0.107	0.136	0.483	0.389	0.348	0.470	0.398	-0.293	0.299	0.131	0.319	0.460	-0.727	0.949	0.689	0.819	0.843	0.568	1.000	0.927	0.857	0.774	-0.710	0.972	0.771	0.957	0.926
f5_end	-0.211	0.263	0.180	0.221	0.461	0.361	0.412	0.453	0.405	-0.223	0.268	0.220	0.347	0.480	0.467	0.865	0.847	0.914	0.782	0.545	0.927	1.000	0.772	0.723	-0.459	0.883	0.915	0.945	0.930
f5_sum	-0.352	0.355	0.112	0.148	0.670	0.503	0.432	0.653	0.608	-0.383	0.362	0.127	0.385	0.629	-0.766	0.669	0.358	0.549	0.997	0.855	0.857	0.772	1.000	0.970	-0.755	0.724	0.465	0.863	0.954
f5_rep	-0.311	0.330	0.114	0.135	0.639	0.468	0.423	0.619	0.654	-0.349	0.328	0.130	0.333	0.622	-0.674	0.563	0.306	0.479	0.972	0.912	0.774	0.723	0.970	1.000	-0.671	0.615	0.411	0.762	0.927
f6_min	0.298	-0.228	-0.045	0.062	-0.416	-0.301	-0.204	-0.384	-0.307	0.308	-0.224	-0.030	-0.161	-0.334	0.989	-0.613	-0.135	-0.213	-0.767	-0.537	-0.710	-0.459	-0.755	-0.671	1.000	-0.665	-0.210	-0.782	0.927
f6_max	-0.248	0.253	0.082	0.100	0.364	0.316	0.261	0.354	0.261	-0.250	0.261	0.100	0.259	0.336	-0.686	0.993	0.734	0.835	0.706	0.387	0.972	0.883	0.724	0.615	-0.665	1.000	0.798	0.905	0.817
f6_end	-0.071	0.137	0.197	0.188	0.214	0.176	0.283	0.211	0.175	-0.070	0.136	0.228	0.209	0.252	-0.205	0.804	0.985	0.926	0.453	0.244	0.771	0.915	0.465	0.411	-0.210	0.798	1.000	0.778	0.688
f6_sum	-0.299	0.332	0.135	0.252	0.564	0.461	0.433	0.565	0.461	-0.313	0.350	0.173	0.444	0.554	-0.589	0.881	0.687	0.881	0.839	0.585	0.957	0.945	0.863	0.782	-0.562	0.905	0.778	1.000	0.946
f6_rep	-0.307	0.337	0.146	0.203	0.621	0.474	0.462	0.608	0.571	-0.332	0.343	0.180	0.400	0.615	-0.649	0.777	0.586	0.743	0.949	0.775	0.926	0.920	0.954	0.927	-0.636	0.817	0.688	0.946	1.000

Table 5-5 Combinations of correlation values and feature sets

# features combinations	0.70	0.65	0.60	0.55	0.50	
4	23,751	4,516	2,952	1,895	1,328	706
5	118,755	6,701	3,177	1,585	823	164
6	475,020	5,100	1,592	670	208	16
7	1,560,780	1,654	308	120	16	-
8	4,292,145	92	-	-	-	-

$$\text{corr}(f_p, f_q) < 0.5$$

Possible set of features

1. $F1_{min}$ minimum difference of article number between two years

If the minimum value is still in a rather large value, such as 3-4 articles, it means that the researcher is a productive one. In average, at least researchers in our experiment data annually had a difference of 1-3 published articles as shown in Table 5-6.

2. $F1_{end}$ difference of article number from the last two years

If the researcher in the last observation years still publishes a rather large number of articles, such as 3-4 articles, it means that the researcher is a productive one.

3. $F2_{rep}$ average-like estimation for the cumulative of article number

If this feature has a rather large number, it means that the researcher is consistent in publishing articles as a sign of a productive one.

4. $F3_{sum}$ sum of differences for cumulative article number with penalty weights
This feature shows the same consistency indicator for productive researchers but with stricter conditions because of the penalty weights.
5. $F4_{min}$ minimum difference of cited article number between two years
This feature shows the same meaning of $F1_{min}$ but for citations.
6. $F5_{end}$ cumulative difference of cited article number from the last two years
This feature shows the same meaning with consistency related features for citations.

After selecting those six features with weaker correlations of less than 0.5, we applied them on the aforementioned topic prediction [36].

5.5.2. Create validation dataset for expertise on topics

After knowing performance indicators for researchers, the problem is to generate ground truth dataset whether a topic is really his or her interest. We used clustering based on fuzzy membership of Fuzzy C-Means (FCM) [63] to determine the formed groups and then randomly analyze some data in each group to set the true-false label.

Table 5-6 Scaling criteria for productivity-dynamicity features

Scale	$F1_{min}$	$F1_{end}$	$F2_{rep}$	$F3_{sum}$	$F4_{min}$	$F5_{end}$
1	0	-4	0.1	<-1.0	0	0
2	1	-3	0.2	-1.0	3	3
3	2	-2	0.3	-0.5	5	5
4	3	-1	0.4	0.0	10	10
5	4	0	0.5	0.5	20	20
6	5	1	0.6	1.0	30	30
7	6	2	0.7	1.5	40	40
8	7	3	0.8	2.0	50	50
9	8	4	0.9	2.5	70	70
10	9	5	≥ 1.0	> 3	≥ 100	≥ 100
Avg.	1.92	5.06	3.49	4.20	2.64	2.47
Std.	0.81	0.75	2.96	1.87	2.12	2.18

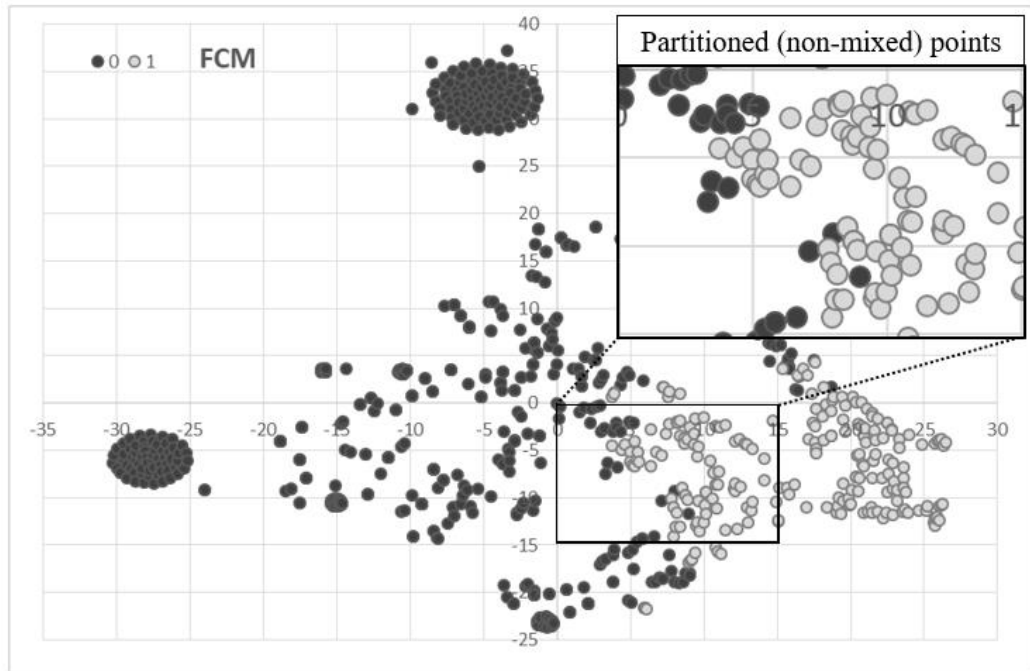


Figure 5-2 t-SNE visualization of scaled data with labels from FCM approach

Since our dataset contains AMiner published articles, we set the positive label or status = 1 which means that the researcher is a specialist on the topic, and the negative label or status = 0 to represent a thriving researcher in the topic. Status = 0 indicates that the researcher has just few articles related to a topic and still on learning phase to be a specialist for certain topic. However, because value ranges for each feature were quite different, we have scaled the feature values according criteria in Table 5-6. Then for visually validation purpose, we transformed researchers with scaled feature values with t-SNE approach. The visualization in Figure 5-2 with FCM results illustrated almost separated data of two groups and validate our approach for generating ground truth dataset. As a comparison we also made t-SNE visualization for unscaled data with FCM labels. Since the visualization gave more mixed results, the scaling process was necessary.

5.5. Summary

Further assessment was conducted to observe the performance of those six features by using them with scaled values and labels from FCM to predict topics of selected AMiner NLP-IE researchers [36]. Table 5-7 displayed the classification results using Python-based Orange toolkit with some standard classifiers of Logistic

Regression, Random Forest and Support Vector Machine (SVM). The results demonstrated four features of $F2_{rep}$ $F3_{sum}$ $F4_{min}$ $F5_{end}$ are more superiors among others.

Table 5-7 Classification accuracies with combinations of productivity-dynamicity features

FCM scaled	$F3_{sum} + F4_{min} + F5_{end}$	$F2_{rep} + F3_{sum} + F4_{min} + F5_{end}$	All six features
Logistic Regression	87.7%	98.4%	97.5%
Random Forest	92.0%	99.5%	99.4%
SVM	89.1%	99.6%	99.1%

Table 5-8 Fuzzy rules generated with FCM labels on scaled data of selected productivity-dynamicity features

Rule	$F2_{rep}$	$F3_{sum}$	$F4_{min}$	$F5_{end}$	Expertise Status
1	<i>large</i>	large	large	large	
2	<i>large</i>	large	medium	large	
3	<i>large</i>	large	medium	medium	1-specialist
4	<i>large</i>	medium	medium	medium	
5	<i>large</i>	medium	small	small	
6	<i>medium</i>	<i>medium</i>	<i>medium</i>	medium	1-specialist
7	medium	medium	<i>small</i>	<i>small</i>	0-thriving
8	<i>small</i>	medium	<i>small</i>	large	
9	<i>small</i>	medium	<i>small</i>	medium	0-thriving
10	<i>small</i>	medium	<i>small</i>	small	

We performed other assessments to observe the feature performance. The assessments applied four features of $F2_{rep}$ $F3_{sum}$ $F4_{min}$ $F5_{end}$ in Table 5-7 because they gave better accuracies in classification experiments. Then, using the same feature values and FCM labels with those four features, we performed the classification to generate fuzzy rules in Table 5-8 using R package of frbs. We also performed experiments on the same data with unscaled values for generating fuzzy rules. However, the results showed fuzzy rules with ambiguities. Thus, we confirmed to apply the data after scaling based on criteria in Table 5-6. With manual observation on Table 5-8, there were three simplified fuzzy rules R_1 , R_2 , and R_3 .

$$R_1: F2_{rep} \text{ large} \xrightarrow{\text{yields}} \text{label}_1$$

$$R_2: F2_{rep} \text{ not. large} \wedge F4_{min} \text{ medium} \xrightarrow{\text{yields}} \text{label}_1$$

$$R_3: F2_{rep} \text{ not. large} \wedge F4_{min} \text{ small} \xrightarrow{\text{yields}} \text{label}_0$$

Based on the resulted Gaussian membership functions with small $f(x, \sigma_{0.175}, \mu_{0.0})$, medium $f(x, \sigma_{0.175}, \mu_{0.5})$, and large $f(x, \sigma_{0.175}, \mu_{1.0})$, then

- R_1 means that a researcher who at least annually publishes two articles on certain topic during 10 years can be stated as a specialist or has expertise on the topic
- R_2 means that a researcher is still a specialist even though has less than two articles on certain topic annually during 10 years, but receives 5-10 citations for his or her articles on the topic
- R_3 means if a researcher does not receive any citation for particular topic in one observed year, then he or she is not specialist

Thus from the original 30 productivity-dynamicity features, we have showed that two features are enough to represent a scholar profile, $F2_{rep}$ for number of published articles and $F4_{min}$ for number of citations.

Since $F2_{rep}$ indicates researcher consistency in publishing scientific articles as a sign of a productive one, we called this feature as a token for dynamicity regarding to publishing behavior. Then $F4_{min}$ which indicating that a researcher should have certain number of citations can be achieved through the published works of his or her students. Thus, $F4_{min}$ also becomes an indirect token for dynamicity regarding to publishing behavior. It should be noted that self-citation by peers does not necessarily means a bad scientist providing the article topics are related to the topics of cited articles [38]. Next section discusses researcher behavior related to publishing articles to find alternatives for performance indicator aside of citation based.

Chapter 6.

EXTRACTING BEHAVIOR FEATURES

Co-authoring with experienced partners can become one of the helping factors in career advancement [64], and it is not surprising to cause topic drift or changes of researchers' interest [25]. A researcher generally prefers to work with other researchers who have high academic level or more expertise and still fitting with his or her own topics. Previous chapters have discussed how to extract the evidence of expertise on specified topics. However, focusing only on the numbers of published articles and received citations could lead to biased scholar profile and debatable expertise. Our contributions to acquire unbiased profile are supported with not only the evidence of productivity-dynamicity of researchers, but also their behaviors, which is observed through relation to others. One typical approach for examining the relations is graph-based analysis with researchers as nodes.

This chapter discussed our approaches for directly analyzing relations of researchers through one-mode (co-author) networks, and their indirect influence through topics as the possibility of interest changes because of others. Thus, the proposed approaches for multi-layered bibliographic networks are the procedures to identify latent topics within article metadata and then make abstraction for one-mode and two-mode networks. We have discussed before about AMiner dataset with NLP-IE researchers and the 30 identified topics. Then, the focus of this chapter is about extracting those preferable aspects as behavior data of researchers related to topics.

Ensuring our assumption about behavior features, we investigated their efficacy in a case of network evolution with different periods. We hypothesized the changes of networks from one period to others is caused by researchers who become their co-authors. We abstracted the networks from article metadata consisting co-author information, texts of title-abstract, and other information such as published year. To test our hypothesis, we have designed a model for network evolution from the view point of each researcher or frequently called as ego network. The model is derived from well-known Stochastic Actor-Oriented Model (SAOM) [65] [66].

6.1. Graph Theories related to Researcher Representation

Some terms related to graph theories are often applied in our approaches for modelling the scholar profile.

a. Dyad, Reciprocity, and Degree

An article written by three researchers a_1, a_2, a_3 is represented as a complete graph g of three nodes where each node connects to the other two nodes. The graph g is undirected since an edge of $a_1 - a_2$ refers to the same edge of $a_2 - a_1$. Because of those three researchers collaborate in co-authoring the article, they have reciprocal relations, which validates undirected edges in the graph.

The edge is often defined as a dyad or a relation between two nodes due to the reciprocity property. There are three dyads of $a_1 - a_2, a_2 - a_3$, and $a_1 - a_3$ in the graph g , which makes each node has a degree value of two. With undirected edges, there is no difference between in-degree and out-degree. In this dissertation, each dyad only has one edge and there is no self-loop. Thus, the degree of a node in an undirected graph represents the number of connected edges or nodes.

b. Transitive Triad

Graph g of three researchers has three dyads and one transitive triad since all researchers are inter-connected as shown in Figure 6-1. In a way, transitive triad relation is a form of cycle. There are other types of triadic relations between three nodes, but we focus on the transitive triad type. Our scholar profile looks on features that appropriately represent the expertise. Those features could be determined by other researchers from past collaborations.

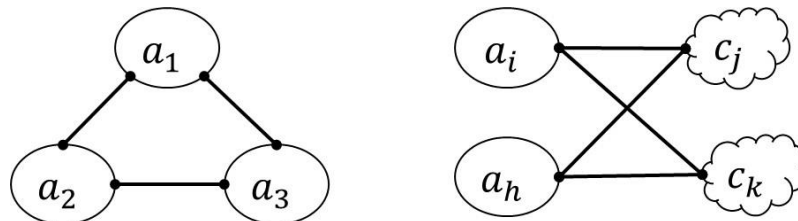


Figure 6-1 Transitive triad relation on a one-mode (co-author) network (left) and cycle relation on a two-mode (author-topic) network (right)

The tendency of a researcher to work with other researchers is higher if they have worked with the clique of the researcher [67]. Thus, the theory of transitive triad supports on the assumption of clique relation. The function of transitive triad applied in this dissertation is described in Appendix 6.

c. Ego network and ego-alter relations

Analyzing past collaborations of a researcher corresponds to understanding the changes of the researcher graphs, which illustrate co-authoring relations to other researchers, on different periods. In a researcher graph, the researcher as a focal point is called as an ego and all connected nodes are called as alters. In case of the previous graph g , each of those three authors makes an ego network. Those three ego networks came from the graph g , i.e. g_1 for researcher a_1 , g_2 for researcher a_2 , and g_3 for researcher a_3 , are identical because the initial graph g is a complete graph.

d. One-mode and Two-mode networks

Analyzing past collaborations of a researcher generally employs on co-author networks in which all nodes within are homogeneous. A co-author network is one-mode type due to the same node type. Our approaches to look on features for the scholar profile employed another type of networks called as two-mode networks that have two node types: author and topic. Relations between researchers based on co-authored scientific article can be abstracted as a two-mode or bipartite networks. Relations between actors in films or students in courses are other forms of bipartite networks. Edges within co-author networks and author-topic networks are undirected.

e. Cycles in two-mode networks

Similar to the node relation of transitive triad in a one-mode of co-author network, there is a cycle relation occurred in a two-mode of author-topic network. The illustration for a two-mode network x includes researcher nodes of a_i and a_h in addition to topic nodes of c_j and c_k as shown in Figure 6-1.

Relations of x_{ij} and x_{ik} show that the researcher a_i has interest on topics of c_j and c_k , while relations of x_{hj} and x_{hk} show that the researcher a_h also has the same interest on c_j and c_k .

Thus, there is a cycle between researchers a_i and a_h through topics c_j and c_k . The function of cycle within author-topic network applied in this dissertation is described in Appendix 6.

6.2. Extracting Exploration and Consistency features

Exploration and consistency features refers to researcher behavior in exploring new topics or exploiting existing topics or called as consistency. Procedures to extract those features are displayed in Figure 6-2 for exploration and Figure 6-3 for consistency. In this dissertation, those values are obtained from publishing experiences during 15 years for AMiner NLP-IE experts. Thus, we did longitudinal data analysis for observing researcher behavior in three waves of five years period, bip_{w1} bip_{w2} bip_{w3} , from bipartite (two-mode) networks of author-topic relations as seen in Figure 3-1. Longitudinal data analysis on bipartite networks was occurred in relations between countries and trade agreements [68], which were applied in this dissertation for relations between researchers based on topics from co-authoring process.

Representations of bipartite networks are matrices bip_{w1} bip_{w2} bip_{w3} with dimensions of researchers as rows and topics as columns. The extracted results of beh_{exp} and beh_{const} have dimensions of researchers as rows and three columns of represented waves. The first column of both matrices have the same values to represent the initialization step. After analyzing the data, we enumerated the levels of feature values as shown in Table 6-1.

Extracting exploration feature basically is obtaining a number of distinct topics from the observed wave which is compared to the initial wave or the first five-years period. The observed wave can be the second or the third wave. Researcher with higher level of exploring behavior, i.e. $beh_{exp}(a_i, 3) = 3$, means that the person likes to keep up to date with current trends of research topics, as illustrated in Figure 6-4.

Table 6-1 Levels for publishing related behavior based feature values

Values of exploring level ...	Values of consistency level ...
1: at most one new topic 2: at most two new topics 3: at least three new topics	1: at most focus on one topic 2: at most focus on two topics 3: at least focus on three topics
... in each year during 5-years period	

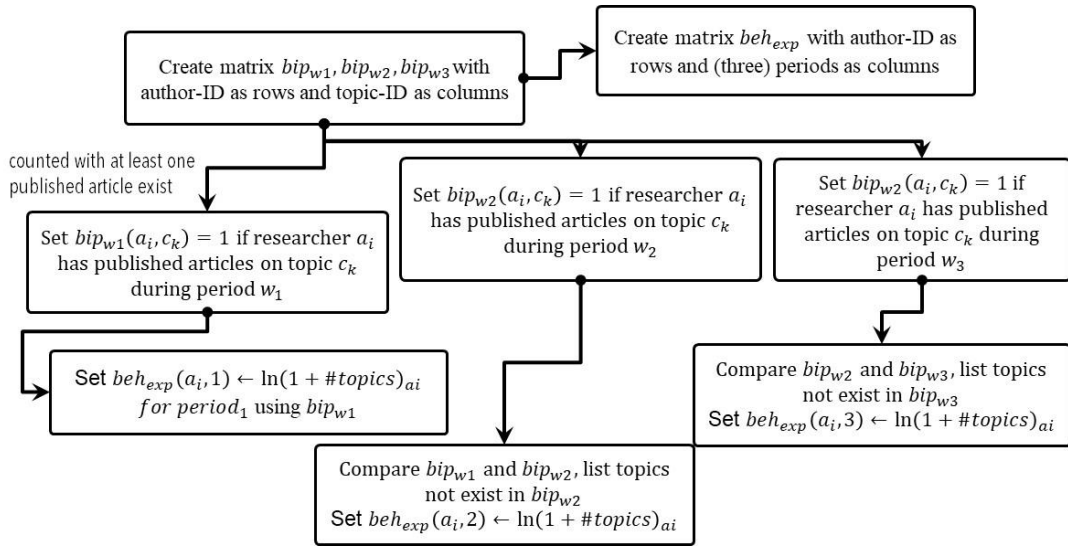


Figure 6-2 Process for extracting Exploration feature

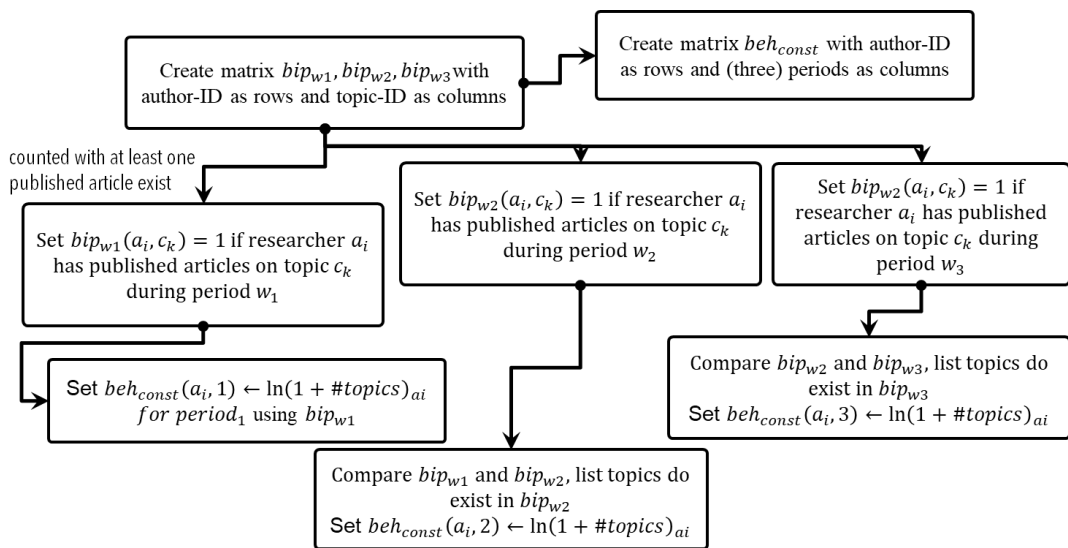


Figure 6-3 Process for extracting Consistency feature

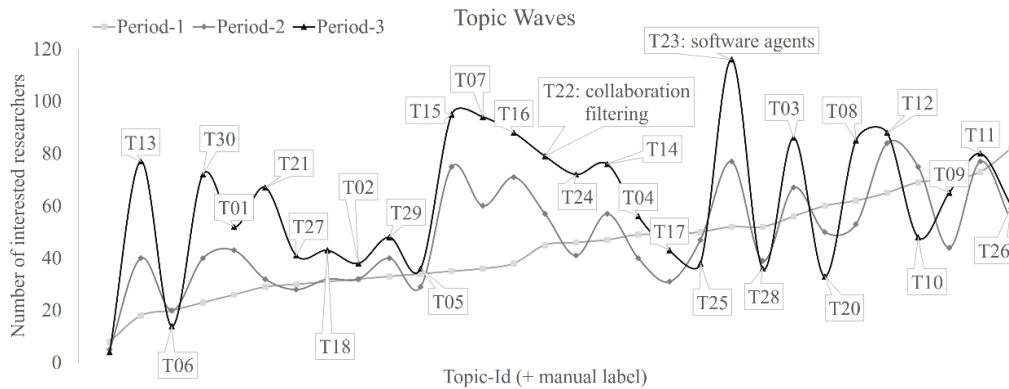


Figure 6-4 Topics with increasing popularities over time from AMiner NLP-IE dataset to illustrate research trends

Illustration in Figure 6-4 suggested the growths of some topics such as “collaboration filtering” (T22) and “software agents” (T23). We validated the topic growth by comparing search results of Google Scholar, in which Topic T22 in Period-1 (1995-2000) had around 10 thousand articles and then increased to ten times more in Period-4 (2010-2015). Another example came from topic T23 with around 400 thousand articles in Period-1 to almost two million articles in Period-4.

Exploration feature observes researcher behavior on widening research interest. In contrast, consistency feature observes researcher behavior on focusing research interest. The next process for extracting consistency feature basically is obtaining a number of same topics from the observed wave which is compared to the initial wave or the first five-years period. The observed wave can be the second or the third wave. Researcher with higher consistency behavior, i.e. $beh_{const}(a_i, 3) = 3$, indicates that the person focuses on more topics compared to keep up with trends. We assumed that the experts are researchers who have higher level values of behavior based features, with the possibilities of: (a) the ones who follow research trends and have higher exploring level, (b) the ones who focus in their works and have higher consistency level, or (c) the ones with conditions somewhat in between.

Checking behavior-based feature of exploring through a model for longitudinal network analysis called as Stochastic Actor-oriented Model (SAOM) becomes the objective in next empirical experiments. Although there are two types of behavior features, we initially investigated whether exploring feature that leads to

interest changes has an influence on the researcher expertise. Other assessments on the behavior feature of consistency are required in further works.

6.3. Experiments Exploration Feature with Stochastic Actor-oriented Model (SAOM)

These experiments aimed to observe behavior-based feature of exploring levels in influence cases by co-authors for the likelihood of interest changes. Since the focal point is researchers as co-authors, we selected a prevalent Stochastic Actor-Oriented Model (SAOM) [65] [66] for modeling the objectives through longitudinal data analysis and did the experiments using SAOM implementation of R package RSIENA (R Simulation Investigation for Empirical Network Analysis) (cran.r-project.org/package=RSiena). SAOM is a multinomial probability model for predicting changes of tie formation in network evolution that requires at least two networks. SAOM through RSIENA models the processes of network change on tie formation and attribute change on researchers' characteristics and behaviors (www.stats.ox.ac.uk/~snijders/siena, descriptions on RSIENA package manual). Thus, the experiments defined input for network change as co-author networks and author-topic networks, in addition to input for attribute change as starting publication year, exploring levels and publishing levels. Both types of networks have been abstracted in Figure 3-1. We defined publishing levels as substitution for consistency features by ignoring topics for feature extraction.

6.4.1. Preparation for RSIENA

Since SAOM is used to observe tie formation, we argued that co-authoring process or forming ties between researchers were depending on some reasons such as exploring levels of co-authors. The experiments also examined other reasons that may influence co-authoring process, such as career age and publishing level. Therefore, we worked on information of co-authors, (latent) topics, and published year within article metadata as listed in Table 6-2. Since longitudinal data analysis requires at least two networks from different periods, we set four periods.

- Period-1 contained any published metadata until 1995 with the earliest year of around 1980.

- Period-2 contained article metadata between 1996 and 2000.
- Accordingly, we set article metadata in Period-3 (2001-2005) and Period-4 (2006-2010).

We also mentioned a list of experts which contained 212 AMiner NLP.IE researchers, as well as a list of topics from clustering. We analyzed the network evolution on two types of networks: co-author and author-topic. Information about co-authors in each article are available from AMiner dataset, but the dataset does not provide topic information.

Table 6-2 RSIENA data input for experiments to observe exploring feature

No	RSIENA data input	Description
1	One mode network data (ND), X size 212 x 212	Co-authoring between researchers in three panel waves binary values, $x_{ij} = x_{ji} = 1$ means that author- i and author- j are co-authors
2	Two-mode (bipartite) network, W size 212 x 30	The relation between researchers and topics in three panel waves binary values, $w_{ih} = 1$ means that author- i has at least one article mapped to topic- h .
3	Individual covariate (IC) size 212 x 1 (career age)	Age for starting in a publication career, constant in all observations, encoded values: 1-5 with the average value is 3. (1: \geq 2010,2:2000-2009,3:1990-1999,4:1980-1989,5: $<$ 1980)
4	Behavior data (BD) publishing level size 212 x 3	Values of publishing level are extracted from articles without topic concern. The graded values of publishing level are: 1: publishes at least one article in a year, 2: publishes at least one article per semester (6 months), 3: publishes at least more than two articles per semester, 4: publishes at least one article in every other month
	exploring level size 212 x 3	Values of exploring level are extracted from articles with topic concern. The graded values of exploring level are: 1: at most one new topic in each year during 5-years period, 2: at most two new topics in each year during 5-years period, 3: at least three new topics in each year during 5-years period The new topic is counted with at least one published article exist.

Aside of network change, individual covariate and behavior data are used to observe whether the attribute change has some influences. Values of career age were assumed from the earliest publication year of articles for each researcher in AMiner dataset. Because of RSIENA guidelines, we enumerated the values of career age.

The fourth input for SAOM is behavior data. We selected exploring behavior for data input and the procedures to extract exploring features. Instead of consistency feature, we examined publishing frequentness of researchers. We aggregated feature $F1$ that shows number of articles published by a_x which are labeled as topic c_i in year t_n based on topic and year to extract publishing features. Using the aggregated value for each period, we computed publishing levels with formula $\ln(1 + \#pubs)$ and rounded to nearest integer values. In short, parameter $\#pubs$ is the aggregated value on topic and year for a_x .

Table 6-3 RSIENA Effects for observing changes in networks and attributes

Effects	Descriptions
<i>Relate to co-author networks (undirected relations between researchers)</i>	
Transitive triads (<i>transtriads</i>)	<ul style="list-style-type: none"> • Positive estimate indicates cyclical pattern among researchers. • Negative estimate indicates co-authorships have hierarchical relations. Thus researchers do not seek co-authors in cyclical pattern.
Knowing the popularity of alters based on degrees (<i>inPop</i>)	In-degree popularity or known as degree of alter is similar to out-degree activity since co-author networks are undirected. Positive estimate supports the Matthew effect of “the richer gets richer” which translated as popular researchers tend to collaborate more.
From topic agreement in the bipartite network (<i>from</i>)	<ul style="list-style-type: none"> • Positive estimate indicates researchers with similar topic interests are most likely having co-authorship relations. • Negative estimate indicates researchers, who have dissimilar interests but possibly related, tend to collaborate. However, it needs further investigation.
Based on the covariate values of career age (<i>simX</i> , <i>egoX</i> , <i>altX</i>)	<ul style="list-style-type: none"> • Positive estimate of <i>simX</i> indicates the researcher tendency to work with co-authors who have the same level of starting publication year. • Positive estimate of <i>egoX</i> indicates senior researchers who have higher values of career age tend to initiate more collaboration. Notes, <i>egoX</i> and <i>altX</i> have similar effects because of undirected co-author networks. In case of <i>altX</i>, it means that senior researchers tend to receive more collaboration.
Based on the behavior values for selection/influence (<i>egoX</i> , <i>altX</i>)	The effects of sender <i>egoX</i> and receiver <i>altX</i> examine selection and influence mechanisms for tie formation in co-author networks based on behavior data of different periods.

6.4.2. RSIENA Scripts

Based on RSIENA guidelines, there are some predefined effects to set the observed model for network change and attribute change that influences the network change. After analyzing our objectives to examine exploring feature among others reasons in tie formation for network change, we listed some effects in Table 6-3. Then, by using RSIENA guidelines with sample R scripts for assigning input from text files, setting the networks and the effects, we followed the rules and specified our model with some snippets of R scripts as showed in Figure 6-5, Figure 6-6, and Figure 6-7. All input in Table 6-2 were formatted as comma separated values (CSV) files. After entering the input files, some RSIENA functions such as “sienaNodeSet” and “sienaDependent” in Figure 6-6 or “includeEffects” in Figure 6-7 were used for assigning networks and effects to generate the observed model (Appendix 1).

```
1 library(RSiena)
2
3 # adjacency matrix (co-author network), size: 212 x 212 scholars
4 # author.w1(i,j) = number of co-authored articles
5 author.w1 <- data.matrix(read.csv("nd2_period1_vtop2.csv", header=FALSE, sep=";"))
6 #Period1:articles from 1969-1995
7 author.w2 <- data.matrix(read.csv("nd2_period2_vtop2.csv", header=FALSE, sep=";"))
8 #Period2:articles from 1996-2000
9 author.w3 <- data.matrix(read.csv("nd2_period3_vtop2.csv", header=FALSE, sep=";"))
10 #Period3:articles from 2001-2005
11
12 # matrix size for bipartite data: 212 authors x 30 topics
13 author.topic.w1 <- data.matrix(read.csv("bip_period1_top2a.csv", header=TRUE, sep=";"))
14 #topic_period1_nd2_top2
15 author.topic.w2 <- data.matrix(read.csv("bip_period2_top2a.csv", header=TRUE, sep=";"))
16 author.topic.w3 <- data.matrix(read.csv("bip_period3_top2a.csv", header=TRUE, sep=";"))
17
18 # covariate data
19 pubyear_mat <- data.matrix(read.csv("author_publish_year.csv", header=TRUE, sep=";"))
20 # behavior data
21 beh_pub_mat <- data.matrix(read.csv("beh_publishing_nd2.csv", header=TRUE, sep=";"))
22 #beh_publishing_nd2
23 beh_exp_mat <- data.matrix(read.csv("beh_exp_top2_efforts_nd1.csv", header=TRUE, sep=";"))
24 )) #beh_exp_top2_efforts_nd2
```

Figure 6-5 Snippet of RSIENA script for assigning input

```
nrauthors <- nrow(author.w1) # 212 authors
authors <- sienaNodeSet(nrauthors,nodeSetName="authors")
nrtopics <- ncol(author.topic.w1) # 30 topics
topics <- sienaNodeSet(nrtopics,nodeSetName="topics")
# oneMode
authorship <- sienaDependent(array(c(author.w1,author.w2,author.w3),
dim=c(nrauthors,nrauthors,3)),type="oneMode", nodeSet="authors")
# bipartite
authortopics <- sienaDependent(array(c(author.topic.w1,author.topic.w2,author.topic.w3),
dim=c(nrauthors,nrtopics,3)),type="bipartite",nodeSet=c("authors","topics"))
# use start_publication as a constant covariate
start_publication <- coCovar(pubyear_mat[, 3 ],nodeSet="authors")
# behavior data
publishing <- sienaDependent(beh_pub_mat, type = "behavior",nodeSet="authors")
exploring <- sienaDependent(beh_exp_mat, type = "behavior",nodeSet="authors")
```

Figure 6-6 Snippet of RSIENA script for assigning networks

```

# 1. Check structural effects
effects_m1 <- includeEffects(effects_m1, transTriads, inPop, name="authorship")

# 2. Check structural of bipartite network effects
effects_m1 <- includeEffects(effects_m1, cycle4, outAct, name="authortopics")

# 3. Between-network: mixed triads
effects_m1 <- includeEffects(effects_m1, from, name="authorship", interaction1 = "authortopics")
effects_m1 <- includeEffects(effects_m1, to, name="authortopics", interaction1 = "authorship")

# 4. Selection mechanisms leading to co-authorship based on start_publication
effects_m1 <- includeEffects(effects_m1, egoX, name="authorship", interaction1 = "start_publication")
effects_m1 <- includeEffects(effects_m1, simX, name="authorship", interaction1 = "start_publication")
effects_m1 <- includeEffects(effects_m1, egoXaltX, name="authorship", interaction1 = "start_publication")

# 5. Selection mechanisms leading to co-authorship based on publishing behavior
effects_m1 <- includeEffects(effects_m1, egoX, name="authorship", interaction1 = "publishing")
effects_m1 <- includeEffects(effects_m1, simX, name="authorship", interaction1 = "publishing")
effects_m1 <- includeEffects(effects_m1, egoX, name="authorship", interaction1 = "exploring")
effects_m1 <- includeEffects(effects_m1, simX, name="authorship", interaction1 = "exploring")

# 6. Influence mechanisms leading to change in publishing behavior of researchers
effects_m1 <- includeEffects(effects_m1, avAlt, name = "publishing", interaction1 = "authorship" )
effects_m1 <- includeEffects(effects_m1, avAlt, name = "exploring", interaction1 = "authorship" )

```

Figure 6-7 Snippet of RSIENA script for assigning effects

6.4.3. RSIENA Results

Results of RSIENA estimates for the observed model with effects in Table 6-3 were shown in Table 6-4. Some specified reasons for changes in co-author networks were endogenous, career age, behavior of publishing-exploring, and mixed effects from the author-topic networks. Endogenous effect meant that next co-author selection was depended on previous co-author selections. Career age meant that co-author selection was depended on seniority level of the candidates, which was also applied on cases caused by behavior of publishing-exploring. Then, mixed effects of author-topic networks meant that selection was depended on topic interest of the candidates which lead to interest changes. The model also specified co-evolution behavior to examine whether alters can influence the behavior of an ego.

RSIENA guidelines suggest the estimation results of a specified model should have convergence ratio ≤ 0.25 . The model in Table 6-4 had convergence ratio value 0.19 which is less than 0.25. This meant that our specified conditions could explain the reasons of network evolutions for AMiner researchers. RSIENA guidelines also mention that t-ratios for all estimates of specified effect functions are around 0.1 in absolute value. Those 25 estimates in Table 6-4 satisfied the t-ratio condition although the values of t-ratios were not listed.

Some estimate values verified significant results with various confidence levels. For example, the estimate of “Degree” effect had $|-3.349/0.653| \approx 5.13 \geq 3.5$ which means a strongly significant result.

Table 6-4 RSIENA results for AMiner NLP-IE dataset

Est		par	s.e.	sig.
Co-author (one-mode) networks				
1	Rate period 1 (from Period-1 to Period-2)	1.995	0.326	***
2	Rate period 2 (from Period-2 to Period-3)	2.756	0.616	***
Endogenous effects				
3	Degree, $\beta_{deg}^{coauthor}$	-3.349	0.653	***
4	Transitive triads	2.084	0.360	***
5	In degree Popularity	-0.006	0.048	
Covariate of career age effects				
6	Ego, β_{ego}	0.226	0.232	
7	Similarity, β_{sim}	2.973	1.012	**
8	Ego x Alter, β_{exa}	-0.816	0.246	**
Behavior effects				
9	ego x alter publishing, β_{exa}^{pub}	0.247	0.265	
10	ego x alter exploring, β_{exa}^{exp}	-0.734	0.437	†
Mixed effect				
11	From topic agreement (bipartite)	-0.043	0.355	
Author-Topic (two-mode) networks				
12	Rate period 1 (from Period-1 to Period-2)	29.636	2.40	***
13	Rate period 2 (from Period-2 to Period-3)	44.766	10.88	***
Mixed effect				
14	Out Degree, $\beta_{outdeg}^{bipartite}$	-0.555	0.023	***
15	Co-authorship to topic agreement	0.012	0.033	
Co-evolution behavior: publishing				
16	Rate period 1 (from Period-1 to Period-2)	2.008	0.262	***
17	Rate period 2 (from Period-2 to Period-3)	2.307	0.395	***
Behavior dynamics				
18	Linear, β_{linear}^{pub}	0.083	0.108	
19	Quadratic, β_{quad}^{pub}	-0.012	0.063	
20	Average Similarity, β_{avSim}^{pub}	5.715	2.638	*
Co-evolution behavior: exploring				
21	Rate period 1 (from Period-1 to Period-2)	4.691	1.272	***
22	Rate period 2 (from Period-2 to Period-3)	4.527	0.985	***
Behavior dynamics				
23	Linear, β_{linear}^{exp}	-0.188	0.164	
24	Quadratic, β_{quad}^{exp}	0.136	0.137	
25	Average Similarity, β_{avSim}^{exp}	7.803	5.526	
Convergence Ratio		0.19, all t-ratios $\leq 0.1 $		
1.7 \leq t-statistic < 2.0; highly suggestive significant		† p < 0.1		
2.0 \leq t-statistic < 2.5; weakly significant		* p < 0.05		
2.5 \leq t-statistic < 3.5; moderately significant		** p < 0.01		
t-statistic (stats.) \geq 3.5; strongly significant		*** p < 0.001		
		<i>italic: not significant</i>		
		t-stats. = par / s.e.		

Table 6-5 RSIENA results for descriptive values

No	Symbol	Description	Value	Data Finding
<i>From all researchers</i>				
1	\bar{v}	Mean for covariate value of career age	3.127	Most researchers began to publish after 1990. Thus, most researchers had middle positions of seniorities compared to AMiner experts in the selected dataset
2	\bar{z}_{pub}	Mean for behavior value of publishing level	1.630	Most researchers at least published two articles per year
3	\bar{z}_{exp}	Mean for behavior value of exploring level	1.540	Most researchers at least explored two new topics per year
<i>From co-author networks in all panel waves (Period-1, Period-2, Period-3)</i> With similarity variable = 1 if two researchers of a dyad have the same value				
4	\widehat{sim}^v	Similarity mean for career age	0.735	At least 70% of co-author pairs have similar career age

```

A total of 1 dependent actor variable.

Number of missing cases per observation:
observation      1      2      3      overall
publishing      0      0      0      0      ( 0.0 %)

Means per observation:
observation      1      2      3      overall
publishing      1.410  1.505  1.986  1.634

@2
Reading constant actor covariates.
-----

1 variable, named:
start_publication

A total of 1 non-changing individual covariate.

Number of missing cases:
start_publication 0      ( 0.0 %)

Information about covariates:
|   |   |   | minimum | maximum | mean | centered |
start_publication 1.0  5.0  3.127  Y
The mean value is subtracted from the centered covariate.
    
```

Figure 6-8 Sample of RSIENA output file from the observed model

A similar outcome occurred for the estimate of “Transitive triads” effect with $|2.084/0.360| \approx 5.79 \geq 3.5$. Those two estimates confirmed reasons for tie formation as a part of the network evolution in co-authoring process. With a strongly significant result on the “Degree” effect, we inferred that asking a researcher to become a co-author was not easy due to the negative estimate. Then, with a strongly significant result on the “Transitive triads” effect, we also inferred that asking a friend of a friend to become a co-author was easier due to the positive estimate. Combining those values is a part of the evaluation function.

6.4.4. RSIENA Evaluation Functions

RSIENA estimated the effects between three network waves in our experiments. Some descriptive values in Table 6-5 were obtained in RSIENA output file as shown in Figure 6-8. Those values became the constants in the evaluations functions for co-author selection as listed in Table 6-6. The first function (1) examined the co-author selection because of career age similarity, the second function (2) was on publishing behavior, while the third function (3) on exploring behavior. Formal equations for those functions were following RSIENA guidelines and substituting the constants based on RSIENA estimates in Table 6-4. For constant value of $\Delta_v = 5 - 1 = 4$ was taken from the covariate of career age with values 1...5 as enumerated in Table 6-2. Some evaluations results obtained from those functions were illustrated in Figure 6-9 for (4.1), Table 6-7 for (4.2), and Table 6-8 for (4.3). In a case of ego with career age $v_i^{age} = 4$ and alter with career age $v_j^{age} = 3$, then function (1) yielded to

$$= 0.23(4 - 3.13) + 0.23(3 - 3.13) + 2.97\left(1 - \frac{1}{4} - 0.73\right) - 0.82(4 - 3.13)(3 - 3.13) \approx 0.22$$

With ego-alter for all career age values, function (4.1) had results in Figure 6-9.

There were five graphics of log-odds in Figure 6-9 to demonstrate interaction between egos and alters with different covariate values of career age. The interaction for each ego scenario assumed that a researcher as an ego has co-authors who all of them have the same career age values.

In a case of ego with career age $v_i^{age} = 4$ and alter with career age $v_j^{age} = 3$, the ego was older than the alter, the log-odds value was $e^{0.22} \approx 1.24$.

Table 6-6 RSIENA evaluation functions to observe co-author selection

Eq.	Function based on RSIENA effects for co-author selection
(4.1)	career age similarity: ego, alter, similarity, ego×alter $f_i^{sel}(x, cov_{start.pub}) = \beta_{ego}(v_i - \bar{v}) + \beta_{alter}(v_j - \bar{v})$ $+ \beta_{sim} \left(1 - \frac{ v_i - v_j }{\Delta_v} - \widehat{sim}^v \right) + \beta_{exa}(v_i - \bar{v})(v_j - \bar{v})$ $= 0.23(v_i - 3.13) + 0.23(v_j - 3.13)$ $+ 2.97 \left(1 - \frac{ v_i - v_j }{4} - 0.73 \right)$ $- 0.82(v_i - 3.13)(v_j - 3.13)$
(4.2)	publishing behavior: ego×alter $f_i^{sel}(x, beh_{pub}) = \beta_{exa}^{pub}(z_i^{pub} - \bar{z}_{pub})(z_j^{pub} - \bar{z}_{pub})$ $= 0.25(z_i^{pub} - 1.63)(z_j^{pub} - 1.63)$
(4.3)	exploring behavior: ego×alter $f_i^{sel}(x, beh_{exp}) = \beta_{exa}^{exp}(z_i^{exp} - \bar{z}_{exp})(z_j^{exp} - \bar{z}_{exp})$ $= -0.73(z_i^{exp} - 1.54)(z_j^{exp} - 1.54)$

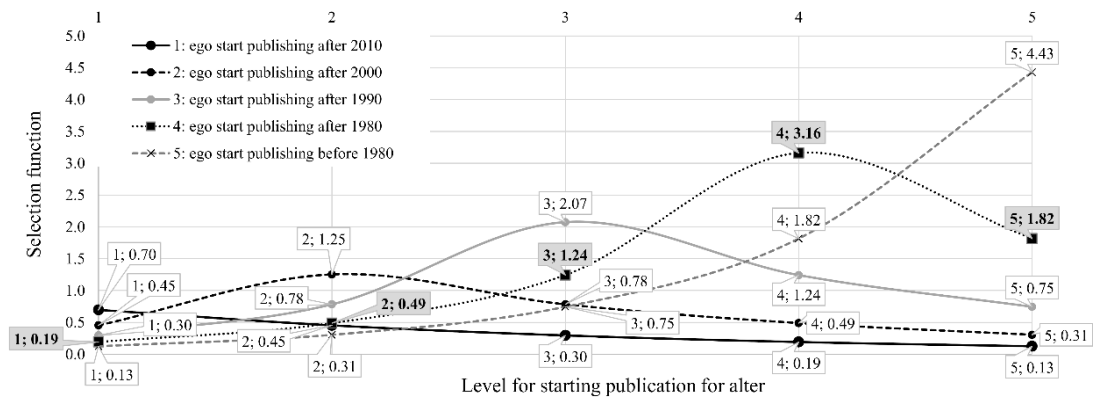


Figure 6-9 Log-odds plot for co-author selection based on career age

Then, in a case of ego with career age $v_i^{age} = 4$ and alter with career age $v_j^{age} = 5$, the ego was younger than the alter, the log-odds value was $e^{0.60} \approx 1.82$.

Thus, the probability of ego $v_i^{age} = 4$ to work with older co-author $v_j^{age} = 5$ was higher than to work with the younger one $v_j^{age} = 3$. However, if the alter with career age $v_j^{age} = 2$, the log-odds value was $e^{-0.72} \approx 0.49$. This meant that the probability

of the ego to work with much younger one was less likely happened or reduced by half.

These interpretations were following the rules defined in RSIENA guidelines. Thus, Figure 6-9 indicates a stronger preference for researchers to be linked with peers who have similar career age with the difference of $\pm 3-7$ years, or seniors who have more writing experience ≥ 10 years. That indication supported knowledge transfer or academic mentoring in collaborations, i.e. supervision activities [69]. Since career age has demonstrated certain level of influence in co-author selection, we argued that career age could be another feature in the scholar profile. In the next chapter, we applied career age value as a feature for determining expertise rank of a researcher.

For co-author selection function related to publishing behavior, with a case of ego_1 and alt_1 , the function (4.2) in Table 6-6 yielded to $= 0.25(1 - 1.63)(1 - 1.63) \approx 0.10$ as shown in Table 6-7.

An ego who publishes less, $z_i^{pub} = 1$, with the values of ego_1 , had lower attraction to productive alters with higher publishing level. The probability value of $ego_1 \times alt_2 = e^{-0.06} \approx 0.94 < 1$ showed that co-authoring between $ego_1 \times alt_2$ is less likely happened. However, the collaboration chance was higher if both authors stand on the same stage such as $ego_1 \times alt_1 = e^{0.10} \approx 1.11$. The probability of middle experts ($z_i^{pub} = \{2,3\}$) to co-author with researchers who have more experience was higher, i.e. $ego_2 \times alt_4 = e^{0.22} \approx 1.25$. Middle experts are researchers who have behavior to publish more frequently.

Therefore, the finding suggests either ego or alter is at least publishing more than one article in a year ($z_i^{pub} \geq 2, z_j^{pub} \geq 2$), i.e. $ego_2 \times alt_2 = e^{0.03} \approx 1.03$.

The probability to connect for an ego $z_i^{pub} = 3$ with more active alter of $z_j^{pub} = 4$ is $ego_3 \times alt_4 = e^{0.80} \approx 2.23$ times or more than twice as high as the probability of no forming ties at all.

For co-author selection function related to exploring behavior, with a case of ego_1 and alt_2 , the function (3) in Table 6-6 yielded to $= -0.73(1 - 1.54)(2 - 1.54) \approx 0.17$ as shown in Table 6-8.

Table 6-7 RSIENA evaluation results based on publishing behavior

Based on publishing behavior		<i>Selection attractiveness</i>			
An alter with z_j^{pub}		alt_1	alt_2	alt_3	alt_4
An ego publishes at least ... (z_i^{pub})					
one article in a year	ego_1	0.10	-0.06	-0.21	-0.37
one article per semester (6 months)	ego_2		0.03	0.13	0.22
more than two articles per semester	ego_3			0.46	0.80
one article in every other month	ego_4				1.39

Table 6-8 RSIENA evaluation results based on exploring behavior

Based on exploring behavior		<i>Selection attractiveness</i>		
An alter with z_j^{exp}		alt_1	alt_2	alt_3
An ego explores ... (z_i^{exp})				
at most one new topic in each year	ego_1	-0.29	0.17	0.63
at most two new topics in each year	ego_2		-0.10	-0.37
at least three new topics in each year	ego_3			-1.38

An ego who explores less, $z_i^{pub} = 1$, with the values of ego_1 , had lower attraction to the alter peers with the same exploring level. The probability value of $ego_1 \times alt_1 = e^{-0.29} \approx 0.75 < 1$ showed that co-authoring between $ego_1 \times alt_1$ is less likely happened.

With cases of $ego_1 \times alt_2$ and $ego_1 \times alt_3$, the findings concluded that researchers will get more benefit if their co-authors have different gap in exploring level.

In a case of ego with $ego_1 \times alt_2$, the log-odds value was $e^{0.17} \approx 1.19$.

In a case of ego with $ego_1 \times alt_3$, the log-odds value was $e^{0.63} \approx 1.88$.

There are two reasonable situations for researchers with less exploring behavior levels.

- Assuming the researchers are the junior ones, then exploring fewer topics means that they still explore candidate topics to become their main interest and make the researchers to have more responsibilities in experimental works.
- Assuming the researchers are the senior ones, then exploring fewer topics means that they are already experts who have decided their main interest.

Both situations support a case of mentoring process [69].

Evaluation function for forming ties from RSIENA estimate of “Degree” (Est-3 in Table 6-4) indicated preference to connect with existing co-authors. By considering RSIENA estimates as costs for forming ties, there was a negative cost -3.35 of “Degree” effect, for co-authoring with researchers who never collaborate before. However, tie formation gave a positive cost 2.08 from RSIENA estimate of “Transitive” (Est-4 in Table 6-4). Evaluation forming function based on previous ties (from “Degree” and “Transitive” effects) gave final cost $-3.35 + 2.08 \approx -1.27$ as a negative value. It meant that the tie formation needed other aspects aside of previous ties, since the forming probability was rather low with $e^{-1.27} = 0.281 < 1$.

By considering career age, a positive value of the function was obtained from a case of $ego_5 \times alt_5$ with $-3.35 + 2.08 + 1.54 \approx 0.27$, the forming probability $e^{0.27} = 1.31$. This finding indicated career age or experience in publishing articles was not significant reason for researchers in co-authoring process. However, the estimates of effects related to career age have quite moderately significant results, especially on “Similarity” and “Ego x Alter”. Therefore, we apply the career age value as one of the features in expertise rank as described in the next chapter.

Table 6-9 RSIENA evaluations related to the collaboration dynamics of co-authors

Function based on RSIENA effects		Confirmed hypothesis
Co-author Selection	a. career age similarity: ego, alter, similarity, ego \times alter $f_i^{sel}(x, cov_{start.pub})$	H1: “Bipartite author-topic networks based on topic interests demonstrate transitive closure and researcher preferences in forming cliques”. H2: “Behavior values from bipartite author-topic networks based on topic interests are associated with experience such that researchers incline to form ties with others in looking for supervision aspect”.
	b. publishing behavior: ego \times alter $f_i^{sel}(x, beh_{pub})$	
	c. exploring behavior: ego \times alter $f_i^{sel}(x, beh_{exp})$	
Co-author Influence	a. publishing behavior $f_i^{inf}(x, beh_{pub})$	
	b. exploring behavior $f_i^{inf}(x, beh_{exp})$	

The evaluation function for publishing (4.2) and exploring (4.3) behavior in Table 6-6 required combinations of ego and alter, and the difference between highest

and lowest values of the functions. In case of exploring behavior, the difference is taken from $|ego_1 \times alt_3| + |ego_3 \times alt_3| = 0.63 + 1.38 = 2.01$ using values in Table 6-8. Thus, the evaluation function for co-author selection with “Degree” and “Transitive” effects in addition to the behavior effect gave a positive result of $-3.349 + 2.084 + 2.01 = 0.745$. This finding indicated exploring behavior was significant reason for researchers in co-authoring process. Therefore, we also apply the exploring behavior values to the features in expertise rank in the next chapter.

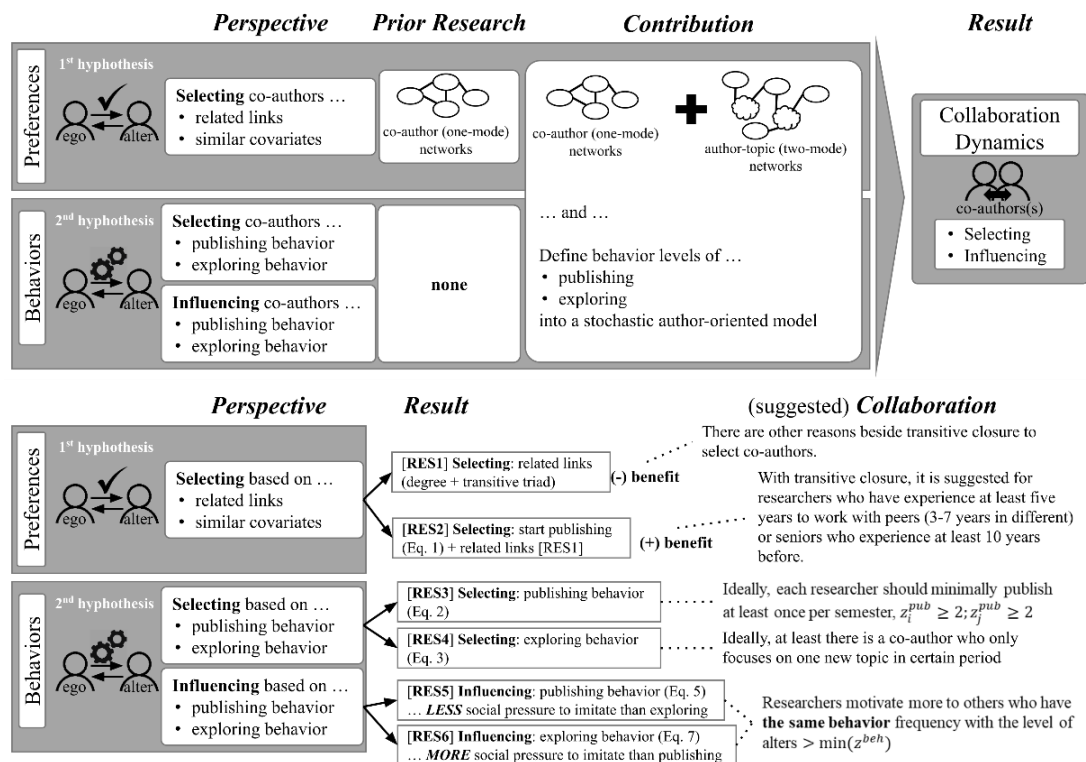


Figure 6-10 Hindsight on co-authoring collaborations from AMiner NLP.IE experts

We also performed further observations as hypothesized in Table 6-9 [40]. The observations were related to the collaboration dynamics between co-. Those hypothesis are still related to whether behavior especially exploring influences the researchers in their publishing works, which eventually has some parts in their career advancing. Summaries for those observations about preferences and behaviors of researchers in co-authoring process are displayed in Figure 6-10.

6.4. Summary

Table 6-9 listed a sample of AMiner NLP.IE researcher used who still active in 1995 until 2015. The number difference of published articles and received citations 2015 could indicated there was a positive influence based on learning from behaviors of other productive researchers. Those hindsight suggest university management or the government to define research policies especially for funding or research grants. For example, each research proposal grant in national level must have a minimal number of lecturers with junior academic rank (one “Asisten Ahli” and one “Lektor”) to construct a good environment for promoting research motivation. Another policy is asking lecturers who have not produced an accredited journal as the first author after some years must be included in the research team. Then, those lecturers should publish an accredited journal article in the next year as the first author.

Table 6-10 Number of articles and citations in a sample of AMiner NLP.IE experts

Scopus ID	Expert Name	± 1995		On 2015	Accumulated after 2015		
		#Docs	#Cites	#Cites	h-index	#Docs	#Cites
7202745471	Kevin Knight	1	1	204	24	94	2425
6603963324	Kristina Lerman	1	9	313	31	148	3321
6602712741	Philip Stuart Resnik	1	1	281	21	74	3143
6602721887	Ellen Riloff	1	4	199	21	50	2329
6603954639	Marti A. Hearst	3	7	427	33	100	6214
7003940794	Luis Gravano	3	19	283	33	81	4681
16410214900	David Eric Yarowsky	1	2	158	19	39	1185
24604968400	Alexander Gelbukh	5	5	209	22	288	1859

We have described procedures to extract features from publishing related behaviors especially on exploration and consistency on topic interest. The behaviors were extracted as longitudinal analysis on longer observation period which is 15 years in this dissertation. Experiment findings showed that the career age and exploration features did matter in co-authoring process which eventually influences the expertise of researchers.

Chapter 7.

EXPERTISE RANK USING SCHOLAR PROFILE

Previous chapters (Chapter 5 and Chapter 6) describe extracting features related to mapped topics of researchers, in which there are two productivity-dynamicity features, in addition to the six behaviors of exploring and consistency features for some periods. This chapter discusses the usage of those features with some additional ones for rank expertise. The additions include the spreading level of topics because of interest changes through graph analysis (Section 7.1.1), and some schemes to acquire features related to citation number (Section 7.1.2). However, our approaches still consider the quality of citations to avoid biased citations through grouping articles based on received citations before counting the articles.

We investigated expertise score of specified topics using the assumption that all evidences have similar weights. Thus, the experiments were performed on a linear model. Some empirical settings were related to give weights to each feature with heuristically stepwise (Section 7.2) and approximate the weights by fitting the feature values (Section 7.3). We observed some variants of linear model, from Gaussian to model error distribution until general assumptions and boosting model. Our observations are included any feature combinations to obtain the expertise scores. Since topic information is required, we manually analyzed which topics being frequently mapped to researchers as their interest in our dataset. Those selected topics became different queries to generalize the empirical settings in our experiments.

Although AMiner gave list of researchers based on their expertise of specified topics, there is no information about the expertise scores. Motivated by previous studies [41] we compared the expertise scores obtained from the proposed scholar profile with existing scores of researchers, which is h-index through correlation analysis. Those h-index scores of Scopus assumed that all researchers have same level of expertise in the listed of Scopus subject areas.

Then, we summarize the results and emphasize the findings to illustrate its possible implementation in real problem related to researchers (Section 7.4).

Chapter 8.

CONCLUSIONS AND FUTURE WORKS

This dissertation introduces the need for a scholar profile with respect to the possibilities of interest changes and focused less on citations to avoid inflating h-index of researchers to show their expertise. The problem also mentioned conditions with no predefined topics, which is valuable for mapping topics to articles and then researchers. Those issues are often occurred in conditions of unrestricted policies like not assessing the article quality of researchers. Thus, this dissertation investigated on modeling a scholar profile with article metadata, which is easy to retrieve especially because of the Internet growth and its information abundance effect. The main contribution for modeling a scholar profile is to acquire credible and less-biased information of researchers throughout productivity-dynamicity and behavior aspects.

We have performed analytical and experimental works to obtain the following findings. Clustering approach to obtain topics has been presented by considering word embedding for representing context relations between words, especially title texts that showed more coherence words within the topics. Then mapping topics to articles and researchers have supported extracting productivity-dynamicity features as well as behavior features of researchers as evidence of their research expertise. However, feature selection with correlation and applying on predicting has validated two notable features on publishing articles and received citations. Those features are sufficient for representing productivity-dynamicity of researchers. Then, the efficacy of behavior features have been confirmed using a network evolution model to ensure that exploration and exploitation of researchers are correlated to their expertise.

We also demonstrated scores for the expertise of researchers on specified topics by using the contributed features. However, we have completed the features with the currentness of researchers in terms of publishing article and citations, and also topics relatedness. To measure the performance, we performed correlation analysis on our expertise score and h-index values of researchers. The findings demonstrated that some features without citations had similar performance compared to all features in rank expertise.

There are further research works related to this dissertation could be performed, i.e. dataset scope, expertise variations of researchers since the current works still focused on the experts and thriving ones. Therefore, the next works should be on the researchers with less or even much less expertise for establishing more generalization on the findings. The following approaches are recommended to perform more observations.

1. Experiments on AMiner dataset with different domains, or a dataset for Indonesian researchers. Then, extend the datasets by snowball sampling to add more variation of expertise levels of the researchers.
2. Experiments on different period length to shorten the longitudinal analysis since a cold-start condition may cause inadequate article metadata.
3. Complement the dataset with funding information to generate scholar profile with the ripple effect on subject domains because funding may encourage research works on certain topics. The phenomenon of research topic burst may indicate repetitiveness and help management or government in designing research policy.

REFERENCES

- [1] A. Lowrie and P. J. McKnight, "Academic Research Networks:," *Eur. Manag. J.*, vol. 22, no. 4, pp. 345–360, Aug. 2004.
- [2] J. Tang, "AMiner: Toward Understanding Big Scholar Data," in *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, 2016, p. 467.
- [3] S. Lin, W. Hong, D. Wang, and T. Li, "A survey on expert finding techniques," *J. Intell. Inf. Syst.*, vol. 49, no. 2, pp. 255–279, Oct. 2017.
- [4] K. Balog, Y. Fang, M. de Rijke, P. Serdyukov, and L. Si, "Expertise Retrieval," *Found. Trends® Inf. Retr.*, vol. 6, no. 2–3, pp. 127–256, 2012.
- [5] X. Kong, Y. Shi, S. Yu, J. Liu, and F. Xia, "Academic social networks: Modeling, analysis, mining and applications," *J. Netw. Comput. Appl.*, vol. 132, pp. 86–103, 2019.
- [6] H. Deng, I. King, and M. R. Lyu, "Formal Models for Expert Finding on DBLP Bibliography Data," in *2008 Eighth IEEE International Conference on Data Mining*, 2008, pp. 163–172.
- [7] J. Tang *et al.*, "Topic level expertise search over heterogeneous networks," *Mach. Learn.*, vol. 82, no. 2, pp. 211–237, Feb. 2011.
- [8] Y. Tu, N. Johri, D. Roth, and J. Hockenmaier, "Citation Author Topic Model in Expert Search," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, 2010, pp. 1265–1273.
- [9] D. Fiala, "Time-aware PageRank for bibliographic networks," *J. Informetr.*, vol. 6, no. 3, pp. 370–388, 2012.
- [10] F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang, "MVCWalker: Random Walk-Based Most Valuable Collaborators Recommendation Exploiting Academic Factors," *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 364–375, Sep. 2014.
- [11] M. Nykl, M. Campr, and K. Ježek, "Author ranking based on personalized PageRank," *J. Informetr.*, vol. 9, no. 4, pp. 777–799, 2015.
- [12] M. Franceschet and G. Colavizza, "TimeRank: A dynamic approach to rate scholars using citations," *J. Informetr.*, vol. 11, no. 4, pp. 1128–1141, 2017.
- [13] Y. Xu, X. Guo, J. Hao, J. Ma, R. Y. K. Lau, and W. Xu, "Combining social network and semantic concept analysis for personalized academic researcher recommendation," *Decis. Support Syst.*, vol. 54, no. 1, pp. 564–573, 2012.
- [14] K. Balog, L. Azzopardi, and M. de Rijke, "A language modeling framework for expert finding," *Inf. Process. Manag.*, vol. 45, no. 1, pp. 1–19, 2009.
- [15] J. Liu, B. Jia, H. Xu, B. Liu, D. Gao, and B. Li, "A TopicRank Based Document Priors Model for Expert Finding," in *Advanced Computational Methods in Life System Modeling and Simulation*, 2017, pp. 334–341.
- [16] C. Moreira and A. Wichert, "Finding academic experts on a multisensor approach using Shannon's entropy," *Expert Syst. Appl.*, vol. 40, no. 14, pp. 5740–5754, 2013.
- [17] N. Torkzadeh Mahani, M. Dehghani, M. S. Mirian, A. Shakery, and K. Taheri, "Expert finding by the Dempster-Shafer theory for evidence combination," *Expert Syst.*, vol. 35, no. 1, p. e12231, 2018.
- [18] L. Bornmann and H.-D. Daniel, "The state of h index research," *EMBO Rep.*, vol. 10, no. 1, pp. 2–6, Dec. 2008.
- [19] A.-W. Harzing, S. Alakangas, and D. Adams, "hIa: an individual annual h-index to accommodate disciplinary and career length differences," *Scientometrics*, vol. 99, no. 3, pp. 811–821, 2014.
- [20] A. Daud, J. Li, L. Zhou, and F. Muhammad, "Temporal expert finding through generalized time topic modeling," *Knowledge-Based Syst.*, vol. 23, no. 6, pp. 615–625, 2010.
- [21] X. Kong, H. Jiang, W. Wang, T. M. Bekele, Z. Xu, and M. Wang, "Exploring dynamic research interest and academic influence for scientific collaborator recommendation,"

Scientometrics, vol. 113, no. 1, pp. 369–385, Oct. 2017.

- [22] M. Neshati, S. H. Hashemi, and H. Beigy, “Expertise Finding in Bibliographic Network: Topic Dominance Learning Approach,” *IEEE Trans. Cybern.*, vol. 44, no. 12, pp. 2646–2657, Dec. 2014.
- [23] G. A. Wang, J. Jiao, A. S. Abrahams, W. Fan, and Z. Zhang, “ExpertRank: A topic-aware expert finding algorithm for online knowledge communities,” *Decis. Support Syst.*, vol. 54, no. 3, pp. 1442–1451, 2013.
- [24] G. Panagopoulos, G. Tsatsaronis, and I. Varlamis, “Detecting rising stars in dynamic collaborative networks,” *J. Informetr.*, vol. 11, no. 1, pp. 198–222, 2017.
- [25] T. Amjad, A. Daud, and M. Song, “Measuring the Impact of Topic Drift in Scholarly Networks,” in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 373–378.
- [26] M. D. Siciliano, E. W. Welch, and M. K. Feeney, “Network exploration and exploitation: Professional network churn and scientific production,” *Soc. Networks*, vol. 52, pp. 167–179, 2018.
- [27] J. G. Foster, A. Rzhetsky, and J. A. Evans, “Tradition and Innovation in Scientists’ Research Strategies,” *Am. Sociol. Rev.*, vol. 80, no. 5, pp. 875–908, 2015.
- [28] C. Bartneck and S. Kokkelmans, “Detecting h-index manipulation through self-citation analysis,” *Scientometrics*, vol. 87, no. 1, pp. 85–98, Apr. 2011.
- [29] B. R. Martin, “Whither research integrity? Plagiarism, self-plagiarism and coercive citation in an age of research assessment,” *Res. Policy*, vol. 42, no. 5, pp. 1005–1014, 2013.
- [30] T. Yu, G. Yu, and M.-Y. Wang, “Classification method for detecting coercive self-citation in journals,” *J. Informetr.*, vol. 8, no. 1, pp. 123–135, 2014.
- [31] K. Moustafa, “Aberration of the Citation,” *Account. Res.*, vol. 23, no. 4, pp. 230–244, 2016.
- [32] X. Bai, F. Xia, I. Lee, J. Zhang, and Z. Ning, “Identifying Anomalous Citations for Objective Evaluation of Scholarly Article Impact,” *PLoS One*, vol. 11, no. 9, pp. 1–15, 2016.
- [33] X. Bai, I. Lee, Z. Ning, A. Tolba, and F. Xia, “The Role of Positive and Negative Citations in Scientific Evaluation,” *IEEE Access*, vol. 5, pp. 17607–17617, 2017.
- [34] A. Abbasi, K. S. K. Chung, and L. Hossain, “Egocentric analysis of co-authorship network structure, position and performance,” *Inf. Process. Manag.*, vol. 48, no. 4, pp. 671–679, 2012.
- [35] D. Purwitasari, C. Fatichah, S. Sumpeno, and M. H. Purnomo, “Ekstraksi Ciri Produktivitas Dinamis untuk Prediksi Topik Pakar dengan Model Discrete Choice,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 4, 2018.
- [36] D. Purwitasari, C. Fatichah, A. G. Sooai, S. Sumpeno, and M. H. Purnomo, “Productivity-based Features from Article Metadata for Fuzzy Rules to Classify Academic Expert,” in *The 10th International Conference on Awareness Science and Technology (iCAST 2019)*, 2019.
- [37] D. Purwitasari, C. Fatichah, I. K. E. Purnama, S. Sumpeno, and M. H. Purnomo, “Inter-departmental research collaboration recommender system based on content filtering in a cold start problem,” in *2017 IEEE 10th International Workshop on Computational Intelligence and Applications, IWCIA 2017 - Proceedings*, 2017, vol. 2017-Decem.
- [38] D. Purwitasari, A. B. Ilmi, C. Fatichah, W. A. Fauzi, S. Sumpeno, and M. H. Purnomo, “Conflict of Interest based Features for Expert Classification in Bibliographic Network,” in *2018 IEEE International Conference on Computer Engineering, Network and Intelligent Multimedia, CENIM 2018 - Proceedings*, 2018.
- [39] D. Purwitasari, R. Alamsyah, D. A. Navastara, C. Fatichah, S. Sumpeno, and M. H. Purnomo, “Visualizing Academic Experts on a Subject Domain Map of Cartographic-alike,” in *4th International Conference on Computer, Communication and Computational Sciences (IC4S2019)*, 2019.
- [40] D. Purwitasari, C. Fatichah, S. Sumpeno, C. Steglich, and M. H. Purnomo, “Identifying

Collaboration Dynamics of Bipartite Author-Topic Networks with the Influences of Interest Changes,” *Springer Int. J. Sci.*, 2020.

- [41] X. Kong, L. Liu, S. Yu, A. Yang, X. Bai, and B. Xu, “Skill ranking of researchers via hypergraph,” *PeerJ Comput. Sci.*, vol. 5, p. e182, Mar. 2019.
- [42] L. Guo, X. Cai, F. Hao, D. Mu, C. Fang, and L. Yang, “Exploiting Fine-Grained Co-Authorship for Personalized Citation Recommendation,” *IEEE Access*, vol. 5, pp. 12714–12725, 2017.
- [43] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient Estimation of Word Representations in Vector Space,” *CoRR*, vol. abs/1301.3, 2013.
- [44] K. Balog, L. Azzopardi, and M. de Rijke, “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*, 2006, pp. 43–50.
- [45] D. Liu, W. Xu, W. Du, and F. Wang, “How to Choose Appropriate Experts for Peer Review: An Intelligent Recommendation Method in a Big Data Context,” *Data Sci. J.*, vol. 14, no. 0, p. 16, May 2015.
- [46] A. Skupin, “A Cartographic Approach to Visualizing Conference Abstracts,” *IEEE Comput. Graph. Appl.*, vol. 22, no. 1, pp. 50–58, Jan. 2002.
- [47] X. Jiang and J. Zhang, “A Text Visualization Method for Cross-Domain Research Topic Mining,” *J. Vis.*, vol. 19, no. 3, pp. 561–576, 2016.
- [48] J. Zhang, C. Chen, and J. Li, “Visualizing the Intellectual Structure with Paper-Reference Matrices,” *IEEE Trans. Vis. Comput. Graph.*, vol. 15, no. 6, pp. 1153–1160, 2009.
- [49] H. Deng, J. Han, M. R. Lyu, and I. King, “Modeling and Exploiting Heterogeneous Bibliographic Networks for Expertise Ranking,” *Proc. ACM/IEEE Jt. Conf. Digit. Libr.*, pp. 71–80, 2012.
- [50] A. Suominen and H. Toivanen, “Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification,” *J. Assoc. Inf. Sci. Technol.*, vol. 67, no. 10, pp. 2464–2476, 2016.
- [51] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 2010, pp. 45–50.
- [52] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” *J. Comput. Appl. Math.*, vol. 20, pp. 53–65, 1987.
- [53] T. Hofmann, “Probabilistic Latent Semantic Indexing,” in *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1999, pp. 50–57.
- [54] H. Zhang and G. Zhong, “Improving short text classification by learning vector representations of both words and hidden topics,” *Knowledge-Based Syst.*, vol. 102, no. Supplement C, pp. 76–86, 2016.
- [55] A. Nurilham, D. Purwitasari, and C. Fatichah, “Ekstraksi Frasa pada Pelabelan Kelompok Artikel Ilmiah dengan Penggabungan Klaster berdasarkan MaximumCommonSubgraph,” *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 7, no. 3, 2018.
- [56] C. Zhang, Z. Li, and J. Zhang, “A survey on visualization for scientific literature topics,” *J. Vis.*, vol. 21, no. 2, pp. 321–335, Apr. 2018.
- [57] A. Skupin, “The world of geography: Visualizing a knowledge domain with cartographic means,” *Proc. Natl. Acad. Sci.*, vol. 101, no. Supplement 1, pp. 5274–5278, 2004.
- [58] K. Hu *et al.*, “Understanding the topic evolution of scientific literatures like an evolving city: Using Google Word2Vec model and spatial autocorrelation analysis,” *Inf. Process. Manag.*, vol. 56, no. 4, pp. 1185–1203, 2019.
- [59] L. van der Maaten and G. Hinton, “Visualizing Data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.
- [60] J. E. Hirsch, “An index to quantify an individual’s scientific research output,” *Proc. Natl.*

Acad. Sci., vol. 102, no. 46, pp. 16569–16572, 2005.

- [61] J. Mueller and A. Thyagarajan, “Siamese Recurrent Architectures for Learning Sentence Similarity,” in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016, pp. 2786–2792.
- [62] Z. Yang, J. Tang, B. Wang, J. Guo, and J. Li, “Expert2Bólè : From Expert Finding to Bólè Search,” *Proc. 15th ACM Conf. Knowl. Discov. data Min.*, pp. 1–4, 2009.
- [63] M.-Y. Chen and D. A. Linkens, “Rule-base self-generation and simplification for data-driven fuzzy models,” *Fuzzy Sets Syst.*, vol. 142, no. 2, pp. 243–265, 2004.
- [64] H. Iglíč, P. Doreian, L. Kronegger, and A. Ferligoj, “With whom do researchers collaborate and why?,” *Scientometrics*, vol. 112, no. 1, pp. 153–174, Jul. 2017.
- [65] T. A. B. Snijders, “The Statistical Evaluation of Social Network Dynamics,” *Sociol. Methodol.*, vol. 31, no. 1, pp. 361–395, 2001.
- [66] T. A. B. Snijders, G. G. van de Bunt, and C. E. G. Steglich, “Introduction to stochastic actor-based models for network dynamics,” *Soc. Networks*, vol. 32, no. 1, pp. 44–60, 2010.
- [67] A. Ferligoj, L. Kronegger, F. Mali, T. A. B. Snijders, and P. Doreian, “Scientific collaboration dynamics in a national scientific system,” *Scientometrics*, vol. 104, no. 3, pp. 985–1012, Sep. 2015.
- [68] M. S. Manger, M. A. Pickup, and T. A. B. Snijders, “A Hierarchy of Preferences: A Longitudinal Network Analysis Approach to PTA Formation,” *J. Conflict Resolut.*, vol. 56, no. 5, pp. 853–878, 2012.
- [69] S. Shibayama, “Sustainable development of science and scientists: Academic training in life science labs,” *Res. Policy*, vol. 48, no. 3, pp. 676–692, 2019.
- [70] D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himmelboim, “Chapter 3 - Social network analysis: Measuring, mapping, and modeling collections of connections,” in *Analyzing Social Media Networks with NodeXL (Second Edition)*, Second Edi., D. L. Hansen, B. Shneiderman, M. A. Smith, and I. Himmelboim, Eds. Morgan Kaufmann, 2020, pp. 31–51.
- [71] N. Bhushan and K. Rai, *Strategic Decision Making: Applying the Analytic Hierarchy Process*. Springer-Verlag, 2004.

Appendix 1. RSIENA SCRIPTS FOR EXAMINING EXPLORING FEATURE

```
1 #-----
2 # created by: Diana Purwitasari (January-March 2019)
3 # Institut Teknologi Sepuluh Nopember, Indonesia
4 #-----
5
6 # Analyze collaboration dynamics of bipartite co-author networks
7 # H1: Bipartite co-authorship networks based on interests demonstrate transitive closure
8 # for researcher preferences to their cliques
9 # H2: The influencing likelihood is associated with experience such that
10 # researchers with particular expertise rank incline to form ties among themselves
11 # and others who have certain level difference
12
13 # behavior data menggunakan articles dengan co-authors tdk hanya dari 212 experts
14 # 1588 papers yang setidaknya 2 author nya ada di 212 experts
15 # 9192 papers yang setidaknya 1 author nya ada di 212 experts
16 # menggunakan i7-8700 3.2GHz 16 core
17 # structural effects:
18 # transTriads,
19 # Between-network
20 #   from,name="authorship",interaction1 = "authortopics"
21 #   to,name="authortopics",interaction1 = "authorship"
22 # Selection mechanisms
23 #   egoX,simX,egoXaltX,name="authorship",interaction1 = "start_publication"
24 #   egoXaltX,name="authorship",interaction1 = "exploring"
25 #   egoXaltX,name="authorship",interaction1 = "publishing"
26
27 library(RSienna)
28
29
30 setwd("D:/my_rcodes/siena_apr2019_cs")
31 # adjacency matrix (co-author network), size: 212 x 212 scholars
32 # author.w1(i,j) = number of co-authored articles
33 author.w1 <- data.matrix(read.csv("nd1_period1.csv", header=FALSE, sep=""))#Period1:articles from
1969-1995
34 author.w2 <- data.matrix(read.csv("nd1_period2.csv", header=FALSE, sep=""))#Period2:articles from
1996-2000
35 author.w3 <- data.matrix(read.csv("nd1_period3.csv", header=FALSE, sep=""))#Period3:articles from
2001-2005
36
37 # matrix size for bipartite data: 212 authors x 30 topics
38 author.topic.w1 <- data.matrix(read.csv("bip_period1_top2a.csv", header=TRUE, sep=";"))
39 author.topic.w2 <- data.matrix(read.csv("bip_period2_top2a.csv", header=TRUE, sep=";"))
40 author.topic.w3 <- data.matrix(read.csv("bip_period3_top2a.csv", header=TRUE, sep=";"))
41
42 # covariate data
43 pubyear_mat <- data.matrix(read.csv("author_publish_year.csv", header=TRUE, sep=";"))
44 # behavior data
45 beh_pub_mat <- data.matrix(read.csv("beh_publishing.csv", header=TRUE, sep=";"))
46 beh_exp_mat <- data.matrix(read.csv("beh_exp_top2_efforts.csv", header=TRUE, sep=";"))
47
48 # encode author.w1(i,j) into binary values
49 author.w1[author.w1>0] <- 1
50 author.w2[author.w2>0] <- 1
51 author.w3[author.w3>0] <- 1
```

RSIENA (R package in Simulation Investigation for Empirical Network Analysis) module requires data input for networks of actors which author in this dissertation and their characteristics on publishing articles and exploring topics. The above scripts were about setting those input as matrices from text files of CSV (comma separated values) for a number of observation periods.

```

52 # author.topic.w1(i,j) = 1 ... in Period1, author-i published at least one article of topic-j
53 author.topic.w1[author.topic.w1>0] <- 1
54 author.topic.w2[author.topic.w2>0] <- 1
55 author.topic.w3[author.topic.w3>0] <- 1
56
57 nrauthors <- nrow(author.w1) # 212 authors
58 authors <- sienaNodeSet(nrauthors,nodeSetName="authors")
59 nrtopics <- ncol(author.topic.w1) # 30 topics
60 topics <- sienaNodeSet(nrtopics,nodeSetName="topics")
61
62 # oneMode
63 authorship <- sienaDependent(array(c(author.w1,author.w2,author.w3),
64   dim=c(nrauthors,nrauthors,3)),type="oneMode", nodeSet="authors")
65 # bipartite
66 authortopics <- sienaDependent(array(c(author.topic.w1,author.topic.w2,author.topic.w3),
67   dim=c(nrauthors,nrtopics,3)),type="bipartite",nodeSet=c("authors","topics"))
68
69 # use start_publication as a constant covariate
70 start_publication <- coCovar(pubyear_mat[, 3 ],nodeSet="authors")
71 # behavior data
72 publishing <- sienaDependent(beh_pub_mat, type = "behavior",nodeSet="authors")
73 exploring <- sienaDependent(beh_exp_mat, type = "behavior",nodeSet="authors")
74
75 data_ml <- sienaDataCreate(authorship, authortopics, start_publication, publishing, exploring, nodeSets
76 =list(authors, topics))
77 print01Report(data_ml,modelname="Model_M1")
78 effects_ml <- getEffects(data_ml)
79
80 # 1. Check structural effects
81 # transTriads ... indicates the cyclical pattern among researchers
82 # inPop ... supports the Matthew effect --> popular researchers tend to collaborate more
83 effects_ml <- includeEffects(effects_ml,transTriads,inPop,name="authorship")
84
85 # 2. Check structural of bipartite network effects
86 # cycle4 ... If a pair of researchers has one topic in common, they will get more topics in common,
87 # or keep several common interests if already exist
88 # outAct ... Topics attracting much attention will continue get even more attention
89 effects_ml <- includeEffects(effects_ml,cycle4,outAct,name="authortopics")
90
91 # 3. Between-network: mixed triads
92 # from ... indicates researchers with similar interests are most likely having co-authorship
93 effects_ml <- includeEffects(effects_ml,from,name="authorship",interaction1 = "authortopics")
94 effects_ml <- includeEffects(effects_ml,to,name="authortopics",interaction1 = "authorship")
95
96 # 4. Selection mechanisms leading to co-authorship based on start_publication values
97 # egoX ... researchers with higher values (seniors) tend to collaborate more
98 # simX ... researchers with similar values tend to collaborate more
99 # egoXaltX ... interaction between ego x alter that give more collaboration
100 effects_ml <- includeEffects(effects_ml,egoX,name="authorship",interaction1 = "start_publication")
101 effects_ml <- includeEffects(effects_ml,simX,name="authorship",interaction1 = "start_publication")
102 effects_ml <- includeEffects(effects_ml,egoXaltX,name="authorship",interaction1 = "start_publication")
103
104 # 5. Selection mechanisms leading to changing interest based on start_publication values
105 # egoX ... researchers with higher values (seniors) tend to have more varied interest
106 effects_ml <- includeEffects(effects_ml,egoX,name="authortopics",interaction1="start_publication")
107
108 # 6. Selection mechanisms leading to co-authorship based on publishing behavior
109 # egoX ... researchers with higher values (frequent publishing) tend to collaborate more
110 # simX ... researchers with similar publishing habit tend to collaborate more
111 effects_ml <- includeEffects(effects_ml,egoX,name="authorship",interaction1 = "publishing")
112 effects_ml <- includeEffects(effects_ml,simX,name="authorship",interaction1 = "publishing")
113 effects_ml <- includeEffects(effects_ml,egoX,name="authorship",interaction1 = "exploring")
114 effects_ml <- includeEffects(effects_ml,simX,name="authorship",interaction1 = "exploring")
115
116 # 7. Influence mechanisms leading to change in publishing behavior of researchers
117 # avAlt ... attractiveness from average behavior of co-authors make ego prefers to have same behavior
118 effects_ml <- includeEffects(effects_ml,avAlt,name = "publishing",interaction1 = "authorship" )
119 effects_ml <- includeEffects(effects_ml,avAlt,name = "exploring",interaction1 = "authorship" )
120
121 proj_model_ml <- sienaAlgorithmCreate(projname='proj_model_ml', seed=123)
122 results_ml <- siena07(proj_model_ml,data=data_ml, effects=effects_ml,prevAns=results_ml,batch=FALSE,
123   verbose=FALSE,useCluster=TRUE,initC=TRUE,nbrNodes=4)
124 siena.table(results_ml, type="html", tstatPrint=TRUE, sig=TRUE, d=3)

```

RSIENA was used for investigating a number of explanations that cause the network evolution of co-author relations during the observed periods. All probable causes based on graph theory or specifically social network analysis approach have been established in RSIENA. The above scripts were observing some of them.

Effect	par.	(s.e.)	t stat.
Network Dynamics			
constant authorship rate (period 1)	1.995	(0.326)	.
constant authorship rate (period 2)	2.756	(0.616)	.
authorship: degree (density)	-3.349***	(0.653)	-5.132
authorship: transitive triads	2.084***	(0.360)	5.791
authorship: degree of alter	-0.005	(0.048)	-0.114
authorship: start_publication ego	0.226	(0.232)	0.974
authorship: start_publication similarity	2.973**	(1.012)	2.937
authorship: start_publication ego x start_publication alter	-0.816***	(0.246)	-3.321
authorship: publishing ego x publishing alter	0.247	(0.265)	0.930
authorship: exploring ego x exploring alter	-0.734†	(0.437)	-1.681
authorship: from authortopics agreement	-0.043	(0.355)	-0.121
constant authortopics rate (period 1)	29.635	(2.402)	.
constant authortopics rate (period 2)	44.766	(10.880)	.
authortopics: outdegree (density)	-0.555***	(0.023)	-24.610
authortopics: authorship to agreement	0.012	(0.033)	0.357
Behaviour Dynamics			
rate publishing (period 1)	2.007	(0.262)	.
rate publishing (period 2)	2.307	(0.395)	.
publishing linear shape	0.083	(0.108)	0.768
publishing quadratic shape	-0.012	(0.063)	-0.193
publishing average similarity (authorship)	5.715*	(2.638)	2.167
rate exploring (period 1)	4.691	(1.272)	.
rate exploring (period 2)	4.527	(0.985)	.
exploring linear shape	-0.188	(0.164)	-1.145
exploring quadratic shape	0.136	(0.137)	0.997
exploring average similarity (authorship)	7.803	(5.526)	1.412
† p < 0.1; * p < 0.05; ** p < 0.01; *** p < 0.001;			
all convergence t ratios < 0.11.			
Overall maximum convergence ratio 0.2.			

The above estimates were obtained after approximating values that following the defined effects in the RSIENA scripts. Basically, the explanations that cause network evolution of co-author networks with respect to author-topic networks are categorized into the dynamics of network and behavior as illustrated in the estimate

results. Some of interpretations for the estimates related to the issues in this dissertations have been discussed.

Appendix 2. SAMPLE DATA OF AMINER EXPERTS

4	5	scopus id	idx_intel	expert_name	Domain	start_pub	career_age	num_articles							tercatat di scopus 2015	
								1995	2000	2005	2010	2015	h-index	#Docs	citations	
6	6602721887	https://www.scopus.com/authid/detail.uri?authorid=6602721887	1016882	Ellen Riloff	a00	NLP	1991	1	12	12	13	11	8	21	50	2329
7	6602745471	https://www.scopus.com/authid/detail.uri?authorid=6602745471	1102663	Kevin Knight	a01	NLP	1989	2	11	13	29	42	13	24	94	2425
8	6602901918	https://www.scopus.com/authid/detail.uri?authorid=6602901918	1132741	Mark Steedman	a02	NLP	1987	2	18	5	9	9	10	25	82	2699
9	35589184400	https://www.scopus.com/authid/detail.uri?authorid=35589184400	1139547	Ian H. Witten	a03	NLP	1980	3	44	41	38	42	12	47	268	11328
10	6603954639	https://www.scopus.com/authid/detail.uri?authorid=6603954639	1225980	Marti A. Hearst	a05	NLP	1991	1	20	24	29	23	24	33	100	6214
11	56017852800	https://www.scopus.com/authid/detail.uri?authorid=56017852800	1287627	Rosie Jones	a08	IE	1999	1	0	3	10	22	5	17	29	1795
12	7201586314	https://www.scopus.com/authid/detail.uri?authorid=7201586314	1294148	Mitchell P. Marcus	a09	NLP	1975	3	23	6	4	4	5	8	21	221
13	7006427633	https://www.scopus.com/authid/detail.uri?authorid=7006427633	1299391	Mark Craven	a10	IE	1994	1	1	10	13	13	3	23	64	2949
14	7003773569	https://www.scopus.com/authid/detail.uri?authorid=7003773569	1318540	Andrew McCallum	a11	IE	1990	2	2	12	25	58	23	52	139	10931
15	7004146387	https://www.scopus.com/authid/detail.uri?authorid=7004146387	1351785	Jude Shavlik	a13	IE	1985	2	37	15	19	33	15	28	129	3242
16	24604968400	https://www.scopus.com/authid/detail.uri?authorid=24604968400	1415644	Alexander Gelbukh	a14	IE	1999	5	2	12	36	56	32	22	288	1859
17	55408819600	https://www.scopus.com/authid/detail.uri?authorid=55408819600	1453295	Kathleen F. McCoy	a15	NLP	1982	2	12	11	10	29	14	12	56	555
18	16410214900	https://www.scopus.com/authid/detail.uri?authorid=16410214900	1456214	David Eric Yarowsky	a16	NLP	1992	1	6	6	25	12	6	19	39	1185
19	7404170783	https://www.scopus.com/authid/detail.uri?authorid=7404170783	1469484	Richard Schwartz	a17	IE	1977	3	47	11	11	11	3	32	154	5431
20	55484284000	https://www.scopus.com/authid/detail.uri?authorid=55484284000	1480651	Robert C. Moore	a18	NLP	1973	3	34	9	13	12	6	10	25	1159
21	6603264513	https://www.scopus.com/authid/detail.uri?authorid=6603264513	1490750	Eugene Agichtein	a19	IE	2000	1	0	1	7	41	35	30	101	4529
22	7501394856	https://www.scopus.com/authid/detail.uri?authorid=7501394856	1492785	Joongmin Choi	a20	IE	1993	1	1	1	7	16	3	9	36	211
23	8938639100	https://www.scopus.com/authid/detail.uri?authorid=8938639100	1509640	Jerry R. Hobbs	a21	NLP	1977	5	52	3	5	13	9	16	71	2117
24	6602730140	https://www.scopus.com/authid/detail.uri?authorid=6602730140	1602810	Robert Gaizauskas	a24	IE	1993	1	1	18	11	16	7	20	106	1504
25	57035492800	https://www.scopus.com/authid/detail.uri?authorid=57035492800	1621619	David Lewis	a25	NLP	1988	5	16	9	4	5	2	12	106	546
26	6603661388	https://www.scopus.com/authid/detail.uri?authorid=6603661388	1631385	Peter Norvig	a26	NLP	1983	2	16	6	2	5	6	7	23	1354
27	6603035586	https://www.scopus.com/authid/detail.uri?authorid=6603035586	1684021	Christopher K. Riesbeck	a27	NLP	1972	3	13	4	5	4	0	7	23	151
28	7006914090	https://www.scopus.com/authid/detail.uri?authorid=7006914090	180345	Lenhart K. Schubert	a32	NLP	1975	3	27	16	6	8	5	11	48	603
29	7003940794	https://www.scopus.com/authid/detail.uri?authorid=7003940794	229728	Luis Gravano	a34	IE	1991	1	15	19	29	14	8	33	81	4681
30	6603963324	https://www.scopus.com/authid/detail.uri?authorid=6603963324	260885	Kristina Lerman	a36	IE	1990	2	4	12	23	21	25	31	148	3321
31	8957737400	https://www.scopus.com/authid/detail.uri?authorid=8957737400	27202	Fabio Ciravegna	a38	IE	1992	1	4	3	16	18	11	15	124	1623
32	6603651046	https://www.scopus.com/authid/detail.uri?authorid=6603651046	346466	Stephen Soderland	a41	IE	1994	1	4	2	5	12	5	20	39	4025
33	35609548400	https://www.scopus.com/authid/detail.uri?authorid=35609548400	347802	Bonnie J. Dorri	a42	NLP	1987	2	36	16	24	27	5	22	86	2024
34	57190866445	https://www.scopus.com/authid/detail.uri?authorid=57190866445	354019	Roger Schank	a43	NLP	1969	4	61	8	5	14	1	18	66	1576
35	55980465200	https://www.scopus.com/authid/detail.uri?authorid=55980465200	384963	Yorick Wilks	a44	IE	1969	4	55	26	26	17	3	20	141	10931
36	7201536920	https://www.scopus.com/authid/detail.uri?authorid=7201536920	393265	Stanley Peters	a45	NLP	1969	4	4	2	8	11	0	12	39	680
37	6603885504	https://www.scopus.com/authid/detail.uri?authorid=6603885504	419410	Barbara J. Grosz	a49	NLP	1977	3	32	9	13	18	11	17	25	2045
38	7202924370	https://www.scopus.com/authid/detail.uri?authorid=7202924370	504588	William Cohen	a47	IE	1985	2	25	25	15	46	29	39	186	6460
39	12793699800	https://www.scopus.com/authid/detail.uri?authorid=12793699800	532397	Dekai Wu	a52	NLP	1986	2	13	7	13	11	10	11	45	906
40	6602712741	https://www.scopus.com/authid/detail.uri?authorid=6602712741	539646	Philip Stuart Resnik	a53	NLP	1990	2	8	6	25	26	14	21	74	3143
41	7406425278	https://www.scopus.com/authid/detail.uri?authorid=7406425278	544526	James F. Allen	a54	NLP	1975	3	49	13	5	5	7	24	123	8586

Those data were samples of AMiner NLP.IE experts applied for the experiments in this dissertation. Some column values were available in the dataset, but further manually collecting was necessary such as the numbers of published articles and received citations after 2015 since AMiner data is collected by using web harvesting mechanisms.

Appendix 3. WEIGHTS FOR EXPERTISE RANK WITH R PACKAGE DECISIONANALYSIS

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1		F1	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15
2	0.43	0.10	0.10	0.10	0.05	0.05	0.05	0.05	0.05	0.05	0.15	0.10	0.05	0.05	0.05
3	ALT1	avgdist	dyn_artc	dyn_cite	exp_w1	exp_w2	exp_w3	con_w1	con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3
4			F3	F4	F5				F9	F10	F11	F12	F13	F14	F15
5			0.15	0.15	0.10				0.10	0.10	0.15	0.10	0.05	0.05	0.05
6	0.47		dyn_artc	dyn_cite	exp_w1				con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3
7			F3	F4	F5				F9	F10	F11	F12	F13	F14	F15
8			0.15	0.15	0.10				0.15	0.05	0.05	0.10	0.10	0.10	0.05
9	0.52		dyn_artc	dyn_cite	exp_w1				con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3
10			F3	F4	F5				F9			F12	F13	F14	F15
11			0.15	0.15	0.20				0.20			0.10	0.10	0.10	
12	0.46		dyn_artc	dyn_cite	exp_w1				con_w2			artc_num	cite1	cite2	
13			F3	F4	F5				F9	F10	F11	F12	F13	F14	F15
14			0.20	0.20	0.10				0.10	0.05	0.05	0.05	0.10	0.10	0.05
15	0.48		dyn_artc	dyn_cite	exp_w1				con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3
16			F3	F4	F5				F9	F10	F11	F12	F13	F14	F15
17			0.15	0.15	0.10				0.10	0.05	0.05	0.10	0.10	0.10	0.10
18	0.54		dyn_artc	dyn_cite	exp_w1				con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3
19			F3	F4	F5	F6	F7	F8	F9	F10	F11	F12			
20			0.20		0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10		
21	0.45		dyn_artc		exp_w1	exp_w2	exp_w3	con_w1	con_w2	con_w3	start_pub	artc_num			
22			F3		F5	F6	F7	F8	F9	F10	F11	F12			
23			0.20		0.15	0.10	0.10	0.10	0.15	0.05	0.05	0.10			
24	0.43		dyn_artc		exp_w1	exp_w2	exp_w3	con_w1	con_w2	con_w3	start_pub	artc_num			
25			F3		F5				F9	F10	F11	F12			
26			0.25		0.20				0.20	0.10	0.10	0.15			
27	0.43		dyn_artc		exp_w1				con_w2	con_w3	start_pub	artc_num			
28			F1	F3	F5				F9	F10	F11	F12			
29			0.10	0.25	0.15				0.15	0.05	0.15	0.15			
30	0.351	avgdist	dyn_artc		exp_w1				con_w2	con_w3	start_pub	artc_num			
31			F1	F3	F5				F9	F10	F11	F12			
32			0.15	0.20	0.15				0.20	0.05	0.10	0.15			
33	0.348	avgdist	dyn_artc		exp_w1				con_w2	con_w3	start_pub	artc_num			
34			F1	F3	F5				F9	F10	F11	F12	F13	F14	F15
35			0.15	0.15	0.15				0.10	0.10	0.10	0.10	0.05	0.05	0.05
36	0.398	avgdist	dyn_artc		exp_w1				con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3

	A	B	C	D	E	F	G	H	I	J	K	L
1		F1	F3	F5	F6	F7	F8	F9	F10	F11	F12	
2	ALT7	0.15	0.20	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.449
3		avgdist	dyn_artc	exp_w1	exp_w2	exp_w3	con_w1	con_w2	con_w3	start_pub	artc_num	
4	ALT_V2_01		0.050	0.100	0.150	0.150	0.150	0.150	0.150	0.050	0.050	0.471
5	ALT_V2_02		0.100	0.100	0.150	0.150	0.100	0.150	0.150	0.050	0.050	0.494
6	ALT_V2_03		0.050	0.100	0.175	0.150	0.100	0.175	0.150	0.050	0.050	0.500
7	ALT_V2_04		0.050	0.100	0.200	0.150	0.100	0.200	0.150	0.025	0.025	0.498
8	ALT_V2_05		0.050	0.075	0.175	0.175	0.075	0.175	0.175	0.050	0.050	0.534
9	ALT_V2_06		0.050	0.075	0.200	0.175	0.075	0.200	0.175	0.025	0.025	0.532
10	ALT_V2_07		0.050	0.075	0.200	0.150	0.075	0.200	0.150	0.050	0.050	0.524
11	ALT_V2_08		0.100		0.200	0.200		0.200	0.200	0.050	0.050	0.593
12	ALT_V2_09		0.050		0.225	0.200		0.225	0.200	0.050	0.050	0.596
13	ALT_V2_10		0.050		0.225	0.225		0.225	0.225	0.025	0.025	0.601
14	ALT_V2_11		0.050		0.225	0.225		0.225	0.225	0.050		0.583
15	ALT_V2_12	0.050	0.050		0.225	0.200		0.225	0.200	0.025	0.025	0.575

Empirical experiments on Section 7.2 with weighted-sum approach requires the weight values. In those experiments, we heuristically performed stepwise approach as listed in the above tables. Selected combinations of weights with represented results that support the issues in this dissertation were illustrated in Section 7.2.

Appendix 4. SAMPLE DATA FOR EXPERTISE RANK FOR QUERY T2

1	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10	F11	F12	F13	F14	F15	F16	F17	
2			0.30					0.30						0.40				
3	1.00	0.10	0.10	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.10	0.10	0.04	0.04	0.04	0.04	0.04	
4	auth_idx	avgdist	trans	dyn_artc	dyn_cite	exp_w1	exp_w2	exp_w3	con_w1	con_w2	con_w3	start_pub	artc_num	cite1	cite2	cite3	cite4	cite5
5	a00	0.203	0.960	0.001	0.001	2	2	2	2	2	2	1	48	6	22	19	8	1
6	a01	0.161	0.984	0.001	0.001	2	2	3	2	2	3	2	95	15	44	30	18	1
7	a02	0.202	1.000	0.001	0.001	2	1	2	2	2	1	2	41	16	19	11	4	1
8	a03	0.148	0.938	1.000	0.300	3	2	2	1	4	3	3	165	43	75	41	11	7
9	a05	0.158	0.884	0.600	1.000	2	3	3	2	2	2	1	96	29	47	25	15	4
10	a08	0.178	0.960	0.100	0.400	0	1	2	0	0	1	1	35	6	13	15	4	2
11	a09	0.244	0.895	0.100	0.100	2	1	1	2	1	1	3	37	10	17	7	6	2
12	a10	0.155	0.995	0.001	0.001	1	2	2	1	0	2	1	37	10	16	11	3	0
13	a11	0.119	0.961	0.001	0.001	1	2	2	1	2	3	2	97	15	34	46	22	3
14	a13	0.161	0.908	0.300	0.400	2	2	2	2	2	2	2	104	38	57	20	4	0
15	a14	0.131	0.993	0.001	0.001	1	2	3	1	1	3	5	106	66	66	6	0	0
16	a15	0.206	0.914	0.900	0.400	2	2	2	2	2	2	2	62	24	37	15	0	0
17	a16	0.161	1.000	0.001	0.001	1	1	2	1	2	3	1	49	4	24	18	8	1
18	a17	0.171	0.942	0.400	0.500	3	1	2	1	2	2	3	80	22	41	17	2	1
19	a18	0.185	0.905	0.001	0.001	3	2	2	1	2	2	3	68	20	26	24	3	1
20	a19	0.185	0.830	0.100	0.100	0	1	2	0	0	1	1	49	23	31	23	4	3
21	a20	0.172	0.866	0.100	0.200	1	1	2	1	0	0	1	25	14	14	0	0	0
22	a21	0.200	0.975	0.001	0.001	3	1	2	1	1	1	5	73	28	32	16	6	0
23	a24	0.181	0.988	0.001	0.001	1	3	1	1	0	2	1	46	13	28	11	1	0
24	a25	0.159	0.900	1.000	0.400	2	1	1	2	2	1	5	34	3	14	10	6	3
25	a26	0.202	0.821	0.500	0.100	2	2	1	2	1	1	2	29	14	12	5	1	3
26	a27	0.245	0.785	0.200	0.100	2	1	1	2	1	2	3	26	13	11	1	1	0
27	a32	0.140	0.984	0.001	0.001	3	2	1	1	3	2	3	57	18	28	16	0	0
28	a34	0.156	0.876	0.300	0.100	2	2	2	2	2	3	1	77	16	23	28	17	1
29	a36	0.206	0.831	1.000	0.400	2	3	3	2	0	2	2	60	23	37	24	1	0
30	a38	0.163	1.000	0.001	0.001	2	1	2	2	1	2	1	41	12	33	3	4	0
31	a41	0.208	0.882	0.001	0.001	1	0	2	1	1	0	1	23	2	10	8	6	2
32	a42	0.149	0.946	0.200	0.500	2	1	3	2	3	3	2	103	32	55	20	1	0
33	a43	0.161	0.925	0.200	0.100	3	0	0	1	2	2	4	88	39	39	9	3	0
34	a44	0.156	0.944	0.900	0.200	3	1	3	1	3	3	4	124	44	66	17	1	0
35	a45	0.253	0.704	0.200	0.300	2	1	2	2	0	1	4	25	8	15	2	0	0
36	a47	0.206	0.889	0.900	0.300	2	2	2	2	2	1	3	72	30	35	13	2	3
37	a49	0.147	0.874	0.100	0.100	2	2	1	2	3	3	2	111	35	57	33	14	1
38	a52	0.227	0.906	0.001	0.001	2	1	1	2	2	2	2	44	16	24	11	2	1
39	a53	0.214	0.812	0.100	0.100	2	2	3	2	0	2	2	65	10	38	22	8	1
40	a54	0.168	0.908	0.200	0.200	3	2	1	1	2	2	3	72	29	32	13	3	2
41	a55	0.194	0.856	0.300	0.100	2	2	3	2	2	1	2	72	32	36	16	2	0

The above values listed some examples from our experiments data with 17 extracted features for author a00-a55 (auth_idx column for author index) in Chapter 7. The values in ...

- columns of avgdist and trans were extracted with procedures in Section 7.1.1
- columns of dyn_artc and dyn_cite were extracted with procedures in Section 5.4
- columns of exp_w1 ... exp_w3 and con_w1 ... con_w3 were extracted with procedures in Section 6.2
- columns of cite1 ... cite5 were extracted with procedures in Section 7.1.2

The feature value of `start_pub` was exist in AMiner dataset with an assumption as mentioned in Section 6.3. Then, the last feature value of `artc_num` was collected manually for all experts used in the experiments through Scopus data.

Appendix 5. SAMPLE RESULTS OF EXPERTISE RANK

	T2	Rank	T4	Rank	T6	Rank	T10	Rank	T13	Rank	T21	Rank	T29	Rank
1	a58	0.844	a58	0.794	a58	0.844	a58	0.817	a58	0.794	a58	0.840	a58	0.813
2	a59	0.741	a59	0.740	a59	0.729	a03	0.717	a03	0.726	a62	0.738	a59	0.726
3	a62	0.738	a03	0.724	a03	0.726	a59	0.708	a62	0.715	a03	0.732	a03	0.725
4	a03	0.736	a05	0.703	a62	0.688	a62	0.688	a59	0.711	a05	0.695	a62	0.707
5	a05	0.695	a62	0.688	a05	0.672	a01	0.681	a05	0.699	a59	0.695	a05	0.699
6	a44	0.674	a01	0.672	a01	0.669	a44	0.667	a56	0.676	a56	0.670	a56	0.680
7	a56	0.669	a42	0.661	a56	0.657	a05	0.656	a01	0.662	a42	0.639	a01	0.635
8	a01	0.635	a14	0.642	a44	0.653	a56	0.653	a42	0.655	a01	0.635	a14	0.635
9	a42	0.630	a56	0.630	a42	0.628	a14	0.632	a44	0.642	a44	0.626	a44	0.626
10	a61	0.608	a44	0.626	a11	0.620	a42	0.625	a49	0.619	a34	0.613	a11	0.623
11	a34	0.592	a61	0.620	a49	0.618	a49	0.619	a14	0.618	a11	0.607	a49	0.618
12	a14	0.592	a49	0.605	a61	0.617	a61	0.613	a11	0.596	a14	0.592	a34	0.618
13	a49	0.585	a11	0.580	a14	0.607	a11	0.611	a61	0.590	a61	0.590	a42	0.612
14	a11	0.580	a34	0.568	a34	0.568	a66	0.573	a13	0.568	a49	0.575	a61	0.590

h-index		num tpcs	start_pub	#Docs	citations
24	a01	8	1989	94	2,425
47	a03	6	1980	268	11,328
33	a05	7	1991	100	6,214
52	a11	7	1990	139	10,931
22	a14	5	1999	288	1,859
33	a34	2	1991	81	4,681
22	a42	3	1987	86	2,024
20	a44	1	1969	141	10,931
39	a49	5	1985	186	6,460
61	a56	8	1978	282	16,672
34	a58	3	1989	208	4,290
33	a59	4	1984	205	3,488
21	a61	1	1970	88	1,834
36	a62	3	1970	93	5,412

Those results were obtained from weighted-sum approach on the combination of feature.

$$\begin{aligned}
 &rank_{wsm}(a_i, c_k)_{filter\ a_i\ with\ c_k} \\
 &= 0.05g(F3_{a_i}) + 0.225g(F6_{a_i}) + 0.225g(F7_{a_i}) + 0.225g(F9_{a_i}) \\
 &+ 0.225g(F10_{a_i}) + 0.025g(F11_{a_i}) + 0.025g(F12_{a_i})
 \end{aligned}$$

Those results were repeated on some selected topics of T2, T4, T6, T10, T13, T21 and T29 to find out the experts on certain subjects. Most of the rank results

showed the same set of experts with different positions, which indicated that the researchers could have different focus.

Appendix 6. MATHEMATICAL FUNCTIONS FOR EVALUATING NETWORK EVOLUTION

RSIENA as a computer program uses estimation techniques following a Markov process, which assumes future changes in a network state are based on the current state of the complete network. The estimation includes repeated simulation measures of social networks according to SAOM and tests the parameters to produce a probabilistic network evolution that brought the observations from each wave to the next. The changes of networks and behavior of nodes are in small steps, which means a change occurs in only one tie value or one behavioral variable. Behaviors of actors, or researchers in this case, affect the network structure of co-author networks and author-topic networks. Then, the network structure also has the possibility to affect the behavior values.

Let the initial network in the first wave is denoted as x^0 . The evaluation function for author i on a network in the next wave x is denoted $f_i(x)$. Then, the probability for the occurrence of the next network x is given by a function that contains some exponential functions $\exp()$ of $p(x^0, x) = \frac{\exp(f_i(x) - f_i(x^0))}{\sum_{x' \in C} \exp(f_i(x') - f_i(x^0))}$

where C denotes a set of all possible networks that can be obtained as a result.

Each evaluation function $f_i(x)$ for researcher a_i is defined as $f_i(x) = \sum_k \beta_k s_{ik}(x)$ with the value β_k as the estimate and $s_{ik}(x)$ is an effect function for all specified k effects.

1. An evaluation function for co-author selection that considers career-age $f_i^{sel}(x, cov_{start.pub})$ has the effects of ego $s_{i1}(x)$, alter $s_{i2}(x)$, similarity $s_{i3}(x)$, ego \times alter $s_{i4}(x)$. There are different combinations of effects specified for different purposes that should be analyzed according to each hypothesized assumption.

To appraise the possibility f_i^{sel} of an ego researcher a_i who has the career-age v_i to work with an alter a_j with the value v_j , the considerations are on:

- the weight of ego $\beta_{ego}(v_i - \bar{v})$ and the weight of alter $\beta_{alter}(v_j - \bar{v})$ which using the same estimate to the ego because of the reciprocity between ego-alter
- the weight of similarity between ego-alter $\beta_{sim} \left(1 - \frac{|v_i - v_j|}{\Delta_v} - \widehat{sim}^v \right)$
- the weight of interaction between ego-alter $\beta_{e \times a}(v_i - \bar{v})(v_j - \bar{v})$

Thus, the evaluation function is defined as

$$f_i^{sel}(x, cov_{start.pub}) = \beta_{ego}(v_i - \bar{v}) + \beta_{ego}(v_j - \bar{v}) + \beta_{sim} \left(1 - \frac{|v_i - v_j|}{\Delta_v} - \widehat{sim}^v \right) + \beta_{e \times a}(v_i - \bar{v})(v_j - \bar{v})$$

2. The same approaches applies to evaluation functions for publishing and exploring behaviors that consider ego \times alter interaction.

$$f_i^{sel}(x, beh_{pub}) \text{ and } f_i^{sel}(x, beh_{exp})$$

3. Some effects for network changes are about transTriads s_{i9}^{net} , inPop s_{i24}^{net} , and cycle4 s_{i11}^{net} , which codified according to RSIENA.

$$s_{i9}^{net}(x) = \sum_{j,h} x_{ij}x_{jh}x_{hi}$$

The “transTriads” effect represents the tendency to co-author with researchers who are mutually linked or indirectly tied because of previous collaborations.

$$s_{i24}^{net}(x) = \sum_j x_{ij} \left(\sum_{h \neq i} x_{hj} + 1 \right)$$

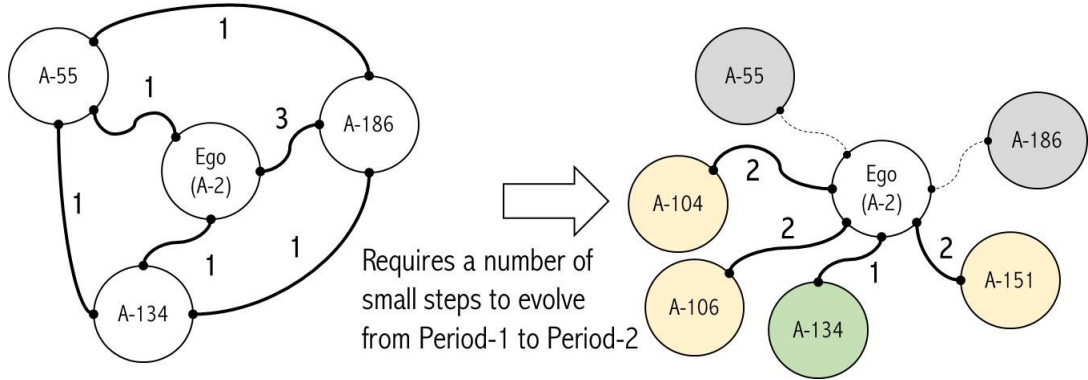
The “inPop” effect represents a situation where popular researchers tend to collaborate more.

$$s_{i11}^{net}(x) = \frac{1}{4} \sum_{j,k,h; all\ different} x_{ij}x_{ik}x_{hj}x_{hk}$$

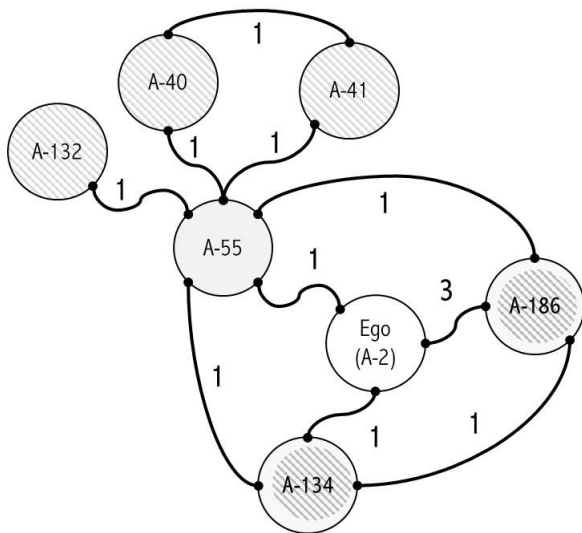
The “cycle4” effect represents a situation where if a pair of researchers has one topic in common, they will get more topics in common. Thus the network x is a two-mode (author-topic) network with author nodes of a_i and a_h , in addition

to topic nodes of c_j and c_k . If there is relations x_{ij} and x_{ik} (means that a_i has interest on c_j and c_k), then there is a likelihood for author a_h who prefers c_j is going to take interest on c_k as well.

Thus $f_i(x) = \beta_{i9}^{net} s_{i9}^{net} + \beta_{i24}^{net} s_{i24}^{net} + \beta_{i11}^{net} s_{i11}^{net}$ among other specified effects in Table 6-3.



In the case of an ego that connected to three co-authors of a_{55} (in one article), a_{134} (in one article), and a_{186} (in three articles) in the Period-1, the estimation procedures observed all possibilities of changes with a number of small steps so the network evolved into the right network in Period-2.



More complete network related to those co-authors in Period-1 showed that a_{55} had other co-authors: a_{40} , a_{41} , and a_{132} . The ego network of a_2 indicated a paper authored by $a_2 a_{55} a_{134} a_{186}$ and two more papers by $a_2 a_{186}$. It seemed that the relations to a_{55} and a_{186} had higher chance to be maintained.

Assuming the evaluation function only considers the network effects, then

$$f_i(x) = \beta_{i.degree}^{net} \sum_j x_{ij} x_{ji} + \beta_{i.transTriad}^{net} \sum_{j,h} x_{ij} x_{jh} x_{hi}$$

With the estimates in Table 6-4, $\beta_{i.degree}^{net} = -3.349$ and $\beta_{i.transTriad}^{net} = 2.084$, the change probabilities a small step for the ego network a_2 are:

1. No change $-3.349 \times 3 + 2.084 \times 3 = -10.05 + 6.25 = -3.80$
2. Drop a_{55} or a_{134} or a_{86} $-3.349 \times 2 + 2.084 \times 1 = -6.70 + 2.08 = -4.62$
3. Add a_{104} $-3.349 \times 4 + 2.084 \times 3 = -13.40 + 6.25 = -7.15$

Given the current state of the network and that evaluation function, ego is most likely to have no change, because that decision maximizes the objective function. However, there are other effects that influence the final objective function, such as the network and behavior effects in Table 6-3 or Appendix 1.

4. Each state of a network is computed to get the likelihood value based on random selections from any probable networks depends on the previous state of the network, which following Markov process.

As an illustration, there are three possible states of a researcher in terms of publishing article. For each year, the researcher does not publish any article (st_N), the researcher writes a draft article but not submitting (st_D), and the researcher submits the draft (st_S).

After observing a period, i.e. 3-5 years, some probability values for state changes of researchers in publishing are:

- If the researcher does not submit any article in current year, there is 25% chance for not publishing in the next year, 50% chance to write a draft, and 25% to submit, $trans - st_N = [0.25 \quad 0.50 \quad 0.25]$.
- If the researcher writes a draft in current year, there is 50% chance still writing a draft in the next year, and 50% to finally submit the draft, $trans - st_D = [0.00 \quad 0.50 \quad 0.50]$.
- If the researcher has already submitted the draft in current year, there is 33% chance for taking a break, 33% chance to only write a draft, and 34% to submit again in the next year, $trans - st_S = [0.33 \quad 0.33 \quad 0.34]$.

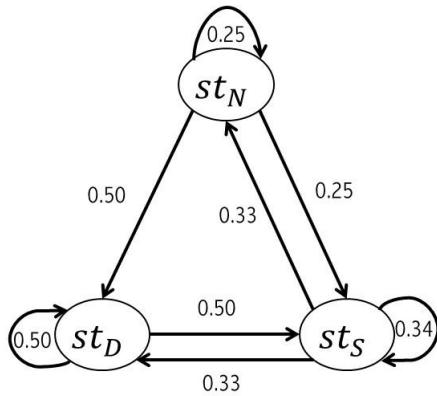
Thus, the publishing transition matrix of a researcher between current year to the next

$$\text{year is } p(y_i, y_{i+1}) = \begin{bmatrix} 0.25 & 0.50 & 0.25 \\ 0.00 & 0.50 & 0.50 \\ 0.33 & 0.33 & 0.34 \end{bmatrix}.$$

Assume that an author a_i has been observed based on the log of publications including writing and submitting processes. The probabilities of $[st_N \ st_D \ st_S]$ for current year $q_0 = [0.0 \ 0.5 \ 0.5]$. The probabilities for the next year q_1 is

$$q_0 p(y_i, y_{i+1}) = [0.0 \ 0.5 \ 0.5] \begin{bmatrix} 0.25 & 0.50 & 0.25 \\ 0.00 & 0.50 & 0.50 \\ 0.33 & 0.33 & 0.34 \end{bmatrix} = [0.165 \ 0.415 \ 0.420]$$

Thus, given that a researcher writes a draft or submits the draft, the possibilities to do at least one of those activities decreases because the researcher may take a break.



Another illustration case: given that a researcher submits an article in current year, in average how many resting years before the researcher submits again?

Based on Markov process, this case requires a computation for $m(st_S, st'_S)$ or mean time to go from state st_S to st_S again.

Calculating $m(st_S, st'_S) = 1 + m(st_N, st_S) \times p_{S,N} + m(st_D, st_S) \times p_{S,D}$ requires:

- $m(st_N, st_S)_t = 1 + m(st_N, st_S)_{t-1} \times p_{N,N} + m(st_D, st_S)_{t-1} \times p_{N,D}$
- $m(st_D, st_S)_t = 1 + m(st_D, st_S)_{t-1} \times p_{D,D}$

Assume as initial values, $m(st_N, st_S)_0 = 0.50$ and $m(st_D, st_S)_0 = 0.50$

Calc.	$m(st_N, st_S)_t$	$m(st_D, st_S)_t$	Error, $thres_{0.01}$
t_1	$1 + 0.50 \times 0.25 + 0.50 \times 0.50 = 1.375$	$1 + 0.50 \times 0.5 = 1.250$	
t_2	$1 + 1.38 \times 0.25 + 1.25 \times 0.50 = 1.969$	$1 + 1.25 \times 0.5 = 0.625$	$1.97 - 1.38 = 0.59$
...
t_{11}	2.66	2.00	

Thus, the mean time value $m(st_S, st'_S) = 1 + 2.66 \times 0.33 + 2.00 \times 0.33 = 2.54$ years. For estimating the network evolution, the transition matrices are not available. The transition values are computed from the objective function $f_i(x)$ for each ego network with a possible state of a network is drawn from numerous possibilities.