การเพิ่มประสิทธิภาพผลการค้นหาโดยการสร้างดัชนีร่วมและการจัดอันดับใหม่แบบปัจเจกบุคคล
สำหรับระบบโซเชียลบุกมาร์กเชิงบรรณานุกรม

นางสาวพิจิตรา จอมศรี

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรดุษฎีบัณฑิต
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยีสารสนเทศ
ภาควิชาคณิตศาสตร์และวิทยาการคอมพิวเตอร์
คณะวิทยาศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2555

ENHANCING THE EFFICIENCY OF SEARCH RESULT USING INTEGRATED INDEXING AND PERSONALIZED RE-RANKING FOR BIBLIOGRAPHIC SOCIAL BOOKMARKING SYSTEMS

Miss Pijitra Jomsri

A Dissertation Submitted in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
Program in Computer Science and Information Technology
Department of Mathematics and Computer Science
Faculty of Science
Chulalongkorn University
Academic Year 2012

Thesis Title	ENHANCING THE EFFICIENCY OF SEARCH RESULT USING INTEGRATED INDEXING AND PERSONALIZED RE-RANKING FOR BIBLIOGRAPHIC SOCIAL BOOKMARKING SYSTEMS
By	Miss Pijitra Jomsri
Field of Study	Computer Science and Information Technology
Thesis Advisor	Associate Professor Peraphon Sophatsathit, Ph.D.
Thesis Co-advisor	Assistant Professor Worasit Choochaiwattana, Ph.D.

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

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

THESIS COMMITTEE

……………………………………………….. Chairman
(Professor Chidchanok Lursinsap, Ph.D.)

……………………………………………….. Thesis Advisor
(Associate Professor Peraphon Sophatsathit, Ph.D.)

……………………………………………….. Thesis Co-advisor
(Assistant Professor Worasit Choochaiwattana, Ph.D.)

……………………………………………….. Examiner
(Assistant Professor Suphakant Phimoltares, Ph.D.)

……………………………………………….. Examiner
(Boonyarit Intiyot, Ph.D.)

……………………………………………….. External Examiner
(Marut Buranarach, Ph.D.)

พิจิตรา จอมศรี : การเพิ่มประสิทธิภาพผลการค้นหาโดยการสร้างดัชนีร่วมและการจัดอันดับใหม่แบบปัจเจกบุคคลสำหรับระบบโซเชียลบุกมาร์กเชิงบรรณานุกรม. (ENHANCING THE EFFICIENCY OF SEARCH RESULT USING INTEGRATED INDEXING AND PERSONALIZED RE-RANKING FOR BIBLIOGRAPHIC SOCIAL BOOKMARKING SYSTEMS)อ.ที่ปรึกษาวิทยานิพนธ์หลัก:รศ.ดร.พีระพนธ์ โสพัศสถิตย์, อ. ที่ปรึกษาวิทยานิพนธ์ร่วม: ผศ.ดร.วรสิทธิ์ ชูชัยวัฒนา, 78 หน้า.

ปัจจุบันการสืบค้นข้อมูล และการใช้ประโยชน์จากระบบเครือข่ายสังคมกลายเป็นส่วนหนึ่งของชีวิตประจำวัน ระบบโซเชียลบุกมาร์กเชิงบรรณานุกรมเป็นตัวอย่างหนึ่งของระบบเครือข่ายสังคมทางวิชาการ ผู้วิจัยตระหนักถึงความสำคัญของการค้นหางานวิจัยผ่านระบบดังกล่าว จึงได้นำเสนอเทคนิคการสร้างดัชนีร่วมและการจัดอันดับใหม่แบบปัจเจกบุคคลสำหรับระบบโซเชียลบุกมาร์กเชิงบรรณานุกรมขึ้น

การทดลองครั้งนี้ได้ทำการสร้างดัชนีจำนวนสามชนิด จากผลการทดลองพบว่าการสร้างดัชนีและโพรไฟล์โดยใช้แท็กอย่างเดียวไม่เพียงพอ การสร้างดัชนีที่เหมาะสมที่สุดสำหรับระบบโซเชียลบุกมาร์กเชิงบรรณานุกรมคือการนำโซเชียลแท็กรวมกับชื่องานวิจัยและบทคัดย่อ (*TTA*) และเทคนิคดังกล่าวยังเหมาะสมที่จะนำมาสร้างโพรไฟล์ของผู้ใช้ เพื่อใช้ในกระบวนการจัดอันดับใหม่แบบปัจเจกบุคคล นอกจากนี้การนำปัจจัยที่สำคัญ คือปีที่งานวิจัยตีพิมพ์ เนื่องจากนักวิจัยส่วนใหญ่นิยมอ่านงานวิจัยที่ถูกตีพิมพ์ในช่วงปีปัจจุบันมาสนับสนุนการจัดอันดับ ทั้งนี้การประมวลกำหนดให้ผู้ทดลองทำการกำหนดคำค้นหาเอง จากผลการทดลองพบว่าการนำปีที่งานวิจัยตีพิมพ์รวมกับการจัดอันดับใหม่แบบปัจเจกบุคคลโดยใช้โซเชียลแท็ก ชื่องานวิจัยและบทคัดย่อมาสร้างโพร์ไฟล์ (***PTTAYRank***) ที่อัตราส่วน 90:10 ถือเป็นเทคนิคที่ดีที่สุดสำหรับการค้นหาเชิงบรรณานุกรม และสามารถนำมาเพิ่มประสิทธิภาพการทำงานของระบบโซเชียลบุกมาร์กเชิงบรรณานุกรมสำหรับผู้ใช้แต่ละคนได้

ภาควิชา คณิตศาสตร์และวิทยาการคอมพิวเตอร์ ลายมือชื่อนิสิต..
สาขาวิชาวิทยาการคอมพิวเตอร์และเทคโนโลยี ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก..........
สารสนเทศ..................................... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์ร่วม..........
ปีการศึกษา 2555.....................................

5173836923 : MAJOR COMPUTER SCIENCE AND INFORMATION TECHNOLOGY
KEYWORDS : PERSONALIZED RE-RANKING / INDEXING / BIBLIOGRAPHIC SOCIAL BOOKMARKING / USER PROFILE

PIJITRA JOMSRI : ENHANCING THE EFFICIENCY OF SEARCH RESULT USING INTEGRATED INDEXING AND PERSONALIZED RE-RANKING FOR BIBLIOGRAPHIC SOCIAL BOOKMARKING SYSTEMS . ADVISOR : ASSOC. PROF. PERAPHON SOPHATSATHIT, Ph.D., CO-ADVISOR : ASST. PROF. WORASIT CHOOCHAIWATTANA, Ph.D., 78 pp.

Currently, searching and utilizing social networking systems are becoming a part of daily life. Bibliographic social bookmarking is an example of academic social networking systems. Researchers realize the importance of searching for papers through bibliographic social bookmarking systems, therefore, the integrated indexing and personalized re-ranking for bibliographic social bookmaking systems were created in this dissertation.

Three indexing types were developed as a part of research experiment. The results showed that only tagging information is inadequate to create efficient indexing and user profile. The integrated indexing making up with social tagging, title, and abstract (TTA) was a more suitable combination. This technique can be applied to buliding user profile and personalized re-ranking. In addition, the year of publication factor can further support the information about recent papers and improve ranking. Evaluation of the proposed technique was carried out with user-formulated queries. The result found that the year of publication and personalized re-ranking ($PTTAYRank$) at 90:10 weighted score provided the best personalized re-ranking and could enhance the performance efficiency of the bibliographic social bookmaking systems.

Department : Mathematics and Computer Science Student's Signature

Field of Study: Computer Science and Information Technology Advisor's Signature

Co-advisor's Signature

Academic Year : 2012

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Associate Professor Dr. Peraphon Sophatsathit and Assistant Professor Dr. Worasit Choochaiwattana who have provided me with a great opportunity to do my research work under their guidance. Their suggestions and comments have motivated me during my Ph.D. study.

I would like to thank dissertation committees, Professor Dr. Chidchanok Lursinsap, Assistant Professor Dr. Suphakant Phimoltares, Assistant Professor Dr. Boonyarit Intiyot, and Dr. Marut Buranarach, for their advices and guidance about research activities. Much of untold appreciations also go to those unaccounted individuals who have lent a helping hand to support this work.

I would like to thank the Suan Sunandha Rajabhat University for providing an academic scholarship that gave me an opportunity to pursue the Ph.D. degree.

Last but not least, I would like to express my sincere gratitude and deep appreciation to my parents, my family, and my friends for constant encouragement, love, and supports throughout my life.

CONTENTS

Page

List of Tables

List of Figures

CHAPTER I

INTRODUCTION

The World Wide Web has expanded rapidly. Web 1.0 is the first generation of the web technology which is static and enables users to have instantaneous access to a large diversity of knowledge items. The second generation of the web technology is Web 2.0 that is dynamic. The principal technology is a defined intersection of web application features that simplify information sharing, user-centered design, interoperability, and collaboration on the website. In addition, Web 2.0 allows users to collaborate and interact with one another. Users in a social media system can create contents in a virtual community, namely, social networking sites, social bookmarking sites, web personalization, and folksonomies. The bibliographic social bookmarking system which is a huge web archive of research papers portrays a typical example of hard to retrieve the desired search results. Thus, some forms of efficient management systems must be devised to assist users in searching and retrieval.

This dissertation deals with indexing and personalized re-ranking with bibliographic social bookmarking. The aim of this dissertation is to enhance search results of research papers with the inclusion of personalized re-ranking criterion and to design mechanisms to solve the search problem. Details on problem identification and motivation, research objectives, scope of the study, definitions of the research terminology, and expected outcomes will be elucidated subsequently.

1.1 Problem Identification and Motivation

At present, trends of using the internet to look for information are increasing, especially, searching which employs search engines and social bookmarking. These can facilitate researchers to explore various related areas of research. Searching on social bookmarking systems is increasingly popular which allow users to share content with one another. However, search results obtained from bibliographic social bookmarking systems are not relevant for user query. The reason may be because most of bibliographic social bookmarking systems use tags only for representing academic paper contents such as indexing and user profile. This can imply that using tag alone may not be sufficient to represent such voluminous information. Consequently, it is interesting to investigate how well a set of tags used as paper indices that link to academic papers on bibliographic social bookmarking can contribute to search results.

This research tries to improve search results for bibliographic social bookmarking by creating user profiles and applying other factors to adjust search results ranking. The main focus of research involves three challenges. First, the use of social tagging to get potential indexing of bibliographic searching is investigated and three heuristic indexing methods are developed: tagging information only indexing method (T), title and abstract indexing method (TA), and tagging with title and abstract (TTA), respectively. Second, combining the query-independent ranking or static ranking of bibliographic social bookmarking system such as priority of paper, posted paper timestamp, year of publication, and number of groups containing the posted paper, with query dependent ranking for more relevant search results. Finally, personalized re-ranking using user profile is created. Each user profile is built from the information being obtained from posted papers such as tag, title, abstract, etc. The system makes all necessary adjustment of ranking search results by matching the terms in the user profile with those in each document search results. Moreover, the performance of the system can be improved by additional factors that may affect the user relevancy.

This research will further enhance the performance of similarity ranking based on potential index of personalized re-ranking with the native bibliographic social bookmarking. Discovering how to improve the competency of these bibliographic searching will help researchers develop bibliographic social bookmarking that meets the users' requirements.

1.2 Research Objectives

The objectives are as follows:

1. To improve relevancy of search results of social bibliographic searching.
2. To design and develop the potential indexing scheme for social bibliographic searching.
3. To enhance the effectiveness of search results for personalized re-ranking.

1.3 Scope of the Study

The scope of work encompasses the following areas:

1. This work concentrates on social bibliographic bookmarking.
2. The data set are collected from CiteULike.
3. The subjects for this experiment are experts in the field of information technology, computer science, and related disciplines.
4. The issue on polysemy is discarded.

1.4 Problem Statements

Problem1: How can one create the potential indexing for research paper searching on bibliographic social bookmarking systems?

Problem2: How can one combine the potential index with query independent ranking?

Problem3: How can one discover the tasks to enhance the effectiveness of personalized re-ranking for research paper obtained from bibliographic social bookmarking systems?

1.5 Research Contribution

The expected outcomes will be as follows:

1. An effective indexing scheme for research paper searching.
2. A framework for personalized re-ranking.

3. Enhancement of the performance efficiency for the personalized re-ranking.

1.6 Related Definitions

Indexing is the process of creating indexes for record collections which is a list of words and associated pointers to where useful material relating to that heading can be found in a document [1].

Static ranking (Query-independent ranking or static features) is one of ranking technique that measure the quality of document. This ranking is independent from the user query and may be used to compute document's static quality score [2-3].

Similarity ranking (Query-dependent ranking or dynamic features) is one of ranking technique that measures a match between user query and document indexing. This ranking is dependent from the user query such as TF-IDF score.

Personalized re-ranking is a one ranking technique that creates user profiles in a second step, after evaluating the corpus ranked through non-personalized scores [4].

1.7 Organization of the Dissertation

The rest of this dissertation is organized in five sections as follows. Chapter 2 recounts some the related work and basic knowledge such as information retriev5al process, top-K retrieval, relevance feedback. Chapter 3 discusses the research processes encompassing problem formulation in mathematical terms, namely, objective function and constraints. Chapter 4 summarizes the experiment and results. Chapter 5 concludes the study and suggests appropriate future work.

CHAPTER II

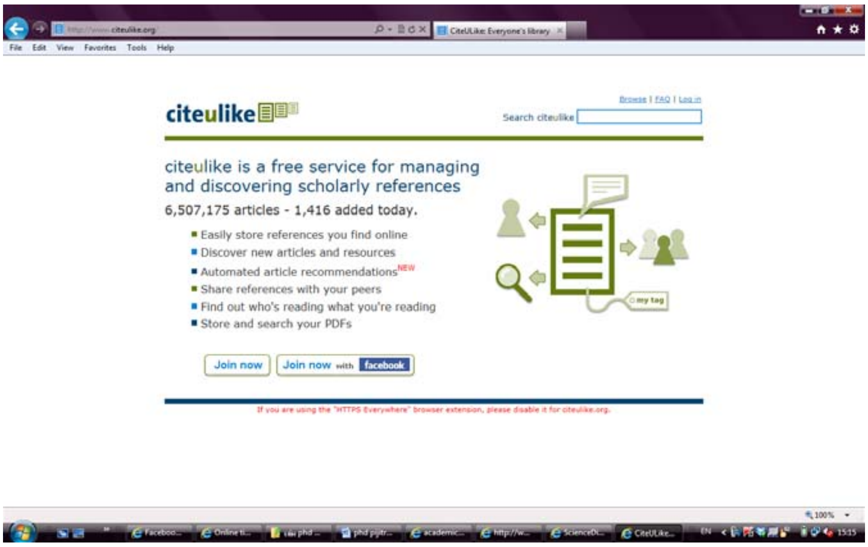
LITERATURE REVIEW

This chapter provides literature review on types of bibliographic social bookmarking systems, personalization ranking techniques, and some basic knowledge in this research area such as information retrieval process, Top-K retrieval, and evaluation methods.

2.1 Related Literature

The attention in online web sharing applications has been increasing with the growth of web 2.0. Social bookmarking might have the most potential as web 2.0 tools can be utilized in academic papers to benefit their users and improve their services. Several services focus on document sharing services. CiteULike [5] is a social bookmarking that helps researchers, scientists, and academic members store, share, get links to academic papers, and organize. Connotea [6] is a discharge online reference management for all scientists and researchers. BibSonomy [7] is a web application system for sharing bookmarks based on publication entries at the same time. Those systems allow users to post research paper that are of interest.

CiteULike is an integrated traditional bibliographic management tool and web-based social bookmarking service. This application has a flexible filing system based on tags [8]. These tags can produce interesting new categories that provide a quick, open, and user-defined classification model. CiteULike is a discharge web service since November 2004. Like many successful application tools, as of November 30, there are approximately 6,485,760 articles, 2012. Figure 2.1 and 2.2 show the interface of CiteULike.


http://www.citeulike.org/
CiteULike: Everyone's library
File Edit View Favorites Tools Help
Browse | FAQ | Log in
Search citeulike
my tag
Join now
Join now with facebook
100%

citeulike

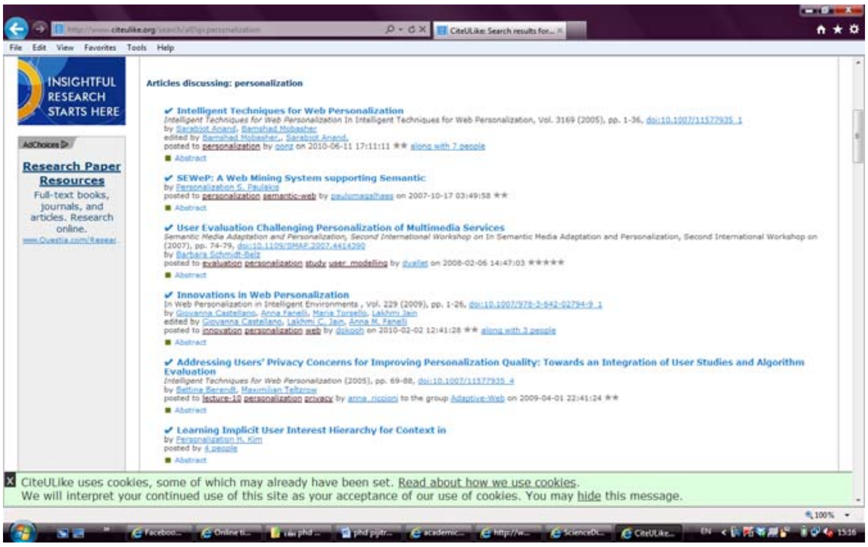
citeulike is a free service for managing and discovering scholarly references

6,507,175 articles - 1,416 added today.

- Easily store references you find online
- Discover new articles and resources
- Automated article recommendationsNEW
- Share references with your peers
- Find out who's reading what you're reading
- Store and search your PDFs

If you are using the "HTTPS Everywhere" browser extension, please disable it for citeulike.org.

Figure 2.1 : The CiteULike homepage.


http://www.citeulike.org/search/all?q=personalization
CiteULike: Search results for...
File Edit View Favorites Tools Help
INSIGHTFUL RESEARCH STARTS HERE
AdChoices
100%

Research Paper Resources
Full-text books, journals, and articles. Research online.
www.Questia.com/Resear

Articles discussing: personalization

✓ Intelligent Techniques for Web Personalization
Intelligent Techniques for Web Personalization In Intelligent Techniques for Web Personalization, Vol. 3169 (2005), pp. 1-36, doi:10.1007/11577935_1
by Sarabjot Anand, Bamshad Mobasher
edited by Bamshad Mobasher., Sarabjot Anand.
posted to personalization by gonz on 2010-06-11 17:11:11 ★★ along with 7 people
Abstract

✓ SEWeP: A Web Mining System supporting Semantic
by Personalization S. Paulakis
posted to personalization semantic-web by paulomagalhaes on 2007-10-17 03:49:58 ★★
Abstract

✓ User Evaluation Challenging Personalization of Multimedia Services
Semantic Media Adaptation and Personalization, Second International Workshop on In Semantic Media Adaptation and Personalization, Second International Workshop on (2007), pp. 74-79, doi:10.1109/SMAP.2007.4414290
by Barbara Schmidt-Belz
posted to evaluation personalization study user_modelling by dvallet on 2008-02-06 14:47:03 ★★★★★
Abstract

✓ Innovations in Web Personalization
In Web Personalization in Intelligent Environments , Vol. 229 (2009), pp. 1-26, doi:10.1007/978-3-642-02794-9_1
by Giovanna Castellano, Anna Fanelli, Maria Torsello, Lakhmi Jain
edited by Giovanna Castellano, Lakhmi C. Jain, Anna M. Fanelli
posted to innovation personalization web by dokooh on 2010-02-02 12:41:28 ★★ along with 3 people
Abstract

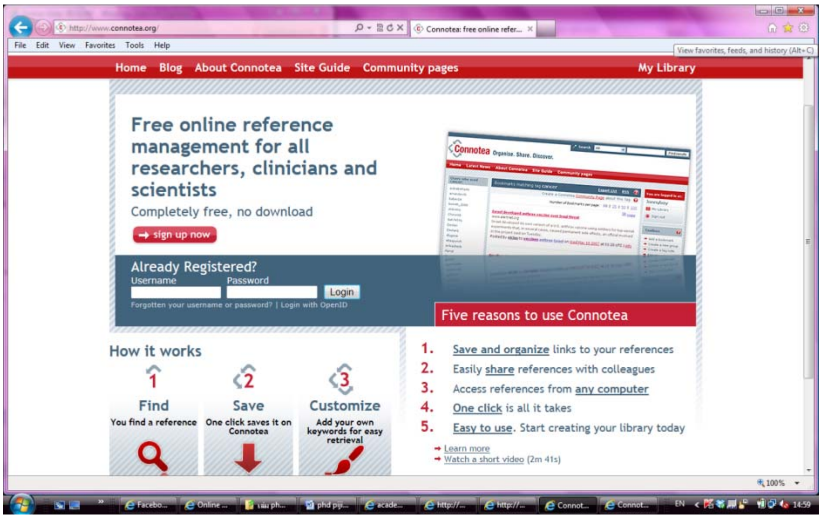
✓ Addressing Users' Privacy Concerns for Improving Personalization Quality: Towards an Integration of User Studies and Algorithm Evaluation
Intelligent Techniques for Web Personalization (2005), pp. 69-88, doi:10.1007/11577935_4
by Bettina Berendt, Maximilian Teltzrow
posted to lecture-10 personalization privacy by anna_riccioni to the group Adaptive-Web on 2009-04-01 22:41:24 ★★
Abstract

✓ Learning Implicit User Interest Hierarchy for Context in
by Personalization H. Kim
posted by 4 people
Abstract

CiteULike uses cookies, some of which may already have been set. Read about how we use cookies.
We will interpret your continued use of this site as your acceptance of our use of cookies. You may hide this message.

Figure 2.2 : The CiteULike search results.

Connotea is a discharge online bibliographic for scientists and clinicians [6] and is one of social bookmarking tools, similar to BibSonomy and CiteULike, where users can save links to their favorite websites. Connotea was created in December 2004. Figure 2.3 and 2.4 show the interface of Connotea.


http://www.connotea.org/
Connotea: free online refer...
File Edit View Favorites Tools Help
View favorites, feeds, and history (Alt+C)
Home Blog About Connotea Site Guide Community pages
My Library
sign up now
Login
Find
Save
Customize
100%

Free online reference management for all researchers, clinicians and scientists

Completely free, no download

Already Registered?

Username Password

Forgotten your username or password? | Login with OpenID

Five reasons to use Connotea

1. Save and organize links to your references
2. Easily share references with colleagues
3. Access references from any computer
4. One click is all it takes
5. Easy to use. Start creating your library today

→ Learn more
→ Watch a short video (2m 41s)

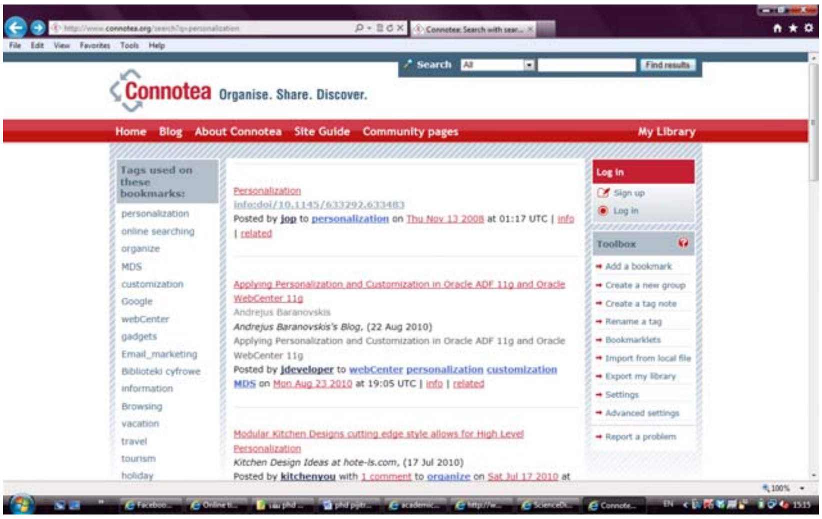
How it works

1 Find: You find a reference

2 Save: One click saves it on Connotea

3 Customize: Add your own keywords for easy retrieval

Figure 2.3 : The Connotea homepage.



Tags used on these bookmarks:

- personalization
- online searching
- organize
- MDS
- customization
- Google
- webCenter
- gadgets
- Email_marketing
- Biblioteki cyfrowe
- information
- Browsing
- vacation
- travel
- tourism
- holiday

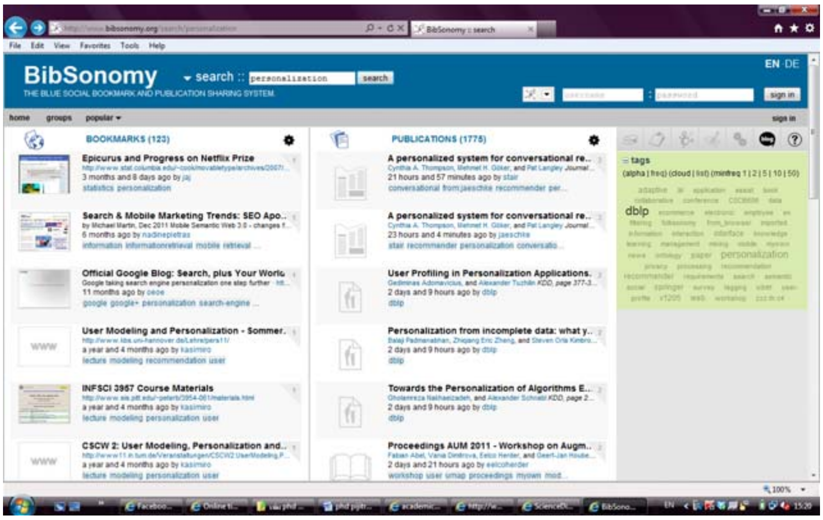
Personalization
info:doi/10.1145/633292.633483
Posted by jop to personalization on Thu Nov 13 2008 at 01:17 UTC | info | related

Applying Personalization and Customization in Oracle ADF 11g and Oracle WebCenter 11g
Andrejus Baranovskis
Andrejus Baranovskis's Blog, (22 Aug 2010)
Applying Personalization and Customization in Oracle ADF 11g and Oracle WebCenter 11g
Posted by jdeveloper to webCenter personalization customization MDS on Mon Aug 23 2010 at 19:05 UTC | info | related

Modular Kitchen Designs cutting edge style allows for High Level Personalization
Kitchen Design Ideas at hote-ls.com, (17 Jul 2010)
Posted by kitchenyou with 1 comment to organize on Sat Jul 17 2010 at

Figure 2.4 : The Connotea search results.

BibSonomy is a publication-sharing system and social bookmarking. This website proposes to combine the features of bookmarking systems as well as team-oriented publication management. This system supports the unification of different communities by offering a social platform for literature exchange and offers users the ability to organize and store their bookmarks. Figure 2.5 and 2.6 show the interface of BibSonomy.



BibSonomy search :: personalization search

THE BLUE SOCIAL BOOKMARK AND PUBLICATION SHARING SYSTEM

BOOKMARKS (123)

Epicurus and Progress on Netflix Prize
3 months and 8 days ago by jaj
statistics personalization

Search & Mobile Marketing Trends: SEO Apo...
by Michael Martin, Dec 2011 Mobile Semantic Web 3.0 - changes f...
6 months ago by nadinepietras
information informationretrieval mobile retrieval ...

Official Google Blog: Search, plus Your World
Google taking search engine personalization one step further
11 months ago by oeoe
google google+ personalization search-engine ...

User Modeling and Personalization - Sommer...
http://www.kbs.uni-hannover.de/Lehre/pers11/
a year and 4 months ago by kasimiro
lecture modeling recommendation user

INFSCI 3957 Course Materials
a year and 4 months ago by kasimiro
lecture modeling personalization user

CSCW 2: User Modeling, Personalization and...
a year and 4 months ago by kasimiro
lecture modeling personalization user

PUBLICATIONS (1775)

A personalized system for conversational re...
Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley Journal...
21 hours and 57 minutes ago by stair
conversational from:jaeschke recommender per...

A personalized system for conversational re...
Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley Journal...
23 hours and 4 minutes ago by jaeschke
stair recommender personalization conversatio...

User Profiling in Personalization Applications...
Gediminas Adomavicius, and Alexander Tuzhilin KDD, page 377-3...
2 days and 9 hours ago by dblp
dblp

Personalization from incomplete data: what y...
Balaji Padmanabhan, Zhiqiang Eric Zheng, and Steven Orla Kimbro...
2 days and 9 hours ago by dblp
dblp

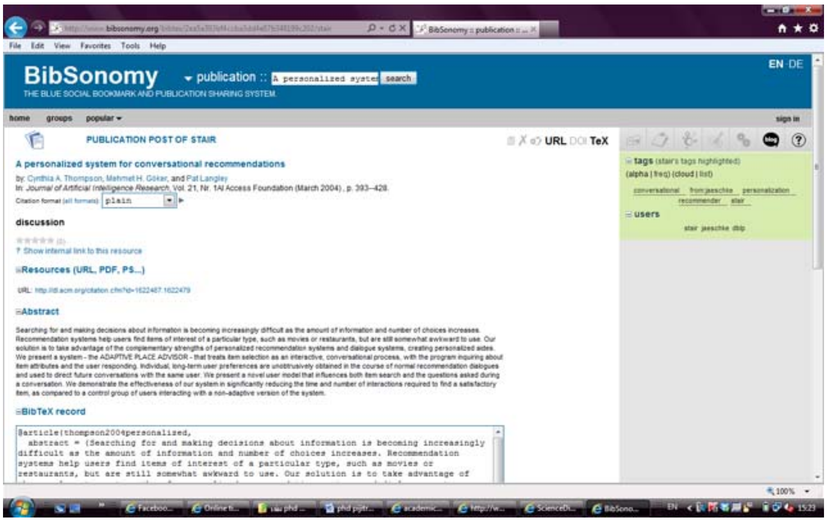
Towards the Personalization of Algorithms E...
Gholamreza Nakhaeizadeh, and Alexander Schnabl KDD, page 2...
2 days and 9 hours ago by dblp
dblp

Proceedings AUM 2011 - Workshop on Augm...
Fabian Abel, Vania Dimitrova, Eelco Herder, and Geert-Jan Hoube...
2 days and 21 hours ago by eelcoherder
workshop user umap proceedings myown mod...

tags
(alpha | freq) (cloud | list) (minfreq 1 | 2 | 5 | 10 | 50)

dblp personalization

Figure 2.5 : The BibSonomy homepage.



File Edit View Favorites Tools Help
BibSonomy :: publication :: ...
EN DE
sign in
home groups popular
URL DOI TeX

BibSonomy publication :: A personalized system search

THE BLUE SOCIAL BOOKMARK AND PUBLICATION SHARING SYSTEM

PUBLICATION POST OF STAIR

A personalized system for conversational recommendations

by: Cynthia A. Thompson, Mehmet H. Göker, and Pat Langley
In: Journal of Artificial Intelligence Research, Vol. 21, Nr. 1AI Access Foundation (March 2004) , p. 393–428.

Citation format (all formats): plain

discussion

? Show internal link to this resource

Resources (URL, PDF, PS...)

URL: http://dl.acm.org/citation.cfm?id=1622487.1622479

Abstract

Searching for and making decisions about information is becoming increasingly difficult as the amount of information and number of choices increases. Recommendation systems help users find items of interest of a particular type, such as movies or restaurants, but are still somewhat awkward to use. Our solution is to take advantage of the complementary strengths of personalized recommendation systems and dialogue systems, creating personalized aides. We present a system - the ADAPTIVE PLACE ADVISOR - that treats item selection as an interactive, conversational process, with the program inquiring about item attributes and the user responding. Individual, long-term user preferences are unobtrusively obtained in the course of normal recommendation dialogues and used to direct future conversations with the same user. We present a novel user model that influences both item search and the questions asked during a conversation. We demonstrate the effectiveness of our system in significantly reducing the time and number of interactions required to find a satisfactory item, as compared to a control group of users interacting with a non-adaptive version of the system.

BibTeX record

```
@article{thompson2004personalized,
  abstract = {Searching for and making decisions about information is becoming increasingly
difficult as the amount of information and number of choices increases. Recommendation
systems help users find items of interest of a particular type, such as movies or
restaurants, but are still somewhat awkward to use. Our solution is to take advantage of
```

tags (stair's tags highlighted)
(alpha | freq) (cloud | list)
conversational from:jaeschke personalization recommender stair

users
stair jaeschke dblp

Figure 2.6 : The BibSonomy search results.

Table 2.1: Service comparison and information of three academic paper sharing systems.

Academic paper sharing	Service				Information			
	community	free tagging	Search	BibTeX	Priority of paper	year	Along with group	Paper posted time
CiteULike	√	√	√	√	√	√	√	√
Connotea	√	√	√	√	-	√	√	√
BibSonomy	√	√	√	√	√	√	√	√

Comparison of services and information provided by the three academic paper sharing system are shown in Table 2.1. Apparently, most of the services and information provided in CiteULike, Connotea, and BibSonomy are similar. However, minor information on paper priority in Connotea system does not exist.

Many researches related to the social bibliographic searching focus on enhancing the capability of academic searching. Thus, users of bibliographic social bookmarking systems can cite the papers of their interest. Researches in the search area that are similar to CiteULike [9 -13] have been carried out extensively.

The primary goal of bibliographic social bookmarking is to serve the tags of each web resource and the needs of individual user to link with academic research papers under a specific circumstance. The systems should also facilitate other users to browse, find items, and categorize. The users then return to the original web page where they can resume working. The tags can also be used to represent academic paper contents, information discovered, shared, and community ranking of the items found. However, search results from these bibliographic social bookmarking are still not good for user relevance. There is the need for users in both social network and social bookmaking systems to post tags and details of user interest for cross-referencing. Systems of collaborative tagging have recently appeared as tools to structure user-generated content and online databases. Several researchers attempt to study the advantages of social tagging which users post such as collaborative tagging systems. Structural analysis of this tagging presents a dynamic model which can predict a stable pattern. Such stability demonstrates that tagged bookmarks may be valuable in aggregate as well as individual applications [14]. As tags become “meaningful” for searching, the tagging process is influenced by tag suggestions [15], which can improve retrieval performance [16], [17]. However, tags cannot only improve the search effectiveness [18], but also support knowledge discovery [19].

Users of bibliographic social bookmarking systems can cite papers of their interest. Further, the system will collect the information of each paper known as a user bookmarking such as bibliographic social bookmarking systems. This information includes year of publication, paper posted time, priority of paper, number of groups contained the posted paper, number of users contained the posted paper, URL, etc. This is interesting to investigate how these factors may affect on improving search result ranking of bibliographic social bookmarking.

The classical Retrieval [20] finds documents corresponding to the user query. The documents with more similar content to the query will be selected as more relevant search results. In other words, their algorithms usually work based on matching words which appear in documents. A prominent example of the content based on ranking algorithms is TF-IDF [21]. Any research communities have hypothesized that focusing on per-user search results by using important parameters or factors to adjust ordering. The ranking function for each user affects result rankings that may improve individual's relevant. User information can be specified by the user (explicitly collected) or can be automatically learned from a user's historical activities (implicitly collected). Bao et al. [22] proposed SocialSimRank and SocialPageRank to improve the web search with data on del.icio.us. Hotho et al. [23] suggested FolkRank that was a ranking algorithm based on PageRank, resources, and tags on del.icio.us altogether, and left the relative importance between resources clear.

Many bibliographic social bookmarking systems have been designed to work using query-dependent ranking (similarity ranking). This ranking focuses on improving the order of search results being returned to the users by measuring match between the content of the web resource and query terms. Various approaches in ranking the results have been studied such as static ranking and indexing method. Query-independent ranking (static ranking) is important to measuring the quality of retrieved documents for a search engine.

In recent years, personalized ranking has been employed in searching area as the result of several researches in the personalized ranking literature. First is automatically learning user preference with requiring implicit user intervention [24-27]. Next, lists of queries can be used to make statements about a user preference given their relative ordering and click-through data in either related queries or the current one

[28]. Zareh et al. [29] enhanced A3CRank method based on click-through, the content, and connectivity. The algorithm outperforms other combination ranking algorithms such as ranking SVM in terms of P@n and NDCG metrics. Dou et al. [30] used click-through data experiment for personalization and found that the use of personalization was highly dependent on ambiguity of the query. If the query was highly specific, then personalization was likely to have a negative effect on the results. This suggested that any deployed personalization system would need to estimate the ambiguity in the query so as to apply personalization only when it was likely to improve the results. However, some researches learn user's preference by requiring explicit answer from the user [31] and applied feature to improve ranking. Berberich et al. [32] suggested T-rank that was a link analysis approach by taking into account the activity and temporal freshness aspects. As such,T-rank results can enhance the quality of ranking.

Most personalized ranking algorithms are based on generated user profiles. A user profile stores approximations of user tastes, preferences and interests.

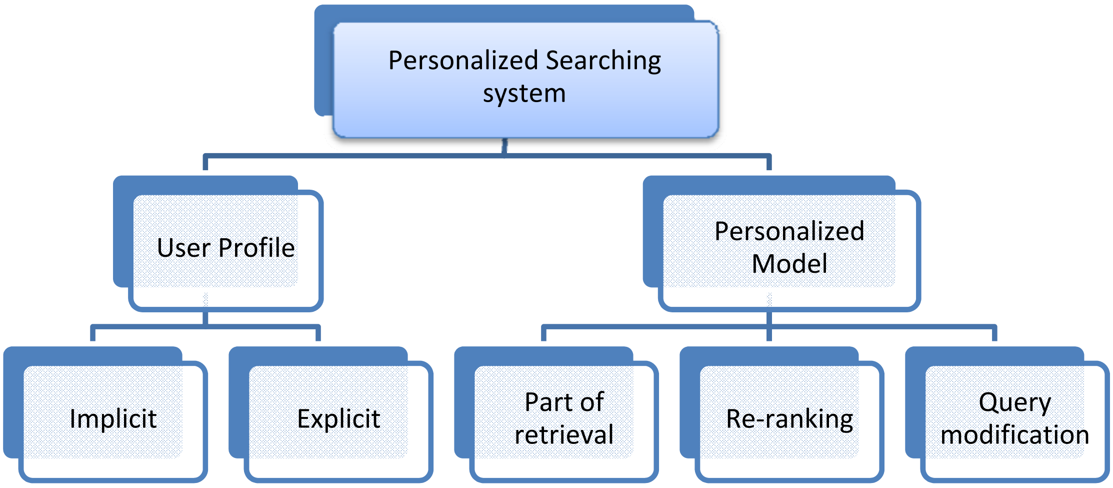


Figure 2.7: Personalized search system.

Techniques for personalized searching systems are shown in Figure 2.7. The systems compose of two parts: the first part is techniques to create the model of personalization and the second part is techniques to construct user profile, aka profiler. Figure 2.8 to 2.10 show the model of personalized search systems [4] consisting of three processes, namely, personalized retrieval process, re-ranking, and query modification.

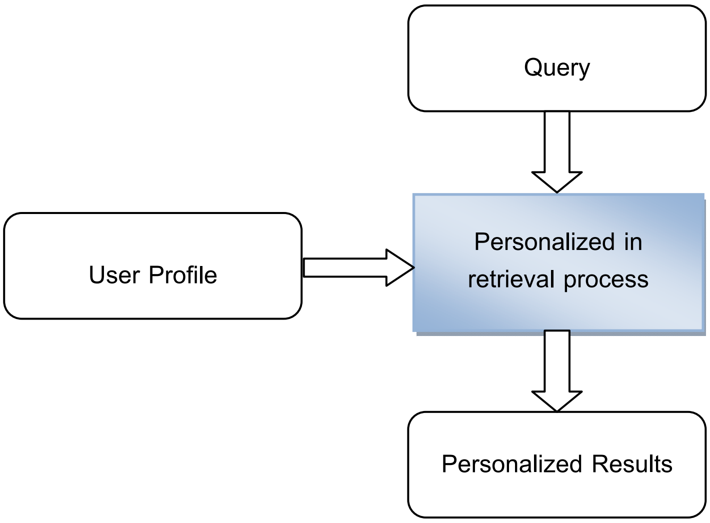


Figure 2.8 : Personalized in part of retrieval process.

Personalized retrieval process: the ranking is a unified process wherein user profiles are employed to score Web contents. The first technique provides the traditional ranking system that can be directly adapted to include personalization.

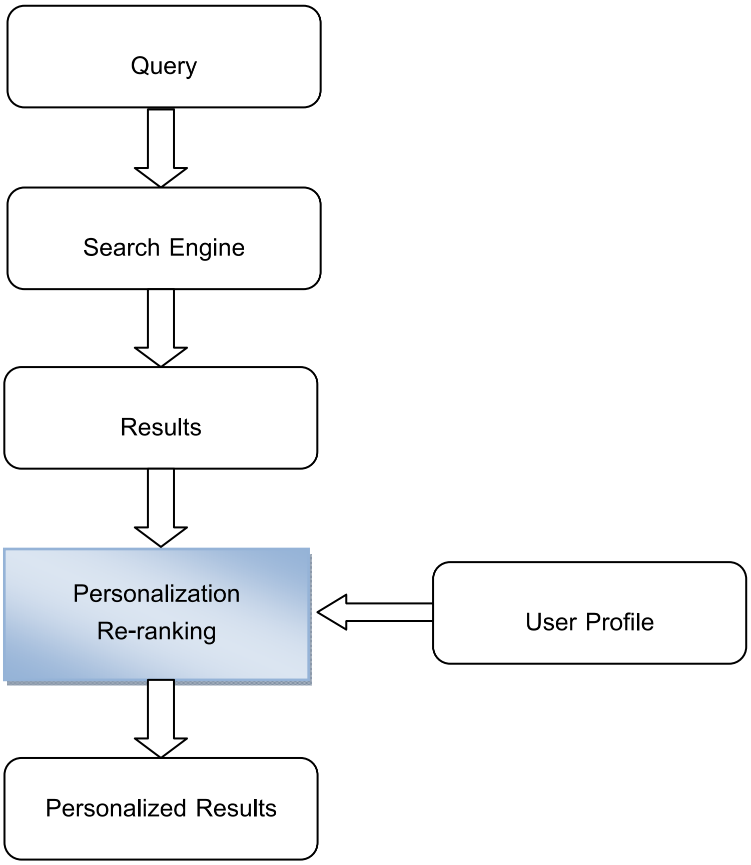


Figure 2.9 : Personalized re-ranking.

Re-ranking: Re-ranking documents as suggested by an external system, such as a search engine, allows the user to selectively employ personalization approaches to increase precision. User profiles take part in the second step after evaluating the corpus ranked via non-personalized scores. Many systems implement this approach on the client-side where the software connects to a search engine and retrieves query results for subsequent local analysis. The analysis is only applied to top ranked resources in the list returned by the search engine to avoid downloading non-candidate documents. The re-ranking approach implemented via client-side software can be considerably slow due to delay accessing the search engine and retrieving the pages to be evaluated. However, suitable representations of user needs can be employed which will improve the personalization performance.

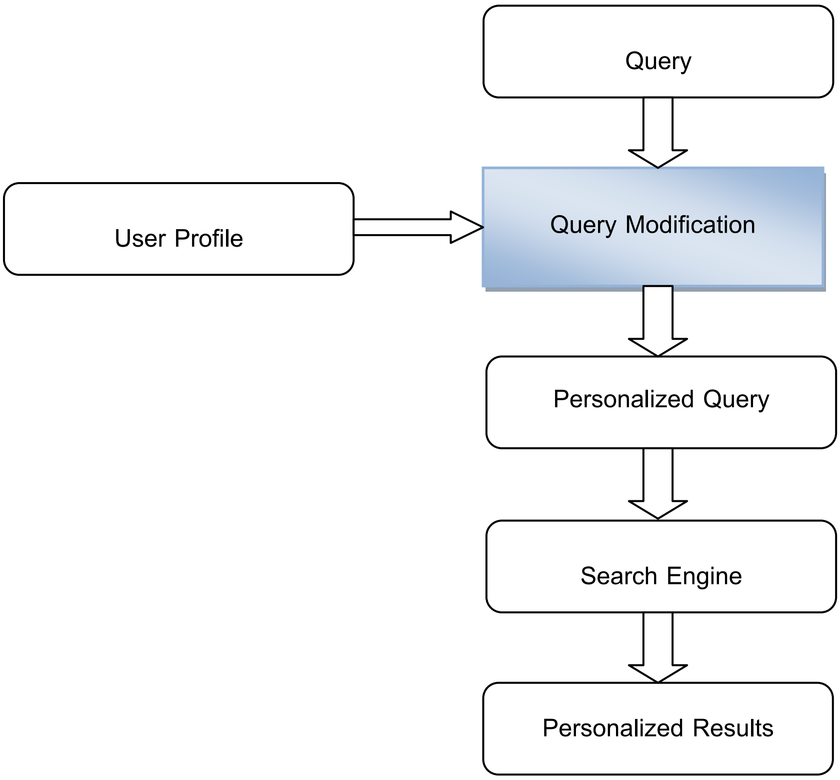


Figure 2.10 : Personalized : query modification.

Query modification: user profiles influence the submitted representation of query or the information needs, augmentation or modification. Retrieval takes place after profiles can modify the representations of the user needs. For instance, the profile may transform them by adding or changing some keywords to better represent the needs in the current profile if the user needs are represented by queries. However, user profiles affect the ranking only by altering the query representations. The advantage of this approach is that the amount of work required to retrieve the results is the same as in the un-personalized scenarios.

Many researchers work with re-ranking process [33] to improve the quality of Web searches on social annotations. Four annotation-based ranking methods were assessed: Query weighted Popularity Count (QWPC), Matched Tag Count (MTC), Popularity Count (PC), and Normalized Matched Tag Count (NMTC) [24], [34]. A general process of re-ranking is to devise efficient mechanisms to re-order the search result ranking using the global importance obtained by personalized ranking criteria.

Presently, there are two major categories of ranking algorithms based on query-dependent ranking (similarity ranking) and query-independent ranking (static ranking). In classical information retrieval [20], the system works to find documents corresponding to the user query. Some researchers examine static features by focusing on relatively simple features of web documents to provide suitable rankings and improve search results for social network systems [32], [35-36].

Personalized ranking has been used to a greater extent in the field of searching as a result of several previous works in the personalized ranking literature. There are many surveys of personalization. Several researchers are working toward automatically learning user preference without requiring explicit user intervention [24], [26], [37-39]. Many personalized ranking algorithms are based on generating user profiles and user behaviors [40-42]. In addition, the essence of social tagging is applied to generate the profile framework [41]. A user profile stores approximation of user tastes, interests, and preferences. Many systems implement this approach on the client-side, e.g., [24], [26], [42-48]. In personalized search systems, user modeling components could affect the search such as link-based personalization score and re-ranking. The personaliztion score is computed by using cosine similarity measure between the most similar concepts of the user profile and each returned document [26], [46]. This score will determine the documents to be returned as search results. The results are then given to the user [49].

2.2 Theoretical Background

Web information search is the process of using a web search engine or social bookmarking to locate documents that are relevant to a certain user query to fulfill the information required. Theoretical models of information retrieval furnish different ways in which the IR problem can be formulated and solved. Details will be elucidated in the sections that follow.

2.2.1 The Information Retrieval Models

Three classical models of information retrieval are discussed, namely, Boolean, Vector Space, and Probabilistic Models.

2.2.1.1 Boolean Model

Boolean Model is one of the oldest and simplest models of Information Retrieval. It is based on set theory and Boolean algebra [50]. In this model, each document is taken as a bag of index terms. Index terms are simply words or phrases from the document that are important to establish the meaning of the document. The query is a Boolean algebra expression using and, or, not connectives. The documents retrieved are the documents that completely match the given query. Partial matches are not retrieved. The retrieved set of documents is not ordered. Hence, the advantages and disadvantages of this model are as follows [51]:

Advantages

- The model is simple, efficient, and easy to implement.
- The model is very precise in nature. The user exactly gets what is specified.
- The model is still widely used in small scale searches like searching emails, files from Local hard drives or in a mid-sized library.

Disadvantages

- The retrieval strategy is based on binary criteria. So partial matches are not retrieved. Only those documents that exactly match the query are retrieved.
- The retrieved documents are not ranked.
- Given a large set of documents, say, at web scale, the model either retrieves too many documents or very few documents.
- The model does not use term weights. If a term occurs only once in a document or several times in a document, it is treated in same way.

2.2.1.2 Vector Space Model

The Vector Space model represents both a query and a document as a vector in a high-dimensional space where each dimension corresponds to a term. Some well-known of vector space model techniques in information retrieval are as follows:

Term-Frequency and Inverse Document Frequency

The term weights in the document and query vector representing the importance of the term for expressing the meaning of the document and query. There are two widely used factors in calculating term weights, namely, term frequency (*tf*) and inverse document frequency (*idf*). The term weights can be approximated by the product of the two factors. This is called the *tf*-*idf* measure.

Let t_i be the i^{th} term of vector x, x be the vector representation of document/query *d,* then tf_{ij} denotes the term frequency, i.e., the number of occurrences of term t_i in document d_j, $freq_{i,j}$ is the number occurrences of term i appeared in document j. N is the total number of documents and n_i is the number of documents in which the term i occurs, idf_i is the inverse document frequency:

$$tf_{i,j} = \frac{freq_{i,j}}{\sum_l freq_{i,q}} \tag{2.1}$$

$$idf_i = log\frac{N}{n_i} \tag{2.2}$$

Cosine Similarity

Cosine similarity is a measure of similarity between two vectors which usually is a task of retrieving documents that match a query. Cosine Similarity is expressed by the formula:

$$similarity(D,Q) = \frac{D \cdot Q}{\|D\|\|Q\|} \tag{2.3}$$

where Q is a query vector and D is a document vector. Then the similarity can be computed as follows:

$$sim(D,Q) = \frac{\sum_{i=1}^{n} D_i \times Q_i}{\sqrt{\sum_{i=1}^{n} D_i^2 \times \sum_{i=1}^{n} Q_i^2}} \tag{2.4}$$

The cosine similarity of two documents will range from 0 to 1. The cosine similarity can be seen as a method of normalizing document length during comparison. Generally, for text matching, the attribute vectors D and Q are usually the term frequency vectors of the documents. The advantages and disadvantages of this model are as follows [51]:

Advantages

- The model is a matching measurement basic method.
- Cosine similarity measurement returns values in the range of 0 to 1.
- Ranking of the retrieved results according to the cosine similarity score is possible.

Disadvantages

- Index terms are considered to be mutually independent. Thus, this model does not capture the semantics of the query or the document.

- Vector Space model cannot denote the “clear logic view” like Boolean model [52].

The Nutch Ranking

Nutch ranking is an open-source search engine platform based on Lucene java. Nutch is a fully edged web search engine that supports crawling, indexing, and ranking. It also has a link graph database and other document formats. The indexing and ranking components are based on apache Lucene. It is developed by Apache Software foundation [53].

The ranking is done in two steps. An initial set of document is retrieved using Boolean model. Then it ranks the initial set using vector space model. The similarity of a query q and a document d is given by:

$$score(q,d) = \sum_{t \in q} (tf(t \in d) \times idf(t)^2 \times boost(t.field \in d) \times lengthNorm(t.field \in d)) \\ \times (queryNorm(q) \times coord(q,d)) \qquad (2.5)$$

The sum term gives the numerator of cosine similarity if each term is assumed to occur once in the query. The normalization factor given in the denominator of cosine similarity is given by the terms like $queryNorm(q)$ and $lengthNorm(t.field \in d)$. The terms in the above expression are explained below:

idf(t) is inverse document frequency of term t.

tf(t $\in$ d) is term frequency of term t in document d.

boost (t.field $\in$ d) is the importance of a term appeared in a document.

lengthNorm(t.field$\in$ d) This factor is calculated using the following expression:

$$norm(t,d) = doc.getBoost() \times lengthNorm(field) \times \prod f.getBoost() \qquad (2.6)$$

doc.getBoost() captures the importance of the document in the collection of documents. It is calculated using a Link Analysis algorithm named Online Page Importance Calculation (OPIC).

lengthNorm(field) captures the normalization factor that depends on the length (number of index terms) of the document.

f.getBoost() captures the importance of a particular field. The product term captures whether the term appears more in the important part of the document than in non-important part.

queryNorm(q) is a normalizing factor used to make scores between queries comparable. This factor does not affect document ranking as it depends only on the query.

coord(q,d) is a score factor based on the number of query terms which are found in the specified document.

Advantages

- The Nutch ranking is a very popular search engine library
- Easy configuration for managing fields to be indexed.
- Open source

Disadvantages

- No assured availability of training or other professional services to fulfill specific software needs or assist with building an application

2.2.1.3 Probabilistic Model

Probabilistic Model endeavors to capture the information retrieval problem within a probabilistic framework. This model assumes that this probability of relevance depends on the document representations and query. Probabilistic Model estimates the probability of document d_j being relevant to a query q. In addition, a portion of all documents that is preferred by the user as the answer set of query q is assumed. Such an ideal answer set is called R and should maximize the overall probability of relevance to user. The predictions in set R are relevant to the query, while documents not present in the set are non-relevant [54].

Formally, given the document vector d and query vector q, the documents are ranked according to the probability of the document being relevant. Mathematically, the scoring function is given by:

$$sim(d_j, q) = \frac{P(R|d_j, q)}{P(\overline{R}|d_j, q)} \quad (2.7)$$

The ratio P (d_j relevant-to q)/ P(d_j non-relevant-to q) which computes the odds of the document d_j being relevant to query q [50],[55].

BM25 is based on the probabilistic retrieval framework, a method based on language modeling and divergence from randomness. This model is a ranking function used by search engines to rank matching documents according to their relevance to a given search query.

$$score(Q, D) = \sum\nolimits_{t \in Q} \frac{(r_i + 0.5)/(R - r_i + 0.5)}{(n_i - r_i + 0.5)/(N - n_i - R + r_i + 0.5)} \times \frac{(k_1 + 1) f_i}{K + f_i} \times \frac{(k_2 + 1) qf_i}{k_2 + qf_i} \quad (2.8)$$

where the summation is now over all terms in the query. N is the total number of documents in the collection, R is the number of relevant documents for this query, n_i is the number of documents containing term i , r_i is the number of relevant documents containing tern i. qf_i is the frequency of term i in the query.k_1, k_2, and K are parameters whose values are set empirically.

Advantages

- Probabilistic model can be easily extended and embedded in more complicated models

Disadvantages

- Not a well-defined generative model
- Many free parameters
- likely to over-fitting

2.2.2 Top-K Rank Retrieval

Top-K rank retrieval is the process of retrieving the K documents that match a given query the most. For each query term, a list of all the documents that contain the term is present. This list is in descending order according to the weight of the term in the document. The weight could again be based on either *tf-idf* or Nutch ranking or any other weighting schemes [56].

2.2.3 Evaluation of Search Results

Evaluation of search results is the major assessment step in building better search engine. Most measures assume binary relevancy, i.e., the document is either relevant or completely irrelevant. These measures include precision and recall. In a real system, recall value is hard to calculate. The method summarizes the ranking by averaging the precision values, a.k.a. Mean Average Precision (MAP), from the rank position where a relevant document was retrieved. Yet another measure called Normalized Discounted Cumulative Gain (DCG) [58] is a measure that gives more weight to highly ranked documents and allows users to consolidate different relevance levels. Details on each technique are described below.

2.2.3.1 Precision and Recall

The basic measures that formally capture the attitudes of valuable query results are Precision and recall.

Precision

Precision is defined as the proportion of retrieved documents that are relevant. Let R is the relevant set of documents for the query. A is the retrieved document. Then

$$Precision = \frac{|R \cap A|}{|A|} \qquad (2.9)$$

Recall

Recall is defined as the proportion of relevant documents that are retrievled. [59].Then

$$Recall = \frac{|R \cap A|}{|R|} \qquad (2.10)$$

In general search engines, it is not clear how the list of ranking should be chosen. Therefore, Average Precision tries to address this problem by combining precision values at all possible recall level.

Average Precision

$$AveP = \frac{\sum_{k=1}^{N} (P(k) \times rel(k))}{|R|} \qquad (2.11)$$

where R is the number of relevant documents, k is the rank, and $P(k)$ is the precision at a given cut-off rank k. $rel(k)$ is the relevance of a given rank k, N is the number of documents retrieved.

Mean Average Precision (MAP)

Mean average precision is the mean of the average precision scores for each query.

$$MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q} \qquad (2.12)$$

where Q is the number of queries.

The main advantage of these measures is that they are simple and commonly used. The main disadvantage of these measures is that they only take into account binary relevance ratings and are not able to cope with multi-graded relevance assignments.

2.2.3.2 Normalized Discounted Cumulative Gain (NDCG)

The Normalized Discounted Cumulative Gain [58] is a measure that gives more weight to highly ranked documents and allows the incorporation of different relevance levels. NDCG rewards relevant documents appearing in the top ranked search results and punishes irrelevant document by reducing their contributions to NDCG. Let M_q be a normalization constant, $r(j)$ be an integer representing the relevancy given by the subject so that the perfect ordering would obtain NDCG of 1, and k be a truncation or threshold level. NDCG is defined as follows:

$$NDCG_q = M_q \sum_{j=1}^{k} \frac{\left(2^{r(j)} - 1\right)}{\log\left(1 + j\right)} \qquad (2.13)$$

The advantage of NDCG is that it applies an information-theoretic model for considering multiple relevance levels. Unfortunately, the NDCG measure values depend on the number of reference relevance values of the dataset. Thus, NDCG values computed for different datasets cannot be directly be compared with one another.

2.2.3.3 Significance Tests

Additional performance of different retrieval procedures must be evaluated because only the means of the assessment measures such as NDCG over the assessed queries are not sufficient. Particularly when the sample data are small, the differences between such means on their own might be misleading. In order to measure how statistically significance these differences are, significance tests such as ANOVA, One-way analysis of variance (one-way ANOVA), are called for to test equality of three or more means, for example, ANOVA is done with the H_o: $\mu_1 = \mu_2 = \mu_3 = ...= \mu_k$.

Significant value of F-test is compared with the F-statistic. If the F-value $>$ α, where α =0.05, then it can be concluded that there is insufficient evidence to reject the null hypothesis at the give level of significance. But if the F-value $\leq$ α, then it can be concluded that there is sufficient evidence to reject the null hypothesis at the given level of significance. In other words, at least one of the population means (μ_i) is different from the rest [60].

CHAPTER III

PROPOSED FRAMEWORK

This chapter contains two parts. The first part is about the framework of research methodology. The second part is the experimental setting to solve three problems and the corresponding hypothesis, experimental setting, and evaluation metrics.

3.1. Research Methodology

The framework is divided into two parts, namely, bibliographic social bookmarking and bibliographic searching.

3.1.1 Bibliographic Social Bookmarking

A bibliographic social bookmarking system, such as CiteULike, provides users with new ways to share their research interests. All public systems can also be searched and filtered by tags. This kind of bibliographic social bookmarking allows users to create their own tags or keywords to attach to the posted papers. Users can automatically share all their public entries with others and comment on other papers afterwards. Moreover, user can also discover interesting papers posted by other users who share the same interests.

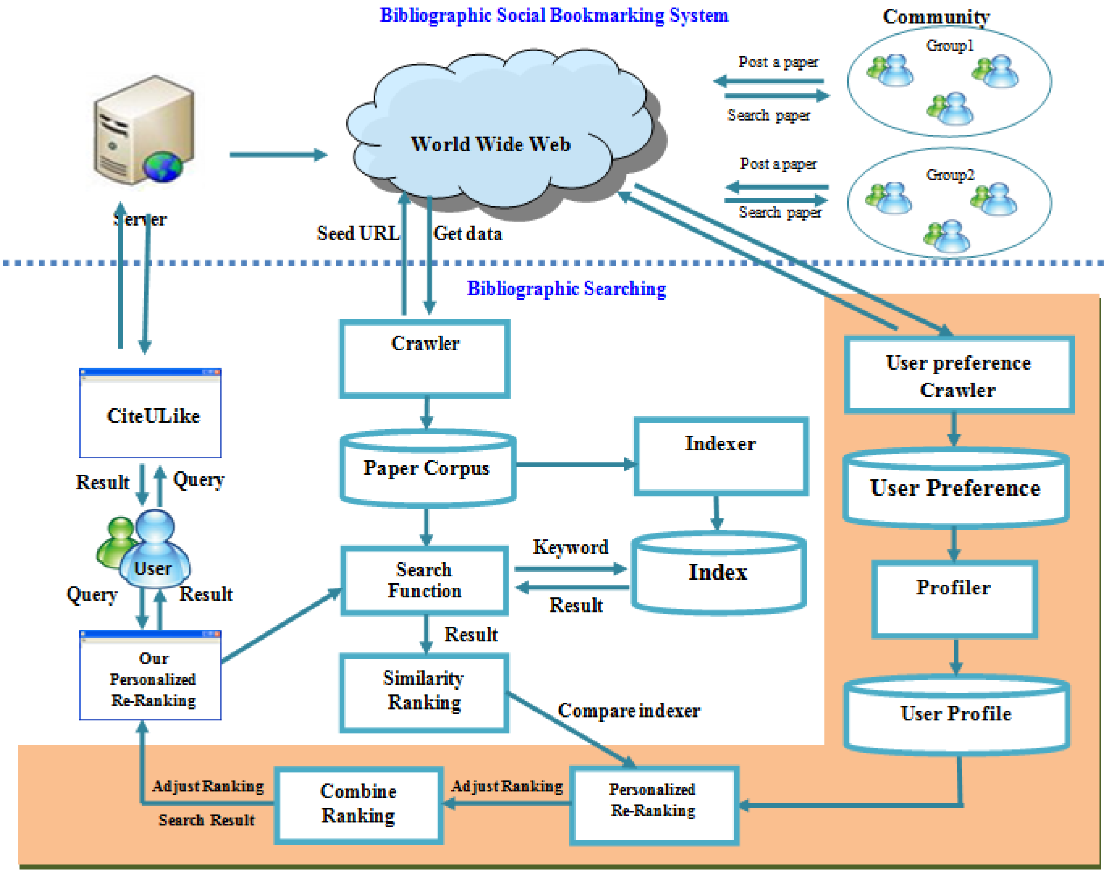


Figure3.1: Framework of community-based on bibliographic search engine.

3.1.2 Bibliographic Searching

This work concentrates on improving bibliographic searching described in nine parts below.

3.1.2.1 Crawler: Crawler is a small computer program that browses directly to the academic paper sharing systems of the internet in a predetermined manner. The academic paper crawler is responsible for gathering research paper information such as tags used, paper title, etc. This useful information helps the system create index for each paper and determine a user's interests.

3.1.2.2 Research Paper corpus: Paper corpus is a collection of academic papers extracted from the academic paper sharing system. A typical crawling process is created using java programming to collect data for the corpus. Each record in the paper corpus contains tag, title name, abstract, and link for viewing full text article, posted time, book title within which the paper is published, paper priority, and posted date.

3.1.2.3 Indexer: All Users in a social bibliographic bookmarking system can cite papers by posting the detail of each paper. Some details of user posts can represent papers such as tag, title, and abstract. This research tries to develop the potential indexer. Therefore, three indexes are created by using 1) *only Tag (T)*, 2) *Title with Abstract (TA),* 3) *Tag, Title, and Abstract (TTA)*. In the process to create the indices which are weights of TF-IDF, representing a statistical measure used to evaluate how important a word is to a document in a collection or corpus. The importance increases proportionally to the number of times a word appears in the document but is offset by the frequency of the word in the corpus. The experimental setting is described in Section 3.2.

3.1.2.4 Search Function: This involves finding the cosine of the angle between two vectors. This measurement is often used to compare documents in text mining. Details of its use in ranking will discuss in the next step.

3.1.2.5 Similarity Ranking: a cosine similarity measurement is used to retrieve and rank search results by comparing the academic paper index with a query. This similarity score is a combination between Boolean model (BM) and Vector Space Model (VSM) of information retrieval. The similarity score of query q for document d is defined in Equation (2.5).

3.1.2.6 User preference crawler and user preference data: This crawler is responsible for crawling user preference data which detail the academic papers posted by individual user, including a set of personally defined data such as tags, title, abstract, paper posted time, year of publication, priority of paper, and number of groups contained the posted paper.

3.1.2.7 A Profiler: User profiling is used to model a user's features or preferences. A profiler is a mechanism that exploits the usage of user self-defined tags, title, and abstract from all posted papers of a user in order to create a user profile. Approaches for profiling users with term vectors are used in the proposed system to create a user profile tagging behavior correctly and accurately [61]. The proposed approach employs the following definitions:

U : Set of users. $U= \{u_1, u_2...,u_n\}$, containing all users in the system.

D : Set of research papers. $D= \{d_1, d_2...,d_m\}$, containing all documents in the document collection.

S : Set of search results. $S= \{s_1, s_2...,s_c\}$, containing documents that are retrieved in the document collection.

UT_{ij} : Set of user profile terms. $UT_{ij}= \{ut_{i1}, ut_{i2}...,ut_{ie}\}$, including all terms of user self-defined tags that have been used by the users u_i , where $i=1,2,...,n$.

ST_{xj} : Set of terms assigned with search result (S) of each research paper. $ST_{xj} = \{st_{x1}, st_{x2}...,st_{xe}\}$, including all terms that have been used by result s_x , where $x=1,2,...,c$.

Each profile specifies user u_i using the term ut_{ij} for document d_k. The user profile is defined below.

Definition [User Profile]: For a user u_i, $i=1,.., n$, let UP_i be the relationship between u_i and item set, $UP_i = \{< u_i ,ut_{ij} >| ut_{ij} \in UT_{ij}, u_i \in U,$ and is specified user u_i using the term ut_{ij} for document $d_k \}$.

Two common user profiler types are user profiler based on tag and user profiler based on tag, title, and abstract. The experiment setting is described in Section 3.2.4.

3.1.2.8 User profile: A profile refers to the explicit digital representation of a person's identity. A prototype of the system and preliminary results are presented. Therefore, a user profile can be considered as the computer representation of a user model that is delivering personalized information.

3.1.2.9 Personalized Re-ranking and Combination of Personalized Re-ranking with Static Ranking

The process consists of personalized re-ranking and the combination of personalized re-ranking with static ranking.

3.1.2.9.1 Personalized Re-ranking: A personalized Re-ranking is a measurement of similarity of user terms profile and terms of document retrieval from search query. This is also known as “*PersonalizeRank*.” The frequency of each term is used in the calculation. The ranking of search results are rearranged from the highest similarity score to the lowest similarity score using cosine similarity defined in Equation (2.3) as follows:

$$Personsali\,zeRank = Sim(ut, st) = \frac{ut \bullet st}{\|ut\|\|st\|} \qquad (3.1)$$

3.1.2.9.2 Combination of Personalized Re-ranking with Static ranking: This step concentrates on static ranking which is the important information posted by the users. These factors are combined with *PersonalizeRank* to adjust ranking. Furthermore, weighting score are included in this step. Experimental setting is described in Section 3.2.4.2.

3.1.2.10 Sample Searching

To test the efficiently of personalized re-ranking, search results from a popular bibliographic social bookmarking such as CiteULike is evaluated. However, to prevent bias from users, an evaluation interface of search engine is developed. Subjects submit queries through this interface. Search results are then displayed by title, abstract, and the full text.

3.2 Design of Experimental Settings for the Research Problem

As mention in Chapter 1, three experiments are designed to solve the three main problems of this research. The goal of the experiment is to validate the proposed methodology through the search results in bibliographic social searching by using personalized re-ranking technique. A governing metric is used to gauge the experimental outcomes.

3.2.1 Evaluation Metrics

The NDCG measurement as defined in Equation (2.13) is used to evaluate the performance of each search engine in the experiment.

3.2.2 Experiment 1: The Potential Indexing

Experiment 1 will solve Problem 1 by investigating social tagging to improve bibliographic indexing. The premise is that in bibliographic social bookmarking, only social tagging may not be enough to represent the academic papers of user interests. Expanding the information of each academic paper by adding title and abstract to create indexing might help improve the system. An indexing method using tagging information together with title and abstract of the paper (TTA) is established. Two indexing approaches were compared: tagging information only indexing method (T) and title with abstract indexing method (TA) to evaluate the proposed indexing method.

Figure 3.2 shows a Framework of Experiment 1. Three different heuristic indexers were developed in this experiment and evaluation process was described in Section 3.2.2.2.

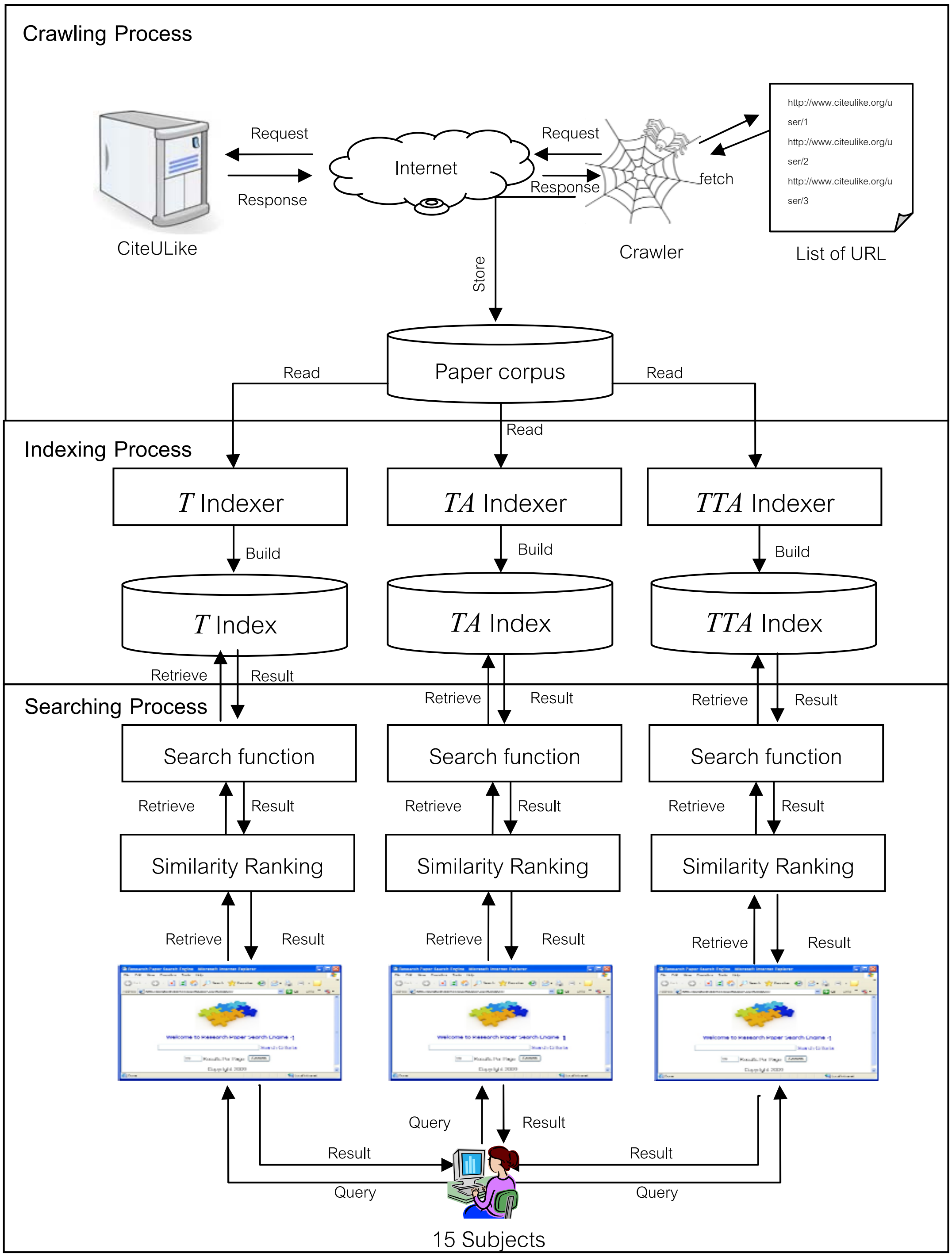


Figure 3.2: Framework of Experiment 1.

Equation (3.2), (3.3), and (3.4) present a modified Term Frequency/Inverse Document Frequency (*tf/idf*) formula for the different indexers, where

T is a set of "Tag only", TA is a set of "Title with Abstract", and TTA is a set of "Tag, Title with Abstract."

Let $|T|$ be the total number of "Tag Only" documents in the corpus, $|t_i \in T|$ be the number of documents where the term t_i appears in tag corpus (that is $n_{ij} \neq 0$), $|TA|$ be the total number of "Title and Abstract" documents in the corpus, $|t_i \in TA|$ be the number of documents where the term t_i appears in title with abstract corpus, $|TTA|$ be the total number of "Tag, Title with Abstract" documents in the corpus TTA, $|t_i \in TTA|$ be number of documents where the term t_i appears in tag, title, and abstract corpus, and $n_{i,j}$ be the number of occurrences of the considered term t_i in document d_j. If the term is not in the corpus, this will lead to a division-by-zero.

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|T|}{|t_i \in T|} \tag{3.2}$$

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|TA|}{|t_i \in TA|} \tag{3.3}$$

$$tfidf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \times \log \frac{|TTA|}{|t_i \in TTA|} \tag{3.4}$$

3.2.2.1 Hypothesis of Experiment 1

1) The null hypothesis:

H_0: There is no statistical difference among the means of NDCG at K=1-5 of the three indexing, TTA, TA, and T.

$(\mu_{\text{TTAindex}} = \mu_{\text{TAindex}} = \mu_{\text{Tindex}})$

2) The alternate hypothesis:

H_1: Not all approaches are equal

$(\mu_{\text{TTAindex}} \neq \mu_{\text{TAindex}} \neq \mu_{\text{Tindex}})$

3.2.2.2 Experimental Setting

Search engines based on three indexers were developed. Equation (3.2) was applied to the first search engine to create the index. Equation (3.3) was applied to the second search engine based on TA indexer, and Equation (3.4) was applied to the last search engine based on TTA indexer. An accompanying interface web page was also developed for the experiment. Fifteen subjects were recruited and each subject performed 3 queries with 3 search engines. The subject specified their search criteria to investigate the results from each search engine deployed. The results could be viewed by title, abstract, and full text, where the numbers of the results displayed per page were defined.

Each subject was instructed to find research papers of their interest and each query contained only one keyword. They formulated their own queries to look for the designated or related research papers. The same queries were subsequently used for each search engine. Then, they were asked to rate the relevancy of the search result set using Likert five-point scale:

Score 0 is not relevant at all.

Score 1 is probably not relevant.

Score 2 is less relevant.

Score 3 is probably relevant.

Score 4 is extremely relevant.

The top 20 search results of each search engine were displayed for relevancy assessment. Since the subjects in this experiment were considered experts in the field, their relevancy ratings were treated to be trust-worthy and perfect.

The relevancy ratings of each resource in the result set were recorded and used to rank the result set which in turn were used as the normalization constants for NDCG computation using K=1-5. One-way ANOVA was applied to measure the mean difference of NDCG scores at K= 1-5 from the three indexing. If the results from the F-value indicated a significant difference at the 0.05 level, the null hypothesis would be rejected. In addition, Levene's test was used to assess the equality of variances in the samples. If the significance value was greater than 0.05, this meant that the variance was equal. Results from ANOVA showed that there were significant differences among the groups as a whole. The subtle differences among the groups were further amplified by multiple comparisons. Moreover, the Tukey *post-hoc* and LSD tests were employed to test the equality of variances. The Dunnett T3 was used to test whether there would be any differences in the variances.

3.2.3 Experiment 2: Static Ranking and Combined Similarity Ranking with Static Ranking

Experiment 2 will solve Problem 2 by combining similarity ranking of search results obtained from search engines using TTA indexer with static ranking, such as year of publication, priority of paper, paper posted time, and number of groups that contained the posted paper. The score values of the combined two methods are adjusted to fit in the range of 0 to 1.

Figure 3.3 shows a framework of Experiment 2. The process is divided into three parts: the first part describes the detail of each property factor, the second part describes how to calculate $StaticRank$, and the last part describes how to combine a similarity ranking with static ranking [62],[63]. The detail of user evaluation is described in Section 3.2.3.6.

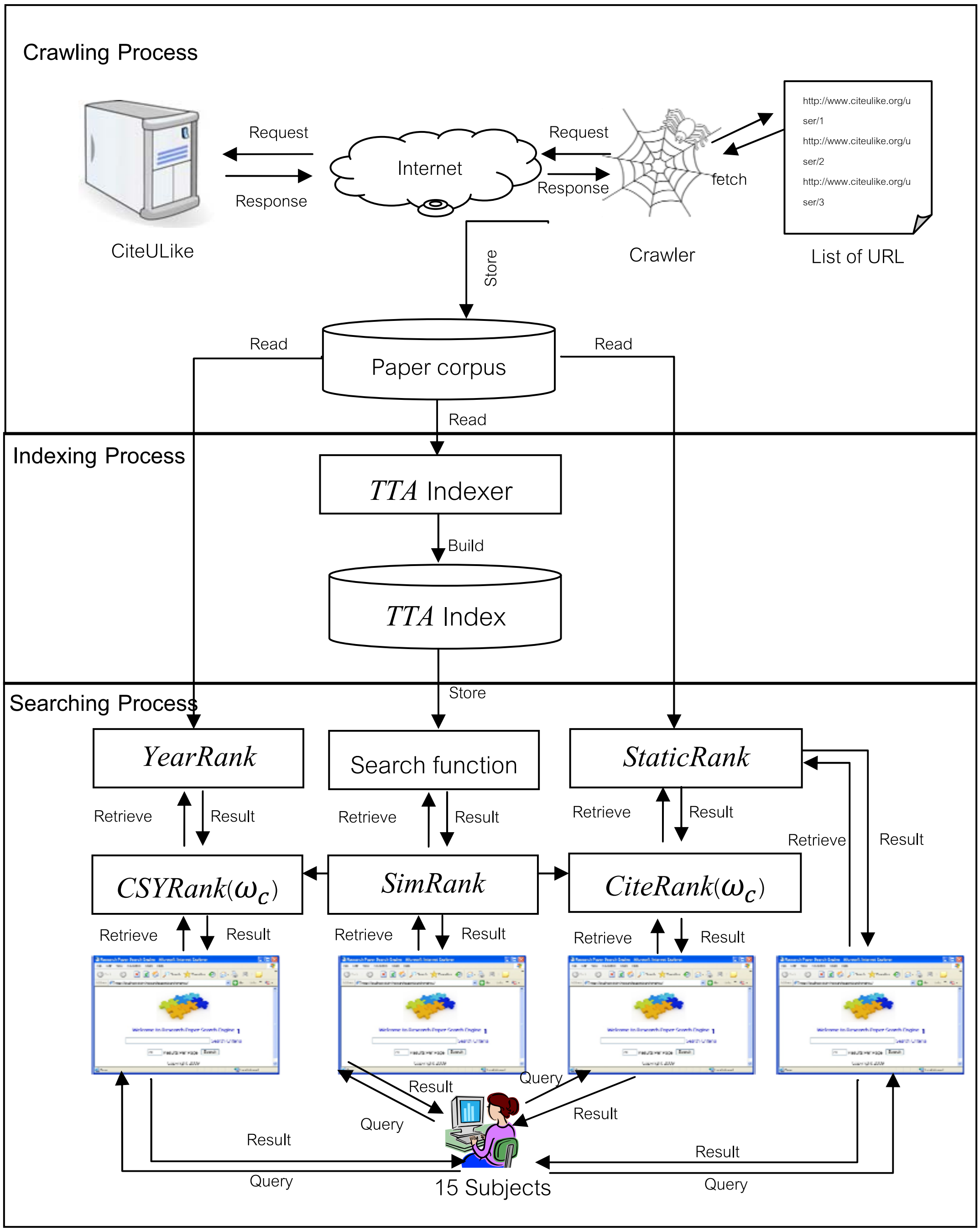


Figure 3.3: Framework of Experiment 2.

3.2.3.1 The Static Ranking Factors

As mentioned earlier, the four principal factors for creating the ranking are 1) year of publication (Y), 2) posted time (C), 3) priority rating (P), and 4) number of groups that contained in the posted paper (G). Details are described below.

*1) Year of publication (**Y**)*

This factor represents current interests of the users, where correctness is utmost important and ranked the highest. Let n denote the recentness of the posted year, where $n_i \in N$, CY be the current year, Y define the score for the publication year, and LY_x be the recentness of publication, where $x=\{1,2,3,4,5,6\}$. The score of the year recentness is calculated before calculating the year of publication as follows:

$$Y=\frac{n_i}{\max_j(n_j)} \qquad (3.5)$$

where, $$N = \begin{cases} 5 & ;\, if\,[LY_1 := CY - 2; \text{range level is } CY\ to\ LY_1\ ] \\ 4 & ;\, if\,[LY_2 := (LY_1 - 1) - 2; \text{range level is } LY_1 - 1\ to\ LY_2] \\ 3 & ;\, if\,[LY_3 := (LY_2 - 1) - 2; \text{range level is } LY_2 - 1\ to\ LY_3\ ] \\ 2; & ;\, if\,[\ LY_4 := (LY_3 - 1) - 2; \text{range level is } LY_3 - 1\ to\ LY_4] \\ 1; & ;\, if\,[LY_5 := (LY_4 - 1) - 2; \text{range level is } LY_4 - 1\ to\ LY_5] \\ 0; & ;\, if\,[LY_6 :\leq LY_5 - 1; \text{range level is less than } LY_5\ ] \end{cases}$$

*2) Paper posted time (**C**)*

Paper posted time information for each paper is composed of posted date and posted time, e.g. “2010-03-15 17:02:45.” Firstly, the paper posted time is sorted based on this information. Then, the posted time score (C) is calculated by the formula in Equation (3.6) C_r denotes the score of the current time rank.

$$C_r=C_{r-1}-0.05 \qquad (3.6)$$

where r =0, 1, 2, …, 19. The initial value is C_0=1.

*3) Priority of paper (**P**)*

In research paper sharing system, the users can rate the importance of their posted papers. This factor represents user's judgement which reflects the level of user interest in the paper. The priority rank score (P) will be calculated by Equation (3.7). Let m be number of priority scale with each paper, where $m_i \in M$.

$$P = \frac{m_i}{\max_{j(m_j)}} \tag{3.7}$$

where $M = \begin{cases} 5; & \text{if priority is Top priority} \\ 4; & \text{if priority is what I really want to read} \\ 3; & \text{if priority is what I will read it} \\ 2; & \text{if priority is what I might read it} \\ 1; & \text{if priority is what I don't really want to read it} \\ 0; & \text{if priority is what I have already read it} \end{cases}$

4) Number of groups contained in the posted paper (G)

This factor represents the number of groups that cites the paper. The number of groups in the paper may be a good indicator to suggest that the paper is popular and interesting. This number of group can be calculated by Equation (3.8). Let G be the score of the number of groups, mg_i is number of group, where $mg_i \in MG$, |g| be the number of groups that appear the same papers, and $\max(mg_j)$ =5.

$$G = \frac{mg_i}{\max_{j(mg_j)}} \tag{3.8}$$

where $MG = \begin{cases} 5; & if\,|g| \geq \max\left(mg_j\right) \\ 0; & if\,|g| = 0 \\ |g|; & otherwise \end{cases}$

3.2.3.2 StaticRank

The static score is a summation of the four factor scores. The value of static score is defined in the range 0-1. Let w be the weighted static score, i be the ranking number, where $\{i = 1, 2, 3, \ldots, n\}$. Equation (3.9) shows an example of static rank formula:

$$StaticRank_i = \omega_1 Y_i + \omega_2 C_i + \omega_3 P_i + \omega_4 G_i \tag{3.9}$$

where $\sum_{j=1}^{4} \omega_j = 1$

3.2.3.3 Combining Similarity Ranking with Static Ranking

Equation (3.11) shows *CiteRank* score using both similarity and static rank [35]. In addition, the weight is applied to each type of rank to find the optimal ranking.

$$CiteRank = (SimRank \times \omega_c) + \left(StaticRank \times (1 - \omega_c)\right) \tag{3.10}$$

Let ω_c be combined weighting score where $\{\omega_c = 0.9, 0.8, \text{and } 0.5\}$. The value of 0.9 means that the combination of similarity and static rank ratio is 90:10, whereas the weight 0.80 represents a 80:20 ratio. Equation (3.11) shows an example of weight ratio 80:20 ($\omega_c = 0.8$).

$$CiteRank = (SimRank \times 0.8) + \left(StaticRank \times (1 - 0.8)\right) \tag{3.11}$$

Weighting complements between similarity and static ranking, whichever is more relevant. Some analysis inferences will be is described in Section 4.3.

3.2.3.4 Combining Similarity Ranking with Year of Publication

Studies on combining similarity ranking with year of publication (*CSYRank)* are attributive to improving search results. The scores on year of publication factor are combined following recent years. Such a paradigm advocates new papers from recent years to get the highest score. Equation (3.12) shows the *CSYRank* scores using both similarity ranking *(SimRank)* and year publication *(YearRank)* scores. In addition, certain weights are applied for each type of rank to find the optimal ranking. Let ω_c be combined weighting scores chosen for the performance evaluation, where

{ω_c =0.9, 0.8, and 0.5}. Any lower rankings whose ω_c<0.5 are unattractive to the users. Equation (3.13) shows an example of weighting ratio 90:10 ratios (ω_c =0.9). The value of 0.9 means that the combination of similarity and year of publication ranking is 90:10. The description is described in Section 4.3.

$$CSYRank=(SimRank \times \omega_c)+(YearRank \times (1-\omega_c)) \tag{3.12}$$

$$CSYRank=(SimRank \times 0.9)+(YearRank \times (1-0.9)) \tag{3.13}$$

3.2.3.5 Hypothesis of Experiment 2

1) The null hypothesis:

H_0: There is no statistical difference among the means of NDCG at K=1- 20 of the ranking methods that are CiteRank (ω_c), CSYRank(ω_c), SimRank , StaticRank.

($\mu_{\text{CiteRank } (\omega c)} = \mu_{\text{CSYRank } (\omega c)} = \mu_{\text{SimRank}} = \mu_{\text{StaticRank}}$)

2) The alternative hypothesis:

H_1: Not all approaches are equal

($\mu_{\text{CiteRank } (\omega c)} \neq \mu_{\text{CSYRank } (\omega c)} \neq \mu_{\text{SimRank}} \neq \mu_{\text{StaticRank}}$)

3.2.3.6 Experimental setting

A total of 15 lecturers and Ph.D. students from Chulalongkorn University, Suansunadha Rajabhat University, and Nakhon Pathom Rajabhat University were recruited to be subjects in the experiments. Each subject was assigned to find research papers using the designated search engines. Ninety queries were asked where each subject performed 6 queries on 4 search engines that were *SimRank StaticRank*, CiteRank(ω_c), and *CSYRank*(ω_c). The subject specified their search criteria to investigate the results from each search program deployed. The result could be viewed by title name, abstract, and hyperlink to full text, where the numbers of results displayed per page were pre-defined.

Each subject was instructed to find research papers of their interest. They formulated their own queries to look for the designated or related research papers. The same queries were subsequently used for each search engine. Then, they were asked to rate the relevancy of the search result set using Likert five-point scale:

Score 0 is not relevant at all.

Score 1 is probably not relevant.

Score 2 is less relevant.

Score 3 is probably relevant.

Score 4 is extremely relevant.

The top 20 search results of each search engine were displayed for relevancy assessment. Since the subjects in this experiment were considered experts in the field, their relevancy ratings were treated to be trust-worthy perfect. Subjects could see the title name to link to full paper download and titleID of the document. However, specific sources of results obtained from each search engine were hidden from the subjects.

The relevancy ratings of each resource in the result set were recorded and used to rank the result set which in turn were used as the normalization constants for NDCG computation using K=1-20. One-way ANOVA was applied to measure the mean difference of NDCG scores at K= 1-20 from the four ranking methods. If the results from the F-value indicated a significant difference at the 0.05 level, the null hypothesis would be rejected. In addition, Levene's test was used to assess the equality of variances in the samples. If the significance value was greater than 0.05, this meant that the variance was equal. Result from ANOVA showed that there were significant differences among the groups as a whole. The subtle differences among the groups were further amplified by multiple comparisons. Moreover, the Tukey *post-hoc* and LSD tests were employed to test the equality of variances. The Dunnett T3 was used to test whether there would be any differences in the variances.

3.2.4 Experiment 3: User Profile and Personalized Re-ranking

The experiment 3 will solve Problem 3 using a personalized re-ranking from the user profiles.

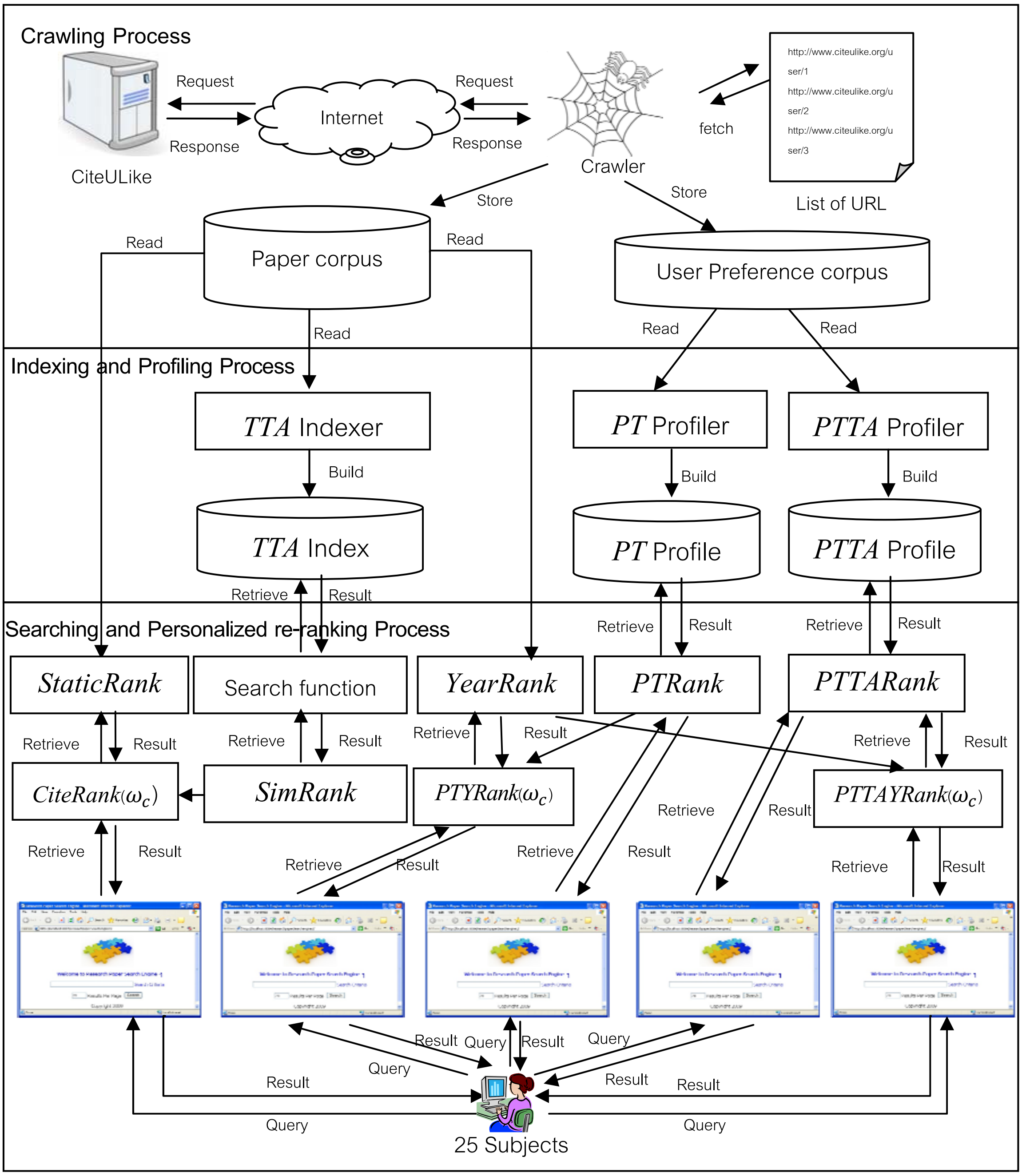


Figure 3.4: Framework of Experiment 3.

The profiles were built from a list of research papers posted on the bibliographic bookmarking system. Figure 3.4 shows a Framework of Experiment 3. This experiment focuses on answering these two questions: 1) what is the technique for creating user profile in bibliographic searching? and 2) how can the query-independent ranking be combined with personalized re-ranking to improve search results? Discovering the answers for the questions could improve the capability of the search.

In the experiment, two profile techniques were developed. The combinations of each profiler with year of publication factor were built. The premise of this combination is that the factor from static ranking can improve user satisfaction. The process is divided into three parts: user profile, combining personalized re-ranking with property factors, and personalized re-ranking with combined ranking. The detail of user evaluation is described in Section 3.2.4.4.

3.2.4.1 User Profile

To test the potential of user profile, two profilers were created: profiling based on user self-defined tags, and profiling based on user defined tags, title, and abstract. The profile requires the following supplementary definitions:

Definition: [T User Profile]: For a user u_i, $i=1,.., n$, let UPT_i be the relationship represented by u_i's tag and item set, $UPT_i = \{< u_i, utt_{ij} > | utt_{ij} \in UTT, u_i \in U,$ and is specified user u_i using the term utt_{ij} for document $d_k \}.$

Definition: [TTA User Profile]: For a user u_i, $i=1,.., n$, let $UPTTA_i$ be the relationship represented by u_i's tag, title, and abstract and item set, $UPTTA_i = \{< u_i, utta_{ij} > | utta_{ij} \in UTTA, u_i \in U,$ and is specified user u_i using the term $utta_{ij}$ for document $d_k \}.$

Table 3.1 shows personalized re-ranking algorithm. The frequency of each term is calculated. Ranking of search results are ordered from the highest to the lowest similarity scores. The cosine similarity is used to compare user profile with document search results of the research papers as shown in Equation (3.14). The personalized re-ranking based on user self-defined tags profile or "*PTRank*" profile is shown in Equation (3.15). Similarly, personalized re-ranking based on user defined tags, title, and abstract, or "*PTTARank*" profile is shown in Equation (3.16).

$$PersonsalizeRank = Sim(ut, st) = \frac{ut \bullet st}{\|ut\|\|st\|} \quad (3.14)$$

$$PTRank = Sim(utt,st) \quad = \frac{utt \bullet st}{\|utt\|\|st\|} \quad (3.15)$$

$$PTTARank = Sim(utta,st) \quad = \frac{utta \bullet st}{\|utta\|\|st\|} \quad (3.16)$$

Table 3.1. Personalized re-ranking algorithm

```
Personalized re-ranking technique
Input :user profile term ($UT_{ij}$) and set of term document search result ($ST_{xj}$)
Output :all rank of the similarity score from $UT_{ij}$ and $ST_{xj}$

Step1: Create a three dimensional array matrix $A$ . The size of matrix $A$ is the number
    of term profile ($UT_{ij}$).
Step2: Set the value of frequency term in matrix $A$.
   i. Get content term of $UT_{ij}$ from user profile corpus and put it in matrix $A$[0][i].
  ii. Get term frequency of $UT_{ij}$ from user profile corpus and put it in matrix $A$[1][j].
 iii. Get the content term of each $ST_{xj}$ such as $ST_{x1}$
      1. Check term matching of $UT_{ij}$ and $ST_{xj}$
      2. If values match, put term frequency from $ST_{xj}$ that match with $UT_{ij}$ in matrix
         $A$[2][k]
      3. If no matching value, do the following steps:
         - Add the length of matrix $A$(length +1)
         - Put content terms that do not match in $A$[0][i+1]
         - Put a value "0" in $A$[1][j+1] if user profile does not have this term
         - Put term frequency that has a search result in $A$[2][k+1]
  iv. Repeat step 2.iii until exhausting all the terms from search result.
Step3: Compute similarity score of each ranked document with the user profile. The
    similarity score and paper ID are kept in matrix $B$.
Step4: Repeated step 2 , and step 3 until finishing all ranks.
Step5: re-rank by rearranging similarity scores from matrix $B$ and return the content to
the user.
```

3.2.4.2 Combination of Personalized Re-ranking with Property Factors

Combination of personalized re-ranking or *CombinePRank* score using both personalized re-ranking and static ranking factor is shown in Equation (3.17). The weight is applied for each rank of rank to find the optimal ranking. Let ω_c be combined weighting score where {ω_c = 0.5, 0.75, 0.8, 0.9}.

$$CombinePRank = (PersonalizedRank \times \omega_c) + (StaticFactor \times (1 - \omega_c)) \quad (3.17)$$

Two type of personalized re-ranking are developed, namely, personalization re-ranking based on tag profiles or *PTRank* and personalization re-ranking based on tag, title, and abstract profiles or *PTTARank*. Evaluation is to compare the performance of these two ranking methods with native bibliographic social bookmarking systems.

3.2.4.3 Hypothesis of Experiment 3

1) The null hypothesis:

H_0: There is no statistical difference among the means of NDCG at K= 1-20 of the ranking methods, i.e., CombinePRank (ω_c), CiteRank (ω_c), bibliographic social bookmarking.

$(\mu_{\text{CombinePRank}(\omega c)} = \mu_{\text{CiteRank}(\omega c)} = \mu_{\text{bibliographic social bookmarking}})$

2) The alternative hypothesis:

H_1: Not all means are equal

$(\mu_{\text{CombinePRank}(\omega c)} \neq \mu_{\text{CiteRank}(\omega c)} \neq \mu_{\text{bibliographic social bookmarking}})$

3.2.4.4 Experiment setting

Seventy five queries were asked by twenty five subjects, who were lecturers and Ph.D. students from Chulalongkorn University, Suansunadha Rajabhat University, and Nakhon Pathom Rajabhat University. Each subject performed 5 queries on all search engines that were *PTRank, PTTARank* , *PTYRank*(ω_c), *PTTAYRank*(ω_c), CiteRank(ω_c), and CiteULike. The queries were formulated according to their own interest. The same queries were used for each search engine. They subsequently rated the relevancy of the search result set using Likert five-point scale:

Score 0 is not relevant at all.

Score 1 is probably not relevant.

Score 2 is less relevant.

Score 3 is probably relevant.

Score 4 is extremely relevant.

The top 20 search results of each search engine were displayed for relevancy assessment. Since the subjects in this experiment were considered experts in the field, their relevancy ratings were treated to be trust-worthy and perfect. Subjects could see the titleID, title name, and link to download full paper. However, specific sources of results obtained from each search engine were hidden from the subjects.

The relevancy ratings of each resource in the result set were collected and used to rank the result set which in turn were used as the normalization constants for NDCG computation using K=1-20. One-way ANOVA was applied to measure the mean difference of NDCG scores at K= 1-20 from the four ranking methods. If the results from the F-value indicated a significant difference at the 0.05 level, the null hypothesis would be rejected. In addition, Levene's test was used to assess the equality of variances in the samples. If the significance value was greater than 0.05, this meant that the variance was equal. Results from ANOVA showed that there were significant differences among the groups as a whole. The subtle differences among the groups were further amplified by multiple comparisons. Moreover, the Tukey *post-hoc* and LSD tests were employed to test the equality of variances. The Dunnett T3 was used to test whether there would be any differences in the variances.

CHAPTER IV

EXPERIMENTAL RESULTS AND DISCUSSION

This chapter describes the results and discusses of the experiments. First, the experimental data and the development of queries used in the experiments are described. Second, experimented methods are described. The chapter concludes with discussion of the research results.

4.1 Data set of the experiment

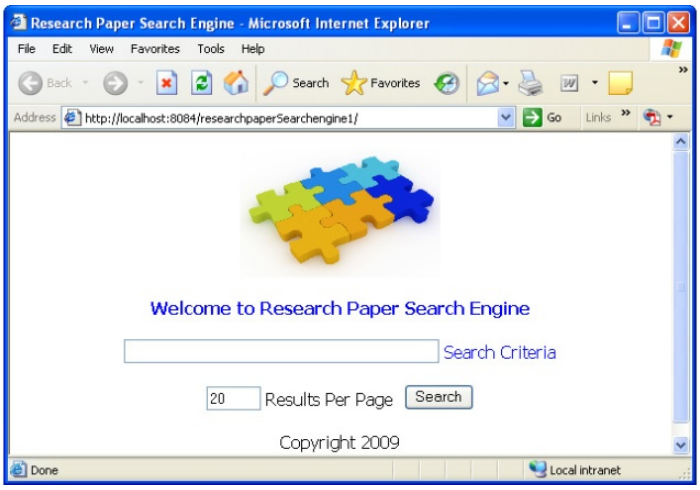
The crawler collected data from CiteULike during March 2009 to February 2012. The collected documents consisted of 175,210 research papers and 5,026 user profiles. Each record of the paper corpus contained tag of each paper, title name, title ID, abstract, and link to view full text article, posted date, posted time, and paper priority. The detail of data set is shown in Table 4.1.

Table 4.1: Detail of data set.

data	Record number
tag	2,136,711
research papers	175,210
user profile	5,026

4.2 Experiment 1: Research Paper Indexing

In this experiment, a search program was developed based on three heuristic indexers. Equation (3.2) was applied to the search program for creating the index. Then the search indexing using Equation (3.3) and (3.4) were created respectively. Figure 4.1 shows the interface web page of the search program. User could specify their search criteria and investigated the results from each search request. Results such as title, abstract, and the full text could be viewed directly.


Research Paper Search Engine - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8084/researchpaperSearchengine1/
Go Links
Done
Local intranet

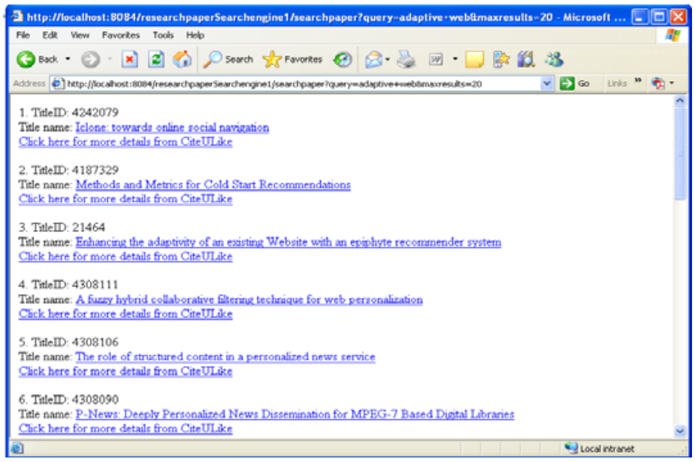
Welcome to Research Paper Search Engine

Search Criteria

20 Results Per Page Search

Copyright 2009

Figure 4.1: Research paper search engine web page.


http://localhost:8084/researchpaperSearchengine1/searchpaper?query-adaptive-web&maxresults-20 - Microsoft ...
File Edit View Favorites Tools Help
Back Search Favorites
Address http://localhost:8084/researchpaperSearchengine1/searchpaper?query=adaptive+web&maxresults=20
Go Links
Local intranet

1. TitleID: 4242079
Title name: Iclone: towards online social navigation
Click here for more details from CiteULike

2. TitleID: 4187329
Title name: Methods and Metrics for Cold Start Recommendations
Click here for more details from CiteULike

3. TitleID: 21464
Title name: Enhancing the adaptivity of an existing Website with an epiphyte recommender system
Click here for more details from CiteULike

4. TitleID: 4308111
Title name: A fuzzy hybrid collaborative filtering technique for web personalization
Click here for more details from CiteULike

5. TitleID: 4308106
Title name: The role of structured content in a personalized news service
Click here for more details from CiteULike

6. TitleID: 4308090
Title name: P-News: Deeply Personalized News Dissemination for MPEG-7 Based Digital Libraries
Click here for more details from CiteULike

Figure 4.2 : Example of search results.

4.2.1 Results of Indexing

The top fifteen ranks of average NDCG scores are shown in Table 4.2.

Table 4.2: The average NDCG scores of top 15 ranks.

Rank	Average of NDCG Score		
no.	T	TA	TTA
1	0.4476	0.6106	0.7187
2	0.4733	0.6305	0.6748
3	0.4909	0.6240	0.6515
4	0.5171	0.6334	0.6329
5	0.5282	0.6274	0.6344

Rank no.	Average of NDCG Score		
	T	TA	TTA
6	0.5211	0.6225	0.6367
7	0.5227	0.6218	0.6281
8	0.5214	0.6197	0.6225
9	0.5279	0.6189	0.6222
10	0.5331	0.6149	0.6267
11	0.5363	0.6075	0.6269
12	0.5417	0.6092	0.6248
13	0.5411	0.6076	0.6239
14	0.5421	0.6131	0.6236
15	0.5401	0.6128	0.6234

Figure 4.3 shows the graph of NDCG average score on three different search indexes. The x-axis denotes the first 15 ranks of the search results, and the y-axis represents the average NDCG score.

The graph shows that the *TTA* indexing method provides the best set of search results among the indexing methods.

One-way ANOVA was applied on NDCG for the top 5 rank respectively to test whether there was any difference in the NDCG average value among the three different indexing methods. As shown in Table 4.3 to 4.5, the test shows that not all of the NDCG averages of the three indexing methods are equal at α=0.05 levels of significance. In other words, the difference in the set of search results returned from the three different indexing methods are statistically significant, i.e., all of them are not equal.

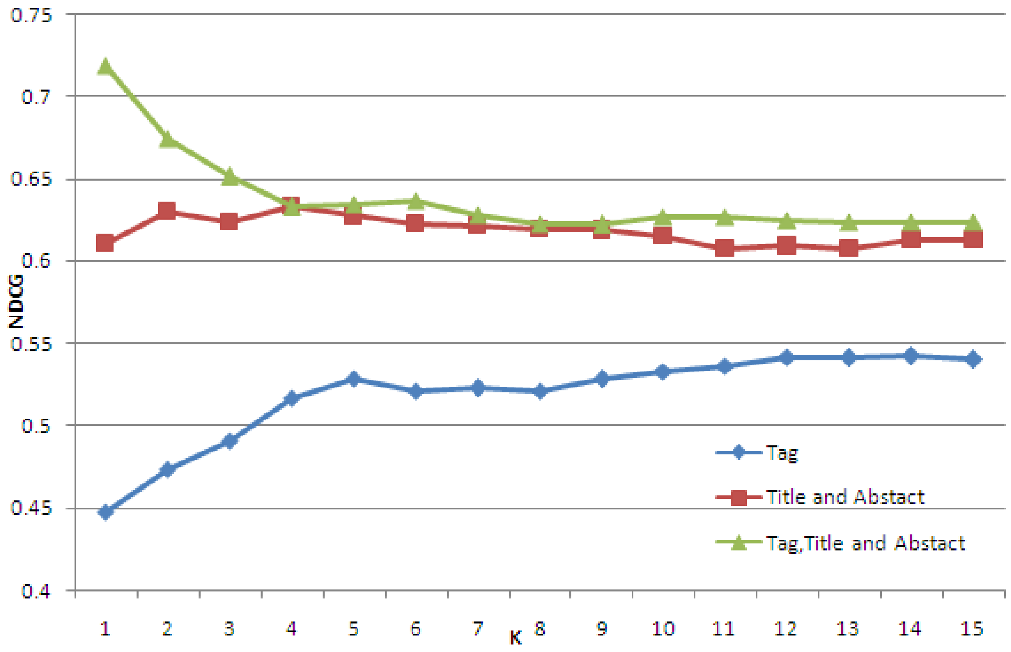


Figure 4.3 : Comparison of the average NDCG on the three indexing methods.

Table 4.3: Test of homogeneity of variances for Experiment 1.

NDCG

Levene Statistic	df1	df2	Sig.
4.204	2	672	.015

Table 4.4: ANOVA for Experiment 1.

NDCG

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3.635	2	1.818	15.533	.000
Within Groups	78.633	672	.117		
Total	82.269	674			

Table 4.5: Robust tests of equality of means for Experiment 1.

NDCG

	Statistic(a)	df1	df2	Sig.
Welch	16.201	2	447.516	.000

a Asymptotically F distributed.

Table 4.6: Results of multiple comparisons means for Experiment 1.

Dunnett T3

(I) TYPE	(J) TYPE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Tindex	TAIndex	-.1337(*)	.03257	.000	-.2117	-.0556
	TTAIndex	-.1709(*)	.03146	.000	-.2463	-.0955
TAIndex	Tindex	.1337(*)	.03257	.000	.0556	.2117
	TTAIndex	-.0373	.03271	.586	-.1157	.0411
TTAIndex	Tindex	.1709(*)	.03146	.000	.0955	.2463
	TAIndex	.0373	.03271	.586	-.0411	.1157

* The mean difference is significant at the .05 level.

The result from Test of Homogeneity of Variances is 0.015 which is less than 0.05. This emphasizes the alternate hypothesis.

The differences among the three indexing methods were performed by multiple comparisons. Table 4.6 shows the results of the multiple comparisons of the three difference indexing methods. Although the previous results suggest that the *TTA* indexing yields a better set of search results compared with the other two methods, multiple comparisons demonstrate that there is no statistically difference between *TTA* indexing and *TA* indexing. However, the mean difference of *TTA* indexing was higher than that of *TA* indexing. Therefore, this can be concluded that *TTA* indexing is a potential indexing choice.

4.3 Experiment 2: Static Ranking and Combination Ranking

From Table 4.7, the NDCG average scores are in the range of 0.5305-0.8671for *SimRank*, *StaticRank*, *CiteRank (50:50)*, *CiteRank (80:20)*.

Figure 4.4 shows the NDCG average scores of six different rankings on *SimRank* and *CiteRank* with 5 different weights. The y-axis denotes the NDCG score and the x-axis represents the first 20 documents of the search results.

Table 4.7: The average of NDCG scores for the first 20 rank of five different rankings.

Rank no.	Average of NDCG Scores				
	Sim Rank	***Static Rank***	***CiteRank (50:50)***	***CiteRank (80:20)***	***CiteRank (90:10)***
1	0.7610	0.5305	0.6996	0.7733	0.7638
2	0.7186	0.5402	0.6988	0.7342	0.7267
3	0.7112	0.5508	0.6862	0.7229	0.7264
4	0.7046	0.5579	0.6886	0.7244	0.7266
5	0.7031	0.5741	0.6906	0.7205	0.7203
6	0.7079	0.5900	0.6913	0.7163	0.7210
7	0.7117	0.5975	0.6968	0.7288	0.7234
8	0.7103	0.6020	0.7047	0.7273	0.7262
9	0.7224	0.6210	0.7090	0.7335	0.7308
10	0.7290	0.6360	0.7153	0.7411	0.7375
11	0.7358	0.6487	0.7256	0.7491	0.7430
12	0.7457	0.6635	0.7363	0.7646	0.7546
13	0.7553	0.6778	0.7499	0.7819	0.7675
14	0.7664	0.6923	0.7639	0.7946	0.7776
15	0.7805	0.7091	0.7770	0.8105	0.7913
16	0.7932	0.7295	0.7890	0.8192	0.8046
17	0.8083	0.7466	0.8035	0.8353	0.8189
18	0.8232	0.7614	0.8209	0.8491	0.8351
19	0.8424	0.7806	0.8380	0.8644	0.8481
20	0.8617	0.8053	0.8562	0.8854	0.8670

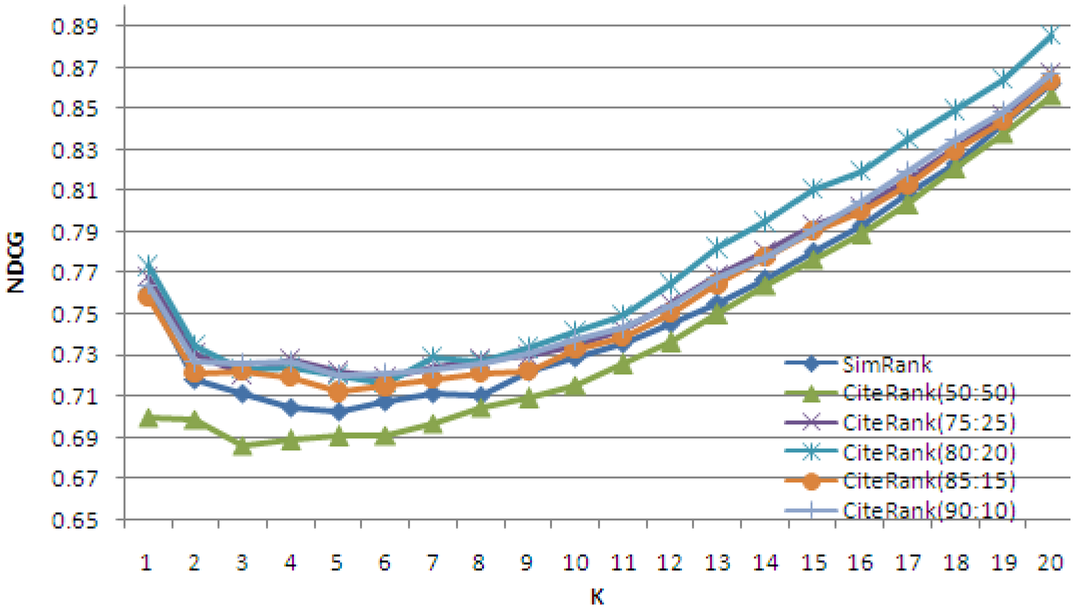


Figure 4.4 : Comparison of the NDCG average score on six ranking methods.

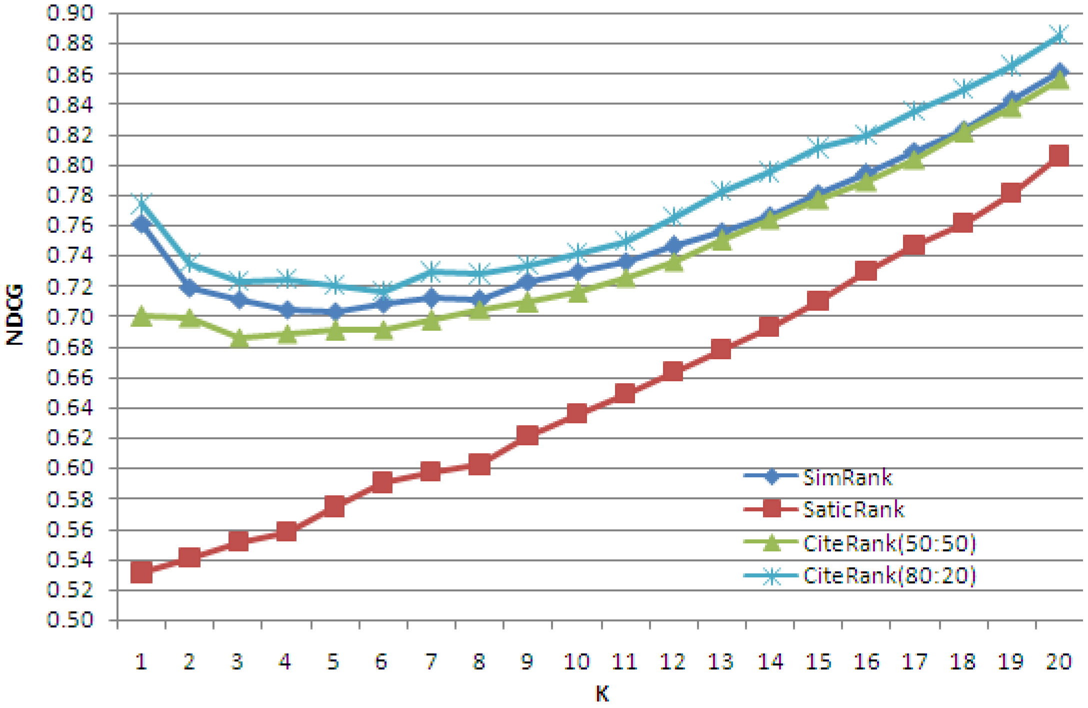


Figure 4.5: Comparison of the NDCG average score on four ranking methods.

Figure 4.5 shows the comparison of the NDCG average score of the four different rankings. It could be easily seen from the graph that *CiteRank (80:20)* had higher NDCG average scores than other ranking. Based on the combined similarity ranking with year of publication experiment, the average of the NDCG scores for

CSYRank with 3 different weight scores, namely, *CSYRank(50:50)*, *CSYRank(80:20*), and *CSYRank(90:10*), are shown in Table 4.8.

Table 4.8: The average of NDCG scores for the first 20 rank of three different rankings:*CSYRank*.

Rank no.	Average of NDCG Scores		
	CSYRank (50:50)	***CSYRank (80:20)***	***CSYRank (90:10)***
1	0.7208	0.7550	0.7646
2	0.6897	0.7157	0.7372
3	0.6728	0.7027	0.7301
4	0.6580	0.7010	0.7188
5	0.6565	0.7009	0.7160
6	0.6648	0.7017	0.7227
7	0.6691	0.7069	0.7189
8	0.6739	0.7073	0.7245
9	0.6867	0.7152	0.7320
10	0.6994	0.7227	0.7359
11	0.7100	0.7343	0.7421
12	0.7183	0.7442	0.7534
13	0.7311	0.7597	0.7653
14	0.7458	0.7737	0.7767
15	0.7626	0.7854	0.7919
16	0.7796	0.7989	0.8040
17	0.7982	0.8169	0.8159
18	0.8159	0.8311	0.8324
19	0.8343	0.8503	0.8472
20	0.8530	0.8687	0.8677

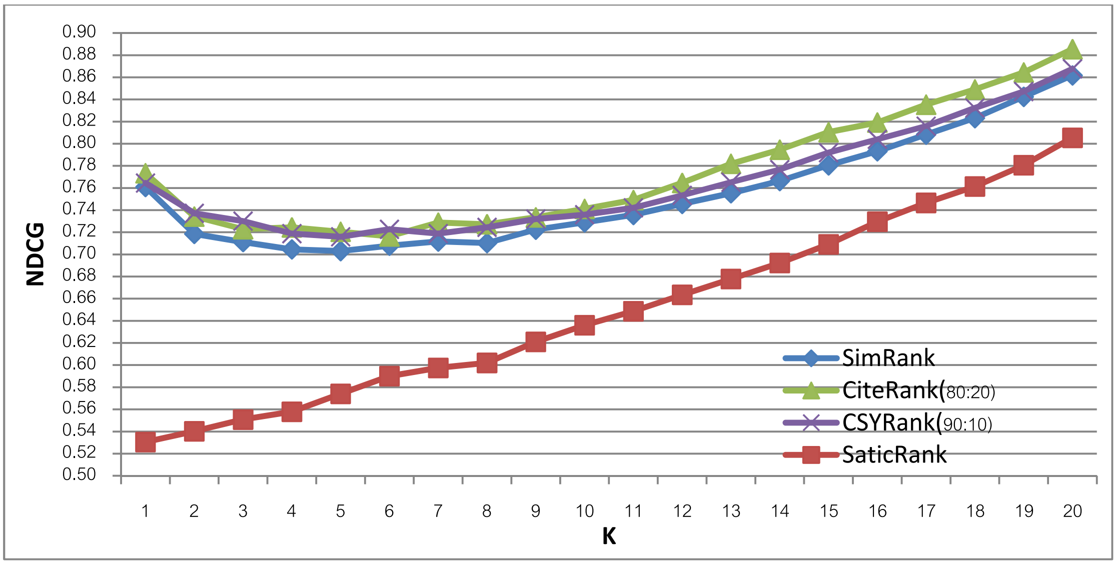


Figure 4.6: Comparison of the NDCG average score of four ranking methods.

Figure 4.6 shows the comparison of the NDCG average scores of the four different rankings. The x-axis represents the first 20 ranks of the search results and the y-axis denotes the NDCG score. The diamond symbol indicates *SimRank*, the triangle symbol indicates *CiteRank (80:20)*, the cross symbol indicates *CSYRank(90:10)*, and the square symbol on the graph indicates *StaticRank*. It could be easily seen from the graph that *CiteRank (80:20*) had higher NDCG average scores than other rankings.

Table 4.9: Test of homogeneity of variances for Experiment 2.

NDCG

Levene Statistic	df1	df2	Sig.
42.803	3	7211	.000

Table 4.10: ANOVA for Experiment 2.

NDCG

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	17.805	3	5.935	118.537	.000
Within Groups	361.040	7211	.050		
Total	378.845	7214			

Table 4.11: Robust tests of equality of means for Experiment 2.

NDCG

	Statistic(a)	df1	df2	Sig.
Welch	102.733	3	3994.386	.000

a Asymptotically F distributed.

Table 4.12: Results of multiple comparisons means for Experiment 2

Dependent Variable: NDCG
Dunnett T3

(I) TYPE	(J) TYPE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
SimRank	StaticRank	.103880(*)	.0077355	.000	.083518	.124242
	CiteRank (80:20)	-.019203	.0073855	.055	-.038643	.000237
	CSY (90:10)	-.010329	.0069295	.584	-.028569	.007910
StaticRank	SimRank	-.103880(*)	.0077355	.000	-.124242	-.083518
	CiteRank (80:20)	-.123083(*)	.0079437	.000	-.143992	-.102173
	CSY (90:10)	-.114209(*)	.0075216	.000	-.134008	-.094410
CiteRank (80:20)	SimRank	.019203	.0073855	.055	-.000237	.038643
	StaticRank	.123083(*)	.0079437	.000	.102173	.143992
	CSY(90:10)	.008873	.0071611	.767	-.009976	.027723
CSY(90:10)	SimRank	.010329	.0069295	.584	-.007910	.028569
	StaticRank	.114209(*)	.0075216	.000	.094410	.134008
	CiteRank (80:20)	-.008873	.0071611	.767	-.027723	.009976

* The mean difference is significant at the .05 level.

The result from Test of Homogeneity of Variances is 0.000 which is less than 0.05. This emphasizes the alternate hypothesis.

One-way and Welch test were applied on NDCG for the top 20 ranks to test whether there was a difference among the NDCG means of the four different rankings. The evidence was found that the difference in the set of search results returned from four different ranking approaches were statistically significance. In other words, not all of the means of NDCG of the four ranking are equal at α=0.05 levels of significance.

The multiple comparisons preformed to find the differences among four ranking methods. Table 4.12 shows the results of multiple comparisons of the four different rankings. The results from the multiple comparisons indicated that the set of mean difference search results provided by the *CiteRank (80:20)* combined weight ranking approach was statistically different from the set of search results provided by the *StaticRank* at K=1-20, and was statistically difference from the set of search results provided by the *CSY(90:10*) at K=1-20. Although the *CiteRank (80:20)* was no statistically difference from the set of search results provided by the *SimRank* method, the mean difference value of *CiteRank (80:20)* was higher than the mean difference value of *SimRank* and *CSY (90:10*).

4.4 Experiment 3: User Profile Combining Personalized Re-ranking with Property Factor

In experiment 3, two types of user profile were created and personalized re-ranking following in Equation 3.15, 3.16, and 3.17 were calculated. The results of the average NDCG score at the first 20 ranks of personalized re-ranking based on “tag only” profiles (*PTRank*) and combination of *PTRank* with *YearRank* are shown in Table 4.13. The six different ranking methods are 1) Personalized re-ranking of Tag User Profile or *PTRank*, 2)*PTYRank(50:50),* 3) *PTYRank(75:25),* 4) *PTYRank(80:20),* 5) *PTYRank (85:15),* and 6) *PTYRank (90:10)*. In addition, the personalized re-ranking based on “tag, title, and abstract” profiles (*PTTARank*) and combination of *PTTARank* with *YearRank* were shown in Table 4.14, based on the following six ranking methods:

1) Personalized re-ranking of the definition of TTA User Profile or *PTTARank*, 2) *PTTAYRank(50:50)*, 3) *PTTAYRank(75:25)*, 4) *PTTAYRank (80:20)*, 5) *PTTAYRank (85:15)*, and 6) *PTTAYRank (90:10)*.

Table 4.13 : The average of the NDCG scores for the first 20 rank of six different rankings: *PTYRank*.

Rank no.	Average of NDCG Scores					
	PT Rank	***PTYRank (50:50)***	***PTYRank (75:25)***	***PTYRank (80:20)***	***PTYRank (85:15)***	***PTYRank (90:10)***
1	0.6426	0.6392	0.6935	0.7238	0.7110	0.6929
2	0.6203	0.6251	0.6441	0.6644	0.6582	0.6475
3	0.6047	0.6145	0.6249	0.6327	0.6432	0.6379
4	0.5877	0.6045	0.6098	0.6317	0.6389	0.6252
5	0.5919	0.5949	0.6111	0.6230	0.6349	0.6207
6	0.5882	0.5869	0.6131	0.6119	0.6263	0.6176
7	0.5826	0.5799	0.6024	0.6043	0.6121	0.6075
8	0.5749	0.5747	0.5938	0.5990	0.6080	0.6024
9	0.5743	0.5718	0.5908	0.5932	0.6026	0.5994
10	0.5737	0.5711	0.5848	0.5892	0.5962	0.5952
11	0.5767	0.5689	0.5873	0.5920	0.5931	0.5934
12	0.5742	0.5670	0.5836	0.5871	0.5895	0.5885
13	0.5716	0.5647	0.5792	0.5850	0.5846	0.5873
14	0.5725	0.5626	0.5756	0.5846	0.5817	0.5864
15	0.5715	0.5621	0.5727	0.5754	0.5764	0.5807
16	0.5710	0.5617	0.5724	0.5749	0.5746	0.5790
17	0.5687	0.5602	0.5716	0.5747	0.5734	0.5770
18	0.5678	0.5623	0.5689	0.5727	0.5735	0.5770
19	0.5683	0.5616	0.5679	0.5725	0.5720	0.5740
20	0.5686	0.5626	0.5686	0.5733	0.5728	0.5747

Table 4.14 : The average of the NDCG scores for the first 20 rank of six different rankings: *PTTAYRank*.

Rank no.	Average of NDCG Scores					
	PTTA Rank	***PTTAY Rank (50:50)***	***PTTAY Rank (75:25)***	***PTTAY Rank (80:20)***	***PTTAY Rank (85:15)***	***PTTAY Rank (90:10)***
1	0.6769	0.6224	0.6482	0.6910	0.6863	0.7339
2	0.6546	0.6212	0.6396	0.6674	0.6512	0.6927
3	0.6466	0.6081	0.6231	0.6402	0.6342	0.6653
4	0.6323	0.5908	0.6077	0.6430	0.6273	0.6535
5	0.6215	0.5848	0.6060	0.6331	0.6262	0.6503
6	0.6101	0.5781	0.6070	0.6278	0.6162	0.6462
7	0.6087	0.5722	0.6007	0.6228	0.6076	0.6409
8	0.6043	0.5704	0.5947	0.6177	0.6040	0.6389
9	0.5991	0.5708	0.5947	0.6148	0.6014	0.6387
10	0.5962	0.5697	0.5877	0.6114	0.5973	0.6357
11	0.5942	0.5711	0.5881	0.6116	0.5979	0.6340
12	0.5926	0.5684	0.5872	0.6060	0.5946	0.6277
13	0.5908	0.5673	0.5839	0.6020	0.5910	0.6262
14	0.5884	0.5672	0.5818	0.6000	0.5876	0.6240
15	0.5860	0.5686	0.5806	0.5994	0.5853	0.6211
16	0.5868	0.5699	0.5808	0.5990	0.5843	0.6212
17	0.5850	0.5698	0.5805	0.5978	0.5825	0.6208
18	0.5847	0.5728	0.5792	0.5993	0.5844	0.6225
19	0.5844	0.5727	0.5794	0.5995	0.5830	0.6208
20	0.5841	0.5736	0.5801	0.6009	0.5840	0.6222

The highest NDCG average score of each personalized re-ranking technique from Table 4.13 and Table 4.14 were selected and shown in Table 4.15. Four different ranking methods are 1) Combination ranking of similarity ranking with static ranking *(CiteRank)* in 80:20 ,2) *PTYRank (80:20)* 3) *PTTAYRank (90:10)*, and 4) CiteULike.

Table 4.15 : The average of the NDCG scores for the first 20 rank of four different rankings.

Rank no.	Average of NDCG Scores			
	CiteRank (80:20)	***PTYRank (80:20)***	***PTTAYRank (90:10)***	CiteULike
1	0.7098	0.7238	0.7339	0.4284
2	0.6914	0.6644	0.6927	0.4246
3	0.6785	0.6327	0.6653	0.4099
4	0.6674	0.6317	0.6535	0.4254
5	0.6565	0.6230	0.6503	0.4271
6	0.6420	0.6119	0.6462	0.4294
7	0.6420	0.6043	0.6409	0.4332
8	0.6316	0.5990	0.6389	0.4332
9	0.6207	0.5932	0.6387	0.4358
10	0.6169	0.5892	0.6357	0.4365
11	0.6120	0.5920	0.6340	0.4371
12	0.6137	0.5871	0.6277	0.4306
13	0.6130	0.5850	0.6262	0.4302
14	0.6072	0.5846	0.6240	0.4292
15	0.6059	0.5754	0.6211	0.4284
16	0.6023	0.5749	0.6212	0.4311
17	0.5990	0.5747	0.6208	0.4296
18	0.5954	0.5727	0.6225	0.4264
19	0.5928	0.5725	0.6208	0.4249
20	0.5936	0.5733	0.6222	0.4276

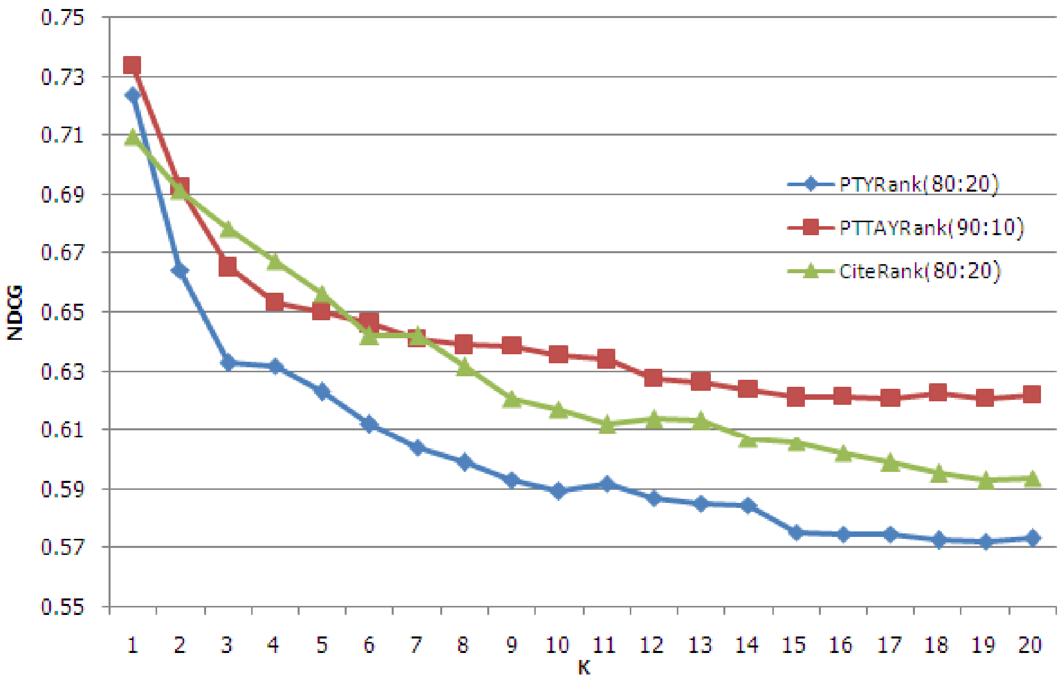


Figure 4.7: Comparison of the NDCG average score among three ranks method.

The results of weighting score show that the combination of personalized re-ranking and year of publication at 90:10 ratio or *PTTAYRank(90:10)* has higher score than *PTYRank(80:20)* and outperforms other ranking methods. Surprisingly, the NDCG score from CiteULike is the lowest. To test the difference of *CiteRank(80:20), PTYRank (80:20)*, and *PTTAYRank (90:10)*, the hypotheses are established as follows:

H_0: There is no statistical difference among the means of the NDCG at K=1-20 of the *CiteRank (80:20), PTYRank (80:20)*, and *PTTAYRank (90:10)*.

$(\mu_{\text{CiteRank}} = \mu_{\text{PTYRank 80:20}} = \mu_{\text{PTTAY 90:10}})$

H_1: Not all means are equal

$(\mu_{\text{CiteRank}} \neq \mu_{\text{PTYRank 80:20}} \neq \mu_{\text{PTTAY 90:10}})$

Table 4.16 : Test of homogeneity of variances for Experiment 3.

NDCG

Levene Statistic	df1	df2	Sig.
.302	2	4197	.739

Table 4.17 : ANOVA for Experiment 3.

NDCG

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.994	2	.497	7.489	.001
Within Groups	278.523	4197	.066		
Total	279.517	4199			

Table 4.18: Results of multiple comparisons means for Experiment 3.

Dependent Variable: NDCG

Type	(I) TYPE	(J) TYPE	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
Tukey HSD	PTYRank (80:20)	PTTAYRank (90:10)	-.0365(*)	.00974	.001	-.0593	-.0137
		CiteRank (80:20)	-.0263(*)	.00974	.019	-.0491	-.0035
	PTTAYRank (90:10)	PTYRank (80:20)	.0365(*)	.00974	.001	.0137	.0593
		CiteRank (80:20)	.0102	.00974	.547	-.0126	.0330
	CiteRank (80:20)	PTYRank (80:20)	.0263(*)	.00974	.019	.0035	.0491
		PTTAYRank (90:10)	-.0102	.00974	.547	-.0330	.0126
LSD	PTYRank (80:20)	PTTAYRank (90:10)	-.0365(*)	.00974	.000	-.0556	-.0174
		CiteRank (80:20)	-.0263(*)	.00974	.007	-.0454	-.0072
	PTTAYRank (90:10)	PTYRank (80:20)	.0365(*)	.00974	.000	.0174	.0556
		CiteRank (80:20)	.0102	.00974	.295	-.0089	.0293
	CiteRank (80:20)	PTYRank (80:20)	.0263(*)	.00974	.007	.0072	.0454
		PTTAYRank (90:10)	-.0102	.00974	.295	-.0293	.0089

* The mean difference is significant at the .05 level.

The results from Test of Homogeneity of Variances table show significance value equals 0.739 which is more than 0.05. Therefore, the two variances are significantly equal.

The results of One-Way ANOVA show that the significance value is 0.001. There is a significant difference between the two groups (the significance is less than 0.05). Therefore, the results demonstrate a significant difference between the three ranking methods.

The experimental results describe some implications of the proposed personalized re-ranking method, *PTTAYRank (90:10)*. The novel technique exploits social tagging for creating user profiles combined with static ranking: year of publication to retrieve more recent papers. The combination increases the efficiency of personalized re-ranking for academic bookmarking systems. The reasons are that the method utilizes the information of user's behavior and the important factors of academic papers that affect user interest, hence yielding better search results. The experiment suggests that *PTTAYRank (90:10)* outperforms *PTYRank (80:20)* from mean difference value. Social tagging combined with title and abstract are suitable for creating user profile since the amount of data is not large. However, the content of the paper or *TTA* for this particular study is still important for large data such as indexer creation. Moreover, the chosen experimental profiles and factors can help the system adjust the ranking and improve search results of academic paper. The results are better than CiteULike because the existing web is filtering system based only on tags which might not obtain as recent papers as the proposed method.

4.5 Discussion

The three problems to improve relevant search results of social bibliographic searching can be further discussed as follows:

There are some indications that the result from the proposed heuristic indexing method *TTA* index is a good potential indexing owing to the use of “Tag, Title, and Abstract” as indexing method. In other words, the information from social tagging with title and abstract of each paper can represent the content of paper better than using tag alone.

The results of the proposed combined ranking methods, i.e., *CiteRank* and *CSYRank* can improve bibliographic search engines*.* This might be because the method utilizes the information of user’s behavior. Especially *CiteRank(80:20),* a combination of the similarity ranking 80% and static ranking 20%, seemed to outperform other weight ratios. This was attributed to the fact that many researchers prefer to read more recent paper or just-posted paper. They also loved to read a classical paper, which was posted across different user groups or communities. The users could get a good paper from priority rating of the user posting. Moreover, *CSYRank(90:10)*, the combination of the similarity ranking 90% and *YearRank* 10% gave a second of highest scores. This was resulted from year of publication that helped improve ranking of search results. At any rate, the content of the paper, or *TTA* is still important for create indexing creation.

The experiment on user profile and personalized re-ranking experiment revealed that personalized re-ranking created from a profile by using *TTA* and year of publication at 90:10 ratio or *PTTAYRank (90:10)* returned the highest NDCG score than other rankings and CiteULike.

The result of personalized re-ranking implied that the important part of searching on bibliographic social bookmarking for each user was *TTA* which served as the best indexing and user profiling. The proper weight score (ω) ranged from 0.5 to 0.9. This value was biased for similarity ranking because user relevance results were more important than static ranking. Therefore, the weighting score of *PTTARank* at 90% meant that matching between user profile terms and search result terms from user query could affect the matching of user interest. The small weighting score of year of publication or *YearRank* at 10% implied that factors from query-independent such as year of publication could enhance personalized re-ranking. Therefore, both *PTTARank* and *YearRank* could improve the relevance of search result considerably.

CHAPTER V

CONCLUSION AND FUTURE RESEARCH

This research has examined how personalized re-ranking can be applied to the domain of bibliographic social bookmarking. More specifically, the task of item personalization was investigated based on the needs for research exploration. Some contributions and implications of this research are described, along with future work to be pursued.

5.1 Contributions and Implications of this Research

This research investigates the usefulness of different algorithms and information sources for personalized re-ranking of bibliographic social bookmarking systems. Three experiments were conducted. The results suggested that social tagging was suitable to create user profiles. The information from “Tag, Title, and Abstract” played an important rule to develop an indexer. Moreover, year of publication was found to enhance personalized re-ranking. Two type of personalized re-ranking were developed, namely, personalization re-ranking based on tag profiles or *PTRank* and personalization re-ranking based on tag, title, and abstract profiles or *PTTARank*. Futher investigation unveiled that combined personalized re-ranking methods led to efficient search results. The *YearRank* was combined with *PTRank* and *PTTARank* by weight score at 90:10 ratio yielding better performance than other rankings and tradition bibliographic social bookmarking systems.

It is apparent that the three research questions are answered. First, *TTA* is the solution to creating indexing. Second, the social tagging with Title and Abstract works well to create indexing and user profile. Third, the weighted score of year of publication and heuristic personalized re-ranking can be combined in the query-independent ranking to enhance search results. The experimental set up of this approach has been applied to academic search and social bookmarking systems.

Search results can be further enhanced by personalization competencies of the user profile which are used to set up indexers for creating personalized re-ranking.

5.2 Future Research

Researchers can apply personalized re-ranking technique that focuses on bibliographic social bookmarking by concentrating on other query independent factors such as user's groups, types of research paper, and names of principal investigator or paper's owner.

Semantic web ought to be incorporated to find the meaning between query with index corpus search terms and user profile terms, as well as synonym and closely related terms. This process will improve search results that match with the meaning of user query rather than just straightforward keyword matching.

Data mining technique can also be used to analyze the data set to enhance personalized re-ranking, especially explicit factors such as user profile and user's groups. Various data mining techniques such as association rule can be applied to determine the relation of item set between user's groups and rearrange ranking order.

REFERENCES

[1] Booth, P.F. Indexing the manual of good practice 2001 K.G. Saur: Indiana University, 2001.

[2] Manning C., Raghavan, P., and Schütze, H. Introduction to Information Retrieval, Cambridge University, 2008.

[3] Richardson, M., Prakash, A., and Brill, E. Beyond PageRank: Machine Learning for Static Ranking. Proceedings of the 15th International World Wide Web Conference, 2006, pp. 707–715, ACM, 2006.

[4] Micarelli, A., Gasparetti, F., Sciarrone, F., and Gauch, S. Personalized Search on the World Wide Web. Proceedings of the Adaptive Web, 2007, pp. 195–230, LNCS, 2007.

[5] Cameron, R. CiteULike. [Online]. 2004. Available from : http://www.citeulike.org [2013,January 29]

[6] Connotea Company. Connotea. [Online]. 2005. Available from: http://www.connotea.org [2013,January 29]

[7] Institute of Knowledge and Data Engineering. BibSonomy. [Online]. 2005. Available from: http://www.bibsonomy.org [2013,January 29]

[8] Emamy, K., and, Cameron, R. Citeulike: A Researcher's Social Bookmarking Service. Web Magazine for Information Professionals [Online]. 2007. Available from: http://www.ariadne.ac.uk/issue51/emamy-cameron/ [2013,January 29]

[9] Capocci, A., and Caldarelli, G. Folksonomies and Clustering in the Collaborative System CiteULike. Journal of Physics A: Mathematical and Theoretical Theory 41 (May 2008) : 1-9.

[10] Farooq, U., Ganoe, C.H., Carroll, J.M., and Giles, C.L. Supporting distributed scientific collaboration: Implications for designing the CiteSeer collaborator. Proceedings of the 40th Annual Hawaii International Conference on System Sciences, USA, 2007, pp.26, IEEE Compute Society, 2007.

[11] Farooq, U., Kannampallil, T.G., Song, Y., Ganoe, C.H., Carroll, John M., and Giles, C. Lee. Evaluating Tagging Behavior in Social Bookmarking Systems: Metrics and design heuristics. Proceedings of the 2007 international ACM conference on Supporting group work (GROUP'07), Sanibel Island, Florida, USA, 2007, pp. 351-360, ACM, 2007.

[12] Bogers, T., and Van, A.D.B. Recommending Scientific Articles Using CiteULike. Proceedings of the 2008 ACM conference on Recommender systems, pp. 287-290, ACM, 2008.

[13] Santos-Neto, E., Ripeanu, M., and Iamnitchi, A. Tracking Usage Attention in Collaborative Tagging Communities. Proceedings of International ACM/IEEE Workshop on Contextualized Attention Metadata: personalized access to digital resources Canada 2007, pp.17-23, IEEE, 2007.

[14] Golder, S. A., and Huberman, B. A. Usage patterns of collaborative tagging systems. Journal of Information Science 32, (April 2006):198–208.

[15] Suchanek, F. M., Vojnovic, M., and Gunawardena, D. Social Tags: Meaning and Suggestions. Proceedings of CIKM'08, 2008, Napa Valley, California, USA: ACM, 2008.

[16] Budura, T.,Michel, S.,Cudre-Mauroux, P., and Aberer, K..To Tag or Not to tag-Harvesting Adjacent Metadata in Large-Scale Tagging Systems. Proceedings of SIGIS'08, Singapore, 2008, pp.733-734, ACM, 2008.

[17] Gelernter, J. A Quantitative Analysis of Collaborative Tags: Evaluation for Information Retrieval—a Preliminary Study. Proceedings of International Conference on Collaborative Computing: Networking, Applications and Work sharing,2007, pp.376-381. New York, IEEE, 2007.

[18] Heymann, P., Koutrika, G., and Garcia-Molina, H. Can social bookmarking improve web search? Proceedings of the international conference on Web search and web data mining (WSDM), 2008, pp.195-206, ACM, 2008.

[19] Xu, S., Bao, S., Fei, B., Su, Z., and Yu, Y. Exploring folksonomy for personalized search. Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, 2008, pp.155-162, ACM, 2008.

[20] Baeza-Yates, R., and Ribeiro-Neto, B. Modern information retrieval. 1st, Addison-Wesley, 1999.

[21] Salton, G., and Buckley, C. Term weighting approaches in automatic text retrieval. Information Processing and Management 24 (1988) : 513–523.

[22] Bao, S., Wu, X., Fei, B., Xue Su, G., Z., and Yu, Y. Optimizing Web Search Using Social Annotations. Proceedings of the 16th international conference on World Wide Web, USA, 2007, pp.501-510, ACM, 2007.

[23] Hotho, A., J¨aschke, R., Schmitz, C., and Stumme, G. Information Retrieval in Folksonomies: Search and Ranking. Proceedings of Proceedings of the 3rd European conference on the Semantic Web: research and applications, 2006, pp.411–426, Springer Berlin / Heidelberg Press, 2006.

[24] Teevan J., Susan T. Dumais, and Horvitz E. Personalizing search via automated analysis of interests and activities. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 449-456, ACM, 2005.

[25] Sugiyama, K., Hatano, K., and Yoshikawa, M. Adaptive web search based on user profile constructed without any effort from users. Proceedings of Proceedings of the 13th international conference on World Wide Web, 2004, pp. 675–684, ACM, 2004.

[26] Sieg, A., Mobasher, B., and Burke, R. Web search personalization with ontological user profiles. Proceedings of the sixteenth ACM Conference on information and knowledge management (CIKM '07), 2007, pp. 525–534, 2007.

[27] Peng, W.-C., and Lin, Y.-C. Ranking web search results from personalized perspective. Proceedings of the 8th IEEE International Conference on E-Commerce Technology and The 3rd IEEE International Conference on Enterprise Computing, E-Commerce, and E-Services, 2006, pp.12, IEEE, 2006.

[28] Radlinski, F., and Joachims, T. Query chains. Learning to rank from implicit feedback. Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining, 2005 239-248, ACM, 2005.

[29] Zareh Bidoki, A.M., Ghodsnia, P., Yazdan,i N., and Oroumchian, F. A3CRank: An adaptive ranking method based on connectivity, content and click-through data. Journal of Information Processing and Management 46 (March 2010) : 159–169.

[30] Dou, Z., Song, R., and Wen, J.-R. A large-scale evaluation and analysis of personalized search strategies", Proceedings of the 16th international conference on World Wide Web, USA, 2007, pp. 581-590, ACM, 2007.

[31] Gemechu, E.F., Yu, Z., and Ting, L. Using Explicit Measures to Quantify the Potential for Personalizing Search. Research Journal of Information Technology 3 (2011) : 24-34.

[32] Berberich, K., Vazirgiannis, M., and Weikum, G. T-Rank: Time-Aware Authority Ranking. In Proceedings of WAW 2004, pp. 131–142, 2004.

[33] Choochaiwattana, W., and Spring, M.B. Applying Social Annotations to Retrieve and Re-rank Web Resources. Proceedings of International Conference on Information Management and Engineering, 2009, pp.215 - 219, IEEE, 2009.

[34] Liu, F., Yu, C., and Meng, W. Personalized web search for improving retrieval effectiveness. IEEE Transactions on Knowledge and Data Engineering, 16,(January 2004) : 28–40.

[35] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. CiteRank: combination similarity and static ranking with research paper searching. International Journal of Internet Technology and Secured Transactions (IJITST). 3 (April 2011) : 161-177.

[36] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. A Combination Ranking Model for Research Paper Social Bookmarking Systems. Proceedings of International Conferences on. Active Media Technology, China, 2011, pp.162–172, Springer-Verlag Berlin, 2011.

[37] Bharat, K., Kamba, T.,and Albers, M. Personalized, interactive news on the web. Journal of Multimedia Systems 6 (September 1998) : 349–358.

[38] Eirinaki, M.,and Vazirgiannis, M. Web mining for web personalization. ACM Transactions on Internet Technology 3 (February 2003) :1–27.

[39] Frias-Martinez, E., Magoulas, G. , Chen, S., and Macredie, R. Automated user modeling for personalized digital libraries. International Journal of Information Management. 26 (June 2006) : 234–248.

[40] Kang H., Joon Yoo, S., Han, D., Jang, H., and Yeon, H. Ranking Model of Medical Institutions based on Social Information and Sentiment Analysis of Reviews. International Journal of Digital Content Technology and its Application 6 (April 2012) : 275-284.

[41] Wang, J., Clements, M., Yang, J., de Vries, A. P., Reinders, M.J.T. Personalization of tagging systems. Information Processing and Management 46 (January 2010) : 58–70.

[42] Wu, Y.K., Wang, Y., Tang, Z.H. A Collaborative Filtering Recommendation Algorithm Based on Interest Forgetting Curve. International Journal of Advancements in Computing Technology 4 (2012) : 149-157.

[43] Icarelli, A., and Sciarrone, F. Anatomy and empirical evaluation of an adaptive web-based information filtering system. User Modeling and User-Adapted Interaction. 14 (June 2004) : 159–200.

[44] Pitkow, J., and others. Personalized Search: A contextual computing approach may prove a breakthrough in personalized search efficiency. Communication of the ACM 45 (2002) :50–55.

[45] Radlinski, F. and Dumais, S. Improving personalized web search using result diversification. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006, pp. 691 – 692, ACM, 2006.

[46] Gauch, S., Chaffee, J., and Pretschner, A. Ontology-based personalized search and browsing. Web Intelligence and Agent Systems 1 (December 2003) :219–234.

[47] You, G., and Hwang, S. Search structures and algorithms for personalized ranking. Journal Information sciences 178 (October 2008) :3925–3942.

[48] Finkelstein, G. and Hille, R.V. Re-Ranking Strategies for Ranking High Precision Information Web Search. International Journal of Engineering and Management Research 1 (2011) : 25-29.

[49] Ma, Z., Pant, G., and Liu Sheng, O. R. Interest-based personalized search. ACM Transactions on Information Systems, 25 (February 2007).

[50] Rijsbergen, C.J. van. Information retrieval. Butterworths, 1979.

[51] Datta, J. Submitted in the partial completion of the course CS 694, 16April 2010.

[52] Gao K., Wang, Y., and Wang, Z. An efficient relevant evaluation model in information retrieval and its application. Proceedings of International Conference on Computer and Information Technology, Los Alamitos, 2004, pp. 845-850, IEEE, 2004.

[53] Hatcher, E., and Gospodnetic, O. Lucene in Action. Manning Publications, 2005.

[54] Robertson, S. E., and Jones, K. S. Relevance weighting of search terms. Journal of the American Society for Information Science 27 (May-June 1976) : 129–146.

[55] N. Fuhr. Probabilistic models in information retrieval. The computer journal, 1992.

[56] Elbassuoni, S. Adaptive Personalization of Web Search. Master's Thesis, Department of Computer Science, Saarland University, 2007.

[57] Fagin,R., Lotem, A., and Naor, M. Optimal aggregation algorithms for middleware. Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, USA, 2001. pp.102–113, ACM Press, 2001.

[58] Jarvelin, K., and Kekalainen, J. IR evaluation methods for retrieving highly relevant documents. Proceedings of the International World Wide Web Conference, 2006, pp.41-48, ACM, 2006.

[59] Büttcher, S., L. A. Clarke, C.,and V. Cormack,G., Information Retrieval: Implementing and Evaluating Search Engines, MIT Press, 2010.

[60] Qualtrics Company. One-Way Analysis of Variance: A Guide to Testing Differences between Multiple Groups. [Online]. 2012. Available from: http://www.qualtrics.com/ university/researchsuite/docs/anova.pdf [2013,January 29]

[61] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. A personalized re-ranking technique for academic paper searching based on user profiles. International Journal of Digital Content Technology and its Applications 6 (2012) : 514-523.

[62] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. Improve Research paper Searching with social tagging - A Preliminary Investigation. Proceedings of the Eight International Symposiums on Natural Language Processing (SNLP2009), Bangkok, 2009, pp.152 - 156 ,IEEE, 2009.

[63] Pijitra Jomsri, Siripun Sanguansintukul, and Worasit Choochaiwattana. A Comparison of Search Engine Using “Tag Title and Abstract” with CiteULike – An Initial Evaluation. Proceedings of the 4th International Conference for Internet Technology and Secured Transactions (ICITST-2009), London, UK, 2009, pp. 1 - 5, IEEE, 2009.

APPENDIX

The appendix presents the list of database tables that are used in our experiment.

Table A.1 Research

Name	Type	Description
titleID	BigINT(20)	Research paper number
titleName	VARCHAR(600)	Title name
userPost	VARCHAR(100)	Name of user post
groupID	VARCHAR(100)	Group number
URLtoPaper	VARCHAR(100)	URL of paper in CiteULike
AUcount	INT(45)	Number of along with User
AGcount	INT(45)	Number of along with Group

Table A.2 Abstracttitle

Name	Type	Description
titleID	BigINT(20)	Research paper number
abstractP	VARCHAR(6000)	abstract
bookTitle	VARCHAR(3000)	Name of book or journal of paper
sFullText	VARCHAR(100)	Group number
postAt	VARCHAR(100)	Time of paper post
StarNum	INT(45)	Priority of paper
year	INT(45)	Year of paper

Table A.3 Along group

Name	Type	Description
listID	BigINT(20)	auto
groupID	integer	Group number
alongPName	VARCHAR(100)	along with User Name
titleID	integer	Research paper number

Table A.4 Author

Name	Type	Description
listID	BigINT(20)	auto
titleID	integer	Research paper number
Author	VARCHAR(100)	Author name

Table A.5 Tag

Name	Type	Description
listID	BigINT(20)	Auto number
titleID	integer	Research paper number
CiteULikeTag	VARCHAR(500)	Tag of paper in CiteULike

Table A.6 Alonggroup user

Name	Type	Description
groupID	BigINT(20)	Group number
alongGName	VARCHAR(100)	along with group name
titleID	integer	Research paper number
alongGID	integer	along with group number

BIOGRAPHY

Name: Ms. Pijitra Jomsri

Date of Birth: 29th January, 1981

Education:

- M.Sc. Program in Computer Science, Faculty of Science, Silpakorn University, Thailand.
- B.Sc. Program in Statistics, Faculty of Science and Technology, Thammasat University, Thailand.

Publications:

- Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W., "A personalized re-ranking technique for academic paper searching based on user profiles" *International Journal of Digital Content Technology and its Applications*, Vol. 6, pp. 514-523, 2012.
- Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W., "CiteRank: combination similarity and static ranking with research paper searching", *International Journal of Internet Technology and Secured Transactions (IJITST)*, , vol. 3 , no.2, pp. 161-177,2011.
- Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W., "A Combination Ranking Model for Research Paper Social Bookmarking Systems", In Proceedings of AMT2011. pp. 162–172, September 7-9, 2011, Lanzhou, China.
- Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W. "A Framework for Tag-Based Research Paper Recommender System: An IR Approach", In Proceedings of AINA 2010, April 20-23, 2010, Perth, Australia.
- Jomsri, P. Sanguansintukul, S., and Choochaiwattana , W. (2009) "A Comparison of Search Engine Using "Tag Title and Abstract" with CiteULike – An Initial Evaluation" In Proceedings of ICITST-2009. November 9-12 ,2009. London ,UK.
- Jomsri, P., Sanguansintukul, S., and Choochaiwattana, W.(2009) ' Improve Research paper Searching with social tagging-A Preliminary Investigation'. , In Proceedings of SNLP2009. *October 20-21, 2009.* Bangkok.