Personalization of Web Search Using Social Information

A Thesis

submitted in partial fulfilment of the requirements for the award of the degree of

Doctor of Philosophy

in

Computer Science and Engineering

Submitted by:

Shubham Goel
(Registration No: 901503035)

Under the guidance of
Dr. Ravinder Kumar
Associate Professor, CSED

Thapar Institute of Engineering and Technology
Patiala-147004, Punjab, India

December 2019

Certificate

I, Shubham Goel, Regn. No. 901503035, hereby certify that the work, which is being presented in the thesis, entitled **Personalization of Web Search Using Social Information**, in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy submitted to the Computer Science and Engineering Department at Thapar Institute of Engineering and Technology, Patiala, Punjab, India is an authentic record of my own work carried out under the guidance of Dr. Ravinder Kumar. The matter presented in this thesis has not been submitted to any other institute for the award of any other degree.

Place: Patiala

Date: 4.8.2020

Shubham Goel

Shubham Goel
Regn. No. 901503035

This is to certify that the statement made above by the candidate is correct to the best of my knowledge.

Ravinder Kumar

(Dr. Ravinder Kumar)
Associate Professor
Computer Science Engineering Department
Thapar Institute of Engineering and Technology
Patiala-147004
Punjab (India)

Abstract

In the recent years, there has been a rapid proliferation in the size of internet as the advent of web 2.0 made an end-user capable of generating various kinds of data by means of Interactive Internet Applications (IIAs), *i.e.*, Facebook, Instagram, twitter, etc. This, in turn, posed a big challenge for the web search platforms to assist the web users in obtaining their desired information. The implication which further made the problem critical for search platforms is diversity in user's outlook towards the same thing. Therefore, to maintain the efficiency of search platforms, user's position must also be strengthened in web search through personalization in content generation. The work, presented in this thesis, proposes a new model for personalization of web search with a focus on the selection of composition corresponding to various supporting modules of an efficient personalization system. Collaborative tagging is one of the applications of IIA that facilitates a web user to annotate a web resource with a tag of interest. It is a first-hand information directly given by a user without any middleman modification, therefore, it is more reliable than any other source. Thus, this collaborative tagging information can be quite helpful in constructing User Interest Profile (UIP) and Resource Illustration Profile (RIP). UIP will provide a complete list of user preferences along with his level of interest in that preference, while RIP enlists the topics about which a web resource describes or is related to, along with the degree of affinity for that topic. But the UIP constructed solely on the basis of user's own information is sparse and needs to be enriched with additional information. However, in the case of resource profile, RIP constructed through collective information from all the users is ambiguous as every user holds a different viewpoint or feeling towards a web resource. The conventional methodologies have failed to redress these problems.

The proposed model focuses on UIP enrichment using two different strategies. First one is clustering of tags based on the concept of semantic relatedness between two tags in the real-world. This has been measured using Word2vec model. The second one is the utilization of user's real society relationship network. It is believed that the present work is the first one to integrate the concept of semantic relatedness for tag clustering. A novel approach has also been designed to handle outlier tags which caused ambiguity based on the concept of collaborative filtering. Even a good UIP and RIP alone cannot create an efficient personalization system, they also require a suitable mapping with user's query requirements. Therefore, in the proposed model, the fuzzy satisfaction requirement-based novel mapping functions have been designed to measure query relevance score and user interest relevance score for a web resource. These scores have been further used to calculate the post-relevance score of a web resource after a suitable trade-off. Unlike

conventional methodologies, the proposed model calculates the trade-off parameter value after ascertaining which user has issued the query, *i.e.*, the value of trade-off parameter varies from one user-query pair to another. It is believed that the present work is the first of its kind to choose a trade-off parameter value based on user-query pair. A series of experiments have been conducted on a *del.icio.us* dataset to evaluate the effectiveness of the proposed model using different evaluation metrics. The results prove that the proposed personalization model has outperformed each and every baseline in relation to complete and efficient personalized web page ranking hierarchy.

Keywords: Personalized search, Collaborative tagging, User profiling, Resource profiling, Ranking, Web search

Acknowledgements

First and foremost, I would like to thank the Almighty who gave me enough strength and courage to overcome all the obstacles arsing during the period of this research work. I shall be lacking in sincerity and regard, if I do not acknowledge with thanks the sincere efforts of all those people who helped me to complete this work successfully.

I would like to express my deep gratitude to my supervisor, **Dr. Ravinder Kumar**, Associate Professor, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala (India), for his erudite guidance, constant advice and encouragement at every step of my Ph.D. program. Without his unfailing support, this thesis would not have been possible. His contribution to this thesis goes well beyond his role as an academic supervisor and includes regular support on a personal level without which this journey may never have been completed. And for this, I am truly grateful and will always remain indebted to him. Furthermore, I am equally thankful to **Dr. Nidhi Walia** w/o Dr. Ravinder Kumar for her moral support and motherly affection at many phases of my Ph.D.

I am also grateful to **Dr. Maninder Singh**, Professor and Head, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala (India), for providing me the necessary administrative assistance in the completion of this research work. I also wish to thank all the members of the doctoral committee, viz. **Dr. Inderveer Chana**, **Dr. Shalini Batra**, and **Dr. Sunil Kumar Singla** for their valuable inputs during my research work. I am also thankful to my Ph.D. Coordinator **Dr. Sushma Jain** and **Dr. Rinkle Rani**, former Ph.D. Coordinator, for their kind co-operation and support during the entire period of this study. I sincerely thank all the faculty members and the supporting staff of Computer Science and Engineering Department for their constant help whenever required especially **Dr. Karun Verma** and **Dr. Prashant Singh Rana**. I am also obliged to the Director, **Prof. Prakash Gopalan**, Dean (RSP), **Prof. Rafat Siddique**, and the management of Thapar Institute of Engineering and Technology, who provided me all the necessary resources and facilities to complete my work.

I would also like to acknowledge the valuable guidance and moral support of **Dr. Seema Bawa**, Professor, Computer Science and Engineering Department, Thapar Institute of Engineering and Technology, Patiala (India) extended to me during the whole period of this study. Her meditation sessions always helped me a lot to overcome all my difficulties with confidence. She has truly been a source of great inspiration for me; and I shall always remain thankful to her for showing me the path to have a positive look towards life.

I am also thankful to **Sh. Subhash Chander Rahi** of Punjabi University, Patiala; who have provided valuable suggestions to improve the quality and readability of this thesis.

I am highly indebted to my parents, **Dr. Anil Kumar Goel** and **Smt. Neelam Goel**, for their exceptional support that paved the way for a privileged education for me. They always stood beside me and encouraged me constantly throughout my life. I would also like to express my heartfelt thanks to my younger brother **Mr. Rajat Goel** who cared the parents much in my absence. I may not have been able to complete my research work without his cooperation and good wishes.

I would also like to thank all of my friends and colleagues, especially **Vijay Prakash Soni**, **Ashwani Kumar**, **Sandeep Saharan**, **Ajay Kumar**, **Mohd Abuzar Sayeed**, **Sahil Sharma**, **Sukhandeep Kaur Shergill** , and **Dr. Megha Bhushan** who always lent me the necessary support to face all oddities of life. Their nice co-operation and encouragement would always be remembered.

Date: (**Shubham Goel**)
Thapar Institute of Engineering and Technology
Patiala, 147004
Punjab, India

List of Publications

Journal Publications (SCI/SCIE):

1. Shubham Goel and Ravinder Kumar. "Brownian Motus and Clustered Binary Insertion Sort methods: An efficient progress over traditional methods", **Future Generation Computer Systems (FGCS)**, 86, pp. 266-280, 2018, **(SCIE Indexed, Impact Factor: 5.768)**
2. Shubham Goel and Ravinder Kumar. "Folksonomy-based User Profile Enrichment using Clustering and Community Recommended tags in Multiple Levels", **Neurocomputing**, 321, pp. 425-438, 2018, **(SCIE Indexed, Impact Factor: 4.072)**
3. Shubham Goel, Ravinder Kumar, Munish Kumar, and Vikram Chopra. "An Efficient Page Ranking Approach Based on Vector Norms using sNorm(p) Algorithm", **Information Processing and Management**, 56 (3), pp. 1053-1066, 2019, **(SCIE Indexed, Impact Factor: 3.892)**
4. Shubham Goel and Ravinder Kumar. "SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction", **Expert Systems With Applications**, 141, pp. not assigned till yet, 2019, **(SCIE Indexed, Impact Factor: 4.292)**
5. Shubham Goel and Ravinder Kumar. "Collaboratively Augmented UIP - Filtered RIP with Relevancy Mapping for Personalization of Web Search", **Information Sciences**, 2020. **(Accepted, SCIE Indexed, Impact Factor: 5.910)**

.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Information is the basic necessity of a person to perform the various tasks. In earlier days, libraries were considered to be the knowledge gateways and the prime source of information retrieval. In order to acquire information about a particular Topic of Interest (TOI), a person has to go through various books located in a library section related to a broader topic to which a TOI belongs. The index structure of the books enables a person to quickly find whether his TOI is present in a book or not irrespective of the quality of TOI information. The time and effort devoted by a person to acquire information about a TOI is directly proportional to user knowledge about relationship between the TOI and a broader topic to which a book is dedicated. However, with the passage of time, there has been a tremendous increase in the number of books and the index structure as well which made the manual information retrieval process difficult and inefficient. Thus, it gave rise to the need for designing an automatic search mechanism.

1.1 Motivation

The advent of web 2.0, the web has transitioned from author centric to user centric, *i.e.*, users are not just limited to consume an information provided by some authors, but are able to generate any information by themselves. A large number of Interactive Internet Applications (IIAs) based on the provisions of web 2.0 were built to allow a user to exchange data over web with other users. Many types of content like text, images, videos, graphics, animations, etc. are regularly published over the web by various IIA users. Some of the common and widely used IIAs are blogging, wiki, social networking, e-commerce, etc. As a result of IIAs usage and users' interest to share their knowledge with everyone, there is a rapid proliferation in the size of the web. According to current statistics, the web constitutes at least 5.39 billion web pages, 800 million YouTube videos,

and 20 billion images available only on Instagram at present [1, 2].

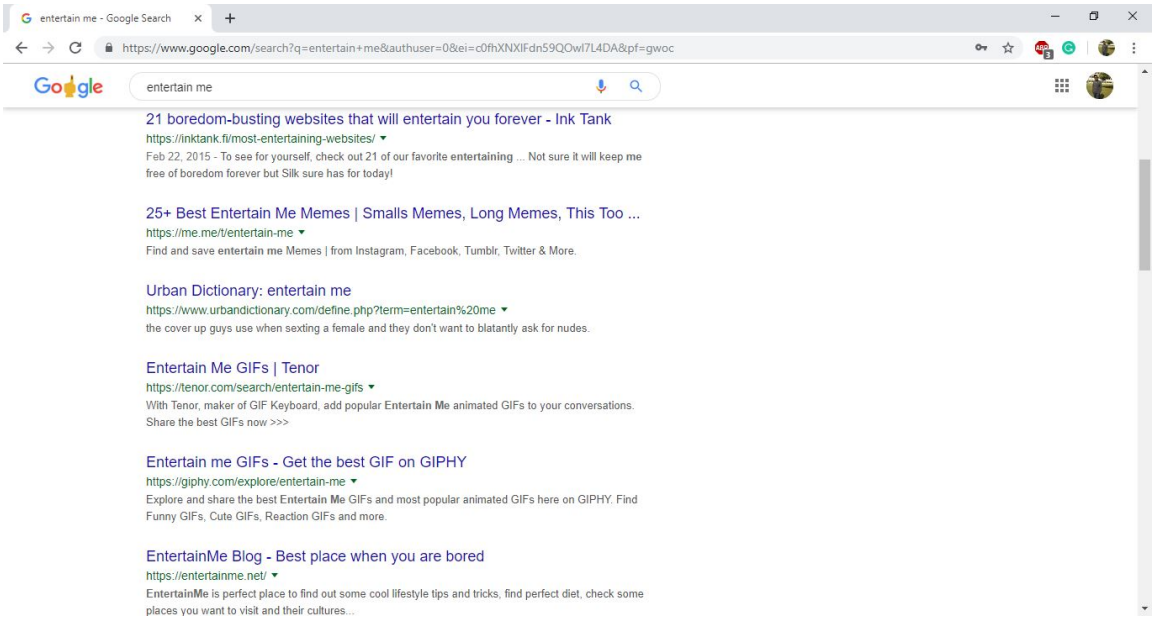
Generally, users make use of web search engines to seek information from the web as manual information retrieval has become infeasible. Some of the popular web search engines are Yahoo, Google, Bing, AOL, Ask.com, etc. According to a traditional perception, search engines are best miners of information from existing sea of information web. However, the enormous volume and unstructured nature of information over the web is causing a big threat to the information retrieval efficiency of search engines. Thousands of web resources are returned by a search engine for a search query, out of which only some are actually relevant to the Topic of Interest (TOI). For most of the users, some of the retrieved results are enough to serve their purpose, while others have no significance for them.

A similar set of web resources is, generally, retrieved by every search engine for the same query, but the ranking hierarchy, *i.e.*, arrangement of web resources in order of their relevance towards a queried TOI is different for all the web search engines. Page ranking algorithms are used as an integral part of the web search engines, so as to perform the ranking task of web resources. The information search quality and user's experience for information retrieval is highly regulated by the ranking efficiency of a page ranking algorithm. In primitive search engines, the ranking of web resources was performed based on the concept of content similarity matching strength which was calculated by counting the number of times a keyword appeared in a web resource. But with the advancement of technology, content-based spamming has become a popular means of fraudulent ranking of spammed web pages which result in failure of content-based ranking and the emergence of a new era of link-based ranking. PageRank [3], HITS [4], and SALSA [5] are some of the well-known link structure analysis-based page ranking algorithms which are used to perform the ranking of retrieved web resources. However, ranking assigned by these page ranking algorithms has not yet been quite efficient and incapable to enlist truly relevant web resources on initial few pages of search results due to the problem of link farming and link spamming. Moreover, the number of computation cycles required are also high. Distributed fashion of relevant web resources across the entire set of retrieved results is of no use for a user as according to web user psychology, a user rarely visits search results beyond the second page. Therefore, to tackle the problem in traditional link structure

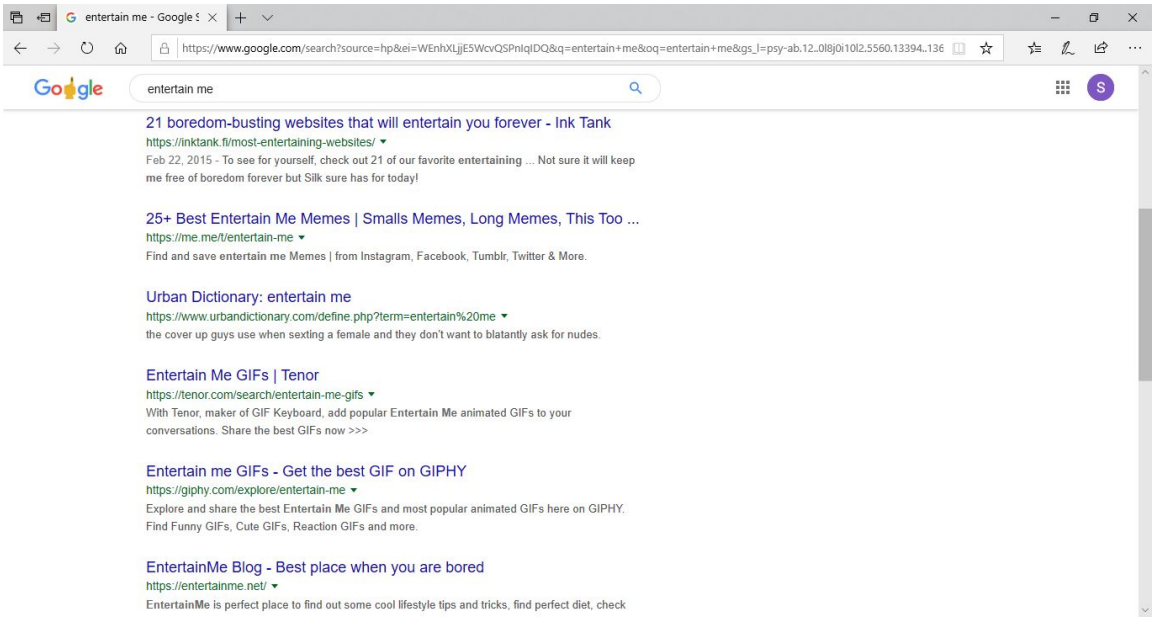
analysis-based page ranking algorithms, Goel et al. [6] devised a variant of SALSA to give $s\text{Norm}(p)$ for the efficient ranking of web pages. Their approach relies on p-Norm from a Vector Norm family for the ranking of web pages as Vector Norms can substantially reduce the impact of low authority weights in the process of computing hub weight for a node.

Presently, with the growing size of web, there has also been a tremendous increase in the number of users, *i.e.*, accounting to about 80 percent during the past 5 years with a current total of 3.578 billion users. Each user on the web exhibits a different viewpoint towards the same information. But most of the web search engines return a common result set with approximately same ranking hierarchy to the different users, irrespective of user preferences, issuing a same query to a particular web search engine. However, in the current scenario of information requirement, the approach of *"one size fits all"*, *i.e.*, similar package of results for everyone is completely undesirable as users have varied requirements for the same query. For instance, a query "entertain me" has been issued by two different users, *i.e.*, u_1 and u_2 to the same web search engine as shown in Fig. 1.1. Out of these two users, u_1 is a movie freak and likes to watch Hollywood movies for entertainment, while u_2 loves to watch comedy shows for the same purpose. Nevertheless of it, both of them will be given a similar result set by the web search engine as shown in Fig. 1.1 irrespective of their preferences.

Apart from these problems, two more problems, *viz.*, polysemy and synonymy have also evolved in current information retrieval scenario. Polysemy means a keyword having multiple meanings when used in a different context. For example, a user has written a query "bank" which can confuse a search engine to decide whether with a reference to query "bank" the user wants to obtain information about a *river bank* or *financial bank*. Usually, the search engine will return a mixture of results, *i.e.*, river and financial bank to a user, prioritizing the ones which are more popular to be ranked at top positions. But executing a web search without any consideration to user context or preference by a search engine will just results in re-formulation of queries, wastage of time, frustration or even void search sessions. Another most important problem of information retrieval is synonymy which results in retrieval of incomplete information. Synonymy means multiple keywords having the same meaning. For example, "amazing" keyword has multiple synonyms like



(a)



(b)

Figure 1.1: Results retrieved by Google for users (a) u_1 and (b) u_2 w.r.t. query "entertain me"

incredible, astonishing, extraordinary, fabulous, etc. Synonymy problem can lead to retrieval of search results only for keyword amazing as query contains keywords "amazing sweets". Rest all synonyms of "amazing" are not considered by the search engine, thus, many relevant results corresponding to user's query are not retrieved which is absolutely undesirable for high level of user satisfaction. All the aforementioned problems faced by current information retrieval system are the prime cause for degradation in the efficiency of search engines. Together these problems lead to an increase in the proportion of ir-

relevant web resources in a result set, inappropriate ranking of relevant web resources, and inconsistency in user satisfaction level. Therefore, in order to effectively deal with the information retrieval problems and demands of information retrieval market, the position of an end-user must also be strengthened in the process of information retrieval by transitioning from a generalized search approach to the personalized one.

In a personalization enabled web search system, the search engine utilizes the information regarding users' preferences or interest to tailor the web search results for them. Basically, personalization is the process of search results customization with respect to user's, *i.e.*, query issuer's preferences. In order to enlist the user preferences, his User Interest Profile (UIP) must be constructed. However, predicting the preferences or interest of every user in a personalized web search system is a very challenging task [7, 8].

Numerous researchers have analyzed various aspects of content generation behaviour or patterns followed by different web users in order to construct their UIP. So, the task of personalizing the web search of these users can be accomplished based on their predicted UIP. Some researchers have used browsing history [9–12], desktop files [13, 14], clickthrough information [15, 16], and social information [17–19] to predict the list of user preferences, *i.e.*, UIP. However, the raw information for UIP construction can be broadly categorized into the following two categories:

(i) *Explicit:* In this type of category, a personalization system asks the users to provide information about themselves [20, 21]. For example, while creating a new Facebook account a user provides every detail about himself like name, education, political interest, family information, work profile, etc.

(ii) *Implicit:* In this category, information about user's interest and behaviour is indirectly inferred from social network, online activities performed or searching patterns [18, 22], *e.g.*, using comments given by user on some online article, browsing history, articles shared online, various tagging actions on collaborative tagging sites, etc.

UIP constructed using implicit information about a user is more commonly used by the researchers in comparison to the explicit one due to certain unavoidable significant deficiencies in its information generation pattern. Its main deficiency is unwillingness of user

to spend extra effort and time to provide his information, *i.e.*, interest and preferences explicitly to the system. The second is inability of the user to accurately formulate the context of his preferences as it is highly dependent on the expertise of a user. Moreover, in implicit information category for UIP construction also, the profiles obtained through social information are found to be more effective than all other methods.

Today, collaborative tagging sites (like MovieLens, Del.icio.us, and Flicker); social network (like LinkedIn, Facebook, and WhatsApp); and microblogging sites (like Friendfeed,Yammer and Twitter) are the prominent platforms under the social information category. People are using these platforms as a means to communicate with each other, comment news, like, share, feedback on web resources (online articles videos and images), etc. The tagging actions performed by various users as a bookmarking service to numerous web resources using the collaborative tagging platforms are a good source to predict the list of user preferences as they are direct word-of-mouth from a user [23–28]. Del.icio.us is used by a user to tag various URLs, Flicker for image tagging, Last.fm for songs, MovieLens for images, etc. These collaborative tasks are very important factors enabling the people, knowingly or unknowingly, to generate more and more information about them on the web. Realising the popularity of collaborative tagging, many other social networking platforms have also started supporting user defined tags to various web contents in order to draw users' attention.

A user profile constructed solely on the basis of user's own activities will result in a sparse and inefficient profile as the performance of UIP is greatly influenced by the amount of user's information taken as input to UIP construction methodology. For some users, abundant information is available, while for others it is very scarce which is absolutely undesirable for effective working of a personalization approach. The scarcity of information does not mean that user has a very limited interest or preferences, but it all depends on "How much a user is active ?" on the social web. It can also be said that amount of user's information available on web is directly proportional to the frequency of social network activities of a user. The performance of UIP directly affects the efficiency of a web search personalization approach as UIP of a user is the backbone of every personalization algorithm designed either for web search or recommender systems [14, 29]. Thus, to avoid injustice on account of an inactive user, some additional information must be

linked with the user's account for the construction of a strong UIP; and the process of linking additional information is known as UIP enrichment.

The researchers have used different strategies for the enrichment of user profile by exploiting the activities performed by user and other activities occurring in the user's social network. Some of them have used the concept of tag clustering [30], community information [7], other tagging action on the same web page [31], sentiments aspect, resource correlations [32], etc. The additional acquired information about a user preferences will help to enrich the user profile and create a good UIP. But in this thesis, it is argued that for the enrichment of user profile a single strategy is not sufficient in today's time. There must be a number of strategies with each using a different set of rules to embed various pattern recognition and learning algorithms for user interest prediction.

Apart from a strong UIP, a personalization approach must also be supported by a good Resource Illustration Profile (RIP) and an effective mapping mechanism to compute the post-relevancy score. RIP of a web resource will provide a summarized viewpoint that different users hold towards it. Basically, RIP provides the information that this particular web resource is related to these topics in a certain manner. The researchers have also analyzed it using collaborative tagging actions performed on a web resource to construct an RIP [33, 34]. The post-relevancy score of a web resource depends on the trade-off of query relevance and user interest relevance scores which are computed using Query-RIP mapping and UIP-RIP mapping respectively. The previous works suffer from the limitations that all the supporting modules of a personalization model have not been considered by most of the researchers. No doubt, UIP is a backbone of a personalization algorithm, but mapping is also equally significant. Some of the researchers who even analyzed every aspect had performed the mapping using keyword matching only which is completely undesirable. As in the present time due to the problem of synonymy, diversity in the level of user interest and resource relevancy for a topic have made the keyword matching an obsolete approach. Therefore, in this thesis, along with every supporting module of a personalization model, the problem of synonymy has also been tackled in a novel way.

Finally, the effectiveness of the devised model is evaluated in the context of web search. In the present scenario, whenever a query is issued to a web search engine under this model,

the originally retrieved list of web pages will be re-ranked. The re-ranking will ensure that the web search results returned to a query issuing user are ranked with most relevant web page, according to user preference and query relevancy at header of list, while other rankings appear as per their relevance values.

1.2 Contribution and Organization of the Work

1.2.1 Main Contribution of the Current Work

This thesis contributes significantly in the following manner:

- Provides a detailed discussion about various factors responsible for transitioning the web search mechanism from generalized to the personalized one. Moreover, various issues involved in degradation of different supporting modules of personalization have also been discussed.
- The study undertakes an extensive literature review regarding various supporting modules of a personalization system. Moreover, the different strategies used by various extant personalization methodologies for performing the task of a dedicated supporting module have also been compared. The selection of certain effective parameters for performance comparison of methodologies under study has been made with due attention.
- A novel model for web search personalization has been proposed after giving special attention to selection of supporting modules configuration.
- A multi-level architecture has been proposed for modeling the UIP of a user. Moreover, a dedicated set of protocols has also been formulated to work on corresponding level in order to mine the user information by utilizing the strategy selected for that level. The impact of both real-world relationship of user own tags and explicitly defined society relationships of a user has also been considered by the proposed UIP modeling methodology.
- A new methodology has also been proposed for modeling the RIP of a web resource as a sole qualitative UIP cannot lead to an efficient personalization model. No doubt, UIP is the backbone of personalization, but still it needs support of other

modules. While modeling the RIP of a web resource, the proposed methodology is capable of outlier detection and filtration to generate a collaboratively filtered RIP.

- This thesis also proposes a dedicated methodology for the calculation of post-relevancy score of a web resource under the proposed personalization model. The final ranking of web resources has been performed on the basis of post-relevancy score. But this score depends on trade-off between two sub-relevancy scores, *i.e.*, query relevancy and user interest relevancy score computed by mapping of Query-RIP and UIP-RIP respectively. Therefore, two separate methodologies have been formulated to compute these mapping scores as a fuzzy satisfaction problem.
- Extensive experiments have been performed on a collaborating tagging dataset of Del.icio.us to validate the personalized ranking of web resources obtained by the proposed personalization model.

1.2.2 Organization of the Work

This section provides a detail about the organization of the current work. The first chapter introduces us to the various aspects and issues of this study. It explains the need for a personalization system and also discusses the various issues involved in its development. However, there are six chapters in all.

Chapter 2 Undertakes the review of various methodologies used for the construction of different supporting modules of personalization. Of all the different supporting modules of personalization, UIP being the most important one, has been discussed in detail. The review covers not only the UIP modeling based on user's own information, but also UIP enrichment as simple UIP suffers from the problem of information sparsity. Apart from it, some widely used RIP modeling methodologies has also been reviewed. Finally, a discussion on different personalization strategies especially result re-ranking and query re-formulation has been presented. On the whole, the literature review covers both the aspects of user's information, *i.e.*, post-search information and social information for configuring different supporting modules.

Chapter 3 proposes a novel web search personalization model which formulates the personalization task as a web resource re-ranking problem. Then, the facts about personalization have been given which explains that it is not a standalone entity, but an

interrelated contribution of multiple supporting modules. After that a description of the dataset used for training and testing the proposed model has been provided. A description about the pre-processing steps used for transformation of unstructured data to structured data and removal of noisy elements has also been given. Some well-known evaluation metrics in the field of information retrieval used in the proposed model for performance quantization have also been described in this chapter. Further, unlike other traditional personalization systems, a specially designed algorithm used to perform sorting task in the proposed personalization model has been clearly described.

Chapter 4 explains the proposed methodology for three-level UIP modeling along with the protocols working at each level for mining user information based on selected strategy. Different state of the art methodologies for UIP modeling have also been described to act as the basis of comparison with the proposed UIP modeling methodology. It not only compare the performance of full-fledged final UIP using experimental evaluations, but also the performance of partial UIPs. Further, it clearly describes the results of different null hypotheses framed for performance evaluation of both partial and full-fledged final UIPs.

Chapter 5, first of all, proposes a novel methodology for RIP modeling taking into consideration the outlier problem faced by traditional resource profiles. Then, an innovative methodology for post-relevancy score calculation has been proposed in order to obtain the personalized ranking of web resources. But a post-relevancy score of a web resource cannot be calculated straightway as it is governed by two sub-relevancy scores, *i.e.*, query relevancy and user interest relevancy score. Therefore, two novel methodologies corresponding to these sub-relevancy scores have been proposed as there is a lot of mapping discrepancies between traditional approach and real-world problem formulation. This chapter also highlights the various experiments conducted to have a comparative analysis of the personalized ranking obtained by the proposed personalization model with other baselines methodologies.

Chapter 6 concludes the work by highlighting the contributions made towards search personalization domain of IR as well as the key findings of this work. Further, it also suggests the scope for future research in the area under study.

Chapter 2

Literature Review

The current research work aims at improving the users' satisfaction by way of personalizing their web search. However, a personalization model is not a single standalone entity, but an interrelated contribution of multiple supporting modules, *i.e.*, UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation. The final ranking of web resources has been performed on the basis of their post-relevancy score values. However, this chapter undertakes to review the various research studies conducted for the accomplishment of aforementioned supporting modules of personalization. Yet before reviewing the supporting modules, a general description about search personalization is provided here for better understanding of the research problem.

2.1 Search Personalization

Generally, every web search engine follows a similar kind of search process in which few words are entered by a user into a dedicated search box, and in return, a long list of search results is provided to a user by a search engine [29]. Although user interaction with search process appears to be simple, *i.e.*, query and corresponding result set as response, but many users make use of search engines to accomplish various complex tasks such as live cooking recipe of a delicious dish or holiday planning [29, 35, 36]. Before initializing a search process, the major challenge in front of a search engine is to predict the information requirement of a user's query which is oftenly short, simple and ambiguous [29, 37]. However, different users issuing the same query may have different information requirements. For example, for the query "pizzas in Delhi", a businessman may require different set of pizza joints than a college student. Therefore, search engines must be personalized in order to tailor the search results for the users by utilizing the information of their interest. The results provided to a user in response to his query by a search engine

will not only be based on the issued query, but also according to the user interest, *i.e.*, User Interest Profile (UIP) [7, 23, 29–31, 34, 35, 38–50].

The sources used to acquire information act as the basis for construction of various supporting modules, every personalization model can be broadly categorized into two categories, *viz.*, post-search information, and social information. Search engine browsing history, session log, click-through data, location, web page dwell time, cursor moments, desktop files, etc. are covered under post-search information, while, social information includes social networks, collaborative tagging, micro-blogging, online feedbacks, etc.

Centroid to every personalization model is User Interest Profile (UIP), where the process of acquiring knowledge about user's preferences by a personalization system is known as UIP. It has been observed that only an efficient and complete UIP can lead to an effective and high performing web search personalization model design. Thus, the next section focuses on extensively reviewing the various competent methodologies for user interest profiling.

2.2 User Interest Profile

User Interest Profile is a requisite technique to perceive the preferences and behaviour of a user. Typically, the process of constructing a UIP of a user can be logically separated into two phases, *i.e.*, information gathering, and information quantization of user's interest [34]. The information gathering phase deals with acquiring as much as information about the user either from user's own activities or activities of others which are, directly or indirectly, linked to a user under consideration. The information quantization phase calculates the degree of interest that a user holds for the preferences predicted through the information gathering phase. Activities of a user and those of others referenced in information gathering phase can be either explicit or implicit.

This thesis significantly shows the diversity in selection of various information methods used by numerous researchers to predict user preferences. It is for this reason that both post-search information and social information categories used in the construction of UIP have been reviewed. The studies on UIP construction have been further divided into two sub-categories, *viz.*, *self activities based user profile*, and *user profile enrichment*. The

former identifies user preferences from activities directly performed by the user himself, while the latter is used in predicting additional information about the user which is not directly related to him, but is quite useful in representing user's potential preferences or interest.

2.2.1 Post-search Information based UIP

Many of today's search engines, either employed in a web search or recommender systems, support personalization of results based on post-search information of an end-user. The information acquired for UIP construction can be either explicit or implicit [21]. Firstly, the studies based on their approach to utilize only self activities of user to construct a UIP have been reviewed. Then the studies based on the use of various strategies for enrichment of a constructed UIP have been reviewed.

Self Activities based User Profile

The most commonly used post-search information is user's search history, *i.e.*, a log of queries issued by a user and corresponding set of results clicked. Several researchers [12, 51–55] have used this search history for the prediction of user's UIP. Speretta and Gauch [51] and Sieg et al. [52] have made use of user's search history classification into topics and concepts for the construction of a UIP. On the basis of semantic similarity strength between constructed UIP and web resources that belongs to result-set, re-ranking of web resources was performed to provide a list of results in order of their relevance to a user. Several language modeling methods on the long-term search history of the user have been compared by Tan et al. [11] for UIP construction and selection of best one to be utilized for web search personalization. Bennett et al. [54] have also studied user's search history by analyzing the documents clicked in the entire search session instead of single query and correspondingly clicked documents. In order to model UIP, Bennett et al. have performed topic modeling of documents using Open Directory Project (ODP) hierarchy. Along with this, both long-term and short-term interests of a user have also been analyzed. Based on the fact that every clicked document in a search session doesn't represent user relevancy as document click can only be due to user curiosity, inability of user to perceive document content from its link description or misleading link description. So, in order to deal this issue, instead of considering every clicked document in the search session, Fox et al. [56] have only analyzed the documents for which minimum web page

viewing time was at least 30 seconds; and it was the final document that was visited by a user during the search session. The impact of dwell time implicit feedback parameter, *i.e.*, document viewing time for UIP modeling was also studied by Liu et al. [57] and Cai et al. [55] using Weibull distribution and 2-level query matching respectively. Moreover, they also claimed that dwell time was a good indicator of web resource being relevant to a query and query issuer. The hot topic analysis based on association rules were used by Chen et al. [58] for making various announcements.

The concept of information scent, introduced by S. Chawla [15, 59] and Chawla and Bedi [60] which was computed using the click-through and page access time information of initial few web search results of current session and previous session logs of the user. Information scent is a dynamic entity which gets updated on each document click of a user in order to refine the listing of subsequent results. The web resources with maximum information scent were more relevant. Moreover, the researcher also optimized the process of information scent calculation and personalization in her work using genetic [15] and ant colony optimization [59] algorithms for better accuracy. Based on the assumptions of Chawla that current search session of user is more important to formulate user information requirement than long-term session, Sugiyama et al. [10] had designed several adaptive approaches for UIP construction. The influence of long-term search session has also been considered by Sugiyama et al., but its degree of influence is less.

Another aspect of post-click behaviour of users was investigated by Guo and Agichtein [61]. They studied the cursor movements and scrolling to determine users' interest which was then used to refine their future search sessions. They designed a Post-Click Behavior model to analyze the examination pattern of relevant and non-relevant documents followed by various users. Similarly, Huang et al. [62] had also utilized cursor activities as implicit parameter for UIP construction. Analyzing the effectiveness of cursor movements behaviour of a user to quantify the relation between a user and his preferences, Buscher et al. [63] had utilized it in their work. The time spent by a user on a displayed segment is measured on the basis of mouse hovering and scrolling pattern. This segment-level displayed time is then utilized to quantify the degree of association between user and topic identified from content modeling of a segment. The results obtained are better than those of Buscher et al. [64] with respect to eye-tracking pattern on segment to quantify

the association of user and topic. Vicente-L'opez et al. [65] proposed six different generic representations of UIP from document collections. The representations of profile are semantic network, weighted keywords, and concepts. An L-topic algorithm was designed by Dou et al. [66] for the construction of topical and click-entropy based UIP for a user from users' query log.

In addition to users' search history and cursor movement, some researchers [13, 14] also used local content information of the users for UIP construction. Teevan et al. [14] used information of e-mails and documents that was either read or created by the user to generate his UIP. They also explained the importance of UIP for personalization of web search. Similarly, files placed on user's desktop screen were utilized by Chiritra et al. [13] to construct user's UIP based on the fact that a person generally keeps most frequently usable or preferable files on the desktop screen. In contrast to using single or double implicit feedback approaches, Balakrishnan and Zhang [67] integrated multiple implicit feedback parameters at one location to monitor the interest of a user in order to personalize his/her web search. The feedback parameters chosen are dwell time, page review, click-through data, and text selection. The individual as well as cumulative impact of these parameters has been studied to conclude that page review and text selection have the lowest and the highest precision respectively.

Along with the utilization of implicit information of users to identify their interest, many researchers have also used explicit information in which preferences are directly provided by a user to the IR system. The information can be either of the form relevancy or irrelevancy of search results or liking-disliking for a particular topic [68]. In order to construct UIP, the searching system can explicitly ask a user to select some topics from the list of database [20] or provide a bunch of documents related to his preferences [69]. Similarly, Micarelli and Sciarrone [70] also asked for interested and non-interested topics from users to construct their UIPs. However, the performance of personalization using explicit information is good, but not widely followed by researchers due to many drawbacks associated with it [71]. To combine advantages of both explicit and implicit information, Hannak et al. [72] designed a dual methodology. The profiles constructed solely on the information extracted from self activities of a user is sparse and inefficient which adversely affects the performance of personalization model. The different amount of information is

available of different users as some users are more active in terms of content generation than others. Thus, in order to prevent injustice on account of less active users, enrichment of their UIPs must be performed.

User Profile Enrichment

Kim and Park [18] and Leung et al. [73] studied the topical preferences of the user determined from user-friendship network, and the semantic relationship of query keywords respectively. These semantic relations are often used for query suggestion in recommender systems. Based on the fact that persons belonging to the same group usually share similar interest and preferences, group relationship of the users is utilized for UIP construction and enrichment [35, 53, 55, 74]. In order to create user groups, Dou et al. [53] and Cai et al. [55] utilized common clicks relationships, while common location relations were used by White et al. [35]. Some other aspects of group formation like demographic information, common interest, occupation, query selection, similarity of desktop files, geographic locations, etc. were analyzed by Teevan et al. [74] for UIP enrichment. The concept of collaborative filtering with browsing history was utilized by Sugiyama et al. [10] for UIP enrichment. Specifically, user-term weight matrix is used to acquire recommendations for enrichment. A detailed explanation of collaborative filtering has been provided in the subsequent section. Liu et al. [9] have devised a methodology to create two types of user profiles and a fusion algorithm for their combination in order to get an enriched UIP. They have used category mapping of query and contextual disambiguation of query words to construct two UIPs. Despite being able to uplift the quality of UIP by enriching it using group relations, none has bothered about difference in the level of relationship strength between the group members. The methodologies constructing UIP on basis the of post-search information of user have rarely or improperly utilized an enrichment strategy as relationship determination without social information is not practical in the present time. This section presents a review of various methodologies of UIP modeling based on post-search information of a user. Table 2.1 summarizes the strengths and weaknesses of various UIP construction methodologies.

Most of the studies have used one or more implicit or explicit parameters to perform the task of user interest prediction, but are unable to create a strong UIP as spending a lot of time on a web page doesn't mean that user is interested in it. The context may have been

Table 2.1: Various Methodologies Constructing UIP based on User's Post-Search Information

UIP Construction Methodology	Information Gathering Approach	UIP Parametric Basis	User Interest Weighting Technique	UIP Enrichment Approach	Experiment Domain
Speretta and Gauch [51]	Implicit	Browsing history classification into topic hierarchy with reference to ODP hierarchy.	-	✗	Google search engine browsing history.
Sieg et al. [52]	Implicit	Mapping of user with ODP hierarchy to obtain ontological user profile.	TF-IDF	✗	Browsing history.
Tan et al. [11]	Implicit	Search history mining using statistical language modeling technique.	TF-IDF	✗	Google and Firefox search history.
Bennett et al. [54]	Implicit	Long-term and short-term user interest identified by topic modeling of clicked document.	-	Recommendations from users with common interest, query selection, etc.	Microsoft Bing search engine's query and correspondingly clicked document log.
Fox et al. [56]	Implicit & Explicit	Baysian & decision tree-based language modeling for UIP. Explicit information for comparing and evaluating UIP.	-	✗	MSN search, Google search logs and Internet explorer-based add-on.
Liu et al. [57]	Implicit	Weilbull distribution-based dwell time modeling and topic of clicked document.	-	✗	Browsing history and corresponding dwell time of clicked document.
Cai et al. [55]	Implicit	Dwell time and topic modeling of clicked document using ODP.	Joint probability distribution	Common click relationship of users	Browsing history and corresponding dwell time of clicked document.

Continued on next page

Table 2.1 (Contd.)

UIP Construction Methodology	Information Gathering Approach	UIP Parametric Basis	User Interest Weighting Technique	UIP Enrichment Approach	Experiment Domain
Chawla [15, 59, 60]	Implicit	Information scent computed using PF.IPF weight and time. Genetic [15] and ant colony optimization [59].	-	✗	Web query session log corresponding to academics, entertainment and sports.
Sugiyama et al. [10]	Implicit	Mining of current search session and UIP incremental updation during next search sessions.	✗	Collaborative tagging of user	Browsing history.
Guo and Agichtein [61] and Huang et al. [62]	Implicit	Post-click behavior model to use cursor movement and scrolling patterns on clicked document	✗	✗	User studies of MIT.
Buscher et al. [63]	Implicit	Segment level display time and cursor movement of clicked document.	TF-IDF	✗	Controlled study of 32 participants.
Vicente-L'opez et al. [65]	Implicit	Content modeling of document collection to give term and subject-based profile.	TF-IDF, diff-Freq	✗	Official document of Andalusian parliament.
Dou et al. [66]	Implicit	L-topic algorithm for topical and click entropy-based modeling of web pages in query log.	-	✗	12 days query log of windows live search.
Teevan et al. [14]	Implicit	Content modeling of email and document created or read by user.	TF	✗	MSN search and user study of 15 participants.
Chiritra et al. [13]	Implicit	Centroid, sentence, lexical compound summarization of desktop files.	TF-IDF, BM25	✗	User studies.
Balakrishnan and Zhang [67]	Implicit	Multiple parameter integration, *i.e.*, dwell time, click through, text selection and page review.	TF-IDF	✗	Controlled simulated data from participants.

Continued on next page

Table 2.1 (Contd.)

UIP Construction Methodology	Information Gathering Approach	UIP Parametric Basis	User Interest Weighting Technique	UIP Enrichment Approach	Experiment Domain
Chiritra et al. [20]	Explicit	Selection of data from ODP topics by user.	✗	✗	User studies.
Li and Zhong [69]	Explicit and implicit	Ontology mining of documents that were explicitly given by user.	✗	✗	User studies.
Micarelli and Sciarrone [70]	Explicit	Interested and not-interested topics asked from user	✗	✗	User studies.
Hannak et al. [72]	Explicit and implicit	User self-defined features and preference identified from user activities.	✗	✗	Google and bing web search log.
Kim and Park [18]	Implicit	Click-through information of user and friends recommendations.	-	Friendship network of social network	Query logs and Facebook.
Leung et al. [73]	Implicit	Web snippets of results corresponding to user query.	-	Cluster of related queries	Google search history.
Dou et al. [53]	Implicit	Click-through information modeling.	-	Grouping of similar click events	12 days MSN query log.
White et al. [35]	Implicit	Browsing history mining and recommendations of similar users.	✗	Grouping of user based on locations.	Browsing history.

Note: The symbol (✗) represents the absence of the feature, while, (-) means no information available regarding the feature in a methodology.

too complex for a user to judge its relevancy. Some users click on many URLs just for the sake of their curiosity or as a habit. Moreover, profile created by this method cannot effectively predict the interest level of a user in a particular preference; and analyzing the browsing history of anyone is also against the privacy norms. So, according to demand of current IR market and construction of a strong and efficient UIP, an alternate approach of UIP construction must be adopted.

2.2.2 Social Information based UIP

In the recent years, research on personalization of web search and recommender systems has witnessed a rapid surge in the usage of users' social information for their UIP construction. There are many social information platforms, out of which collaborative tagging sites (like del.icio.us, MovieLens,flicker), microblogging sites (like friendfeed, tumbler, twitter), and social network sites (like linkedin, snapchat) are the prominent ones. The information about a user acquired from these social platforms is quite useful to construct his UIP. However, among all the sources used to acquire social information, collaborative tagging is most preferable by the researchers, academia and industry. The other social information sources are less preferred due to unavailability of large public datasets, reluctance of users to share their information, and various privacy issues involved in that data. Thus this thesis focuses on constructing a social information based UIP through collaborative tagging as envisaged by some other research studies. Further, some other social information sources have also been studied. Formally, collaborative tagging is represented by Collaborative 3-partite Graph (C3TG) as follows:

Definition 1 *A **Collaborative 3-partite Graph** (C3TG) is a special type of graph denoted by $G_3(V, E)$ such that $V \in (U \cup T \cup W)$ and $E \in R^u_{t,w}$. On the whole, C3TG can be stored into the memory as a quadruple $(u, t, w, R^u_{t,w})$ and $R^u_{t,w} \subset (U \times T \times W)$.*

Where, U, T, and W depict the User, Tag, and Web Resource sets respectively. The ternary relation denoted by $R^u_{t,w}$ depicts that a web resource named w is annotated by a user u using the tag of interest t. An example of collaborative tagging is presented in Fig. 2.1, where a part of the ternary relations of Users: Abu, Ajay, Geeta, and Ravi has been considered. Here, in Fig. 2.1, the ternary relation of Ravi depicts that web resource Carol Hurst has been annotated by Ravi using the tag "Book".

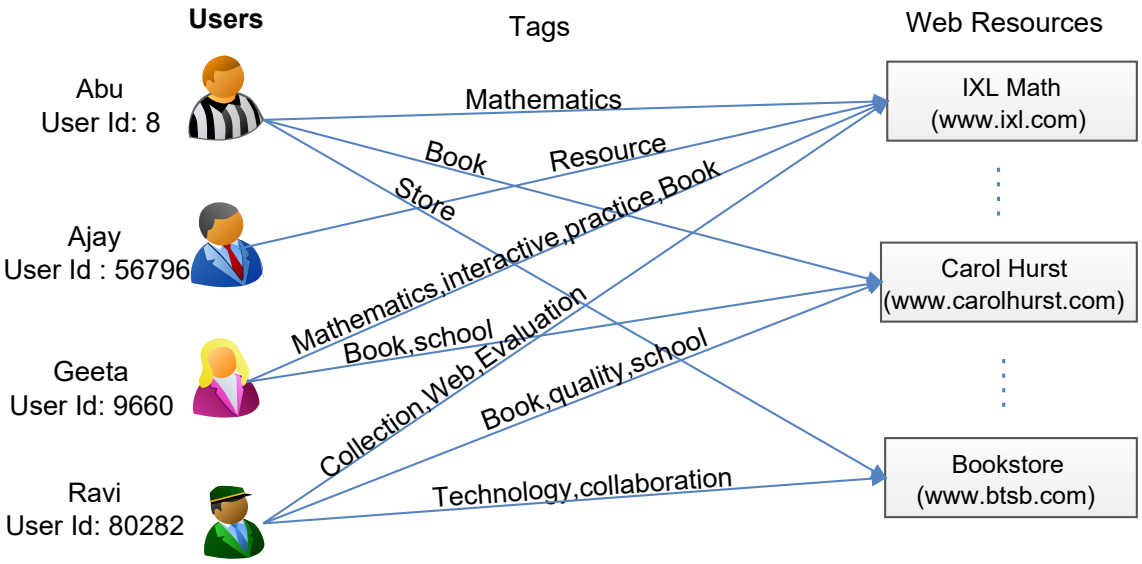


Figure 2.1: Example of Collaborative Tagging

As in the case of post-search information based UIP, firstly the methodologies utilizing user's own activities have been reviewed for UIP construction. Then, a discussion on enrichment of UIP has been undertaken. Some researchers inferred UIP for personalizing the web search [7, 23, 30, 31, 34, 45, 46, 75], while others used it for recommender systems [47–50, 76, 77].

Self Activities based User Profile

Bischoff et al. [78] and Heymann et al. [25] have studied the importance of information provided by tags or social annotations using some real collaborative tagging datasets for inferring the preferences of a web user. Realising the significance and growing popularity of social annotation systems, Maniu and Cautis [79] and Gupta et al. [24] designed a framework for ideal collaborative tagging network, and summarized various characteristics of tagging information respectively. Gupta et al. [24] also summarized the different techniques of tag generation, analysis and visualization. A detailed survey of various tag usage patterns generally followed by users in a collaborative tagging system has been done by Golder and Huberman [80] in their work to help other researchers in designing better UIP construction techniques. In addition to analyzing user's personal information in a folksonomy system, Yeung et al. [81] investigated the accuracy of a collaborative tagging-based UIP.

Two web search personalization algorithms, *viz.*, SocialPageRank (SPR) and SocialSimRank (SSR) were designed by Bao et al. [45], where UIP was created solely on the basis

of annotations made by the respective user for all the web pages. User web search history along with folksonomy information has been exploited by Kumar and Kim [46] to construct a UIP. Tags corresponding to the web pages, clicked by a user in his previous search session, were retrieved from del.icio.us for UIP modeling by Kumar and Kim [46]. Moreover, they also utilized HAC clustering algorithm to remove redundant tags from the constructed UIP of a user. In their work Gemmell et al. [82] also utilized the concept of tag clustering to construct user profile and provide the personalization experience to a user. The naive and co-occurrence approaches were explored by Michlmayr and Cayzer [83] for the construction of a tag-based UIP. Inspite of this, still their methods of UIP construction have some drawbacks like usage of different vocabulary by various users, polysemy and synonymy.

Usually, a user has more than one interest with varying degree of preferences, and prediction of an approximate value for these degree of preferences always remains a cause of great concern for the researchers as the approaches like TF [84], TF-IUF [26], BM25 [85], etc. have their own advantages and disadvantages. Cai et al. [34] studied all these issues and proposed Normalized Term Frequency (NTF) for assigning a suitable weight to an interest as per the user's degree of preference. The results obtained are more satisfactory as compared to other approaches. Bouadjenek et al. [31] studied the impact of embedding the social information of users in index structure of Information Retrieval (IR) systems. The UIP was created using UTF-IUF approach which was an adaptation of famous TF-IDF approach of IR. To evaluate the effectiveness of their formulations in a real-time environment, Bouadjenek et al. [17] have designed LAICOS which is a web search engine for information retrieval. LAICOS performs the function of web search by considering the tags or annotations available with a web page as metadata for the construction of a UIP. The mathematical formulation of some approaches used to assign the degree of preference to a user's interest has been represented by a tag in Table 2.2.

Du et al. [23] designed a user profiling technique which is an intergeneration of user's own tags and ratings on the 1 to 5 scale. Similarly, Zhou et al. [86] and Kumar et al. [30] have modeled UIP based on both annotations that were made by user and various web resources annotated by a user. In order to weight various terms in UIP by Zhou et al., relationship strengths of UIP terms and terms mined from high ranked documents

Table 2.2: Approaches Assigning Degree of Preference to a Tag in UIP

Approach	Mathematical Formulation	Description
TF [84]	tf	where, tf is number of times user used tag x.
UTF-IUF [26]	$tf * log\frac{N}{n_x}$	where, N is total number of users; n_x is number of users using tag x.
BM25 [85]	$\frac{tf*(k_1+1)}{tf+k1*(1-b+b*\frac{ul}{avgul})}$	where, both parameters k_1 and b are usually set to standard values 2 and 0.75.; ul represents the length of user profile; and $avgul$ is average length of user profile.
NTF [34]	$\frac{N_x^{u_i}}{N^{u_i}}$	where, $N_x^{u_i}$ represents the number of instances when user u_i annotates the web resources with tag x; and N^{u_i} is the number of web resources annotated by u_i.
TF-IDF [30]	$\frac{tf}{total_{tag}} * log_2\frac{D}{d}$	where, $total_{tag}$ is total number of tags; D is total number of documents; and d number of documents annotated with tag x.

were measured using a regularizing function. Both Kumar et al. [30] and Shepitsen et al. [87] have measured the degree of relevance for a tag in user's UIP by using TF-IDF approach.

Hannon et al. [48], Kacem et al. [88], and Younus et al. [89] have demonstrated the benefit of using real time web information from micro-blogging web services like twitter for the construction of UIP in a recommender system environment as the collected information is a good source to predict the interest of a user. For constructing user's UIP, Hannon et al. [48] had extracted terms from latest 100 tweets of user and weighted the terms using TF-IDF approach. Kacem et al. [88] have used the concept of TF for the term "weighting" where terms are extracted from post made by users themselves. Still in addition to this, the researchers also studied the influence of time-sensitivity element of user profiles in a persolization system.

A recommender system for events was designed by Horowitz et al. [90], where contextual information was implicitly extracted from user's LinkedIn account and used it to construct UIP. Sometimes, UIP constructed through social information suffers from the problem of tag ambiguity as tag can refer multiple topics at a single time which results in improper recommendations and web page ranking hierarchy by a personalized recommender system

and search engine respectively. So, to avoid this problem, Xu et al. [91] devised the concept of ontological similarity of tags in users UIP. Firstly, the disambiguation of tag information is performed using external domain ontologies, followed by semantic quantification in second step. Similarly, Hawalah and Fasli [92] also explored the ontological concepts of user referred web pages to construct a dynamic UIP, as interest of a person can change over a period of time.

Han et al. [93] analyzed the user annotations in a folksonomy system and designed a algorithm for mapping of these annotations with an already existing domain ontology. The user's tag-based UIP is leveraged to propose two models to enhance the user preference boundaries. First model is used to find latent tag preferences, and another is the latent tag annotation model. To address the issue of tag-ambiguity, Sang et al. [94] suggested tag refinement for which they proposed Ranking-based Multicorrelation Tensor Factorization (RMTF) using various aspects of C3TG. Yang et al. [50] presented a survey on social recommender systems based on the concept of collaborative filtering. The task and working of recommender system based on social and traditional approaches was also stated in the survey. Harpale et al. [95] evaluated UIP construction techniques on CiteData and showed that quality of personalized arrangement of search results was based on the effectiveness and completeness of user profile. As already mentioned that user profile constructed only on user's own information is sparse and ineffective. So, user profile enrichment must be a part of UIP construction methodology.

User Profile Enrichment

Xie et al. [96], in their study, explored many extant user profiling techniques in an attempt to answer two major research questions. The first one is to quantify the number of tags or preferences that are enough to represent a qualitative UIP. Second question relates to identifying the circumstances under which a UIP enrichment strategy must be adopted to enhance the performance of a personalization model. The results of their experimental study reveal that UIP size varies from user to user; and UIP constructed solely on the basis of user's own annotations presents an incomplete profile which fails to list all preferences of the user. Thus, there is a need to incorporate a UIP enrichment strategy at the time of UIP creation. The matrix factorization approach is getting popular these days as it identifies the latent feature relationship between users and items. It is also being used

in many recommender systems for recommending an item to a user, either not visited or seen by a user before.

Two matrix factorization based methods, *viz.*, svdCUIP and modSvdCUIP were devised by Kumar et al. [30] for the enrichment of UIP. As the name suggests, the concept of Singular Value Decomposition (SVD) is used to make clusters of tags that were annotated to different web pages. Each cluster represents a single topic of interest where member tags are syntactically and semantically similar. These clusters help to predict the interest of a user. According to their results, modSvdCUIP is far better than svdCUIP. Factorization made on the basis of SVD can result in both negShamriative and positive elements in a computed matrix. However, if the source matrix is highly sparse, then the matrix obtained by multiplication of factorized matrices will be highly negative. To handle this limitation of matrix factorization, Luo et al. [97] devised a collaborative filtering system based on Non-Negative Matrix Factorization (NMF). The main idea is to shift the non-negative adaptation process from whole feature matrices to each involved feature; and non-negative single element based upgrade rules are designed. Subsequently, tikhonov regularization [98] is also integrated to the system to be named as Regularized Single Element based NMF (RSNMF). Similarly, Shepitsen et al. [87] also used the concept of tag clustering to determine the user's probable interests, *i.e.*, enrichment which are not covered by user's UIP constructed from user's own annotations only. The sentiment aspect of tags was incorporated by Xie et al. [32] for the enrichment of user profile.

The data which is used by recommender system for making recommendations to users is usually of very high dimension which causes these systems to suffer from many problems. Out of these problems, some of the prominent ones are high storage and computational complexity, and slower convergence rate which makes even NMF-based collaborative recommender system unfit for the industrial usage. Luo et al. [99] devised an Alternating Direction method (ADN)-based non-negative latent factor (ANLF) system to solve these problems. The results of the experiment on a large dataset confirmed an increase in the convergence rate and a decrease in complexity.

Markines et al. [100] analyzed the constructed UIP from the semantic content point of view to identify a suitable approach or technique for similarity measure in tags. Similarly, with regard to the concept of Kumar et al. [30], Markines et al. [100] and Shepitsen et al. [87],

where additional information for UIP enrichment was inferred based on latent similarity relationships between terms either measured using SVD or matrix factorization, Zhou et al. [101] used the concept of deep semantics. They devised a model to integrate user's own tag based UIP with topic modeling of high frequency words or phrases from documents in user domain using Latent Dirichlet Allocation (LDA). Mulhem et al. [102] had used the term-relationships, but unlikely Zhou et al. [101] relationships of terms in user's query and UIP were identified for query expansion. They also utilized LDA to perform topic modeling of both user's query and UIP. Similarly, LDA topic modeling concept was used by Varshney et al. [103] to formulate the interest of a user.

The benefit of incorporating the community information of a user into UIP construction process was also analyzed by some researchers based on the fact that behaviour or interest of a person is highly influenced by his friend circle [18,33]. A strategy to integrate the collaborative tagging information into neighbourhood-based model preference recommender system was designed by Luo et al. [49] in order to improve its performance. UIP of a user under consideration was updated using these recommended preferences. Similarly, Liu and Lee [47] also incorporated social network information of a user into collaborative filtering in order to enhance their recommendation efficiency. Social network relationships like diverse neighbour groups, nearest neighbours, etc. were collected from various social network websites.

Valcarce et al. [104] have argued that the neighbourhood-based recommendations for profile construction are far more effective than matrix factorization-based approach to unearth the hidden relation between user and a preference. However, they also highlighted that the performance of neighbourhood-based methods is closely related with a selection of a clustering strategy. Working on the same assumptions, Hunag et al. [105] have used k-NN approach for UIP enrichment in order to make personalized recommendations to a user in tag aware recommender system. The associations between a user and the web resources annotated by a user were also deeply analyzed through an adaptive model. In addition to k-NN approach, both qualitative and quantitative prescriptive of similarity measure were considered before recommending any additional information for user profile.

Hannon et al. [48] later discovered that using only the user's own tweets is not enough to design a good UIP, so they used user's twitter social graphs to enhance the accuracy

of the system named *Twittomender*. The concept of recommendations from users who are either followers or followees of the user for whom UIP is to be constructed has been adopted for UIP enrichment by Hannon et al. [48]. The weight of every recommended term to user is equal to the weights of that term in recommender's UIP. The benefit of constructing a student interest profile which enlists the social and academic preferences, education and social background of the student for a university recommender system has been explored by Kanoje et al. [106]. This recommender system helps the student to find a most suitable university to pursue their education. The information extracted from user's facebook profile, post and shares acts as the basis for UIP construction.

The word-of-mouth recommendation from a friend can also serve as an important source for enriching the profile of a user to personalize his web search. Shafiq et al. [7] analyzed the frienship network for UIP where users related to the same Wikipedia page update history were considered as friends. Trust and relevancy matrix is also created to compute most trustworthy friends. Recommendation of preferences is only taken from friends selected by various social network analysis strategies applied to a network of trustworthy friends. Similarly, trust relationship network of user for profile enrichment was exploited by Wu et al. [107] using the concept of opinion leaders. The researchers devised an algorithm named *OLrs* to identify an opinion leader for user and take recommendations from them.

Heath et al. [108] designed algorithms to analyze the social network of a user and compute the trust metrics. The tagging actions and comments of a user serve as the basis for input information for the algorithms to compute trust levels and recommend additional information to a user. They also considered the recommenders' experience towards the topic by analyzing the reviews of web resources annotated with a particular tag and computing the prevalence of a recommender towards it. The affinity score between two users was computed on the basis of review analysis which was provided to the algorithm in friend-of-a-friend manner. Huang et al. [109] also analyzed the user's friendship network for the modeling of user interest, in which they fused together the frequency, duration, and recency of tags with the neighbours' information from social friends network. The resultant profile was then used for making collaborative recommendations.

The researchers have followed yet another concept of investigating the group interest of

user's social group for refinement of users search task. Zeng et al. [110] have provided the framework for user interest models and social network-based group interest model to predict the interest of groups that a user is involved in. The authors have refined the web search by utilizing a semantic dataset together with refinement at the individual user interest level and group interest level of the groups to which the user belongs. Xie et al. [33] incorporated the concept of random walks to be performed on Multi-Faceted Graph (MFG) to measure user-user similarity and identify latent user communities. Zhang et al. [111] designed a personalized image recommender system on the basis of assumptions made by Xie et al. [33]. The additional information about user's preferences for UIP enrichment was collected by Zhang et al. [111] using correlation strength among various elements of Collaborative 3-partite Graph. Yang et al. [112] devised an algorithm for social network recommendation along with the concept of information aging as a solution to information overdue issue as user preferences might change over a long time. They also designed an algorithm to identify the nearest neighbours of the user under consideration from his trust network to recommend information for the enrichment and better performance of UIP.

The graphs used to identify the latent relationship between the users are usually undirected which result in high dimensional sparse matrices. Many researchers have used these graphs to identify user-user relationships, and made recommendations for UIP enrichment. The Non-negative Latent Factor (NLF) models are very effective to mine different patterns from the graphs, but they cannot perform well on undirected graphs. However, Luo et al. [113] addressed this issue in their work and gave Symmetric and Non-negative Latent Factor (SNLF) model as a solution to it. Rawashdeh et al. [114] adopted another concept of Katz proximity to effectively analyze the path-based proximity measures in C3TG for helping users to find relevant information as per their interest. BM25 was used to assign weights to user preferences.

In their research work, M. Y. H. Al-Shamri [115] discussed the various methodologies used for the construction of a user profile in recommender systems especially for demographic-based recommenders. A detailed description of suitable similarity measuring approaches was provided with their advantages and disadvantages. The importance of constructing a user profile is not just limited to a simple client-server model, but has even extended

beyond it. Saoud and Kechid [116] devised an approach to utilize the social profile of a user to make a personalized search in distributed search environment. Their approach led to improve the performance of source selection and result merging process of distributed search systems which were earlier based on textual information matching. In order to enrich the social profile of a user, Saoud and Kechid [116] explored the various relationships between user's social entities.

This subsection reviews the various methodologies of UIP modeling based on user's social information. Table 2.3 summarizes the strengths and weaknesses of various UIP construction methodologies.

2.2.3 Discussion

The literature review undertaken in this section provides an insightful description of various methodologies adopted by different researchers for the construction of User Interest Profile (UIP) in order to personalize the web search experience of a present day user. Firstly, for constructing a UIP, an attempt is made to gather maximum information about the interest and preferences of a user under consideration, followed by the quantization of gathered information. Nowadays, a normal web user usually has his interest in multiple things which makes quantization necessary to arrange the required information according to user's preference. The information can be collected either explicitly or implicitly from the user, like topic of interest, online activities like web history or collaborative tagging actions respectively. Both explicit and implicit methods used for the collection of information have their own pros and cons as already discussed under Section 2.2.1, but implicit methods are considered better than the other ones. The literature reviewed under this section covers both the information categories, *i.e.*, post-search information and social information used to construct a UIP.

As already explained in Sections 2.2.1 and 2.2.2, the efficiency of a personalization model greatly depends on the effectiveness and completeness of a UIP. As the amount of available information varies from one user to another *i.e.*, some users are found to be more active in terms of content generation than others. So, the profiles constructed solely on the basis of information extracted from self activities of a user is sparse and inefficient. It adversely affects the performance of personalization model. Thus, in order to avoid this drawback,

Table 2.3: Various Methodologies Constructing UIP based on User's Social Information

UIP Construction Methodology	UIP Enrichment Type	UIP Enrichment Approach	User Interest Weighting Technique	UIP Parametric Basis	Relationship Type	UIP Sparsity Handler	Experiment Domain
Bao et al. [45] & Noll and Meinel [84]	✗	✗	TF	User's tags	Ternary	✗	Del.icio.us
Kumar and Kim [46]	✗	✗	TF	Tags annotated to web pages clicked by user in previous search session	Ternary	✗	AOL query log, Del.icio.us
Gemmell et al. [82]	✗	✗	TF-IDF	User's tags	Ternary	✗	Del.icio.us
Michlmayr and Cayzer [83]	✗	✗	-	Naive and co-occurrence approach on user's tags	Ternary	✗	Social bookmarking
Xu et al. [26]	✗	✗	TF-IUF	User's tags	Ternary	✗	Del.icio.us, Dogear
Vallet et al. [85]	✗	✗	BM25	User's tags	Ternary	✗	Del.icio.us
Cai et al. [34, 38]	✗	✗	NTF	User's tags	Ternary	✗	Del.icio.us [34], FMRS [38]
Bouadjenek et al. [31]	✗	✗	UTF-IUF	user's tags and interests in other tag using matrix factorization	Ternary	✓	Del.icio.us
Du et al. [23]	✗	✗	TF-IDF	user's tags	Ternary	✗	MovieLens, Epinion
Kumar et al. [30]	Tag-Tag	Tag clusters using HAC & SVD-based tags similarity	TF-IDF	user's tags and cluster recommendations	Ternary	✓	Del.icio.us, AOL query log
Zhou et al. [86]	✗	✗	UIP and High ranked document term strength	User's tags and tags of resources annotated by user	Ternary	✗	Del.icio.us
Shepitsen et al. [87]	Tag-Tag	Tag clusters using HAC	TF-IDF	User's tags and cluster recommendations	Ternary	✗	Del.icio.us, Last.fm

Continued on next page

Table 2.3 (Contd.)

UIP Construction Methodology	UIP Enrichment Type	UIP Enrichment Approach	User Interest Weighting Technique	UIP Parametric Basis	Relationship Type	UIP Sparsity Handler	Experiment Domain
Hannon et al. [48]	User-User	Follower-Followee relationship frequency	TF-IDF	User's preferences and similar users recommendations without trust measurement	Asymmetric	✓	Twitter
kacem et al. [88]	✗	✗	TF	Terms extracted from user post	Asymmetric	✗	Twitter
Younus et al. [89]	✗	✗	BM25	Language modeling of user's online post	Asymmetric	✗	Twitter,CiteData
Horowitz et al. [90]	✗	✗	-	User's LinkedIn profile Info	Asymmetric	✗	LinkedIn
Xu et al. [91], Hawalah and Fasli [92] & Han et al. [93]	✗	✗	TF-IDF	Mapping of user's tags and ODP ontologies	Ternary	✗	Del.icio.us
Sang et al. [94]	✗	✗	TF-IDF	User's tags and multi-correlation tensor factorization	Ternary	✗	Flicker
Harpale et al. [95]	✗	✗	-	User's tags and probabilistic latent semantic analysis	Ternary	✗	CiteData
Xie et al. [32]	Tag-Tag	Sentiment mapping of tags	NTF	User's tags and recommendations by sentiment context of tags	Ternary	✓	SenticNet, FMRS, MovieLens
Markines et al. [100]	Tag-Tag	Semantic similarity of tags with terms from external data source	-	User's tags and semantically similar data	Ternary	✓	BibSonomy

Continued on next page

Table 2.3 (Contd.)

UIP Construction Methodology	UIP Enrichment Type	UIP Enrichment Approach	User Interest Weighting Technique	UIP Parametric Basis	Relationship Type	UIP Sparsity Handler	Experiment Domain
Zhou et al. [101] and Mulhem et al. [102]	Tag-Tag	LDA-based topic modeling of document in user domain	TF	User's tags and recommended topics	Ternary	✓	Del.icio.us, BibSonomy
Kim and Park [18]	User-User	Topic similarity in topic-based profile	-	user's preferences and credible users recommendations	Symmetric	✓	Facebook, Google query log
Xie et al. [33]	User-User	Latent user community & multi-faceted folksonomy graph	-	User's tags and similar users recommendations without trust measurement	Asymmetric, Ternary	✓	NUS, Flicker
Luo et al. [49]	User-User	Neighbourhood-based model using tagging relations on the same item	TF	User's tags and similar users recommendations without trust measurement	Ternary	✓	MovieLens
Liu and Lee [47]	User-User	Nearest neighbour network	TF	User's preferences and neighbours' recommendations without trust measurement	Symmetric	✓	Users of cyworld
Valcarce et al. [104] and Hunag et al. [105]	User-User	KNN clustering of users	TF-IDF	User's tags and recommendations by nearest neighbours	Ternary	✓	MovieLens
Shafiq et al. [7]	User-User	Network analysis of credible users, fractional cascading	TF	User tags and friends' recommendations	Symmetric	✓	Wikipedia page update history

Continued on next page

Table 2.3 (Contd.)

UIP Construction Methodology	UIP Enrichment Type	UIP Enrichment Approach	User Interest Weighting Technique	UIP Parametric Basis	Relationship Type	UIP Sparsity Handler	Experiment Domain
Wu et al. [107]	User-User	OLrs algorithm to identify opinion leader based on trust relations of user	TF	User's tags and recommendations from opinion leaders	Ternary	✓	Epinions
Heath et al. [108]	User-User	Analysis of user network subject to their trust relation and topic affinity	-	User's tags and recommendations from trusted users for some topics	Ternary	✓	Del.icio.us, Revyu
Huang et al. [109]	User-User	cosine similarity of user profile	TF-IUF	User's tags and similar users' recommendations without trust measurement	Ternary	✓	Del.icio.us
Zeng et al. [110]	User-User	Group interest modeling	TF-IDF	User's tags and group interest of groups to which a user belongs	-	✓	SwetoDBLP dataset
Zhang et al. [111]	User-User	Correlation strength of C3TG component	-	User's tags and recommendations from similar users without trust measurement	Ternary, Asymmetric	✓	Flicker
M. Y. H. Al-Shamri [115]	User-User	Correlation of demography-based UIP	-	User's topics and similar users' recommendations without trust measurement	Symmetric	✓	MovieLens demography data

Note 1: The symbols (✗)and (✓) represent the absence and presence of the feature respectively, while (-) means no information is available regarding the feature in a methodology.

Note 2: As sparsity handling capability of various UIP construction methodologies is different, the table shows only the presence and absence of sparsity handler.

Note 3: Ternary relation type refers to relationship of user, web resource, and annotation, while symmetric and asymmetric related to user-user or tag-tag relationship.

researchers have suggested enrichment of users' UIPs. As a result, various UIP enrichment strategies adopted by different researchers have also been reviewed in this section. These strategies include both post-search and social information based UIP. The methodology proposed in this work for UIP construction has been discussed at length in User Interest Profile (UIP) modeling Chapter of this thesis.

2.3 Resource Illustration Profile

Resource Illustration Profile (RIP) is also an important module of a personalization model. Typically, the RIP corresponding to a web resource provides an insightful description of resource content, topics to which a resource is related, what kind of information requirements can be fulfilled by a web resource, etc. Together with this, RIP also tells that up to which level a web resource can fulfil the information requirement of user for a certain topic or content. Therefore, like a UIP, the process of constructing a RIP corresponding to a web resource can be logically separated into two phases, *i.e.*, information gathering, and information quantization of web resource affinity. The information gathering phase deals with the identification of topics that a web resource describes about or is related to. The information quantization phase calculates the affinity score with which a web resource is related to the topics predicted in information gathering phase. Here also, the literature review covers both the phases of RIP construction simultaneously. The studies selected for review have been classified into two categories, *i.e.*, Content based and Tag based RIP modeling.

2.3.1 Content based RIP Modeling

The content analysis of a web resource is the traditional and most widely used method for the construction of an RIP corresponding to a web resource. Not only the textual component of a web resource, but other components like page title, HTML markup schema, anchor text, user feedback comments, etc. have also been analyzed by many researchers under the content based RIP modeling. Generally, the topics related to a web resource that are listed by a web resource provider are vague and misleading. Thus, content analysis made with the help of special algorithms has become quite important to model an

RIP.

The vector corresponding to RIP of a web resource can be obtained by using a popular technique of bag-of-words representation of a web document, where words or terms are generally weighted by Term-Frequency [117]. There are other weighing factors like TF-IDF and BM25 which can also be used to weight identified terms. But the RIP obtained from bag-of-words is purely based on terms, other than stop-words, used in the document; and it fails to capture the existing semantic relationships between the terms. However, this issue has been tackled by topic modeling, which can capture the words with similar semantics and put them into a same group called topic. In topic-based representation of a web resource, synonym words are considered as the same topic. Moreover, RIP obtained from bag-of-words is insufficient to represent multiple topics to which a web resource may be related. The researchers have analyzed different topic-based representations for RIP [20, 35, 118, 119]. In these methods, topic relativity of a web resource is represented as probability distribution over topics, where degree of affinity for a topic is shown using topic weights [20].

Knowledge about the topics of a web resource can be identified from the concept hierarchies like Open Directory Project (ODP) [20, 118] or topic modeling technique [120]. ODP is the largest and most widely used human generated comprehensive directory of web pages which is maintained and built by a large number of passionate volunteers and editors. An abstract description of ODP with two levels of category hierarchy is given in Fig. 2.2. The web resources can be mapped with topic hierarchies of ODP. Sontag et al. [118] devised a probabilistic generative model based on 2-level categories of ODP to identify the topics to which a web resource belongs. However, modeling RIP of a web resource based on topics obtained from ODP suffers from certain limitations. The ODP has not been updated for quite long time; and many web documents do not appear in the ODP list. Moreover, a large amount of manual effort is involved in categorization of a web document.

In order to address, the issues relating to ODP, T. Hofmann [121], and Blei et al. [122] devised the techniques, *viz.*, Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA) respectively for automatic topic modeling of the documents. Sun et al. [123], and Gracia et al. [124] designed a unified topic modeling framework

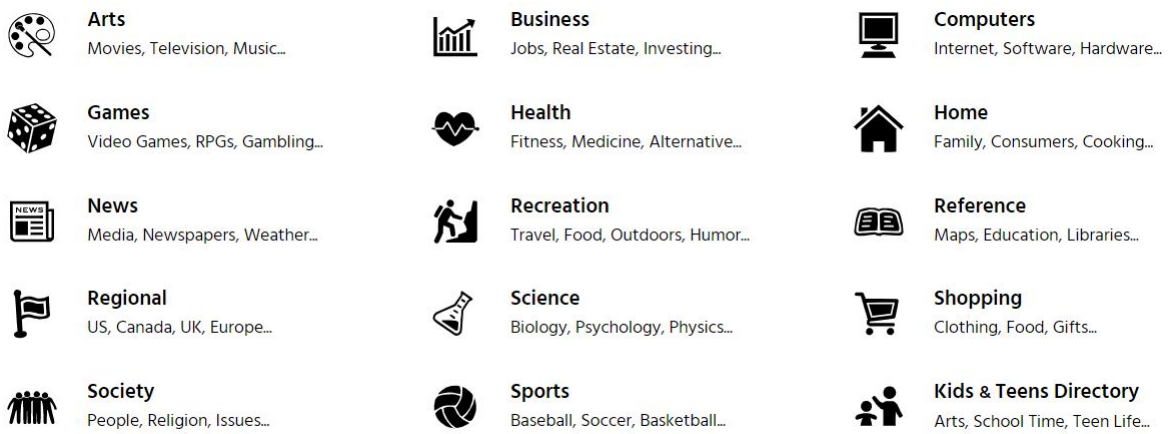


Figure 2.2: Two-level Category Hierarchy of ODP

by utilizing both text and structure of web documents. Sun et al. [123] made estimates of modeled topics by maximization of a log-likelihood of a joint probability distribution function, whereas Gracia et al. [124] used a fuzzy term weighing approach. However, Mei et al. [125] designed a topic modeling technique by utilizing a graph structure based harmonic regularizer for statistical topic model. Based on the principles of LDA, a multi-grain model was designed by Xie and Xing [120] for topic modeling and clustering of similar documents. Like LDA for text document, a latent topic model specifically for hypertext documents was given by Gruber et al. [126]. As we know that a web document describes many topics at different places in a document with different informativity levels; and sometimes, text present in a document is not sufficient to effectively utilize PLSA or LDA based techniques for topic modeling. So, in order to deal with this problem, Cheng et al. [127] devised Biterm topic model (BTM) based on principles of LDA for short text topic modeling.

The techniques such as distinct terms, ODP topic hierarchies, and content based latent topic modeling used to model an RIP have a common deficiency as these represent only the content provider viewpoint, but fail to capture the end-user's viewpoint. Therefore, to capture users' views about a web resource, tag based RIP modeling came into existence.

2.3.2 Tag based RIP Modeling

Tags annotated by various users to different web resources are a very good source for constructing RIP of a web resource, as these tags are direct word-of-mouth from the

users without any middleman modification. Basically, these tags represent the actual thoughts and viewpoints that a user holds for a web resource in reference to the type and level of information fulfilment ability of a web resource. Hence, many researchers have utilized collaborative tagging information of web resources in different ways to model their RIP [23, 34, 45, 81, 83].

Cai et al. [34] devised the RIP of a web resource by utilizing every tag annotated to that web resource, where affinity of a tag in RIP was measured through NTF approach. The authors also studied the limitations of various other approaches like BM25, TF and TF-IRF in predicting the affinity with which a tag in RIP can illustrate a web resource. Similarly, Du et al. [23], Bao et al. [45], and Yeung et al. [81], also used all tags corresponding to a web resource to model RIP, but both Bao et al. and Yeung et al. used TF instead of NTF used by Cai et al. [34]. In order to capture the advantages of both content and tag based RIP, some researchers [86, 102, 116] in their work also clubbed the distinct terms and tags annotated to a web resource at a common place to model RIP of a web resource. The authors followed various approaches for clubbing web page terms and tags. Mulhem et al. [102] used a probabilistic modeling equation, whereas both Zhou et al. [86] and Saoud and Kechid [116] had just used simple union of web document terms and annotated tags. Still the results obtained by RIP constructed through clubbed information of terms and tags are not much different from those of tag based RIP. The sentiment aspect of tags was also studied by many researchers. Xie et al. [32] based the modeling of RIP on the assumption that context of tag can be better understood only by sentiments. On the other hand, Bouadjenek et al. [31] constructed RIP by utilizing the tags of a user issuing the query and other users close to him annotating that web resource. Bouadjenek et al. [128], in their work, examined various social information retrieval policies and numerous RIP modeling techniques.

The construction of RIP by using every tag annotated to a web resource is called collective resource profile which has been used in almost every RIP modeling technique. But the consideration of each and every tag annotated to a web resource by various users cannot produce a qualitative RIP as it suffers from the problem of conflicting annotations which may be either intentional or unintentional. Therefore, to tackle this issue, Xie et al. [129] and Xu et al. [91] devised the concept of social filtering based on topic relevant user

communities and domain communities respectively. To identify these communities, Xie et al. [129] followed different approaches, but cluster-based approach is the most accurate. The concept of LDA is used to perform topic modeling in cluster-based communities. The ambiguity of tags has been handled by Xu et al. [91] through the concept of ontological mapping. Resolving this problem of conflicting tags has widened scope for further research in the field of RIP modeling; and this research work is also a modest attempt in that direction.

Table 2.4: Methodologies Constructing RIP based on Web Resource Representation Approach

RIP Construction Methodology	Content based Modeling			Tag based Modeling	
	Terms	Topics		All Tags	Tag Refinement
		ODP	Latent Topics		
Ustinovskiy et al. [42]	✓	✗	✗	✗	✗
Shen et al. [37]	✗	✓	✗	✗	✗
Yan et al. [41]	✗	✓	✗	✗	✗
White et al. [35]	✗	✓	✗	✗	✗
Hassan and White [39]	✗	✗	✓	✗	✗
Teevan et al. [14, 29]	✓	✗	✗	✗	✗
Liu et al. [57]	✗	✓	✗	✗	✗
Chirita et al. [20]	✗	✓	✗	✗	✗
Dou et al. [53]	✗	✓	✗	✗	✗
Sugiyama et al. [10]	✓	✗	✗	✗	✗
Bennett et al. [54]	✗	✓	✗	✗	✗
Sontag et al. [118]	✓	✗	✗	✗	✗
Bennett et al. [54]	✗	✓	✗	✗	✗
White et al. [130]	✓	✗	✗	✗	✗
Balakrishnan and Zhang [67]	✓	✗	✗	✗	✗
Tan et al. [11]	✗	✗	✓	✗	✗
Speretta and Gauch [51]	✗	✓	✗	✗	✗
Liu et al. [9]	✗	✓	✗	✗	✗
Buscher et al. [63, 64]	✓	✗	✗	✗	✗
Chirita et al. [13]	✓	✗	✗	✗	✗
Sieg et al. [52]	✗	✓	✗	✗	✗
Younus et al. [89]	✓	✗	✗	✗	✗
Hawalah and Fasli [92]	✗	✓	✗	✗	✗
Azad and Deepak [131]	✓	✗	✗	✗	✗
Yeung et al. [81]	✗	✗	✗	✓	✗
Mulhem et al. [102]	✓	✗	✗	✓	✗
Zhou et al. [86]	✓	✗	✗	✓	✗
Saoud and Kechid [116]	✓	✗	✗	✓	✗
Michlmayr and Cayzer [83]	✗	✗	✗	✓	✗
Xu et al. [91]	✗	✗	✗	✓	✓
Gemmell et al. [82]	✗	✗	✗	✓	✓

Continued on next page

Table 2.4 (Contd.)

RIP Construction Methodology	Content based Modeling			Tag based Modeling	
	Terms	Topics		All Tags	Tag Refinement
		ODP	Latent Topics		
Xie et al. [96]	✗	✗	✗	✓	✗
Maniu and Cautis [79]	✗	✗	✗	✓	✗
Rawashdeh et al. [114]	✗	✗	✗	✓	✗
Sang et al. [94]	✗	✗	✗	✓	✓
Shafiq et al. [7]	✓	✗	✗	✗	✗
Kumar et al. [30]	✓	✗	✗	✗	✗
Cai et al. [38]	✗	✗	✗	✓	✗

This section presents the review study of various methodologies of RIP modeling based on the approach used to represent a web resource as different methodologies have different levels of resource illustration ability. The strengths and weaknesses of various methodologies of web resource representations have been highlighted in Table 2.4.

2.3.3 Discussion

This section undertakes review of the various approaches followed by different researchers for RIP modeling in their works for personalizing the users web search. Basically, RIP corresponding to a web resource provides an insightful description of resource content, topics to which a resource is related, what kind of information requirements can be fulfilled by a web resource, etc. Like a UIP, an RIP is also constructed in two phases, *i.e.*, information gathering and information quantization of web resource affinity. Both the information categories of web resource illustration, *i.e.*, content based modeling of RIP and tag based RIP modeling have been reviewed here. As the content based modeling of RIP has many drawbacks, the tag based modeling has been preferred for proposing a novel methodology for RIP modeling and that has been discussed at length in Resource Illustration Profile (RIP) modeling and Personalization of web resources Chapter of this thesis.

2.4 Personalization Methodology

After modeling a strong user interest profile and resource illustration profile, researchers have now utilized the acquired knowledge of UIP and RIP for personalizing the user's web search. Typically, the research community working on information retrieval and

ranking strategies has divided the web search personalization into two categories based on methodology to perform the personalization task. In the first category, personalization is achieved by re-ranking of retrieved results, while query re-formulation acts as a basis of personalization in the second category. A detailed review of the various strategies relating to both these categories is presented in the subsequent subsections.

2.4.1 Web Resource Re-ranking

Re-ranking of web resources, already retrieved as a result of user's query by a web search engine, is the most widely used approach for personalization task. Then, ranking action is performed to re-arrange the web resources in a list of retrieved results which are put into a certain order by determining query issuers's interest. Web resources with greater relevancy to user's interest are accorded higher ranks in the ranking hierarchy, followed by other web resources as per their relevance values. This has been clearly explained by an example in Fig. 2.3. As per the example, the originally retrieved list of web resources by a search engine for a query *Mathematics* issued by user Geeta has been re-ranked according to her UIP. As is evident from Geeta's UIP that she is more interested in viewing study portals of the subject rather than books. However, the originally retrieved list has kept web resources corresponding to subject books like Carol Hurst and Bookstore at higher ranks which is undesirable as per user interest. Therefore, considering the user preferences, personalization of the user's web search has been performed by re-ranking the web resources to place interactive math portals like Math Open Reference, etc. at higher ranks and increase the user satisfaction level. In the personalization work presented in this work, two web resources cannot be assigned the same rank, therefore, a ranking hierarchy is implemented as ranked list of web resources. Both the ranking hierarchy and ranked list has been used interchangeably in this thesis. The search engines like Google, Yahoo, Bing, etc. have been commonly used by the researchers to obtain the original list of web resources to act as input for their re-ranking approaches. Moreover, the entire result set present in the original list is not re-ranked by the researchers, but only the top n number of web resources are utilized for an experiment [7, 33–35, 53, 54].

The approach based on re-ranking was devised by Dou et al. [53] for performing a web search personalization task in their work, where original result list was retrieved from

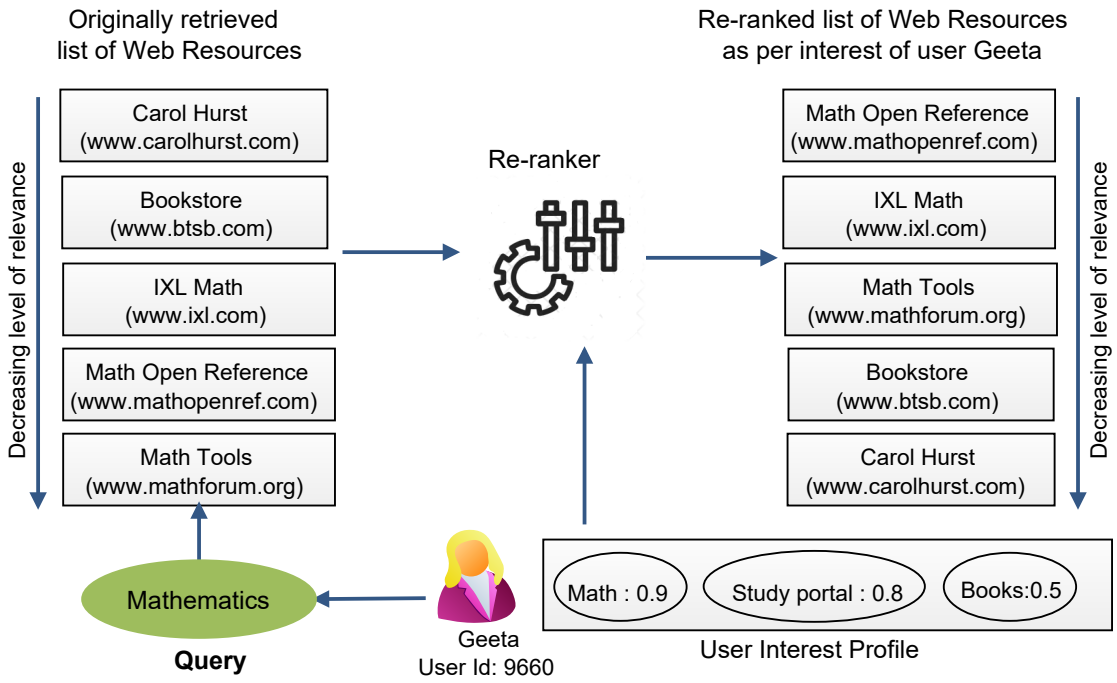


Figure 2.3: Re-ranking of Web Resources according to User Interest

MSN search engine. However, the authors utilized only the first 50 results from the result list, and calculated a personalization score corresponding to each selected web resource. After this, web resources were re-ranked according to their calculated scores. However, still the obtained result list was not the final one as it only depicted the relevance of web resource for a query issuing user, *i.e.*, UIP-RIP mapping. However, the calculation of Query-RIP mapping score still need to be examined. But Dou et al. [53] were unable to get relevance mapping scores of Query-RIP, thus, they used the ranks accorded to web resources in the originally retrieved list by a search engine to depict the relevance of a web resource for the query. Finally, the authors fused together the ranks given on the basis of UIP-RIP mapping with ranks in the original list using rank fusion method of Dwork et al. [132] in order to obtain the final personalized ranking of web resource.

Like Dou et al. [53], Vu et al. [133] also followed re-ranking approach for personalization of web search. However, instead of using MSN, the authors used Bing search engine to the obtain original result list; and the relevancy mapping of UIP-RIP was only performed on top 10 results. The researchers have also used the ranks of web resources in the original result list to get relevance mapping scores of Query-RIP instead of calculating these scores. The rank fusion approach followed by these researchers is better than the one followed by Dou et al. [53]. Some researchers used not only the simple term-based, but

conceptual relations in their re-ranking approach design. Two concept based re-ranking algorithms, *i.e.*, GEW and 3C were devised by Hawalah and Fasli [92] for calculating the mapping score of ontology based UIP and RIP in their personalization framework. Similarly, Shrivastav et al. [134] have also used ontology information to design a framework for searching and ranking multimedia content on web, while and Susan et al. [135] performed personalized web search based on ontology information.

Many researchers also made use of social information present around query issuing users and web resources retrieved by search engine in the original result list for performing re-ranking under the personalization task. Basically, they used social information based user's UIP and tag based RIP of web resources for re-ranking. Vallet et al. [85], Xie et al. [96] and Cai et al. [38] used keyword similarity-based measure on tag based UIP and RIP to calculate the relevancy score of web resources retrieved in the original list by search engine for query issuing user. Vallet et al. [85] prepared the final personalized list of web resources according to user's interest on the basis of UIP-RIP mapping score values without giving any consideration to query relevancy score. On the other hand, Xie et al. [96] and Cai et al. [38] computed Query-RIP mapping scores using keyword similarity measure. The authors then computed the final personalized scores as linear aggregation [96] and multiplication [38] of UIP-RIP and Query-RIP mapping scores to obtain a final ranked list. According to Vallet et al. and Cai et al., the results obtained on the basis of social information are more satisfactory than traditional methods. The linear aggregation function as used by Xie et al. [96] is the most widely used approach for calculating the final personalized score from trade-off of UIP-RIP mapping score and Query-RIP mapping score, either computed by authors or extracted from original result list provided by search engine [17, 23, 31, 88]. In addition to tag vector-based mapping of UIP and RIP, Gemmell et al. [82] used the concept of tag clustering. The association of various users and web resources with the constructed clusters was computed; and on the basis of commonality in cluster associations of user and web resources relevancy score was calculated. Basically, the frequency of tags in shared cluster is used to quantify the relevancy score.

To quantify the relevance of a tag based web resource profile for user, Maniu and Cautis [79] devised the concept of user social networking graph where score was calculated on

the basis of distance or other properties of intermediate neighbours. The re-ranking was then performed by authors to personalize the result list as per user interest. Another graph based relevancy measurement approach, *i.e.*, Katz proximity [136] was used by Rawashdeh et al. [114] for UIP-RIP mapping, where the proximity of two nodes was the weighted sum over the collection of possible paths connecting the nodes. Basically, by proximity, the authors meant to represent the closeness of two nodes, *i.e.*, user or web resource. To decrease the complexity of graph based re-ranking approaches Shafiq et. al. [7] used the concept of fractional cascading. The concept of ontological relationship was also used by researchers like Xu et al. [91] for calculating the relevancy mapping score of tag based UIP and RIP. Firstly, the authors mapped the user and web resource tags with pre-defined ontologies; and then, a semantic similarity was computed between ontologically mapped profiles. The re-ranking of web resources was performed based on their relevancy score values for a user. Hunag et al. [105] also designed a probabilistic inference model for calculating mapping scores of user and web resource profile vectors instead of using a traditional keyword matching.

2.4.2 Query Re-formulation

In an information retrieval system, generally, a set of keywords known as query is used by every user to express his information needs or requirements. But due to diversity in the interest of users and presence of many web resources, same keywords are being used variously by the users in a query for different information requirements. However, the researchers have tried to find the solution of this problem through query re-formulation. It is defined as follows:

Definition 2 *A **Query re-formulation** is the process of transforming the original query Q issued by a user to any web search platform into another query $Q^{'}$, either by reduction or expansion of Q.*

Where, transformation by query reduction [137] means removal of superfluous information from Q to give $Q^{'}$, while expansion of query [138] means enriching Q with some additional information to obtain $Q^{'}$. To the best of our knowledge and recent survey conducted by Bouadjenek et al. [128], there is hardly any contribution towards social information based query reduction, but query expansion is found to be the basis of personalization task in

almost every existing work.

Zhou et al. [86] devised a statistical tag-topic model to extract the terms from query issuing user's UIP which were quite relevant to the terms in user's query. The selected terms for user's UIP were then used for query expansion in an appropriate manner. Similarly, Chirita et al. [139] also directly used the terms of UIP for query expansion in their work. The authors studied the minimum cardinality of terms required for query expansion, and suggested that the decision of cardinality is completely determined by different features of the original query like length, topical purity or topical distinctness, etc. The concept of machine learning has also been used by some researchers like Yin et al. [140] for query re-formulation by identifying the level of similarity between the issued query and query present in user's log history. The authors also suggested to use small snippets associated with a web resource instead of entire text of a web resource. The help of statistical moments and SVM classifier is taken by Singh et al. [141] for designing an image retrieval system based on visual content and annotations at the time of query by a user.

In the recent years, many web search engines have also started providing the facility of query suggestion and automatic completion. Mostly, this facility is based on either location, popularity of certain types of queries or pre-searched query by the same or different users on the search engine servers. In the case of query suggestion, a list of many similar queries is recommended to the users in order to help them to formulate a better query for more qualitative information retrieval. However, instead of utilizing the generalized data for query suggestion or auto-completion, it should be more user's interest-oriented in order to provide a personalized search experience to him [142–145]. Adeyanju et al. [142] analyzed user's UIP, search patterns, and search behaviour of other similar users to determine a suitable recommendation of query list, whereas Cai et al. [146] and Shokouhi [147] used only UIP of a user.

The knowledge extracted from collaborative tagging information has also been explored by many researchers for query expansion as collaborative tagging platforms provide more qualitative information from user's point of view. Lioma et al. [148] devised the technique for query expansion by utilizing logical inference similarity of query terms and various tags used by different users to enrich the query with additional information. The tags collection has also been used by Jin et al. [149] for query expansion, but here the authors prefer to

use co-occurrence similarity approach of query terms and tags for selecting the tags for expansion. In addition to traditional approaches to measure similarity of terms and tags, Lin et al. [150] proposed a machine learning based term ranking to select more relevant tags to query terms. The tag based approaches used to provide an identical expansion to every user are not desirable and satisfactory as relevance of the same web resource varies from user to user. So, query expansion must be personalized [151, 152].

Bertier et al. [153] devised an adaptation of famous PageRank algorithm named TagRank algorithm to identify the highly relevant tags form user's UIP for query expansion for which help of Tagmap matrix was taken by the authors. Similarly, Biancalana et al. [154] proposed Nereau based on co-occurrence relation of query terms and user UIP tags. The concept of ranking the tags of UIP was also used by Bouadjenek et al. [155] in their work for personalized query expansion. The authors considered the semantic similarity of tags and query terms. Moreover, proximity relation of query and query issuing user was also taken into consideration by the authors while designing a tag ranking algorithm for selecting the candidates of query expansion. Some authors like Kumar et al. [30] used the concept of tag clustering and clustered user interest profile to provide a disambiguation to user's query for better topic relevant expansion. Mulhem et al. [102] designed a probabilistic matching framework for UIP based query expansion to provide personalization.

2.4.3 Discussion

This section undertaken here includes the approaches used by different researchers for finally personalizing the user's web search by applying the knowledge about user preferences and web resource information types and affinity levels. Typically, there are two categories of web search results personalization, *i.e.*, results re-ranking and query re-formulation. Basically, re-ranking of web search results list retrieved by a search engine for the original query is a post-search personalization as per user preferences and web resource RIP. In comparison to this, query re-formulation is the pre-search personalization where query is re-formulated or re-designed as soon as query is issued by the user to a search engine. However, this research work aims at designing a novel result re-ranking approach based on post-relevancy score of a web resource. The step by step procedure used for computing the post-relevancy score of a web resource along with UIP-RIP and Query-RIP

mapping scores has been described in Resource Illustration Profile (RIP) modeling and Personalization of web resources Chapter of this thesis.

2.5 Research Questions

The generic research question which forms the basis of this thesis is:

How can we design a personalization model for improving the user satisfaction towards the web search ?

The various terms involved in the question need to be defined explicitly. Firstly, the model for web search is *personalized*, *i.e.*, the web search results returned to a query issuing user follow a customization approach; not the *"one size fits all"* approach. Secondly, *user satisfaction* can be considered as the quality of web search results returned to the user in response to a query issued by him to a search engine. The most feasible meaning to quality in terms of user satisfaction is the ranking of search results with respect to user preferences and query relevancy, starting with most relevant one at the header of a ranking hierarchy, followed by others as per their relevance values. This scenario for personalization will not only increase the user satisfaction, but also help to save user's time that got wasted in navigating through the links to irrelevant web resources, and movement from one result page to another.

A personaliztion model is not a single standalone entity, but a interrelated contribution of multiple supporting modules, *i.e.*, UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation. The final ranking of web resources is performed on the basis of their post-relevancy score values. Firstly, the current research work concentrates on how to utilize the collaborative tagging information of a user in order to construct his UIP. It addresses the problem of predicting an approximate value for the level of interest that a user may hold for a particular tag, and answers the following question:

RQ 1 *How to construct a user's UIP using his collaborative tagging information ?*

To answer this question, various approaches [26,31,34,84,85] were investigated to quantify the user's level of interest for a particular tag. The selected approach was then utilized for UIP construction. Recent studies [7,30–32] on various methods of user profile construction have indicated that the profile constructed by utilizing only the information generated by a

user himself is very weak and sparse. The problem identified is not with the approach used to model user profile from user's own tagging information, but with the quantity of that information. For some users, abundant information is available, while for others, it is very scarce as it all depends on activeness of a user on the social web. Therefore, to tackle this problem, various researchers have suggested the employment of a UIP enrichment strategy for a strong user profile. However, they all neglected the fact that in today's time a single UIP enrichment strategy would not be able to create a strong UIP. Moreover, selection of a strategy for UIP enrichment also matters a lot; it should be more real-world approximation of the problem. This problem associated with UIP enrichment has been addressed through the following question:

RQ 2 *How can we perform the user profile enrichment for construction of a strong UIP?*

The purpose has been attained by studying the scope of multiple strategies, with each using different learning and pattern recognition algorithms for UIP enrichment. All the strategies also take into account the real-world relationships of user's information along with information relating to others for UIP enrichment. This research work proposes a new methodology for the construction of a UIP through collaborative tagging information of the user. Both implicit and explicit methods have been used to collect the required information at different points of UIP construction.

The user interest profiling has always been the backbone of a personalization model; and this research topic has been studied most extensively for web search personalization. In addition to it, there are some other supporting modules of personalization also which could not get due attention. However, these modules do not make as much impact as a UIP, but still they can significantly impact the performance of a personalization model. One such supporting module is Resource Illustration Profile (RIP) which provides a summarized viewpoint of all users who have tagged that web resource. Therefore, this type of RIP is called as collective RIP. Some recent studies [33,34] on RIP construction have unfolded the presence of outliers in collective RIP. Basically, outliers are missaligned user viewpoints, either intentional or unintentional, which degrade the ranking of a relevant web resource and are completely undesirable in effective web search personalization methodology. Thus, the problem associated with RIP construction has been addressed through the following question:

RQ 3 *How to construct a RIP of a web resource in collaborative tagging system ?*

To address this question, the concept of social filtering has been incorporated in the work presented in this thesis. The social filtering will help to detect outlier tags responsible for missaligned viewpoints and depurate the RIP.

For each user and web resource, the corresponding UIP and RIP were created respectively, which were then utilized by web search engine for personalization of search results. A relevancy mapping of a user's UIP and RIP of every web resource is performed by search engine in order to rank web resources according to their relevance for the user. Along with UIP-RIP mapping, the RIP of a web resource must also be mapped with user's query because it represents the current need of the user, while a UIP represents the information which a user usually prefers. Although, there are some inherited deficiencies like synonymy problem in both UIP-RIP and Query-RIP mapping which have been generally neglected. In most of the studies, there is hardly any well-defined formulation for mapping functions; and even those presenting the mapping functions have just used keyword matching only. Therefore, this work addresses the problem associated with mapping task through the answer to the following question:

RQ 4 *How can we perform query relevancy and user interest relevancy mapping for a web resource ?*

This problem has been addressed by proposing suitable mapping functions for performing the relevancy mapping of a web resource with respect to user and query. The proposed functions are more real-world approximation of the mapping problems in comparison to the existing ones. The research then focuses on the final supporting module of web search personalization, *i.e.*, ranking of web search results. The post-relevancy score of a web resource, computed on the basis of trade-off between query relevancy and user interest relevancy score, has been used to perform the final ranking of web resources. As per the studies on web search personalization, trade-off parameter value can be fixed for the entire personalization system irrespective of query and user issuing that query [7, 31, 34]. But trade-off parameter value should not be fixed; and this has been explained through the following question:

RQ 5 *How to compute a suitable trade-off parameter value for every instance in a per-*

sonalization system ?

To answer this question, the work presented in this thesis proposes a new methodology for query and user dependent trade-off parameter value computation. After addressing all the aforementioned research questions, the question arises whether the proposed personalization methodology helps to improve the user satisfaction towards the web search. This concern has been addressed with the help of each supporting module of personalization which finally ranked the web resources in order of their relevancy level; and then the quality of ranking has been evaluated by using various evaluation metrics.

2.6 Thesis Objectives:

Thesis Objectives: On the basis of research questions as mentioned above, the objectives formulated for this work are as follows:

1. To review the literature concerning contributions made by different Social Information Retrieval (SIR) models on the basis of their information retrieval policies.
2. To perform pre-processing of the collected social information and form a structured dataset.
3. To design a novel methodology for the personalization of web search.
4. To validate and verify the proposed approach using user studies.

Chapters 3 to 5 addresses the aforementioned research questions. Moreover, all the chapters of the thesis, individually or collectively, fulfil the corresponding objectives. The summary provided at the end of each chapter describes the answer corresponding to the research question presented in that chapter. The last chapter of this thesis summarizes the answer to each research question. The subsequent sections highlight the main contribution of this research work, and also present the organization of this research work.

Chapter 3

Personalization model and experimental methodology

The web search personalization model proposed in this thesis formulates the personalization task as a web resources re-ranking problem as per user's interest [7, 14, 17, 23, 31, 33, 34, 96]. However, the methodology followed by proposed model to accomplish the task is quite different. As per the analysis of literature, a web search personalization is not a single standalone entity but an interrelated contribution of multiple supporting modules. The key modules to support personalization as listed by various researchers in the literature are UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation[1]. On the basis of computed post-relevancy score, final personalized ranking of web resources is performed. However, different researchers have concentrated on improvement and efficiency enhancement of different modules, but there was hardly any research work that have considered all the modules. Therefore, to provide a relevant personalized ranking of web resources to query issuing user, a novel personalization model has been proposed inconsideration to every key supporting module of web search personalization. An effective and efficient web search personalization increases not only the user satisfaction level towards a web search engine, but also decreases the number of required computation cycles which are generally wasted due to repeated query reformulation in case of irrele-

[1]The contents of this chapter are partly published in:

- Shubham Goel and Ravinder Kumar. "Brownian Motus and Clustered Binary Insertion Sort methods: An efficient progress over traditional methods", Future Generation Computer Systems (FGCS), 86, pp. 266-280, 2018.
- Shubham Goel and Ravinder Kumar. "Folksonomy-based User Profile Enrichment using Clustering and Community Recommended tags in Multiple Levels", Neurocomputing, 321, pp. 425-438, 2018.
- Shubham Goel and Ravinder Kumar. "SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction", Expert Systems With Applications, 141, pp. not assigned till yet, 2019.
- Shubham Goel and Ravinder Kumar. "Collaboratively Augmented UIP - Filtered RIP with Relevancy Mapping for Personalization of Web Search", Information Sciences, 2020. (Accepted)

vant information retrieval. The consolidated picture of the interrelated contribution of key supporting modules considered under proposed model for web search personalization is represented in Fig. 3.1. However, methodologies proposed for constructing the different supporting modules as depicted in Fig. 3.1 will be covered in upcoming chapters in a one-by-one fashion.

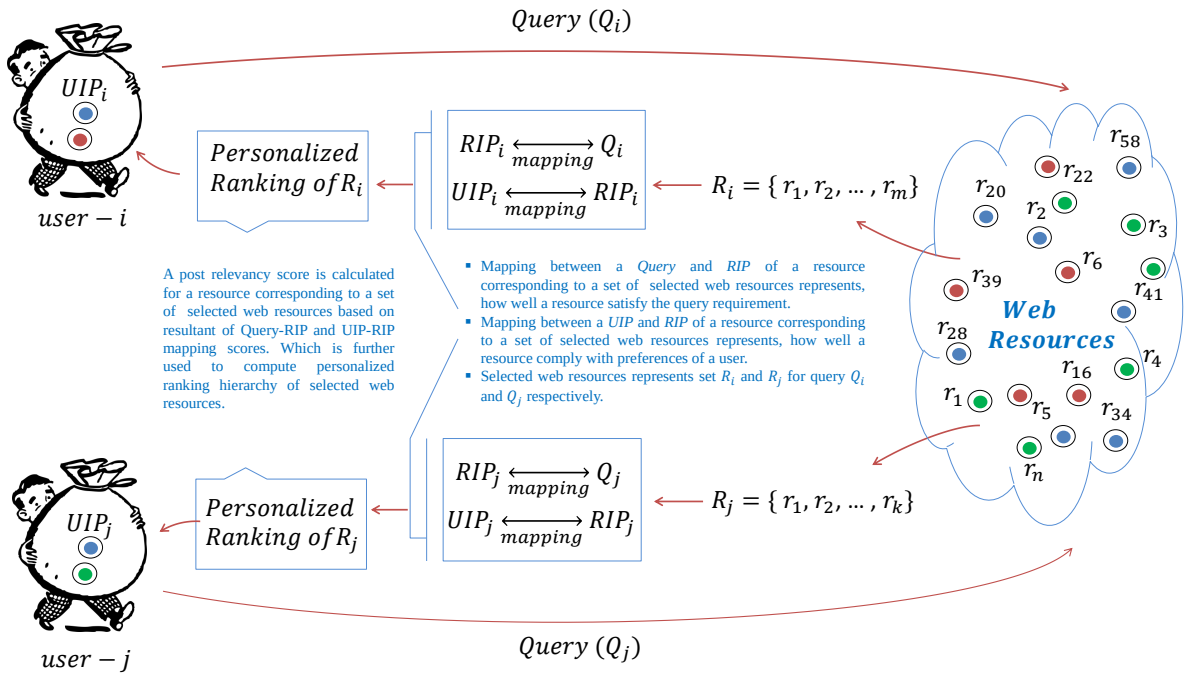


Figure 3.1: Web search personalization model.

Basically, in simple terms, it can also be said that personalization model is a type of ranker which ranks web resources according to user interest, by mapping their informativity level for user and query requirements. The web resources which act as input to the personalization model as depicted by R_i and R_j in Fig. 3.1 is the originally retrieved ranked result list of web resources corresponding to the user's query by a search engine utilizing it's inbuilt traditional web resource ranker.

To demonstrate the effectiveness of the proposed model for personalization of user's web search, extensive experiments were performed on a dataset of voluminous size. As already discussed, that performance of a personalization model is governed by the methodology adopted for supporting modules construction. Therefore, the experimentation presented in this work is not limited to only final personalization task, *i.e.*, re-ranking of web

resources, but also other various supporting modules considered in proposed web search personalization model.

This chapter provides a detailed description of the datasets used, evaluation metrics, and state of the art methodologies for the purpose of a comparison. Moreover, the performance of constituent supporting modules of proposed model is also compared with the one used by other extant personalization systems. In contrast to the general approach of implementation setup used by traditional systems which consist of only dataset and evaluation metrics, the proposed model also accommodates the algorithm used for performing various sorting jobs. To implement all the algorithms of the proposed model Python 2.7.10 Shell has been used. The training and testing hardware used to obtain the results of experiments consisted of Dell Workstation T5600- with 2.6 GHz Intel Xenon e5 2650 CPU and 8GB 1600MHz DDR3 RAM, running windows platform.

3.1 Dataset Description

Two different types of datasets have been used to conduct the experiment. First one is the folksonomy data collected from del.icio.us[1]; and second is the entire collection of web pages of Wikipedia[2]. In the del.icio.us dataset, there are about 1867 users, 53388 tags out of which there are 10358 unique tags, 69226 web pages, and 437593 folksonomy relations. The user-tag frequency data has nearly 179203 records, where each row denotes the frequency of a particular tag used by a user. On an average, each user has used 24 tags and annotated 65 web resources. For more details about del.icio.us dataset, refer Table 3.1.

Table 3.1: Detailed description of del.icio.us dataset

Dataset features	Quantity
Folksonomy Relations	437593
Users	1867
Tags before pre-processing	53388
Tags after pre-processing	38937
Web Pages	69226
Bidirectional User-User Relations	7668

[1]https://http://del.icio.us.com
[2]https://meta.wikimedia.org/wiki/Data_dumps

The dataset collected from del.icio.us cannot be directly used in the experiment as it has some deficiencies due to the presence of noisy elements. Conducting the experiment without resolving the influence of noisy elements can result in lower efficiency. Therefore, to tackle the impact of noisy elements, regular expressions and python scripts are used to perform pre-processing tasks. Some of the pre-processing tasks are as follows:

1. Removal of tags which are only a concatenation of integers or some special symbols.
2. Alphanumeric tags are processed to leave only the textual part.
3. Tokenization of tags like "virtual#literacy_social.networking#a#wave#of_learning "or "informationvisualization"as these can be easily understood by human eyes, but difficult for machine.
4. Porter's algorithm [156] was used to perform stemming operation on some tags.
5. Meaningless and too personal tags are removed.
6. Some tags are just stopwords like "ourself "or "between ", etc. which are completely undesirable in the final dataset.
7. Finally, every Non-English tag has also been removed.

English Wikipedia dataset have been selected for building a large corpus of word vectors using Word2vec model. The dataset is public, thoroughly described and analyzed by Wikimedia Foundation [157]. As dataset is available in XML format, it cannot be directly used for Word2vec model training. Therefore, it has been converted into text format. Finally, after the completion of training process, a trained model, a large corpus of unique words and their corresponding word vectors have been obtained. The corpus has nearly a collection of 880802 word vectors. In addition to word-vectors, a total of 261 topics of interest were selected from Open Directory Project (ODP) to be utilized for community modeling and filter out the outlying tags from web pages by Intelligent Collaborative Filtering (ICF).

The del.icio.us dataset was used to evaluate the proposed methodologies corresponding to various supporting modules of proposed model and validate using 5-fold cross-validation. At the training stage, User Interest Profile (UIP) and Resource Illustration Profile (RIP) were modeled using corresponding methodologies designed under proposed model and

state of the art models, whereas the testing set was used to formulate the user queries. The evaluation metrics selected for the purpose of performance comparison of personalization models have been discussed in the Section 3.3. However, state of the art models with which comparison of proposed model is made have been discussed in the upcoming chapters in a supporting module by module fashion.

3.2 Sorting Algorithm

In most of the personalization work, studied in the literature, researchers have used standard sorting algorithms for performing various sorting tasks in different supporting modules of personalization. For example, in UIP the preferences of users are sorted according to the level of interest user have in the corresponding tags. But the tags in unsorted version of UIP are usually in partially sorted order that can not be efficiently sorted using standard sorting algorithms as this will just increase the time complexity of the system. So to handle this issue in proposed model, the Clustered Binary Insertion Sort (CBIS) algorithm designed by Goel and Kumar [158] specially to sort partially sorted data with less time complexity is used in this thesis. The CBIS algorithm is as follows.

Clustered Binary Insertion Sort (CBIS) algorithm is an improvement over Insertion sort (IS) and its famous variant Binary Insertion Sort (BIS) algorithms. In CBIS, firstly, a comparison of the element at Current Pointer (COP) is made with the element at Position Pointer (POP), and then, with elements either in left or right subparts. Here, after the decision of subpart, a binary search is performed on elements in the selected subpart. Function $binary\_loc\_finder$ is used to perform binary search operation and location identification to insert an element at COP into either left or right selected subpart. POP is updated to the latest location value. Insertion and shifting operation is handled by $place\_inserter$ function. Final outcome after completion of all iterations is a sorted list of elements. This name for the proposed algorithm CBIS has been chosen for the reason that after the decision of subpart, a cluster of elements is formed ranging $[0, POP - 1]$ for the left subpart and $[POP + 1, COP - 1]$ for the right subpart. Therefore, due to the use of binary search logic on a cluster of elements, it is named as Clustered Binary Insertion Sort. CBIS is much more efficient than IS, BIS and Brownian Motus Insertion Sort (BMIS) in terms of a number of comparisons required to identify correct location

Algorithm 1: Clustered Binary Insertion Sort

Input : A list of uniformly distributed elements *a_list*
Required : A Sorted list
Initialize : $POP \leftarrow 0$ ▷ *POP is position pointer*
for $i \leftarrow 1$ *to* $length(a\_list) - 1$ **do** ▷ *a_list start at index 0*
  $COP \leftarrow i$ ▷ *COP is current pointer*
  $key \leftarrow a\_list[COP]$
  if $key \geq a\_list[POP]$ **then** ▷ *left or right movement decision*
    $place \leftarrow binary\_loc\_finder(a\_list, POP+1, COP-1, key)$ ▷ *right movement*
  else
    $place \leftarrow binary\_loc\_finder(a\_list, 0, POP-1, key)$ ▷ *left movement*
  $position \leftarrow place$ ▷ *POP is updated*
  $a\_list \leftarrow place\_inserter(a\_list, place, current)$ ▷ *Insert COP in sorted list*
  $i \leftarrow i + 1$
Function $binary\_loc\_finder(a\_list, start, end, key)$**:**
  if $start == end$ **then**
    if $a\_list[start] > key$ **then**
      $loc \leftarrow start$
      return loc
    else
      $loc \leftarrow start + 1$
      return loc
  if $start > end$ **then**
    $loc \leftarrow start$
    return loc
  else
    $middle \leftarrow \lfloor \frac{start+low}{2} \rfloor$
    if $a\_list[middle] < key$ **then**
      return $binary\_loc\_finder(a\_list, middle+1, end, key)$
    else if $a\_list[middle] > key$ **then**
      return $binary\_loc\_finder(a\_list, start, middle-1, key)$
    else
      return $middle$

Algorithm 2: Shifter

Function $place\_inserter(a\_list, start, end)$**:**
  $temp \leftarrow a\_list[end]$
  for $k \leftarrow end$ *to* $start$ **do**
    $a\_list[k] \leftarrow a\_list[k-1]$
    $k \leftarrow k - 1$
  $a\_list[start] \leftarrow temp$
  return a_list

for the element at COP. For step by step explanation of CBIS, refer Algorithm 1; and for *place_inserter* function, refer Algorithm 2. Moreover, for details regarding complexity

analysis and performance of CBIS refer Goel and Kumar [158].

3.3 Evaluation metrics

In an Information Retrieval (IR) system, evaluation is a type of quantification process where a dedicated metric is associated with the results provided by an IR system in response to various queries issued by a same or different users [159]. Basically, evaluation metrics accounts the relevance of retrieved results for a query issuing user. In this thesis, performance evaluation using different metrics only meant the quantification of quality and relevancy of retrieved results not the physical performance of the system, *i.e.*, processing speed of user queries by IR system. The metrics selected for evaluation will be discussed from two viewpoints firstly for quality of user's profile and then for final personalized ranking of web resources by a personalization system. The separate discussion according to UIP is only provided to make concept more clear as UIP is a backbone module of every personalization system.

Four metrics have been used to evaluate the efficiency and effectiveness of the methodology proposed for UIP construction under proposed model of personalization. The first one is Precision@K ($P@K$) presented by Xie et al. [33]. This is the commonly used evaluation metric to measure the accuracy of obtained UIP as follows:

$$P@K = \sum_{i=1}^{n} \frac{rel(t_i)}{n} \tag{3.1}$$

$$rel(t_i) = \begin{cases} 0, & if\ t_i \notin Top\_K \\ 1, & if\ t_i \in Top\_K \end{cases} \tag{3.2}$$

Where, $rel(t_i)$ denotes the relevancy of tag t_i of testing set for the obtained UIP. $P@K$ is the percentage of tags in the testing set present among K most favourable tags in obtained UIP of a user. More the value of $P@K$, greater would be the accuracy of UIP construction methodology.

The second metric is Mean Reciprocal Rank (MRR), measured as a multiplicative inverse of the rank of a target tag t_i of testing set, averaged across the number of target tags in the testing set. It is the most widely used metric in the field of information retrieval by

search engines and recommender systems. Formally, MRR is defined as:

$$MRR = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{r_i} \tag{3.3}$$

Where, r_i is the rank of target tag t_i into UIP.

The third metric is *imp* presented by Shepitsen et al. [87]. It is measured as an improvement in the rank of a target tag t_i of the testing set. Formally, *imp* is defined as:

$$imp(t_i) = \frac{1}{r_i^p} - \frac{1}{r_i^s} \tag{3.4}$$

$$imp = \frac{1}{n}\sum_{i=1}^{n} imp(t_i) \tag{3.5}$$

Where, r_i^p and r_i^s are the rank of target tag t_i into UIP constructed by proposed and state of the art methodology respectively under corresponding personalization models ; and n is the number of tags in the testing set.

The fourth metric is *completeness*, a very appropriate metric to define the effectiveness of proposed methodology of UIP construction. It is measured as the percentage of target tags of the testing set present in the final constructed UIP, no matter what is the position of these tags in UIP ranking hierarchy. Formally, *completeness* is defined as:

$$completeness = \frac{|Test_t \cap UIP_t|}{|Test_t|} \tag{3.6}$$

Where, $Test_t$ and UIP_t are testing and UIP tag set respectively. Basically, the target tags of testing set are also the actual tags which are used to annotate web pages by a user. Due to division of user actual tags into training and testing sets for evaluation purpose, there are chances that some tags assigned to testing set do not have any instance in training set. Therefore, this metric helps to evaluate how many of these types of tags are successfully predicted in final constructed UIP. If all the tags get predicted, then *completeness* value will be 100% which means UIP covering each and every preference of the user, but it is the ideal case which doesn't exist till now.

In order to evaluate the final ranking by a personalization model the metrics as discussed above have been used, *i.e.*, $P@K$, MRR , Relative Improvement in Location (RIL) and

Completeness (*comp*) but viewpoint corresponding to personalized ranking. Each metric analyze different aspects of the ranking. Table 3.2 provides the mathematical formulation and description of the selected evaluation metrics. In each metric, Q refers to the set of queries and U_t set of users formulated using data in a testing set; and q_i represents a target query of the testing set.

Table 3.2: Metrics for experimental evaluation of proposed web search personalization model

Metric	Mathematical Formulation	Description
Precision@K ($P@K$)	$\frac{1}{\|Q\|} \sum_{q_i \in Q} L_i$	where, L_i depicts the location of target web page among the result set retrieved by a personalization model for a query q_i. If L_i is among topmost K web pages returned to user, then, L_i is equal to 1, otherwise 0.
Mean Reciprocal Rank (MRR)	$\frac{1}{\|Q\|} \sum_{q_i \in Q} \frac{1}{L_i}$	where, L_i depict the location of first relevant web page among the web pages retrieved as response to a user by personalization model for a query q_i.
Relative Improvement in Location (RIL)	$\frac{1}{\|Q\|} \sum_{q_i \in Q} (\frac{1}{L_i^p} - \frac{1}{L_i^b})$	where, L_i^p and L_i^b depict the location of first relevant web page among the web pages retrieved as response to a user by the proposed and baseline model respectively for a query q_i.
Completeness (*comp*)	$\frac{1}{\|U_t\|} \sum_{u_i \in U_t} \frac{\|W_{test}^{u_i} \cap W_{recom}^{u_i}\|}{\|W_{test}^{u_i}\|}$	where, $W_{test}^{u_i}$ and $W_{recom}^{u_i}$ depict set of web pages that relevant to user u_i in testing set and set formed by the web pages that are recommended as relevant to user u_i respectively by personalization model.

3.4 Summary

This chapter provides an outline description of the model proposed for personalization of web search. An abstract role of various supporting modules, considered in proposed model, representing their interrelated contribution towards personalized search is also discussed. In particular, the datasets collected for implementation of proposed model and its constituent modules is presented in this chapter. The evaluation metrics along with mathematical formulation chosen for performance evaluation of proposed UIP modeling methodology under proposed model and final ranking allotted by proposed model to various web resources are also discussed. Moreover, the chapter also discussed about the algorithm used to perform the sorting at various instances in different supporting modules of proposed personalization model. In the next chapter, a methodology used to perform

the task of UIP modeling, under the proposed personalization model, has been proposed.

Chapter 4

User Interest Profile (UIP) modeling

The User Interest Profile (UIP) of a user enlists the preferences of a user along with the degree of interest in each preference, where the process of acquiring and quantization of information related to user preferences is known as UIP modeling. According to research works studied in the literature, UIP is the centroid, *i.e.*, main supporting module of any personalization model. In the simple layman terms, it can also be said that UIP is a backbone of every personalization model. Any claim of designing a personalization system without the knowledge of user's interest, *i.e.*, absence of UIP construction or learning model is baseless as presenting the information to a user according to his/her preference is the only goal of personalization. Moreover, it can also not be denied that only an efficient and complete UIP can lead to an effective and high performing web search personalization methodology design. Therefore, in this chapter the proposed methodology for the modeling of User Interest Profile and its enrichment has been described, as UIP based on only user's self information is sparse and incomplete[1]. Along with proposed methodology for UIP modeling, the performance comparison of proposed methodology for UIP and state-of-the-art methodologies is also present in subsequent sections.

[1]The contents of this chapter are partly published in:

- Shubham Goel and Ravinder Kumar. "Brownian Motus and Clustered Binary Insertion Sort methods: An efficient progress over traditional methods", Future Generation Computer Systems (FGCS), 86, pp. 266-280, 2018.
- Shubham Goel and Ravinder Kumar. "Folksonomy-based User Profile Enrichment using Clustering and Community Recommended tags in Multiple Levels", Neurocomputing, 321, pp. 425-438, 2018.
- Shubham Goel and Ravinder Kumar. "SoTaRePo: Society-Tag Relationship Protocol based architecture for UIP construction", Expert Systems With Applications, 141, pp. not assigned till yet, 2019.

4.1 UIP Modeling

In today's time, the interest of a user cannot be determined solely on the basis of a single strategy. There must be multiple strategies using different pattern recognition and learning algorithms. Three strategies have been selected and deployed in the current work. Based on these strategies, a three-level UIP, *i.e.*, Folksonomy-based UIP (FBUIP), Cluster Recommended UIP (CRUIP), and Friends Recommended UIP (FRUIP) has been proposed. Formally, UIP is defined as:

Definition 3 *A User Interest Profile (UIP) is a list of vectors corresponding to FBUIP, CRUIP, and FRUIP of a user i and denoted by* $\overrightarrow{U}_i$ *as:*

$$\overrightarrow{U}_i = (\overrightarrow{U}_i^{fb}, \overrightarrow{U}_i^{cr}, \overrightarrow{U}_i^{fr})$$

Where, $\overrightarrow{U}_i^{fb}$, $\overrightarrow{U}_i^{cr}$ and $\overrightarrow{U}_i^{fr}$ represent the vectors corresponding to FBUIP, CRUIP and FRUIP respectively.

In order to predict user interest (direct-indirect) without any ambiguity regarding the formulation of vectors corresponding to different levels of three-level UIP, there must be clearly defined set of rules. So a dedicated protocol has been formulated for each level, *i.e.*, FBUIP, CRUIP and FRUIP. The first protocol has been used to obtain user own tags with appropriate weights, and also serves as a foundation ground for the remaining protocols, therefore, it is named as *Base protocol* (B_p). The remaining two protocols, *i.e.*, *Guild protocol* (G_p) and *Congregation protocol* (C_p) have been used to recommend tags from society relationships of user and group relations of user tags respectively. The B_p will work for FBUIP, whereas G_p and C_p will work for FRUIP and CRUIP respectively. The direct interest of a user has been identified by B_p protocol, whereas for prediction of indirect interest, *i.e.*, UIP enrichment both G_p and C_p are responsible. Formally, in terms of protocols based architecture, UIP can be re-defined and represented by a vector as shown in Eq. (4.1).

$$\overrightarrow{U}_i = F_{trade}(\overrightarrow{U}_i^{B_p}, \overrightarrow{U}_i^{G_p}, \overrightarrow{U}_i^{C_p}) \tag{4.1}$$

Where, $\overrightarrow{U}_i^{B_p}$, $\overrightarrow{U}_i^{G_p}$ and $\overrightarrow{U}_i^{C_p}$ represent the strategy vectors, *i.e.*, tags with their suitable weights obtained from B_p, G_p and C_p respectively. A strategy vector provides an insight into UIP corresponding to strategy followed by the dedicated level of a three-level UIP. The

trade-off of strategy vectors is depicted by a function F_{trade}. Now onwards, both strategy vector and UIP vector will be used interchangeably. Here, for the current research work, a single dedicated protocol is formulated for each level of a three-level UIP where a different strategy has been used at different level, for direct-indirect interest prediction. However, the proposed methodology is not only limited to these *three-protocols* or *three-levels*; it can be extended upto *n-protocol* working at *m-level* architecture depending on the number of strategies selected for direct-indirect interest prediction. Each protocol will enlists the rules of implementing the corresponding strategy. The description of the entire model is divided into two sub-modules: first explains the process of direct interest identification, whereas, second describes the process of indirect interest prediction of a user.

The generic framework of the proposed *protocol based architecture model* of UIP is represented with the help of a flow diagram in Fig. 4.1. The numbers marked into a circle

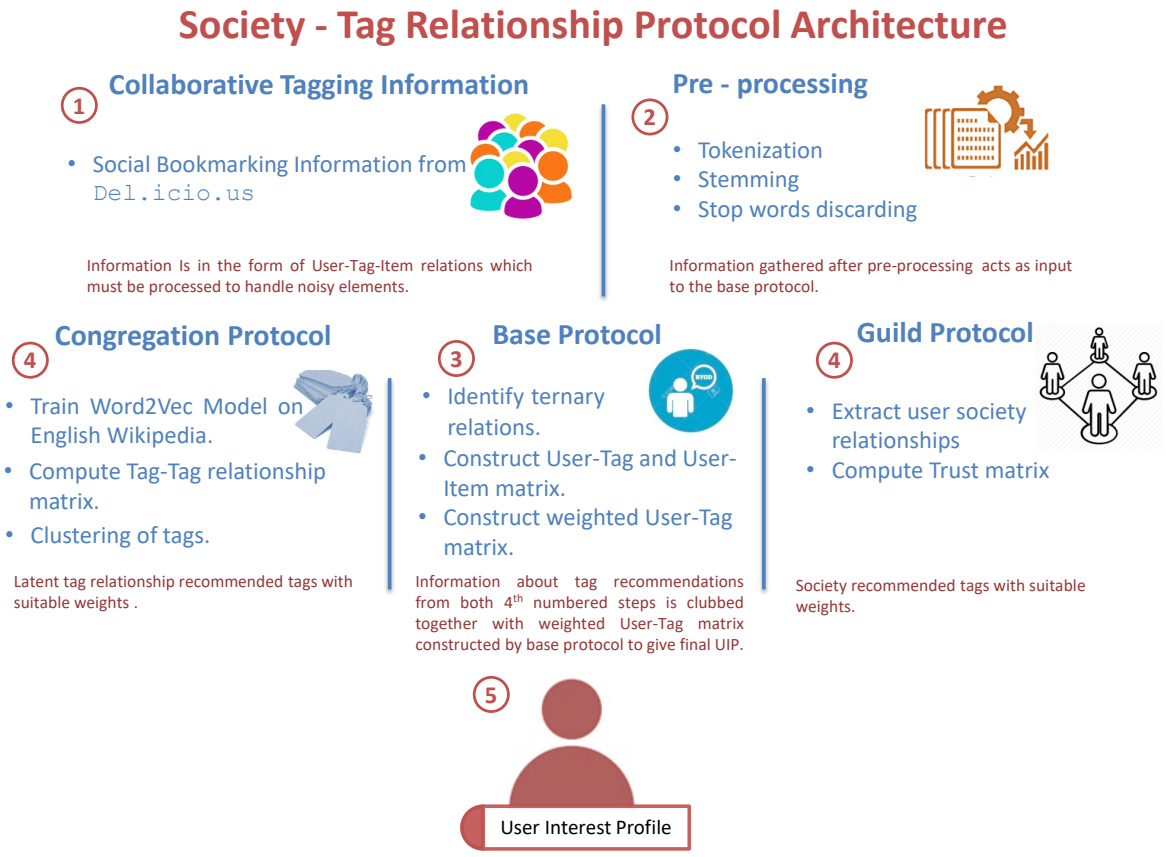


Figure 4.1: Flow diagram of UIP modeling

depicts the sequence in which that process must be performed to construct UIP in step 5. In case of Guild and Congregation protocols both steps are same numbered, *i.e.*, 4, which denotes that both can be performed in parallel. A detailed description of each protocol

is provided in subsequent subsections.

4.1.1 Direct Interest Identification

Direct interest of a user represents the user preferences which are explicitly defined by a user and identified from his/her social network activities. With the common goal to ultimately refine and personalize the user's web search, researchers from both academia and industry are extensively analyzing user generated tagging-based profile construction techniques [34, 46]. The foremost step in B_p is to identify ternary relations, *i.e.*, $R^u_{t,w}$ of a user from the collaborative tagging information as shown in Fig. 4.1. The ternary relations represent "which user has annotated which web resource with which tag ". A simplified view of the network formed by ternary relations for the whole dataset is presented in Fig. 4.2 as the complete view of the network is very complex. The tags collectively used by a user provide a valuable and precise description of user's direct interest. But as per studies of human nature and mind, a person cannot like each and every item by equal amount of interest. There always exist some difference in the degree of interest. Therefore, appropriate weight must be assigned to the tag in accordance with the degree of interest a user holds for that tag. Here, collaborative tagging information has been obtained from del.icio.us; and tags represent the activities performed by a user. The tags used by a user himself for annotating various web pages and the degree of interest assigned to them by a user constitute the strategy vector corresponding to B_p which can be formally represented as follows:

Definition 4 *Let* $\{t_1^{u_i}, t_2^{u_i}, ..., t_n^{u_i}\}$ *and* $\{\theta_1^{u_i}, \theta_2^{u_i}, ..., \theta_n^{u_i}\}$ *are the sets of tags used by user* i *himself and the corresponding degree of interest he holds for those tags respectively. For target user* i*, UIP vector obtained by* ***Base Protocol*** *(*B_p*) is represented by* $\overrightarrow{U}_i^{B_p}$ *as:*

$$\overrightarrow{U}_i^{B_p} = (t_1^{u_i} : \theta_1^{u_i}, t_2^{u_i} : \theta_2^{u_i}, ..., t_n^{u_i} : \theta_n^{u_i})$$

Where, n depicts the cumulative count of tags used by a user i for annotating various web pages; and $\theta_j^{u_i}$ is the degree of interest in tag $t_j^{u_i}$. The technique of Normalized Tag Frequency (NTF) has been used for calculation of $\theta_j^{u_i}$ based on the assumption made in the work by [34]. According to the assumption, degree of interest that a user holds for a particular tag must be inconsideration to the clause that if a user utilize a tag more frequently in comparison to other tags in order to annotate various web resources than

user is more interested in that tag. Thus, merely calculating the tag frequency for degree of interest will result in biasness towards an active user who annotates web resources very frequently. The mathematical formulation for calculation of $\theta_j^{u_i}$ is as follows:

$$\theta_j^{u_i} = \frac{c_j^{u_i}}{c^{u_i}} \tag{4.2}$$

Where, $c_j^{u_i}$ depicts the count of dataset records for which user u_i has used tag t_j to annotate the web pages; and c^{u_i} is the count of web pages annotated by u_i. Larger the value of degree of interest, *i.e.*, $\theta_j^{u_i}$, more would be the probability of t_j being user preferable tag of user u_i.

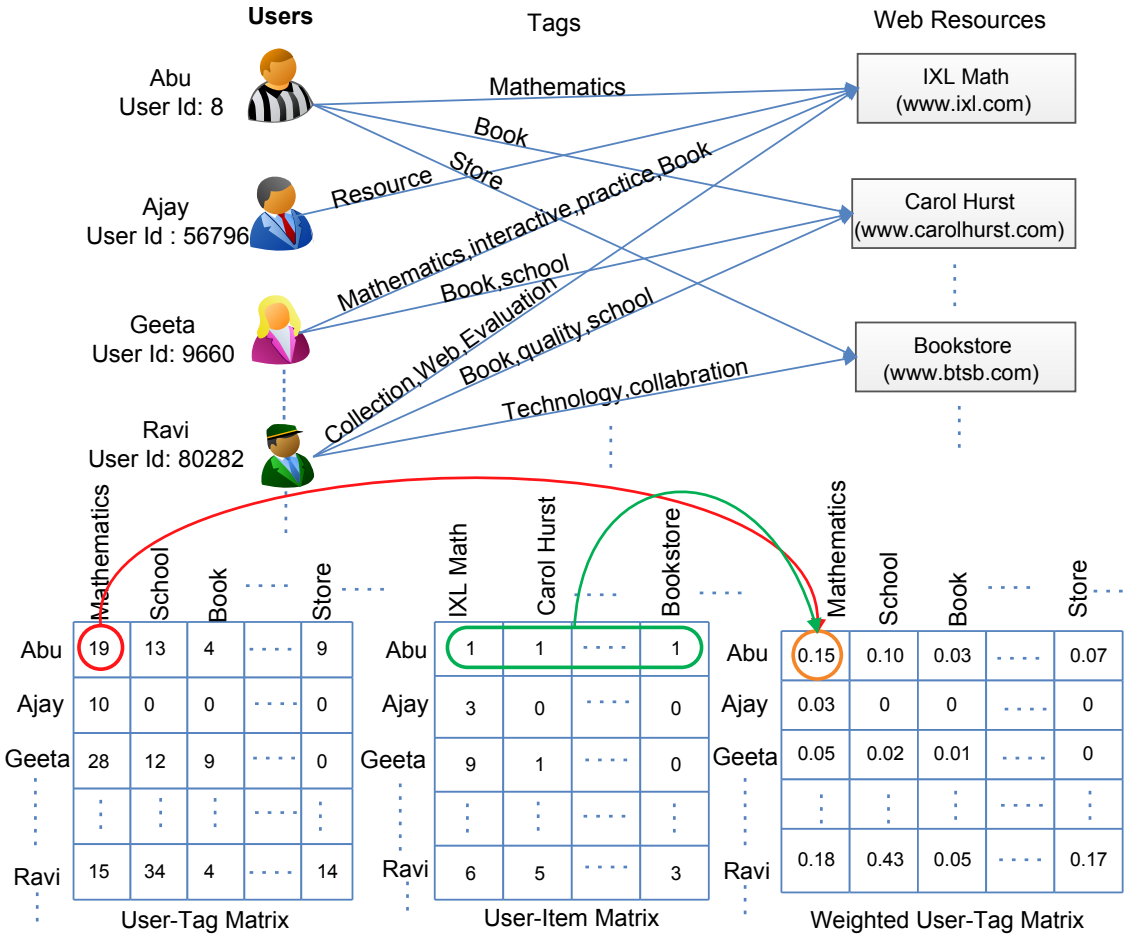


Figure 4.2: Illustration of UIP constructed using base protocol

For instance, an example of B_p based profile construction process is explained in Fig. 4.2, where a part of the ternary relations of Users: Abu, Ajay, Geeta, and Ravi is considered. The User-Tag and User-Item matrices are constructed from these ternary relations. How-

ever, the matrices shown here are the actual matrices constructed for the entire ternary relation schema of the users under consideration on the actual dataset. Each cell ij^{th} of User-Tag matrix will represent the $c_j^{u_i}$ in Eq. (4.2), whereas in the case of User-Item matrix each cell represents the number of times u_i has annotated the web resource w_j. The summation of all the values in a row corresponding to the user's User-Item matrix represents the c^{u_i} in Eq. (4.2). Each cell ij^{th} of Weighted User-Tag matrix in Fig. 4.2 represents the degree of interest $\theta_j^{u_i}$ a user u_i has in tag t_j calculated using Eq. (4.2). In the case of user *Abu*, the degree of interest he has in tag $Mathematics$ is 0.15, where the value of $c_j^{u_i}$ is 19 and c^{u_i} is 122 (summation of values in row corresponding to Abu in User-Item matrix). For visual representation, different colored codes and shapes have been used, *i.e.*, brown colored circle for degree of interest; red colored circle and arc for $c_j^{u_i}$; and green colored oval and arc for c^{u_i}. Similarly, remaining degree of interest values for Abu, Ajay, Geeta and Ravi can be calculated, where c^{u_i} values for Ajay, Geeta, and Ravi is 312, 497, and 79 respectively.

4.1.2 Indirect Interest Identification

An appropriate description of a user's interest is only provided by his own tags, but the construction of UIP using solely these tags is inefficient and incomplete. It all depends on the frequency of social network activities of a user because the amount of user's information available on web is directly proportional to the measure of user's activities. Therefore, to avoid injustice on the account of an inactive user, some additional information must be linked with the user's account for construction of an UIP. This additional information is not something new, it is already available around the user, but in a latent form, *i.e.*, not explicitly stated. G_p and C_p protocols are used to infer the user's additional information from his real-world society and implicit tag relationships respectively using various pattern and learning algorithms. The information predicted by G_p and C_p contributes to the indirect interest of a user.

4.1.2.1 Society Relationship Network

Prediction of a qualitative additional information for UIP enrichment without considering society relationships is impossible. It is an universally accepted fact that behaviour of a person is strongly influenced by the society he keeps; and any change in a person's

behaviour has a direct impact on his likings/dislikings. Society of a person constitutes friends, neighbours, colleagues, relatives or any person who is, directly or indirectly, related to a person under consideration. Thus, it is essential to analyze this relationship network in order to unearth the information for enrichment a user's profile.

In the physical world, every person has a different type of social relationships with various persons in his society; some of them are taken into full confidence, while others remain mere acquaintances. The people in the first category are considered more trustworthy in comparison to others. Therefore, generally, a person likes to share everything with people in his inner private circle and also takes their advice. Considering this fact, society relationship types and Trust Matrix (TM) have been fused together into G_p for predicting additional information. A simplified view of the society relationships and corresponding trust matrix for the whole dataset is presented in Fig. 4.3 as the entire view of the network is quite complicated. Formally, TM is defined as:

Definition 5 *A **Trust Matrix** for user set U describes the extent of trust one user has over another. An adjacency matrix is used to denote the Trust Matrix (TM):*

$$TM_{i,j} = T_{r_{i,j}}$$

Where, $TM_{i,j}$ depicts the extent of trust u_i has over u_j, which is measured by trust score $T_{r_{i,j}}$. The magnitude of $T_{r_{i,j}}$ has been calculated using Eq. (4.3):

$$T_{r_{i,j}} = \begin{cases} 1, & \exists u_j \in IPC_{u_i} \\ \dfrac{1}{|IPC_{u_i}|} * |u_x|, & \exists u_x \in IPC_{u_i} \wedge \exists u_y \in IPC_{u_j}(u_x = u_y) \\ 0, & \text{otherwise.} \end{cases} \tag{4.3}$$

Where, IPC_{u_j} and IPC_{u_i} represent the set of users in the inner private circle of users u_j and u_i respectively. In G_p protocol, trust score of only those users has been calculated who are members of the first two-domains of society relationship network for a user under consideration. IPC_{u_i} constitutes direct social relatives of u_i and have the same extent of trust, *i.e.*, equal to 1. However, in the G_p protocol only friends of a user is considered but other social relatives can also be accommodated based on their trust levels and information availability.

Definition 6 *Let* $\{t_1^{gp_i}, t_2^{gp_i}, t_3^{gp_i}.....,t_n^{gp_i}\}$ *and* $\{\omega_1^i, \omega_2^i, \omega_3^i.....,\omega_n^i\}$ *are sets of tags recommended to user i by* G_p *protocol; and the corresponding degree of interest that the user i may hold for those tags respectively. For target user i, UIP vector obtained by* ***Guild Protocol*** *(*G_p*) is represented by* $\overrightarrow{U}_i^{G_p}$ *as:*

$$\overrightarrow{U}_i^{G_p} = (t_1^{gp_i} : \omega_1^i, t_2^{gp_i} : \omega_2^i, ..., t_n^{gp_i} : \omega_n^i)$$

Where, n depicts the cumulative count of tags recommended to user i by G_p protocol; and ω_j^i is the degree of interest that user i may hold for tag $t_j^{gp_i}$. In order to make tag recommendations to u_i and assigning a degree of interest to the tags Algorithm 3 has been used by G_p. For more details, refer Algorithm 3.

First Hypothesis: Given the results of various parameters, does the partial UIP constructed based on B_p and G_p protocol information is more efficient than UIP's corresponding to the state of the art methodologies?

$$H_0 : E_{(B_p \cup G_p)} \geq E_{state\ of\ the\ art\ methodologies}$$

Where,

H_0 *is the null hypothesis*

$E_{(B_p \cup G_p)}$ = efficiency of partial UIP constructed using information generated by B_p and G_p protocol

$E_{state\ of\ the\ art\ methodologies}$ = efficiency of UIP constructed using state of the art methodologies

For instance, an example of G_p based UIP construction is explained in Fig. 4.3, where a part of the social relationship network of user *Ajay* has been considered. The Trust Matrix is constructed from Eq. (4.3), and Weighted User-Tag Matrix is obtained from B_p protocol as shown in Fig. 4.2. Here, Algorithm 3 has been used to make tag recommendations with a suitable degree of interest to *Ajay*. Firstly, users having trust score equal to or more than 0.8 have selected from social relatives of *Ajay*; and then top 10 tags from the sorted version of B_p based UIPs corresponding to selected users have been chosen for the recommendation. There might be a possibility that same tag is recommended to *Ajay* by his multiple social relatives. Therefore, before making any recommendation, records

Algorithm 3: Guild protocol (G_p) recommended UIP

Input : Trust Matrix TM, user set U and B_p protocol based UIP $\overrightarrow{U}_i^{B_p}$
Output: Tags recommended by G_p and their corresponding degree of interest
Initialize :
$index1, 2 \leftarrow 1$ ▷ *index*1 *and index*2 *are index pointers*
$Caller\_B_p \leftarrow \overrightarrow{U}_i^{B_p}$
for $m \leftarrow 1$ *to* $ncols(Caller\_B_p)$ **do** ▷ *ncols() count number of columns in matrix*
  $Caller\_tags \leftarrow Caller\_tags.append(Caller\_B_p(1, m))$
for *each user* j *in* U **do**
  if $TM(i, j) \neq 0$ **then** ▷ *trust score must not equal to zero*
    $Relative\_Tscore(1, index1) \leftarrow j$
    $Relative\_Tscore(2, index1) \leftarrow TM(i, j)$
    $index1 \leftarrow index1 + 1$
$Relative\_Tscore \leftarrow sort(Relative\_Tscore)$ ▷ *sort w.r.t trust score in decreasing order*
for $k \leftarrow 1$ *to* $ncols(Relative\_Tscore)$ **do**
  if $Relative\_Tscore(2, k) \geq 0.8$ **then**
    $user\_relative \leftarrow Relative\_Tscore(1, k)$
    $B_p\_user \leftarrow$ fetch B_p based UIP of user $user\_relative$
    for $l \leftarrow 1$ *to* $ncols(B_p\_user)$ **do**
      if $(B_p\_user(1, l)$ ***NOT IN*** $Caller\_tags\ )$ ***AND*** $(l \leq 10)$ **then**
        $Caller\_G_p(1, index2) \leftarrow B_p\_user(1, l)$
        $degree\_of\_interest \leftarrow Relative\_Tscore(2, k) * B_p\_user(2, l)$
        $Caller\_G_p(2, index2) \leftarrow degree\_of\_interest$
        $index2 \leftarrow index2 + 1$
$Caller\_G_p \leftarrow sort(Caller\_G_p)$ ▷ *sort w.r.t degre_of_interest in a decreasing order*
$temp(1, 1) \leftarrow Caller\_G_p(1, 1)$
$temp(2, 1) \leftarrow Caller\_G_p(2, 1)$
$index1 \leftarrow 2$
for $r \leftarrow 2$ *to* $ncols(Caller\_G_p)$ **do**
  $flag \leftarrow 0$
  for $z \leftarrow ncols(temp)$ *to* 1 **do**
    if $Caller\_G_p(1, r) == temp(1, z)$ **then**
      $flag \leftarrow 1$
      break
  if $flag == 0$ **then**
    $temp(1, index1) \leftarrow Caller\_G_p(1, r)$
    $temp(2, index1) \leftarrow Caller\_G_p(2, r)$
    $index1 \leftarrow index1 + 1$
$Caller\_G_p \leftarrow temp$

corresponding to duplicate tags are removed keeping only those records which have a high degree of interest value. After removing the duplicate records, User-Tag matrix is updated for *Ajay* using recommended additional information by G_p and information by B_p as shown in Fig. 4.3. Similarly, for other users also, G_p based tag recommendation can be made, but here only *Ajay* is considered just for making visualization simple and

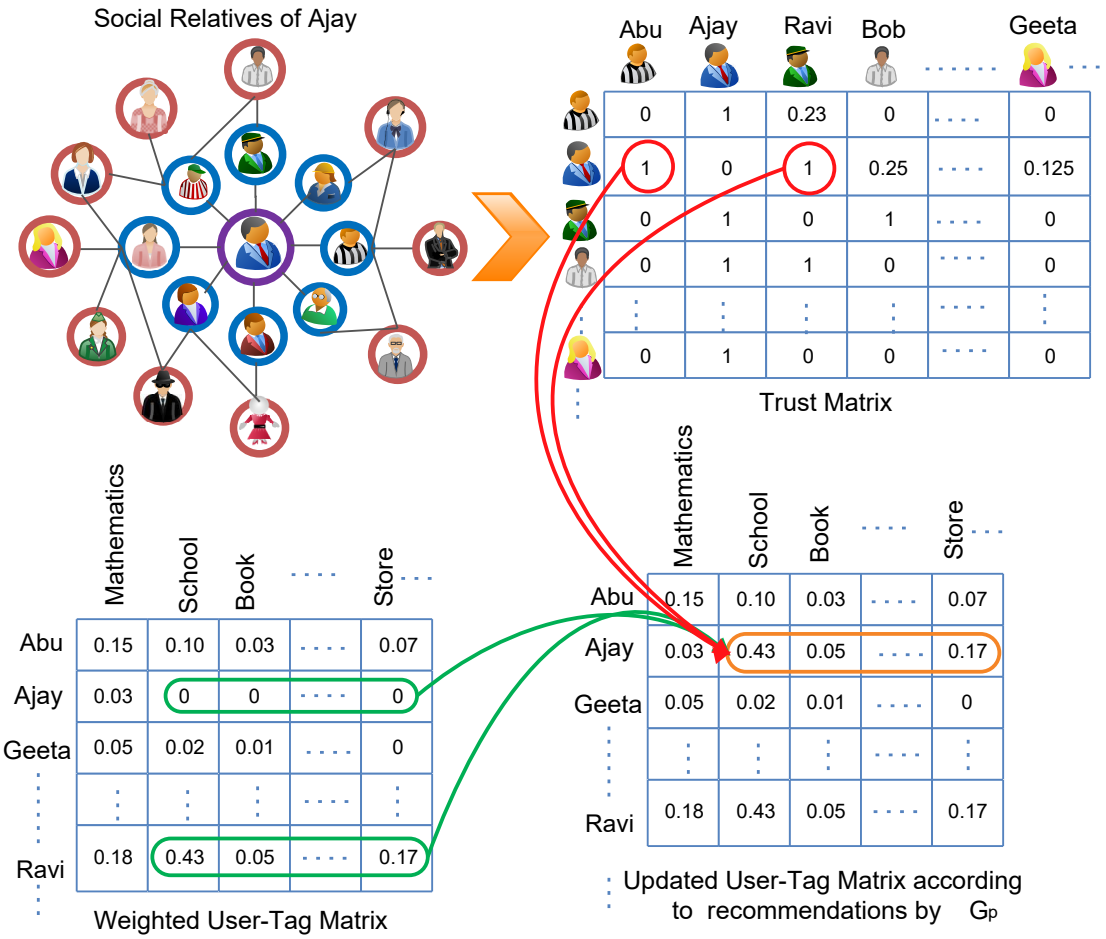


Figure 4.3: Illustration of partial UIP constructed using base and guild protocol

maintaining the consistency between all examples. Here, to describe the example, different colored codes and shapes have been used, *i.e.*, brown colored circle shows the degree of interest in recommended tags. Red colored circle and arc represents the trust score; and green colored oval and arc stand for B_p based UIP of selected social relatives. Each i^{th} row of updated User-Tag matrix corresponds to partial UIP based on B_p and G_p for a user u_i. For details regarding the process of calculating cell values in updated User-Tag matrix, refer Algorithm 3.

4.1.2.2 Tag Relationship Network

Mining of additional information only on the basis of tags recommended by social relatives of a user, for the enrichment of his UIP, is also not sufficient. The construction of UIP using this additional information and user generated tag information can represent user's interest to some extent, but still UIP remains incomplete. Therefore, some amount of supplementary information about the user is still required to further enlarge his preference

boundaries. The analysis of real-world tag relationships can unearth various hidden facts about a pair of tags and can prove to be a valuable input for UIP enrichment. Therefore, these relationships have been incorporated in the proposed methodology and identified by C_p protocol. Firstly, a Tag-Tag Relationship matrix (TTR_m) is constructed in which each cell measures the semantic relatedness level between the corresponding tags. Secondly, after computing TTR_m, clustering is performed on its row vectors.

Semantic relatedness is capable of entertaining any type of relationship between the tag pairs, whereas other primitive methods like syntactic, co-occurrence and semantic similarity can take care of only one type of relationship like "*is a*" relationship. For example, both car and bus are semantically similar, but car and driving are semantically related. Miniature depiction of semantic relatedness is presented in Fig. 4.4, where length of the line segment represents the measure of semantic relatedness. Lesser is the length of line segment between two tags, higher is the semantic relatedness of two tags. Like car and driving is more semantically related then car and bus. On the basis of semantic relatedness, probability that a person, who likes the car may also like bus is much less than the probability of a person interested in both car and driving. Therefore, C_p can also entertain the tags which are neither semantically nor syntactically similar, but are semantically related to each other in the real-world.

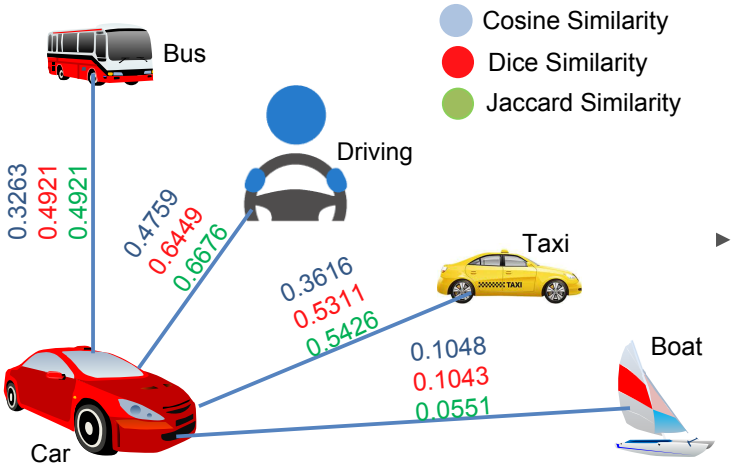


Figure 4.4: Semantic Relatedness

The preliminary requirement for computation of TTR_m is the tag-vectors which are obtained with the help of Word2vec model developed by team of researchers from google under the guidance of Mikolov et al. [160]. Word2vec is a composition of several interrelated neural-network models, which cooperatively produce the multi-dimensional word-

vectors for every distinct word in a large corpus of text. Each unique word in the corpus is represented as a vector on the vector space representation of entire corpus. Vectors corresponding to the words that belong to a related context in the corpus are positioned in the close proximity in vector space. Each image in Fig. 4.4 represents the actual word vector obtained from Word2vec model. Here, car shares a more proximal space with driving and taxi as compare to bus and boat, so comparatively car is more related to driving. Word2vec comes with two different architectures, Continuous Skip-Gram Model (CSGM) and Continuous Bag-of-Words (CBOW). Current word is used in CSGM to predict the window of contextual words in the surrounding. However, in CBOW the window of contextual words in the surrounding are used to predict the current word. According to Google, CSGM is slower than CBOW, but more efficient to obtain accurate vector representation of words. In order to further optimize CSGM, hierarchical softmax and negative sampling are also used. Out of several predictive modeling methods, Word2vec is computationally more efficient than its peers. For this reason, CSGM architecture of Word2vec has been used in C_p.

The word-vectors corresponding to each and every tag in a collaborative tagging dataset under consideration was selected from the collection of word-vectors created by Word2vec model. Though before making any selection, stemming of all the tags was performed in order to avoid different vector representations of two similar tags. For example, both books and book are two similar tags in the real-world, but they are considered as two different tags, if stemming is not performed. In C_p, semantic relatedness between two tags is computed using a similarity measures [31] discussed in Table 4.1 between their word vectors .

Where, $\overrightarrow{t_i}$ and $\overrightarrow{t_j}$ are corresponding word vectors for tags t_i and t_j respectively, n is the length of a vector and w_{ki},w_{kj} are vector elements. In Fig. 4.4, the value of semantic relatedness between two tags is clearly depicted on line segment joining the tags with different color codes for different similarity measures discussed in Table 4.1. More the value of similarity measure between the word vectors of t_i and t_j, higher is the probability of t_i being semantically related to t_j in the real-world. The TTR_m will provide a valuable assistance in obtaining the fine clusters with precise boundaries and minimum inter-cluster similarity.

Table 4.1: Similarity measures summarization (i.e.,$TTR_m(t_i, t_j)$)

Similarity measure	Mathematical Formulation
Cosine	$cos(\overrightarrow{t_i}, \overrightarrow{t_j}) = \frac{\sum_{k=1}^{n} w_{ki} * w_{kj}}{\sqrt{\sum_{k=1}^{n} w_{ki}^2 * \sum_{k=1}^{n} w_{kj}^2}}$
Dice	$dice(\overrightarrow{t_i}, \overrightarrow{t_j}) = \frac{2 * \sum_{k=1}^{n} w_{ki} * w_{kj}}{\sum_{k=1}^{n} w_{ki}^2 + \sum_{k=1}^{n} w_{kj}^2}$
Jaccard	$jaccard(\overrightarrow{t_i}, \overrightarrow{t_j}) = \frac{\sum_{k=1}^{n} w_{ki} * w_{kj}}{\sum_{k=1}^{n} w_{ki}^2 + \sum_{k=1}^{n} w_{kj}^2 - \sum_{k=1}^{n} w_{ki} w_{kj}}$

After computing the Tag-Tag Relationship matrix, clustering of the tags in a collaborative tagging dataset is performed using the row vectors corresponding to those tags in TTR_m. A cluster comprises of various tags that are associated with each other in some context, but their strength of association varies from one tag pair to another. Moreover, it can also not be denied that a cluster is a composition of multiple related contexts. Out of several eminent clustering algorithms, Hierarchical Agglomerative Clustering (HAC) is incorporated into C_p. The key factors behind its incorporation are its ability to accommodate a large number of unevenly sized clusters and a voluminous amount of data without any degradation of scalability and efficiency. The additional information identified by C_p using TTR_m and tag clusters is quiet advantageous in the enrichment of a user's UIP.

Definition 7 *Let* $\{t_1^{cp_i}, t_2^{cp_i}, t_3^{cp_i}....., t_n^{cp_i}\}$ *and* $\{\rho_1^i, \rho_2^i, \rho_3^i....., \rho_n^i\}$ *are sets of tags recommended to user* i *by* C_p *protocol and the corresponding degree of interest that the user* i *may hold for those tags respectively. For target user* i*, UIP vector obtained by* ***Congregation protocol*** *(*C_p*) is represented by* $\overrightarrow{U}_i^{C_p}$ *as:*

$$\overrightarrow{U}_i^{C_p} = (t_1^{cp_i} : \rho_1^i, t_2^{cp_i} : \rho_2^i, ..., t_n^{cp_i} : \rho_n^i)$$

Where, n depicts the cumulative count of tags recommended to user i by C_p protocol; and ρ_j^i is the degree of interest that user i may hold for tag $t_j^{cp_i}$. In order to make tag recommendations to u_i and assigning a degree of interest to the tags, Algorithm 4 has been used by C_p. For more details, refer Algorithm 4.

Algorithm 4: Congregation protocol recommended UIP

Input : Set of clusters $Sclus$, Tag-Tag Relationship matrix TTR_m and B_p protocol based UIP $\overrightarrow{U}_i^{B_p}$.
Output : Tags recommended by C_p and their corresponding degree of interest
Initialize :

```
index1, 2 ← 1                                   ▷ index1 and index2 are index pointers
for each clus in Sclus do                        ▷ clus is the cluster in Sclus
    index1 ← 1
    Caller_B_p ← U_i^{B_p}
    for k ← 1 to ncols(Caller_B_p) do    ▷ ncols() count number of column in matrix
        user_tag ← Caller_B_p(1, k)
        for each tag j in clus do
            if user_tag == j then
                tag_list ← tag_list.append(j)        ▷ list of user's tags present in clus
    for each tag j in clus do
        total_rel ← 0
        for each tag l in tag_list do
            if j == l then
                total_wt ← total_wt + Caller_B_p(2, l)
                 ▷ Total weight of user's tags present in clus
                break
            else
                total_rel ← total_rel + TTR_m(l, j)
                 ▷ Total semantic relatedness b/w cluster tag & user's tags
        avg_rel = total_rel/length(tag_list)
        temp_profile(1, index1) ← j
        temp_profile(2, index1) ← avg_rel
        index1 ← index1 + 1
    temp_profile ← sort(temp_profile)     ▷ sort w.r.t. avg_sim in a decreasing order
    avg_tagwt ← total_wt/length(tag_list)
    tag_rec ← length(tag_list)/2                 ▷ number of tags recommended by clus
    for m ← 1 to tag_rec do
        Caller_C_p(1, index2) ← temp_profile(1, m)  ▷ tag recommended by clus to user
        degree_of_interest ← temp_profile(2, m) * avg_tagwt
        Caller_C_p(2, index2) ← degree_of_interest
        index2 ← index2 + 1
    Empty(tag_list)                              ▷ Delete all elements tags_of_tag_list
    total_wt, total_rel ← 0
Caller_C_p = sort(Caller_C_p)     ▷ sort w.r.t. degree_of_interest in a decreasing order
```

Second Hypothesis: Given the results of various parameters, is the partial UIP constructed based on B_p and C_p protocol information is more efficient than UIP's corresponding to the state of the art methodologies?

$$H_0 : E_{(B_p \cup C_p)} \geq E_{state\ of\ the\ art\ methodologies}$$

Where,

H_0 *is the null hypothesis*

$E_{(B_p \cup C_p)}$ = efficiency of partial UIP constructed using information generated by B_p and C_p protocol

$E_{state\ of\ the\ art\ methodologies}$ = efficiency of UIP constructed using state of the art methodologies

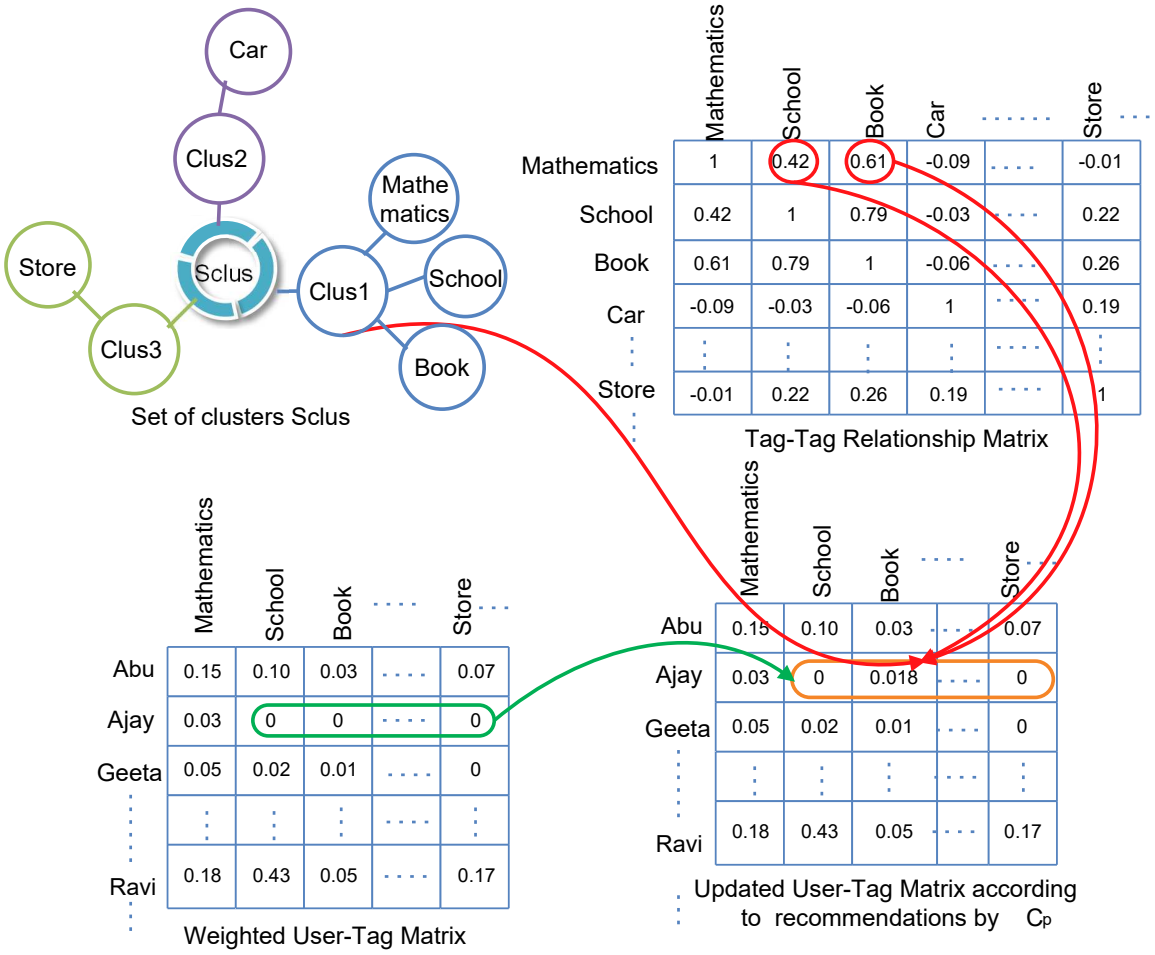


Figure 4.5: Illustration of partial UIP constructed using base and congregation protocol

An example of C_p based UIP construction is explained in Fig. 4.5, where a part of the actual tags dataset is considered. The Tag-Tag Relationship Matrix TTR_m has been constructed from Table 4.1 using vector representation of tags generated by Word2vec model, whereas cluster set *Sclus* has been obtained from TTR_m and HAC. Weighted User-Tag Matrix has been obtained from B_p protocol as shown in Fig. 4.2. Here, Algorithm 4 has

been used to make tag recommendations with suitable degree of interest to *Ajay*. Firstly, the clusters *clus* having the tags generated by *Ajay* have been identified as *clus*1 and the tags of *clus*1 that are not in *Ajay* profile are listed in a list, *i.e.*, $School, Book$. Then, the semantic relatedness of tags listed in the list with each tag of *Ajay* has been measured using TTR_m. Out of all the potential candidates, *i.e.*, $School, Book$ only *Book* is recommended to *Ajay*. User-Tag matrix is updated for *Ajay* using recommended additional information by C_p and information by B_p as shown in Fig. 4.5. Similarly, for other users also, C_p based tag recommendation can be made. Here, to describe the example, different colored codes and shapes have been used, *i.e.*, brown colored circle shows the degree of interest in recommended tags. Red colored circle and arc have been used showing the contribution of TTR_m and $Sclus$, whereas green colored oval and arc stand for B_p based UIP of the user to whom recommendations are made. Each row i of updated User-Tag matrix corresponds to partial UIP based on B_p and C_p for a user u_i. For the process of calculating cell values in updated User-Tag matrix and selection of cluster, refer Algorithm 4.

Lastly, the additional information about the user, generated in FRUIP and CRUIP levels by utilizing G_p and C_p protocols respectively, is clubbed with the one provided by B_p protocol at FBUIP level. The clubbing of information is done with the help of F_{trade} function in order to enrich the user's UIP and construction of a full-fledged final UIP of a user. The information lies in the form of tag recommendations and the corresponding degree of interest that may hold for those tags. Full-fledged final UIP of a user u_i has been constructed by F_{trade} using Eq. (4.4).

$$U_{i,j} = \begin{cases} \theta_j^{u_i}, & \exists\ t_j \in \overrightarrow{U}_i^{B_p} \\ \omega_j^{u_i}, & (\exists\ t_j \in \overrightarrow{U}_i^{G_p} \wedge \exists\ t_k \in \overrightarrow{U}_i^{C_p} (t_j = t_k))\ \vee (\exists\ t_j \in \overrightarrow{U}_i^{G_p} \wedge \exists\ t_j \notin \overrightarrow{U}_i^{C_p}) \\ \rho_j^{u_i}, & \exists\ t_j \in \overrightarrow{U}_i^{C_p} \wedge \exists\ t_j \notin \overrightarrow{U}_i^{G_p} \\ 0, & \text{otherwise.} \end{cases} \tag{4.4}$$

Where, $U_{i,j}$ depicts the cell value of User-Tag matrix corresponding to full-fledged final UIP for user u_i and tag t_j. $\overrightarrow{U}_i^{B_p}$, $\overrightarrow{U}_i^{G_p}$ and $\overrightarrow{U}_i^{C_p}$ are UIP vectors corresponding to information provided by B_p, G_p and C_p protocol respectively. Here, $\theta_j^{u_i}$, $\omega_j^{u_i}$ and $\rho_j^{u_i}$ are the degrees of interest that user u_i has in the tag t_j as predicted by B_p, G_p and C_p proto-

cols respectively. For u_i, all the tags and their respective degrees of interest in User-Tag matrix, for which degree of interest value is non-zero, constitutes full-fledged final UIP of user u_i, *i.e.*, $\overrightarrow{U}_i$.

Third Hypothesis: Given the results of various parameters, is the constructed full-fledged UIP based on B_p, G_p and C_p protocol information more efficient than UIPs corresponding to state of the art methodologies and partial UIPs ?

$$H_0 : E_{(B_p \cup G_p \cup C_p)} > E_{(B_p \cup G_p)} \textit{ and } E_{(B_p \cup G_p \cup C_p)} > E_{(B_p \cup C_p)} \textit{ and } E_{(B_p \cup C_p)} \geq E_{\textit{state of the art methodologies}}$$

Where,

H_0 *is the null hypothesis*

$E_{(B_p \cup G_p \cup C_p)}$ = efficiency of full-fledged final UIP constructed using information generated by B_p, G_p and C_p protocol.

$E_{(B_p \cup G_p)}$ = efficiency of partial UIP constructed using information generated by B_p and G_p protocol.

$E_{(B_p \cup C_p)}$ = efficiency of partial UIP constructed using information generated by B_p and C_p protocol.

$E_{\textit{state of the art methodologies}}$ = efficiency of UIP constructed using state of the art methodologies.

4.2 State of the art methodologies

To prove the effectiveness of methodology proposed for UIP construction, it has been compared with four state of the art methodologies of UIP construction in the folksonomy *a.k.a.* collaborative tagging system. Some of the methodologies construct UIP only on the basis of user's own tags; and no enrichment strategy is used, while in some methodologies both the user's own tags and enrichment are employed. Special attention has been given to the selection of each state of the art methodology to be used for having a comparison with the proposed one. These state of the art methodologies are summarized and described in Table 4.2.

Table 4.2: Summary of State of the art Methodologies for UIP construction

Methodologies	Description
Methodology-1 (UTF-IUF)	It was presented by Bouadjenek et al. [31], in which the level of interest in a tag used by a user is measured by UTF-IUF values.
Methodology-2 (NTF)	It was presented by Cai et al. [34], in which a user's level of interest in the tag is calculated using NTF values.
Methodology-3 (TF-IDF & Clustering)	It was given by Kumar et al. [30], in which degree of interest for a user in a tag is calculated using TF-IDF values. Not only the tags used by the user for tagging different web resources are used in UIP construction, but a UIP enrichment strategy is also used. For this, clustering of tags is performed to group semantically similar tags in one cluster. As a result, they designed two UIP's: svdCUIP and modSvdCUIP out of which modSvdCUIP is more efficient than svdCUIP as per their own experiments. Here, only modSvdCUIP has been compared with the proposed methodology.
Methodology-4 (UTF & Latent Friendship Network)	It was given by Shafiq et al. [7], where the degree of interest in a tag for a user depends on the UTF, *i.e.*, frequency that tag has been used by the user. Only the UTF factor was not responsible for UIP construction, but also the latent friendship network of the user.

4.3 Results and Comparisons

The performance of proposed methodology is compared with state of the art methodologies described in Table 4.2 using the del.icio.us dataset as mentioned in Table 3.1. The comparison is not only performed with full-fledged UIP of a user but also with intermediate UIP's of a user, as the proposed methodology doesn't construct full-fledged UIP all at once but in multiple levels. Each level employs a different protocol working on particular strategy for UIP construction and therefore, final obtained UIP is a joint venture of multiple strategies. The analysis of contribution made by different strategies to the performance of proposed methodology is performed by comparing intermediate UIP's with state of the art methodologies, which will help to select the high performing contributors to full-fledged UIP and eradicate the least ones.

Analyzing the impact of the different number of tag clusters in an input data set is essential to measure the contribution of Cluster Recommended UIP (CRUIP) to final obtained UIP. Therefore, Number of Clusters, *i.e.*, N_{clus} parameter is varied from two to forty at an interval of two in the experiments corresponding to both the proposed and TF-

IDF plus clustering-based methodology. The effect of three different similarity measures used to calculate tag-tag semantic relatedness as discussed in Table 4.1, is also visualized together with the impact of N_{clus} parameter. Evaluation metrics defined in Section 3.3 are used to quantify the comparison. The results of comparison have been visualized with the help of two types of plots, *i.e.*, bar-plots and line plots, with separate color codes for UIP constructed by state of the art methodologies and proposed methodology.

Fig. 4.6 illustrates the comparison of proposed and state of the art methodologies on the basis of evaluation metric MRR. It is the most prominent metric used in the field of information retrieval. It emphasizes the significance of placing the most favorable tags of the user close to the head of the ranking hierarchy of tags inside user's UIP. Tags inside UIP are arranged according to the level of interest user hold for them from most interested to least interested. Firstly, the comparison of three different CRUIP's constructed w.r.t. *Cosine, Dice, and Jaccard* similarity measures use to calculate semantic relatedness between two tags, is performed in Fig. 4.6(a). The bar-plot represents the difference in performance of *Cosine, Dice and Jaccard* similarity measures based on CRUIP in enriching the user's FBUIP, *i.e.*, user's own tag-based profile. On the whole, each bar is a measure of MRR value corresponding to the intermediate UIP which is a contribution of NTF, *i.e.*, FBUIP and CRUIP. Here, for the sake of convenience and simplicity, the MRR values are averaged (Y-axis) over all the users in a dataset. Together with similarity measure variance the N_{clus} parameter (X-axis) is also varied to analyze its impact on MRR values of respective CRUIP's. The Fig. 4.6(a) clearly highlights that the jaccard similarity based intermediate UIP leads the race among its peers nearly for every value of parameter N_{clus}. However, for N_{clus} equal to 34 jaccard similarity based intermediate UIP has achieved the highest average MRR value of 0.198136 among all CRUIP and FBUIP based intermediate UIP's.

Based on the results pattern shown in Fig. 4.6(a), jaccard similarity based intermediate UIP is chosen for comparison with TF-IDF plus clustering based state of the art methodology in Fig. 4.6(b). The impact of N_{clus} parameter (X-axis) on the average MRR values (Y-axis) of methodologies under consideration is also analyzed in Fig. 4.6(b) but similarity parameter is kept constant. The obtained results are visualized with help of bar-plot in Fig. 4.6(b) clearly shows that the intermediate UIP surpass the state of the

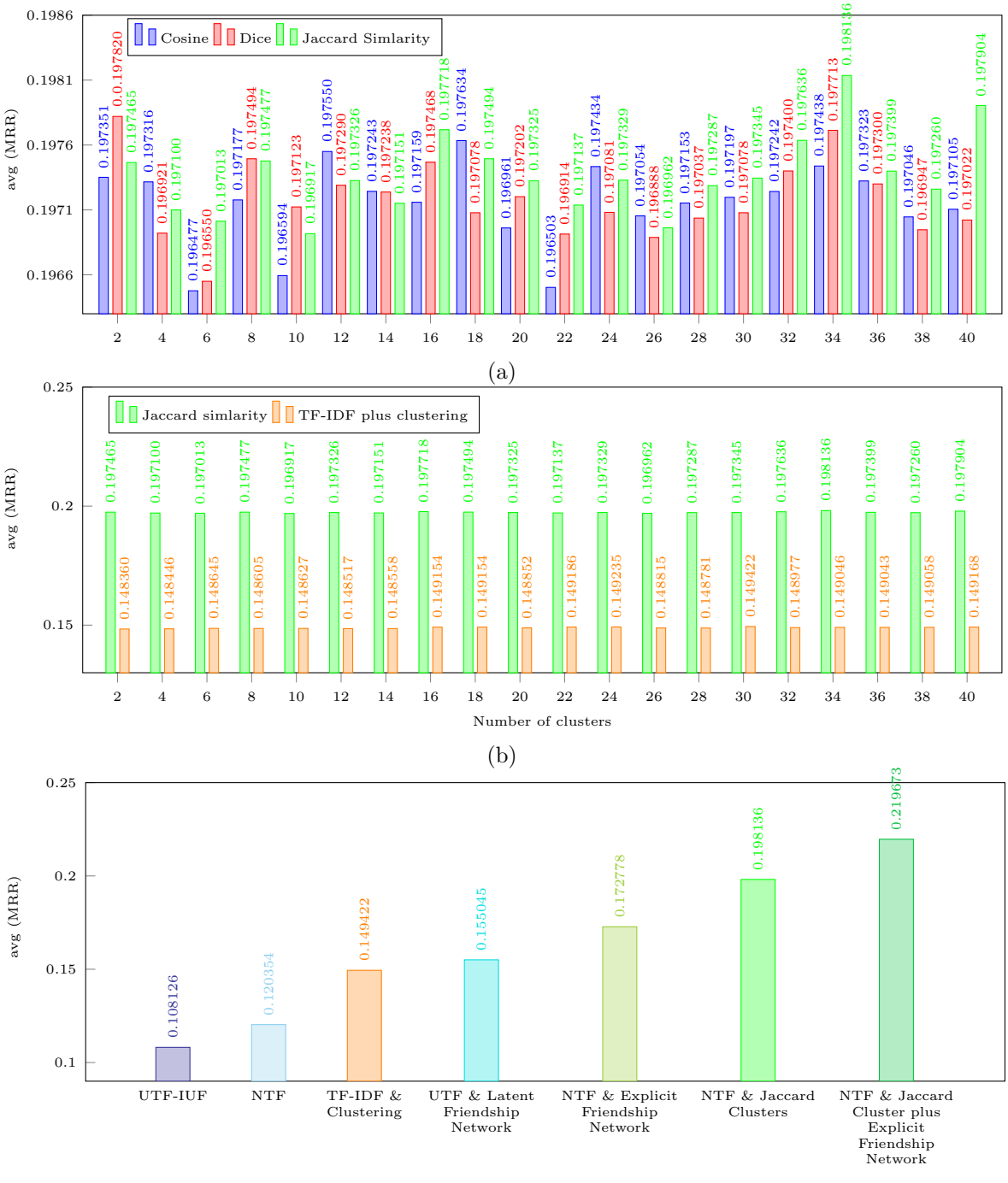


Figure 4.6: Comparative analysis of average MRR value for (a) CRUIP based intermediate UIP's corresponding to different similarity measures (b) Jaccard similarity based intermediate UIP and TF-IDF plus clustering based UIP over different values of N_{clus} (c) State of the art methodologies based UIP and proposed methodology based intermediate UIP's and Final UIP.

art methodology for each value of N_{clus} parameter with a good margin, *i.e.*, 32.83 % at max.

In Fig. 4.6(c), performance of two intermediate UIP's and full-fledged final UIP of the proposed methodology is compared with all state of the art methodologies, mentioned in

Table 4.2, on basis of MRR evaluation metric. One intermediate UIP is corresponding to the joint venture of FBUIP and jaccard similarity based CRUIP, named as NTF & Jaccard Clusters as NTF approach used to compute FBUIP. Similarly, another intermediate UIP is corresponding to the joint venture of FBUIP and FRUIP, named as NTF & Explicit Friendship Network. The full-fledged final UIP obtained as the final product of proposed methodology is named as NTF & Jaccard Clusters plus Explicit Friendship Network. Both intermediate UIP's and full-fledged final UIP of proposed methodology outperforms each and every state of the art methodology by a considerable amount. NTF & Explicit Friendship Network based UIP outperforms the dominant state of the art methodology by 11.43%, and the least performer, *i.e.*, UTF-IUF by 59.79% whereas, NTF & Jaccard Clusters by 27.79% and 83.24% respectively. In the whole, full-fledged final UIP outperform outperforms the dominant state of the art methodology by 41.68%, and the least performer, *i.e.*, UTF-IUF by 103.16% which is a very high margin. One more thing can be noticed from Fig. 4.6(c) is that based on average MRR value, contribution of CRUIP based strategy is more than FRUIP in affecting the enrichment of UIP with the most favorable tags of a user and arranging them higher in ranking hierarchy. Greater the value of MRR, higher would be the accuracy of UIP.

The comparison of proposed and state of the art methodologies made on the basis of *imp* evaluation metric is shown in Fig. 4.7. Firstly, the amount of improvement made by intermediate UIP in comparison to TF-IDF plus clustering based state of the art methodology while predicting the rank of target tags in the testing set is visualized with help of line-plot in Fig. 4.7(a). The N_{clus} parameter (X-axis) is also varied to analyze change in *imp* value of intermediate UIP's. Similar to MRR evaluation metric case *imp* are also averaged over the users in the dataset for sake of simplicity. As depicted by Fig. 4.7(a) jaccard similarity measure based intermediate UIP causes maximum improvement in comparison to its counterparts with highest value 0.274085 at N_{clus} is 34.

In Fig. 4.7(b) the performance of two intermediate UIP's and full-fledged final UIP of proposed methodology is compared with all state of the art methodologies on basis of *imp* evaluation metric. There is an improvement of more than 27% in the ranking of tags of the testing set by the final UIP of proposed methodology as compared to state of the art ones. Maximum improvement is recorded for UTF-IUF based UIP, *i.e.*, 39.62 %

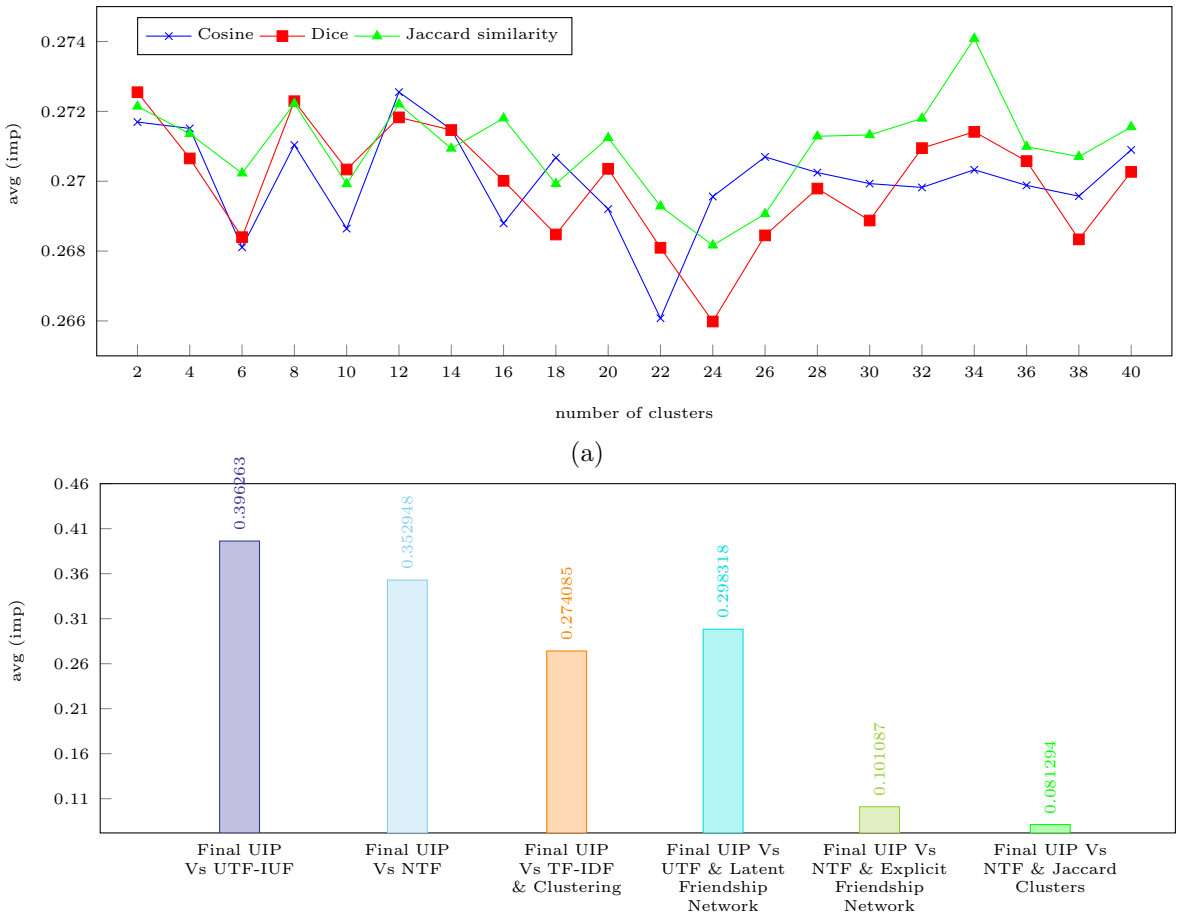
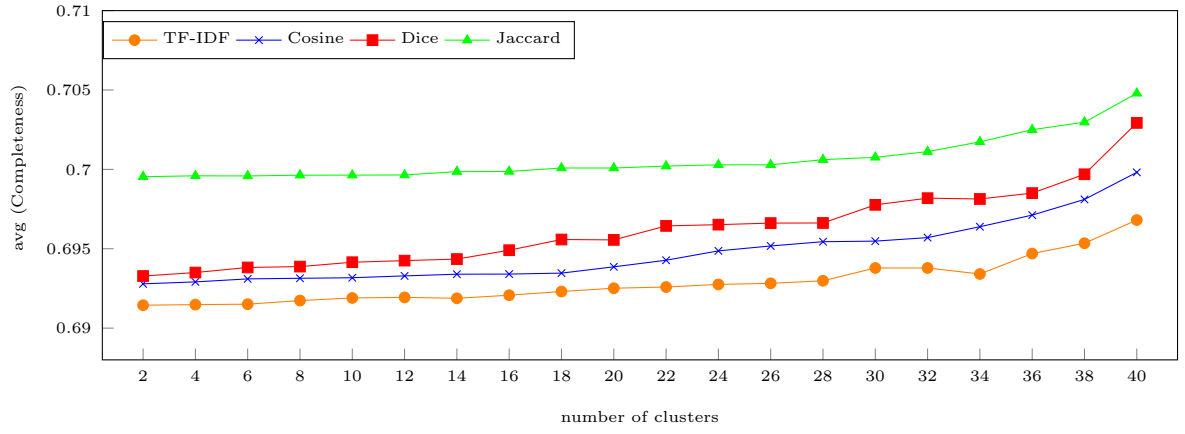


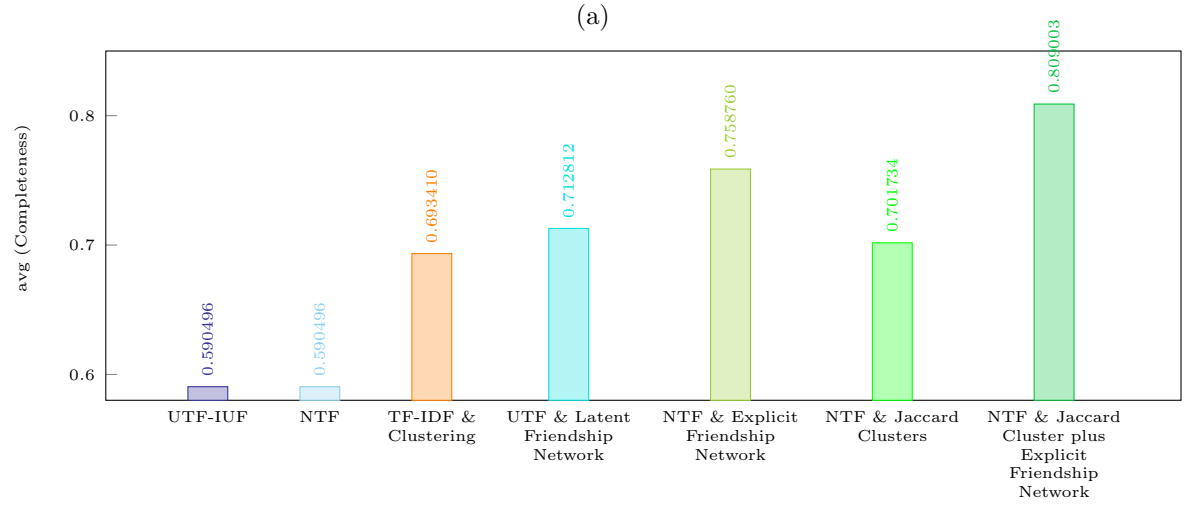
Figure 4.7: Comparative analysis of average improvement in the ranking of target tags by the (a) CRUIP based intermediate UIP's over different similarity measures as compared to TF-IDF plus clustering based UIP (b) Final UIP Vs to State of the art methodologies based UIP and intermediate UIP's of proposed methodology.

and least for TF-IDF plus clustering based UIP, *i.e.*, 27.40 % among the state of the art methodologies. Experiment to measure the improvement made by final UIP in the ranks of target tags in comparison to NTF & Jaccard Clusters and NTF & Explicit Friendship Network intermediate UIP's is also conducted. The result of which shows that final UIP is not dependent on single strategy but it is a joint venture of multiple strategies. The only difference is in the percentage contribution of different strategies towards final UIP. Here, the contribution of CRUIP based strategy is more as the improvement made by final UIP in comparison to NTF & Jaccard Clusters is less than NTF & Explicit Friendship Network. So again the effect of CRUIP is more in the enrichment of UIP which is also confirmed by MRR evaluation metric.

Fig. 4.8, represents the percentage of target tags present in the testing set which are also present in the UIP constructed using respective methodologies. It is a prominent metric



(a)



(b)

Figure 4.8: Comparative analysis of average Completeness value for (a) CRUIP based intermediate UIP's over different similarity measures and TF-IDF plus clustering based UIP corresponding to different values of N_{clus} (b) State of the art methodologies based UIP and proposed methodology based intermediate UIP's and Final UIP.

to evaluate the capacity of UIP enrichment strategy to correctly predict the tags in which user is interested or may be interested no matter the what level on interest user have in the predicted tags. The completeness values are averaged over the users in the dataset to make the visualization process simple as it is a very cumbersome task to represent the completeness value for each and every user in a dataset. Here also, N_{clus} parameter (X-axis) is varied to analyze its effect on average completeness values of UIP's. It can be observed from Fig. 4.8 (a) that all three intermediate UIPs surpassed the state of the art methodology in terms of average completeness (Y-axis) with a good margin for each value of N_{clus}. However, the difference in the amount of average *completeness* values of jaccard similarity measure based intermediate UIP and TF-IDF plus clustering based UIP is very high as shown in Fig. 4.8(a). The highest *completeness* values for both UIP's is recorded

at N_{clus} is 40 but the maximum difference in their *completeness* values is recorded at N_{clus} is 34.

Based on the results shown in Fig. 4.8(a) jaccard similarity measure based CRUIP at N_{clus}, 34 is chosen for contribution to the construction of full-fledged final UIP of the proposed methodology. Similarly to Fig. 4.6(c) and Fig. 4.7(b), comparison of intermediate UIP's and final UIP is performed with all state of the art methodologies as shown in Fig. 4.8(b). The results shows that the NTF & Jaccard Clusters based intermediate has only outperformed the UTF-IUF, NTF, and TF-IDF plus clustering based UIP with a good margin. But NTF & Explicit Friendship Network based intermediate UIP outperformed all the state of the art methodologies by enriching the user's intermediate UIP with 75.87% target tags of the testing set. The full-fledged final UIP has able to predict much more percentage of target tags, *i.e.*, 80.90% which is 13.49 % more than the dominant state of the art methodology and 37% more than the least performing methodology, *i.e.*, UTF-IUF and NTF. However, in contrary to the scenarios followed in Fig. 4.6(c) and Fig. 4.7(b), here FRUIP based strategy is proved to be the prominent contributor in affecting the enrichment of UIP with the most favorable tags of a user. More the value of *completeness*, higher are the chances of user preferences to be present in obtained UIP.

The performance comparison of proposed and state of the art methodologies on the basis of evaluation metric $P@K$ is illustrated in Fig. 4.9 with help of line-plots and bar plots. Similarly to Fig. 4.6(a),Fig. 4.7(a) and Fig. 4.8(a), first the performance comparison of three indeterminate UIP's, designed on the basis of *Cosine, Dice and Jaccard* similarity measures based CRUIP, is performed. The comparison is performed to analyze the effect of similarity measures on the precision value of UIP of a user for the different values of K (X-axis). In contrary to experiments for MRR, *imp* and *completeness* evaluation metrics, here only the similarity measure parameter is varied but N_{clus} parameter is kept constant, *i.e.*, 34. The basis for choosing N_{clus}, 34 is that at this number of clusters maximum value of MRR, *imp* and *completeness* evaluation metrics are achieved by intermediate and final UIP. Another reason for choosing 34 is to make analyzing and visualization process simple as varying both N_{clus} and K parameters at the same time results in a very complex process which is very hard to interpret. Here, also the precision values are

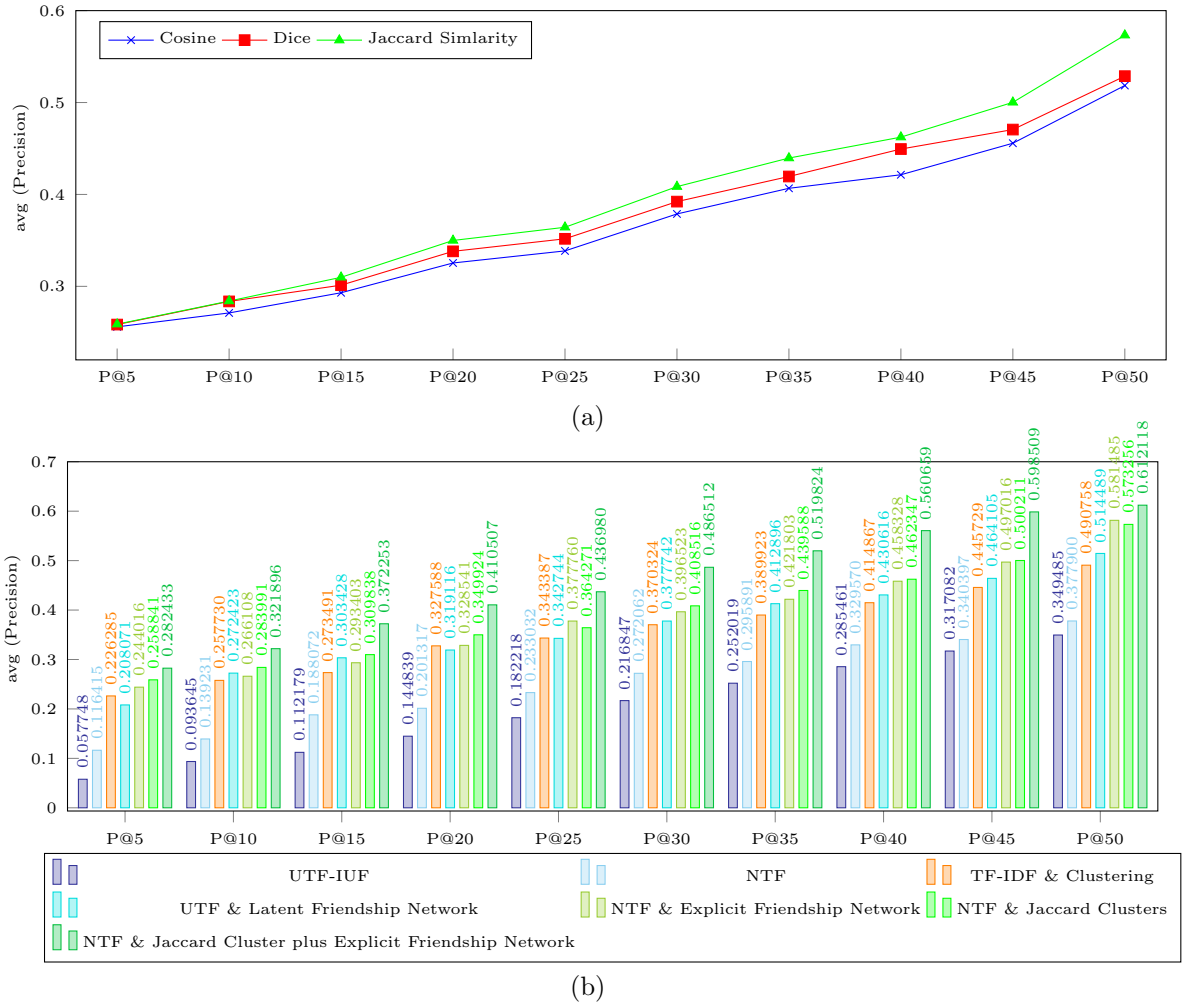


Figure 4.9: Comparative analysis of average precision value over different values of K for (a) CRUIP based intermediate UIP's over different similarity measures (b) State of the art methodologies based UIP and proposed methodology based intermediate UIP's and Final UIP.

averaged over the users in the dataset. According to results shown in Fig. 4.9(a), all three intermediate UIP's have achieved a good precision value for every value of K, but the intermediate UIP corresponding to jaccard similarity measure is the highest performer. The increase in the value of K will also lead to increase in the value of average precision (Y-axis).

The comparison results of intermediate UIP's, final UIP constructed by proposed methodology with all state of the art methodologies is shown in Fig. 4.9(b). The results pattern shows that every UIP either intermediate or final constructed from strategies employed in proposed methodologies outperforms every state of the art methodology with different margins for each value of K. For $K = 50$, *i.e.*, $P@50$, the final UIP can achieve a value of 0.612118, which is almost 18.97 % more than the nearest state of the art methodology

in average precision value for same K. Even as $K = 5$, the final UIP can obtain a precision of 0.282433 which is still 24.81% more than the nearest one. Here, the contribution percentage of CRUIP and FRUIP towards the efficient enrichment of final UIP cannot be clearly judged as for some values of K, CRUIP leads FRUIP and for some FRUIP leads CRUIP. However, their leading margin always remains less no matter what the value of K, which means both are equally valuable for UIP enrichment. So the results of MRR, *imp*, *completeness* and $P@K$ proves the claim that single strategy is not enough for the enrichment of UIP there must be multiple ones.

The testing of protocols based hypotheses discussed in Section 4.1 is performed on the basis of results shown in Figs. 4.6 to 4.9, while Table 4.3 describe these results. The symbol tick (✓) represents that the null hypothesis H_0 is true; and symbol cross (✗) means that the null hypothesis H_0 is false. Whereas ENC means that the experiment has not been conducted to measure the efficiency of intermediate UIP over state of the art methodologies for a particular metric. NA simply depicts that the testing of hypothesis to prove the efficiency of one intermediate UIP over another is not required.

Based on the result patterns as shown in Figs. 4.6 to 4.9 and results of hypotheses testing in Table 4.3, some general observations can be made. Firstly, methodologies, *i.e.*, UTF-IUF and NTF, which do not employ any UIP enrichment method are least performers as it can only predict tags whose instances are in the training set. Secondly, in the case of methodologies TF-IDF plus clustering and UTF-IUF plus latent friendship network, there is a very little difference in performance in terms of MRR, *imp*, *completeness* and $P@K$ values. But in overall scenario latent friendship network strategy for UIP enrichment wins the race. However, both methodologies are far better than UTF-IUF and NTF by a significant margin in performance and support the claim of this work of employing UIP enrichment strategy. But the final UIP constructed by proposed methodology leads the race with a considerably high margin for every evaluation metric. The combined effect of both *tag clustering* and *friendship network* that too explicitly defined by a user, prove the claim of present work for developing a novel methodology for the construction of more accurate and effective User Interest Profile.

Table 4.3: Hypothesis Testing of intermediate and full-fledged final UIP

Metric	**UIP construction methodology**	**First Hypothesis**	**Second Hypothesis**	**Third Hypothesis**
P@K	Methodology-1	✓	✓	✓
	Methodology-2	✓	✓	✓
	Methodology-3	✓	✓	✓
	Methodology-4	✓	✓	✓
	B_p plus C_p	NA	NA	✓
	B_p plus G_p	NA	NA	✓
MRR	Methodology-1	✓	✓	✓
	Methodology-2	✓	✓	✓
	Methodology-3	✓	✓	✓
	Methodology-4	✓	✓	✓
	B_p plus C_p	NA	NA	✓
	B_p plus G_p	NA	NA	✓
imp	Methodology-1	ENC	ENC	✓
	Methodology-2	ENC	ENC	✓
	Methodology-3	ENC	ENC	✓
	Methodology-4	ENC	ENC	✓
	B_p plus C_p	NA	NA	✓
	B_p plus G_p	NA	NA	✓
completeness	Methodology-1	✓	✓	✓
	Methodology-2	✓	✓	✓
	Methodology-3	✓	✓	✓
	Methodology-4	✓	✗	✓
	B_p plus C_p	NA	NA	✓
	B_p plus G_p	NA	NA	✓

4.4 Summary

In this chapter, the methodology for modeling of an efficient and complete User Interest Profile (UIP) has been proposed. Basically a UIP enlists the user preferences in which a user is interested or may be interested in along with the degree of interest in each preference. UIP is also considered as the main supporting module of a personalization system and it can also not be denied that performance of any personalization system is directly proportional to UIP efficiency. The proposed methodology uses a multi-strategy, multi-

level approach for UIP modeling where a different strategy is adopted at a different level. Moreover, in the proposed methodology a dedicated protocol has been designed for every level to extract or mine the user's information from the strategy adopted at that level. However, in the present work only three strategies are selected to give three-level UIP, *i.e.*, Folksonomy-based UIP (FBUIP), Cluster Recommended UIP (CRUIP), and Friends Recommended UIP (FRUIP). The B_p will work for FBUIP, whereas G_p and C_p will work for FRUIP and CRUIP respectively. B_p will utilize user own tags, while G_p and C_p use society relations and tag relations respectively. The responsibility of UIP enrichment lies with G_p and C_p protocols. The proposed methodology is not limited to 3-protocols or 3-levels, it can be extended to n-protocol or m-levels depending on strategies selected for UIP information. An extensive set of experiments are performed to analyze and validate the effectiveness of a UIP created by proposed methodology. The results of experiments are quantified by using various evaluation metrics and provides the answer to some key research questions raised in Literature review chapter of this thesis.

RQ1 *How to construct a user's UIP using his collaborative tagging information ?*

To answer this question, the UIP corresponding to every user in the dataset under consideration has been constructed using the B_p protocol at FBUIP level. The protocol will identify ternary relations, *i.e.*, $R^u_{t,w}$ of a user from his collaborative tagging information. Basically, it will enlist the tags which are used by a user himself for tagging at various instances. For measuring the degree of interest towards each preference, *i.e.*, tag, protocol has used NTF approach. Still the UIP constructed by utilizing only the users own information with help of B_p is sparse and incomplete which must be enriched in order to give effective personalization results.

RQ2 *How can we perform the user profile enrichment for construction of a strong UIP?*

To answer this question, strategies of user's society relationship network and tag relationship network are utilized for mining of additional information about the user with help of G_p and C_p protocols respectively. The recommendations provided by these protocols are in the form of tags along with the probable degree of interest that a user may holds for a recommended tag. The information about user preferences from B_p is clubbed together

with recommended information from G_p and C_p protocols by F_{trade} function to obtain a full-fledged final UIP. Moreover, two intermediate UIPs are also constructed by clubbing the information from G_p with B_p and C_p with B_p to separately analyze the impact of user's additional information as predicted by selected enrichment strategies. Patterns shown by the experimental results in Figs. 4.6 to 4.9 and results of hypotheses testing in Table 4.3 supports the answers to above discussed questions. In next the chapter, discussion regarding the Resource Illustration Profile (RIP) modeling and other related modules of a personalization model is provided.

Chapter 5

Resource Illustration Profile (RIP) modeling and Personalization of web resources

In Chapter 4, firstly a discussion regarding the importance of a UIP towards a personalization model and the methodology proposed for UIP modeling was provided, that is to be followed by the personalization model presented in this thesis. The proposed methodology not only utilizes a strategy to extract knowledge from the user's own collaborative tagging information but also use strategies for UIP enrichment. Mining of additional information about the user is performed on the basis of user's explicitly defined society relationships and real-world tag relationships. UIP constructed based on the proposed methodology is more efficient and complete in comparison to the UIP by other methodologies, as confirmed by the evaluation results of the experiments.

Still for a qualitative and effective personalized ranking of web resources by a personalization model, contribution of other supporting modules of personalization is also necessary. As per discussion in Chapter 3, the key supporting modules renaming after UIP modeling are RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation[1]. Each of these supporting module handles a different task but with a common aim to provide a high performing personalization model. Most of the research works studied in the literature have either ignored the importance of these supporting modules or fails to solve various issues involved in these modules. In the proposed personalization model presented in this thesis, we have taken care of all above mentioned supporting modules along with

[1]The contents of this chapter are partly published in:

- Shubham Goel and Ravinder Kumar. "Brownian Motus and Clustered Binary Insertion Sort methods: An efficient progress over traditional methods", Future Generation Computer Systems (FGCS), 86, pp. 266-280, 2018.
- Shubham Goel and Ravinder Kumar. "Collaboratively Augmented UIP - Filtered RIP with Relevancy Mapping for Personalization of Web Search", Information Sciences, 2020. (Accepted)

the issues which hinders the efficiency of a corresponding module. For each module, a novel methodology have been proposed to perform the dedicated task of that module. The methodologies proposed and issues of corresponding modules will be discussed in subsequent sections in an orderly fashion.

5.1 RIP Modeling

In a collaborative system, the resource profile of a web resource provides a collective viewpoint of all the users annotating that web resource with some tags. Basically, this collective viewpoint illustrates that a particular web resource is related to these topics or items in one way or the other; and anyone who is interested to get some information about these topics or items can refer to it. In addition to the list of topics or items, *i.e.*, tags which illustrate a web resource, RIP also contribute towards the degree of affinity with which a tag can illustrate a web resource. Formally, RIP is represented as follows:

Definition 8 *Let* $\{t_1^{r_i}, t_2^{r_i}, ..., t_m^{r_i}\}$ *and* $\{\alpha_1^{r_i}, \alpha_2^{r_i}, ..., \alpha_m^{r_i}\}$ *are the sets of tags used by various users to annotate web resource* i *and the corresponding degree of affinity with which a web resource* i *is illustrated by those tags respectively. For target web resource* i*, RIP is represented by* $\overrightarrow{R}_i$ *as:*

$$\overrightarrow{R}_i = (t_1^{r_i} : \alpha_1^{r_i}, t_2^{r_i} : \alpha_2^{r_i}, ..., t_m^{r_i} : \alpha_m^{r_i})$$

Where, m depicts the cumulative count of tags used by various users for annotating the web resource i; and $\alpha_j^{r_i}$ is the affinity with which a tag $t_j^{r_i}$ can illustrate a web resource i. The value of $\alpha_j^{r_i}$ is highly influenced by the fraction of users who have annotated the web resource i with tag $t_j^{r_i}$ and all the users who had annotated the web resource i with any tag. A RIP constructed on the basis of each and every tag used to annotate a web resource by different users is known as collective RIP. But due to their own personal reasons the users may hold a different viewpoint towards the same web resource. For example, four users Abu, Ajay, Geeta, and John annotate the web page of vegetarian restaurant, *i.e.*, *Shree Rathnam* with some tags of their interest. Ajay and Geeta share a similar kind of interest in their food habits and are vegetarian. John preferred vegetarian and non-vegetarian food as well, while Abu usually prefer a non-vegetarian food as depicted by their UIPs. The annotations made by Ajay, Geeta and John towards this restaurant are

almost similar and indicate towards a vegetarian food, but Abu had used tags like mutton, chicken, and sushi which are known non-vegetarian foods. There can be any intention of Abu behind making such kind of annotations to the *Shree Rathnam*. No doubt, in collective RIP, affinity of these tags is small, but this will definitely impact the ranking of *Shree Rathnam* web page, whenever a query for a vegetarian restaurant is issued in a search engine. Moreover, these outliers, *i.e.*, tags not in alignment with the purpose of a web resource and put an additional burden on the post-relevancy score calculation process in terms of extra computation cycles. Therefore, in order to overcome these shortcomings of collective RIP, outliers must be detected and not to be considered in RIP construction. The proposed methodology for RIP modeling incorporates an Intelligent Collaborative Filter (ICF) to detect the outliers which is a more real-world approximation of the outlier detection problem than its counterpart [129].

The working of ICF is based on the concept of community modeling where each community corresponds to a Topic of Interest (TOI). It is universally accepted that members belonging to the same TOI community present a similar outlook and have the same feelings. Here, both users and web resources can be the potential members of multiple TOI communities subject to the condition that their degree of membership may be different. Thus, a separate community profile is created for every user and web resource, *i.e.*, User Community Profile (UCP) and Resource Community Profile (RCP). Formally, UCP and RCP can be defined as follows:

Definition 9 *Let* $\{c_1^{u_i}, c_2^{u_i}, ..., c_p^{u_i}\}$ *and* $\{\varrho_1^{u_i}, \varrho_2^{u_i}, ..., \varrho_p^{u_i}\}$ *are the sets of TOI communities to which user* i *belongs and the corresponding degree of membership he holds for those communities respectively. For target user* i*, UCP vector is represented by* $\overrightarrow{C}_i^u$ *as:*

$$\overrightarrow{C}_i^u = (c_1^{u_i} : \varrho_1^{u_i}, c_2^{u_i} : \varrho_2^{u_i}, ..., c_p^{u_i} : \varrho_p^{u_i})$$

Definition 10 *Let* $\{c_1^{r_j}, c_2^{r_j}, ..., c_q^{r_j}\}$ *and* $\{\varpi_1^{r_j}, \varpi_2^{r_j}, ..., \varpi_q^{r_j}\}$ *are the sets of TOI communities to which web resource* j *belongs and the corresponding degree of membership it holds for those communities respectively. For target web resource* j*, RCP vector is represented by* $\overrightarrow{C}_j^r$ *as:*

$$\overrightarrow{C}_j^r = (c_1^{r_j} : \varpi_1^{r_j}, c_2^{r_j} : \varpi_2^{r_j}, ..., c_q^{r_j} : \varpi_q^{r_j})$$

Where, p and q depict the cumulative count of TOI communities to which user i and

web resource j belongs; and $\varrho_k^{u_i}$ and $\varpi_k^{r_j}$ depicts their degree of membership in TOI community $c_k^{u_i}$ and $c_k^{r_j}$ respectively. The value of membership degree for both user and web resource towards a TOI community is predicted from the analysis of real-world Tag-TOI relationships. Therefore, a Tag-TOI Relationship matrix ($TToR_m$) is constructed in which each cell measures the semantic relatedness level between the corresponding tag and TOI.

Semantic relatedness is capable of entertaining any type of relationship between the Tag-TOI pair, whereas other conventional methods like syntactic, co-occurrence, semantic similarity, and Linear Dirichlet Allocation can take care of only one type of relationship. Thus, community modeling in ICF can also entertain the Tag-TOI pairs which are neither semantically nor syntactically similar, but are semantically related to each other in the real-world.

The preliminary requirement for computation of $TToR_m$ is the Tag and TOI vectors, which are obtained in the similar way as tag vectors are obtained from Word2vec model to compute TTR_m in Chapter 4. Word2vec is computationally more efficient than several other predictive modeling methods. Due to this reason, it has been used in ICF. The word-vectors corresponding to each and every tag in a collaborative tagging dataset under consideration and TOIs taken from Open Directory Project (ODP) [161] were selected from the collection of word-vectors created by Word2vec model. Jaccard distance between the vector representation of a tag and TOI has been used to compute the semantic relatedness level between them as explained in Eq. (5.1);

$$TToR_m(\overrightarrow{t_i}, \overrightarrow{t_j}) = \frac{\sum_{k=1}^{n} w_{ki} * w_{kj}}{\sum_{k=1}^{n} w_{ki}^2 + \sum_{k=1}^{n} w_{kj}^2 - \sum_{k=1}^{n} w_{ki} w_{kj}} \tag{5.1}$$

Where, $\overrightarrow{t_i}$ and $\overrightarrow{t_j}$ are vector representations of Tag t_i and TOI t_j respectively, while n is the vector length and w_{ki},w_{kj} are elements of corresponding vectors. Greater the value of jaccard distance between $\overrightarrow{t_i}$ and $\overrightarrow{t_j}$, higher would be the real-world semantically relatedness between t_i and t_j. After computing $TToR_m$, community modeling and collaborative filtering were performed using Algorithm 5 to eliminate outlier annotations from a web resource. In addition to this, ternary relations corresponding to outliers were removed

from C3TG for refinement; and then, a refined RIP is constructed as per description in Definition 8. For more details, refer Algorithm 5.

Algorithm 5: Intelligent Collaborative Filtering

Input : Tag-TOI Relationship matrix $TToR_m$, topic of interest list TOI and Collaborative 3-Partite Graph $C3TG$
Output : Outliers and refined Collaborative 3-Partite Graph
Initialize :
$resource\_list \leftarrow$ Fetch every web resource from $C3TG$
$user\_list \leftarrow$ Fetch every user from $C3TG$
for *each* wr *in* $resource\_list$ **do** ▷ *wr is a web resource*
  $tag\_list \leftarrow$ Fetch the tags from $C3TG$ corresponding to wr
  for *each topic in* TOI **do**
    $sum, m\_deg \leftarrow 0$
    for *each* t *in* tag_list **do** ▷ *t is a tag*
      $sum \leftarrow sum + TToR_m(t, topic)$
    $m\_deg \leftarrow sum/length(tag\_list)$
    $temp\_profile(wr, topic) \leftarrow m\_deg$
    ▷ *membership degree of wr in community cooresponding to topic*
  $temp\_profile \leftarrow sort(temp\_profile)$ ▷ *sort w.r.t. m_deg in a decreasing order*
  $RCP(wr, topic) \leftarrow$ Fetch records corresponding to first η topics from $temp\_profile(wr, topic)$ ▷ *η denote threshold computed using Chebyshev Law*
  Empty(tag_list) ▷ *Delete all element tags of tag_list*
for *each* u *in* $user\_list$ **do** ▷ *u is an user*
  $tag\_list \leftarrow$ Fetch the tags from $C3TG$ corresponding to u
  for *each topic in* TOI **do**
    $sum, m\_deg \leftarrow 0$
    for *each* t *in* tag_list **do** ▷ *t is a tag*
      $sum \leftarrow sum + TToR_m(t, topic)$
    $m\_deg \leftarrow sum/length(tag\_list)$
    $temp1\_profile(u, topic) \leftarrow m\_deg$
    ▷ *membership degree of user u in community cooresponding to topic*
  $temp1\_profile \leftarrow sort(temp1\_profile)$ ▷ *sort w.r.t. m_deg in a decreasing order*
  $UCP(u, topic) \leftarrow$ Fetch records corresponding to first η topics from $temp1\_profile(u, topic)$ ▷ *η denote threshold computed using Chebyshev Law*
  Empty(tag_list) ▷ *Delete all element tags of tag_list*
for *each* $R^u_{t,wr}$ *in* $C3TG$ **do** ▷ *$R^u_{t,wr}$ is a ternary record for (user, tag, web resource)*
  $resource\_community \leftarrow$ Fetch the TOI communities from RCP corresponding to web resource wr
  $user\_community \leftarrow$ Fetch the TOI communities from UCP corresponding to user u
  if $(resource\_community \cap user\_community) == \emptyset$ **then**
    $records\_to\_remove \leftarrow records\_to\_remove.append(IndexOf(R^u_{t,wr}))$
    ▷ *list of indexes corresponding to outliers relations*
$C3TG = Drop(C3TG, records\_to\_remove)$
▷ *Remove all records resulting in outlires to give refined C3TG*

In order to construct an unambiguous RIP, an example of outlier tags' detection and their

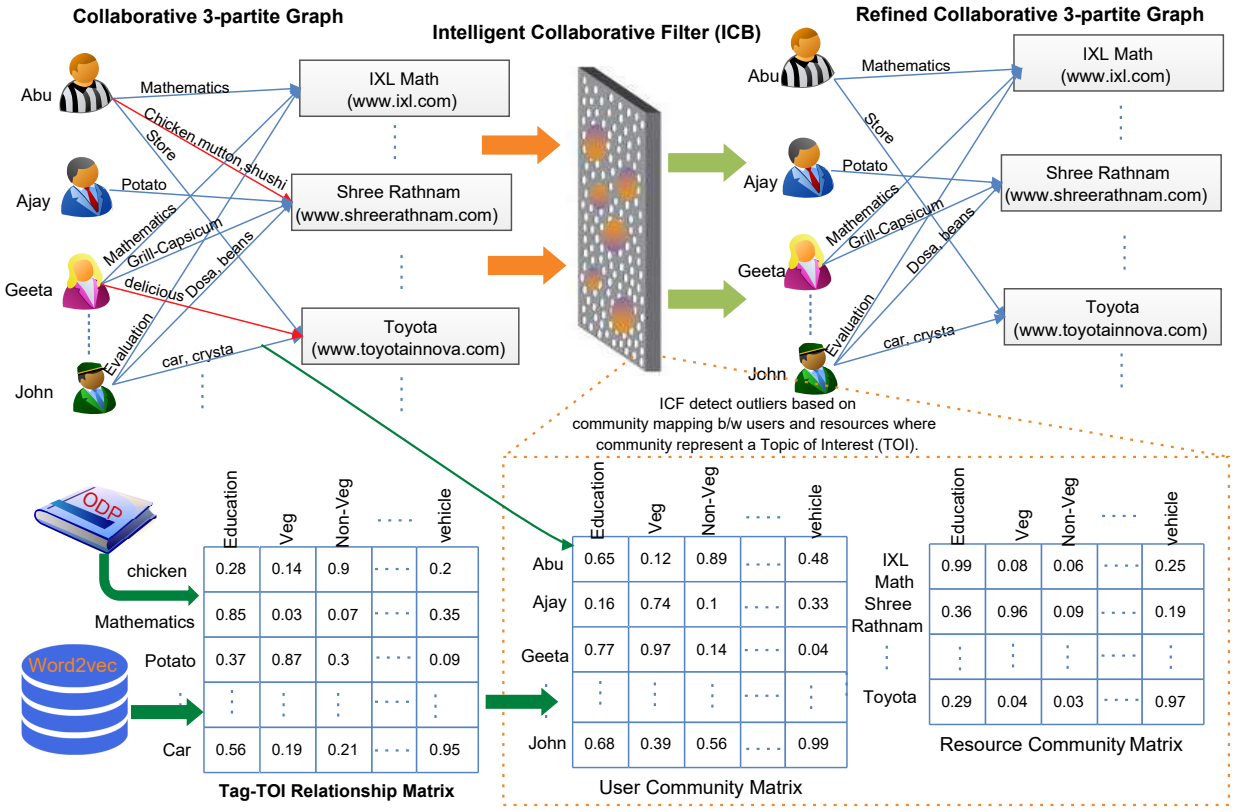


Figure 5.1: Framework for Intelligent Collaborative Filter

removal using ICF is explained in Fig. 5.1, where a part of C3TG has been considered to illustrate the process. Outlier tags have been represented in C3TG using red colored coded relations which get detected and removed from the original C3TG as soon as C3TG passes through ICF to get a refined C3TG. The colored circle present on ICF shows the blockage of outliers. The Tag-TOI Relationship Matrix $TToR_m$ was obtained from Eq. (5.1), while User Community matrix and Resource Community matrix were obtained from $TToR_m$ and C3TG using Algorithm 5.

5.2 Personalization methodology

Personalization of user's web search refers to the ranking of web resources in accordance to user's preferences as every user holds a different interest for various items. It increases not only the user satisfaction level towards a web search engine, but also decreases the number of required computation cycles which are generally wasted due to repeated query reformulation in case of irrelevant information retrieval. In this section, a novel methodology has been proposed for post-relevancy score computation of a web resource. On the basis

of computed post-relevancy score, personalized ranking of web resources is performed. The visual representation of the process for post-relevancy score computation is already presented in personalization model as shown in Fig. 3.1. Now in this section, the mathematical formulation of the proposed methodology have been described for post-relevance score computation.

The post-relevance score of a web resource is a trade-off between two sub-relevancy scores, *i.e.*, query relevancy score and user interest relevancy score. Therefore, firstly, the formulation of sub-relevancy scores is presented, followed by their trade-off mechanism to get post-relevancy score of a web resource. Fig. 3.1 provides more details about the components, and these have been described in the subsequent subsections of this chapter.

5.2.1 Query relevancy mapping

A query issued by a user signifies the need for a particular information and is usually represented by a term vector. Only the terms excluding every stopword are considered into the term vector. Formally, query is defined as:

Definition 11 *Let* $\{t_1^{q_i}, t_2^{q_i}, t_3^{q_i}, ..., t_l^{q_i}\}$ *is the query terms set excluding stopwords corresponding to the query issued by a user. For target query* i*, term vector is represented by* $\overrightarrow{Q}_i$ *as:*

$$\overrightarrow{Q}_i = (t_1^{q_i} : 1, t_2^{q_i} : 1, t_3^{q_i} : 1, ..., t_l^{q_i} : 1)$$

Where, l is the vector length, *i.e.*, the cumulative count of terms in the user's query excluding stopwords. It is assumed that every term other than stopwords in users query is equally important and follows a conjunction relation. Due to this reason, weightage of each term in term vector is taken as 1.

Web search engines consider $\overrightarrow{Q}_i$ as input to the process of retrieving and ranking of web resources in place of users' original query. This process is known as query relevancy mapping. As the name suggests, $\overrightarrow{Q}_i$ is mapped with RIP corresponding to every web resource in the resource set to assign a certain score to them. Basically, this score quantifies the relevance of a corresponding web resource k for $\overrightarrow{Q}_i$, *i.e.*, how well a web resource k can fulfill the information need of the user as per the $\overrightarrow{Q}_i$ among all other web resources retrieved by the search engine. Therefore, this score is called as query relevancy score which

can be formally defined as:

Definition 12 *A query relevancy score is a resultant of mapping between query* $\overrightarrow{Q}_i$ *and RIP corresponding to a web resource in resource set* R*. For a target web resource* k*, query relevancy score is represented by* $S_k^{q_i}$ *as:*

$$S_k^{q_i} = Q_{map}R(\overrightarrow{Q}_i, \overrightarrow{R}_k)$$

$Q_{map}R$ generalized over the entire resource set R is as follows:

$$\begin{bmatrix} t_1^{q_i} \\ t_2^{q_i} \\ \vdots \\ t_a^{q_i} \\ \vdots \\ t_l^{q_i} \end{bmatrix} \Xi \begin{bmatrix} t_1^{r_1} : \alpha_1^{r_1} & t_2^{r_1} : \alpha_2^{r_1} & t_3^{r_1} : \alpha_3^{r_1} & \dots & t_m^{r_1} : \alpha_m^{r_1} \\ t_1^{r_2} : \alpha_1^{r_2} & t_2^{r_2} : \alpha_2^{r_2} & t_3^{r_2} : \alpha_3^{r_2} & \dots & t_m^{r_2} : \alpha_m^{r_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^{r_k} : \alpha_1^{r_k} & t_2^{r_k} : \alpha_2^{r_k} & t_3^{r_k} : \alpha_3^{r_k} & \dots & t_m^{r_k} : \alpha_m^{r_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^{r_d} : \alpha_1^{r_d} & t_2^{r_d} : \alpha_2^{r_d} & t_3^{r_d} : \alpha_3^{r_d} & \dots & t_m^{r_d} : \alpha_m^{r_d} \end{bmatrix} \rightarrow \begin{bmatrix} S_1^{q_i} \\ S_2^{q_i} \\ \vdots \\ S_k^{q_i} \\ \vdots \\ S_d^{q_i} \end{bmatrix}$$

Where, Ξ symbolizes the query relevance mapping function $Q_{map}R$; and d is the cumulative count of web resources in the resource set R. The count of tags in RIP of the web resource k is represented by m; and the value of m can vary from one web resource to another. Further, $t_j^{r_k}$ is the tag used by some user to annotate a web resource k; and $\alpha_j^{r_k}$ represents the degree of affinity with which $t_j^{r_k}$ can illustrate the web resource k. Higher the value of $S_k^{q_i}$, more would be the web resource k relevant to the query i. The value of any $S_k^{q_i}$ can vary from [0,1].

As per reviewed studies, every conventional search engine uses query relevance score $S_k^{q_i}$ for the ranking of web resources which generally compute its value solely on the basis of a keyword or term matching. Higher the proportion of matched terms between $\overrightarrow{Q}_i$ and $\overrightarrow{R}_k$ among all the terms in the $\overrightarrow{Q}_i$, more would be the value of $S_k^{q_i}$. However, in the present time, it is undesirable to depend solely on term matching approach and giving no consideration to affinity of the term for a web resource as there has been a tremendous increase in the number of competent web resources. Almost every web resource in competent group has the same set of terms, however, their affinity varies from one web resource to another. Therefore, the impact of affinity with which a term or tag can illustrate a web resource must be considered in Query-RIP mapping. The assumption or claim contrived for query relevancy-based score computation for a web resource is illustrated through an

example as follows:

Example 1 *Suppose a user, let us say, John issues a query "SUV Crysta"in a search engine for which a number of web pages are retrieved out of that RIPs corresponding three web pages* $\overrightarrow{R}_1$, $\overrightarrow{R}_2$ *and* $\overrightarrow{R}_3$ *are as follows:*

$$\overrightarrow{R}_1 = (SUV : 0.5, Crysta : 0.6, Car : 0.45, \dots)$$
$$\overrightarrow{R}_2 = (SUV : 0.2, Crysta : 0.1, Red : 0.5, \dots)$$
$$\overrightarrow{R}_3 = (SUV : 0.8, Toyota : 0.75, Car : 0.35, \dots)$$

Where, query $\overrightarrow{Q}_1$ *is:*

$$\overrightarrow{Q}_1 = (SUV : 1, Crysta : 1)$$

If the conventional approach is followed, *i.e.*, only keyword matching for the mapping of $\overrightarrow{Q}_1$ with $\overrightarrow{R}_1$, $\overrightarrow{R}_2$ and $\overrightarrow{R}_3$ in order to calculate their $S_1^{q_1}$, $S_2^{q_1}$ and $S_3^{q_1}$ respectively, then a following relationship of query relevancy scores is obtained:

$$S_1^{q_1} = S_2^{q_1} > S_3^{q_1}$$

The relationship clearly indicates towards the drawbacks of conventional approach as both $\overrightarrow{R}_1$ and $\overrightarrow{R}_2$ appears with the same query relevancy score, but in reality as per the affinity of tags in their RIPs, $\overrightarrow{R}_1$ has greater relevance to query rather $\overrightarrow{R}_2$. If the query relevancy mapping is formulated using the formulations of Cai et al. [34], then the relationship of relevancy is scored as follows:

$$S_1^{q_1} > S_2^{q_1} > S_3^{q_1}$$

Here, the results are found to be more reliable than those attained through a conventional approach, and also confirm the claim of this work that it is undesirable to perform Query-RIP mapping solely on the basis of keyword matching. However, the Query-RIP mapping still has not showcased the real-world formulation of the mapping problem, and the absence of which can hamper the validity of a relevancy score. The terms or keywords of a query not present in the RIP of a web resource may be strongly related to that web resource in the real-world. This can be attributed to the reason that a term related to a web resource may not be there in RIP due to problem of synonymy or limited awareness of web resource annotators. For instance, term $Crysta$ is not present in the web resource R_3, but very strongly related to $Toyota$ and Car; as $Toyota$ is a very famous car manufacturer

of which *Crysta* is a popular model. So, the combined influence of tags in R_3 is more informative to query than R_2. Moreover, as illustrated by Cai et al. [34], every term in $\overrightarrow{Q}_i$ is a fuzzy requirement of query issuer which is expected to be fullfilled by a web resource in the best possible manner. Therefore, to analyze the impact of real-world relationship of those query terms which are not present in RIP but related to a web resource in one way or the other, $Q_{map}R$ is formulated as a fuzzy satisfaction problem using Eq. (5.2) in order to compute $S_k^{q_i}$ for web resource k w.r.t. query i.

$$Q_{map}R(\overrightarrow{Q_i}, \overrightarrow{R_k}) = \begin{cases} 1, & \forall t_l^{q_i} \in \overrightarrow{Q_i} \wedge \exists t_j^{r_k} \in \overrightarrow{R_k}((t_l^{q_i} = t_j^{r_k}) \wedge (\alpha_j^{r_k} = 1)) \\ \frac{\sum_{i=1}^{|\overrightarrow{Q_i}|} \alpha_j^{r_k}}{|\overrightarrow{Q_i}|} * \frac{|Q_i \cap R_k|}{|\overrightarrow{Q_i}|} + Z(\overrightarrow{Q_i}, \overrightarrow{R_k}), & \exists t_l^{q_i} \in \overrightarrow{Q_i} \wedge \exists t_j^{r_k} \in \overrightarrow{R_k}((|t_l^{q_i} = t_j^{r_k}| \geq 1) \wedge (\alpha_j^{r_k} \neq 1)) \\ 0, & \text{otherwise.} \end{cases} \tag{5.2}$$

Where, $Z(\overrightarrow{Q_i}, \overrightarrow{R_k})$ can be formulated using Eq. (5.3) as:

$$Z(\overrightarrow{Q_i}, \overrightarrow{R_k}) = \sum_{t_l^{q_i} \in (Q_i \setminus R_k)} \sum_{t_j^{r_k} \in (R_k \setminus Q_i)} \frac{\alpha_j^{r_k}}{1 - rel(t_l^{q_i}, t_j^{r_k})} \tag{5.3}$$

Where, Q_i and R_k represent the set of terms in $\overrightarrow{Q_i}$ and tags in $\overrightarrow{R_k}$ respectively. Further, $\alpha_j^{r_k}$ represents the affinity of a tag j, *i.e.*, $t_j^{r_k}$ in RIP of web resource k; and $t_l^{q_i}$ is a term in query i. The relative complement of R_k in Q_i is denoted by $Q_i \setminus R_k$. Similarly, $R_k \setminus Q_i$ denotes relative complement of Q_i in R_k. The real-world relatedness, *i.e.*, $rel(t_l^{q_i}, t_j^{r_k})$ is measured using the semantic relatedness of vector representations corresponding to $t_l^{q_i}$ and $t_j^{r_k}$ respectively. Word2vec model has been used to obtain these vector representations. Based on $Q_{map}R$ in Eqs. (5.2) and (5.3), following relationship of query relevancy scores is obtained for Example 1.

$$S_1^{q_1} > S_3^{q_1} > S_2^{q_1}$$

The $Q_{map}R$ results of the current work are more satisfactory than those of other similar studies. Both top and bottom cases in Eq. (5.2) represent the boundaries of query relevancy score, whereas middle case represents the real gist of the proposed Query-RIP mapping. The ranking of web resources is performed on the basis of their respective query relevancy scores. The web resource with the highest value of $S_k^{q_i}$ has been ranked at the first position in the ranking hierarchy, while other rankings appear as per their

values.

5.2.2 User interest relevancy mapping

Performing the ranking of web resources solely on the basis of their query relevancy score may fail to satisfy the user information requirement. A web resource considered most relevant as per Query-RIP mapping may not be in alignment with User Interest Profile (UIP) which is against the personalization of web search principles. UIP of a user provides a detailed description about the preferences of a user and his level of interest. Therefore, to design an efficient personalization framework, the web resources considered to be relevant as per users' query must also comply with their UIPs. So, the UIP-RIP mapping comes into the picture to compute the user interest relevancy score for a web resource. Basically, this score quantifies the relevance of corresponding web resource k for $\overrightarrow{U}_i$, *i.e.*, how well a web resource k complies with preferences of a user as per the $\overrightarrow{U}_i$ among all other web resources retrieved by the search engine. Formally, user interest relevancy score can be defined as:

Definition 13 *A user interest relevancy score is a resultant of mapping between UIP of a user $\overrightarrow{U}_i$ and RIP corresponding to a web resource in resource set R. For a target web resource k, user interest relevancy score is represented by $S_k^{u_i}$ as:*

$$S_k^{u_i} = U_{map}R(\overrightarrow{U}_i, \overrightarrow{R}_k)$$

$U_{map}R$, generalized over the entire resource set R, is as follows:

$$\begin{bmatrix} t_1^{u_i} : \delta_1 \\ t_2^{u_i} : \delta_2 \\ \vdots \\ t_a^{u_i} : \delta_a \\ \vdots \\ t_n^{u_i} : \delta_n \end{bmatrix} \Upsilon \begin{bmatrix} t_1^{r_1} : \alpha_1^{r_1} & t_2^{r_1} : \alpha_2^{r_1} & t_3^{r_1} : \alpha_3^{r_1} & \dots & t_m^{r_1} : \alpha_m^{r_1} \\ t_1^{r_2} : \alpha_1^{r_2} & t_2^{r_2} : \alpha_2^{r_2} & t_3^{r_2} : \alpha_3^{r_2} & \dots & t_m^{r_2} : \alpha_m^{r_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^{r_k} : \alpha_1^{r_k} & t_2^{r_k} : \alpha_2^{r_k} & t_3^{r_k} : \alpha_3^{r_k} & \dots & t_m^{r_k} : \alpha_m^{r_k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_1^{r_d} : \alpha_1^{r_d} & t_2^{r_d} : \alpha_2^{r_d} & t_3^{r_d} : \alpha_3^{r_d} & \dots & t_m^{r_d} : \alpha_m^{r_d} \end{bmatrix} \rightarrow \begin{bmatrix} S_1^{u_i} \\ S_2^{u_i} \\ \vdots \\ S_k^{u_i} \\ \vdots \\ S_d^{u_i} \end{bmatrix}$$

Where, Υ symbolizes the user interest relevance mapping function $U_{map}R$. Tag in UIP of user i is denoted by $t_a^{u_i}$ which can be either provided by FBUIP, CRUIP or FRUIP; and δ_a is the level of interest that user i holds in preference denoted by a tag $t_a^{u_i}$. Higher the value of $S_k^{u_i}$, more would be the web resource k relevant to user i. The value of any $S_k^{u_i}$

can vary from [0,1].

Similarly to Query-RIP mapping problem, $U_{map}R$ is also formulated as a fuzzy satisfaction problem using Eq. (5.4) in order to compute $S_k^{u_i}$ for the web resource k with respect to UIP of a user i. Here, each tag in $\overrightarrow{U}_i$ is the fuzzy requirement of user preferences for the web resource to be enlisted in user interest relevant web resource list. In addition to fuzzy satisfaction modeling of UIP-RIP mapping, the influence of user's interest level in preference and affinity with which a tag can illustrate a web resource are also considered while designing user interest relevancy mapping function $U_{map}R$ in Eq. (5.4).

$$U_{map}R(\overrightarrow{U_i}, \overrightarrow{R_k}) = \begin{cases} 1, & \forall t_n^{u_i} \in \overrightarrow{U_i} \wedge \exists t_j^{r_k} \in \overrightarrow{R_k}((t_n^{u_i} = t_j^{r_k}) \wedge (\delta_n > 0) \\ & \wedge(\alpha_j^{r_k} = 1)) \\ \dfrac{\sum_{i=1}^{|\overrightarrow{U_i}|} \delta_n * \alpha_j^{r_k}}{|\overrightarrow{U_i}|} * \dfrac{|U_i \cap R_k|}{|\overrightarrow{U_i}|} + W(\overrightarrow{U_i}, \overrightarrow{R_k}), & \exists t_n^{u_i} \in \overrightarrow{U_i} \wedge \exists t_j^{r_k} \in \overrightarrow{R_k}((|t_n^{u_i} = t_j^{r_k}| \geq 1) \wedge \\ & (\delta_n > 0) \wedge (\alpha_j^{r_k} \neq 1)) \\ 0, & \text{otherwise.} \end{cases} \tag{5.4}$$

Where, $W(\overrightarrow{U_i}, \overrightarrow{R_k})$ can be formulated using Eq. (5.5) as:

$$W(\overrightarrow{U_i}, \overrightarrow{R_k}) = \sum_{t_n^{u_i} \in (U_i \setminus R_k)} \delta_n * \sum_{t_j^{r_k} \in (R_k \setminus U_i)} \frac{\alpha_j^{r_k}}{1 - rel(t_n^{u_i}, t_j^{r_k})} \tag{5.5}$$

Where, U_i and R_k represent the sets of tags in $\overrightarrow{U_i}$ and $\overrightarrow{R_k}$ respectively. The relative complement of R_k in U_i is denoted by $U_i \setminus R_k$. Similarly, $R_k \setminus U_i$ denotes relative complement of U_i in R_k. The influence of tags in user's UIP which do not match with any tag in RIP of a web resource, but are strongly related to RIP and computed using $W(\overrightarrow{U_i}, \overrightarrow{R_k})$. The user interest relevancy scores of web resources obtained from $U_{map}R$ help to rank web resources according to the personal preferences of the user which is a very important step to achieve personalization. The web resource with the highest value of $S_k^{u_i}$ has been ranked at the first position in the ranking hierarchy, while other rankings appear as per their values.

5.2.3 Post-Relevancy Score

A personalized web search system is committed to satisfy the users' information requirements according to their perspective in a best possible way. It fulfills both current and preferred information requirements represented by a query and UIP respectively with an appropriate trade-off. So, a Post-Relevance Scoring function $PRScore(\overrightarrow{U_i}, \overrightarrow{Q_i}, \overrightarrow{R_k})$ is used in personalized web search system to perform the trade-off task which assigns a post-relevancy score, *i.e.*, $P_k^{u_i,qi}$ to a web resource. Basically, $P_k^{u_i,qi}$ is the trade-off of query relevancy score $S_k^{q_i}$ and user interest relevancy score $S_k^{u_i}$ of web resource k. The mathematical formulation of $P_k^{u_i,qi}$ is shown in Eq. (5.6):

$$P_k^{u_i,qi} = \begin{cases} 1, & S_k^{q_i} = 1 \text{ and } S_k^{u_i} = 1 \\ \beta * S_k^{q_i} + (1-\beta) * S_k^{u_i}, & S_k^{q_i} > 0 \text{ and } S_k^{u_i} \geq 0 \\ 0, & S_k^{q_i} = 0. \end{cases} \tag{5.6}$$

Where, β is the trade-off governing parameter that satisfy $0 \leq \beta \leq 1$. A personalized search engine performs the ranking of web resources on the basis of their post-relevancy score values,but a major implication lies in the value of β. Almost every work studied in the literature had selected a single value for β and was fixed for a personalized system irrespective of the query and the user who had issued that query. It can be argued that the selection of a common value for β is undesirable as this approach is not in accordance with the objectives of a personalized system. Therefore, this work proposes a novel approach to calculate the value of β which varies from one user-query pair to another. For mathematical formulation of β, refer Eq. (5.7):

$$\beta = 1 - \frac{\sum_{t_l^{q_i} \in Q_i} \sum_{t_n^{u_i} \in U_i} rel(t_l^{q_i}, t_n^{u_i})}{|Q_i| * |U_i|} \tag{5.7}$$

Basically, the level of alignment between an issued query Q_i and UIP of a user issuing that query decides the value of β. The impact that alignment level incurs on the value of β is visually represented in Fig. 5.2, where the alignment level is represented by the distance between the centers of blue and brown circles. More the alignment between a query and UIP, lesser would be the value of β and greater the contribution of $S_k^{u_i}$ towards

a $P_k^{u_i,qi}$ for a web resource k.

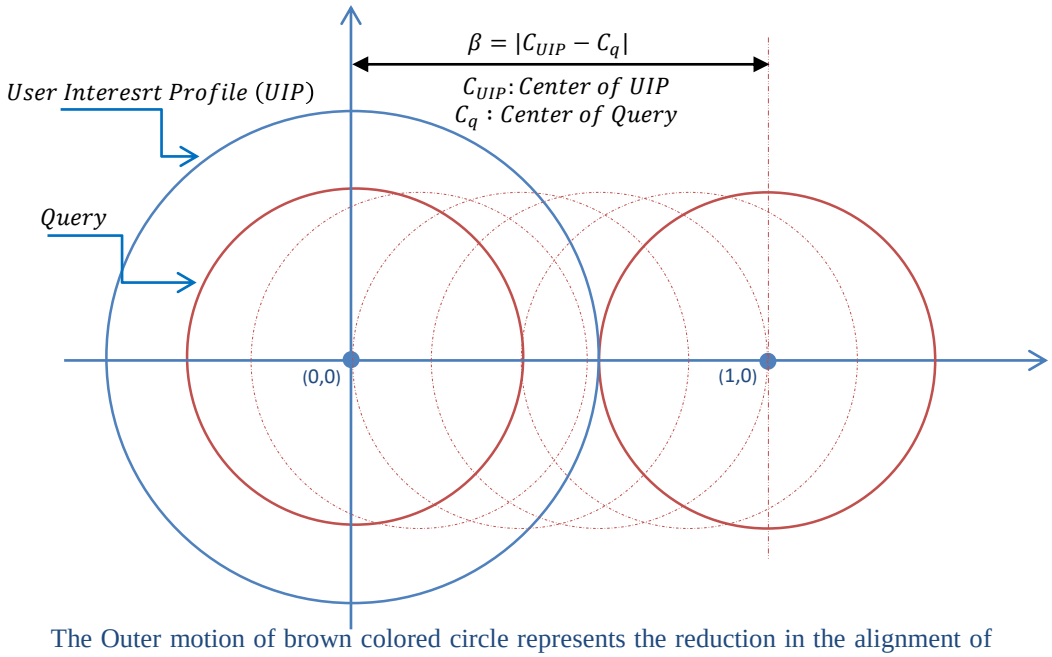


Figure 5.2: Query-UIP Alignment

The interrelated contribution of the various supporting modules, *i.e.*, UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation of the personalization model proposed in Chapter 3 will lead to a more qualitative and relevant personalized ranking of web resources. A special attention has been given on the selection of configuration corresponding to various supporting modules, which is described with help of methodologies proposed for designing of the supporting modules. In Chapter 4, methodology proposed for UIP modeling is described while in this chapter methodology corresponding to remaining modules, *i.e.*, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation is presented. In the next sections, performance of personalized ranking obtained by joint venture of all supporting modules working under proposed personalization model is analyzed.

5.3 Baseline methodologies

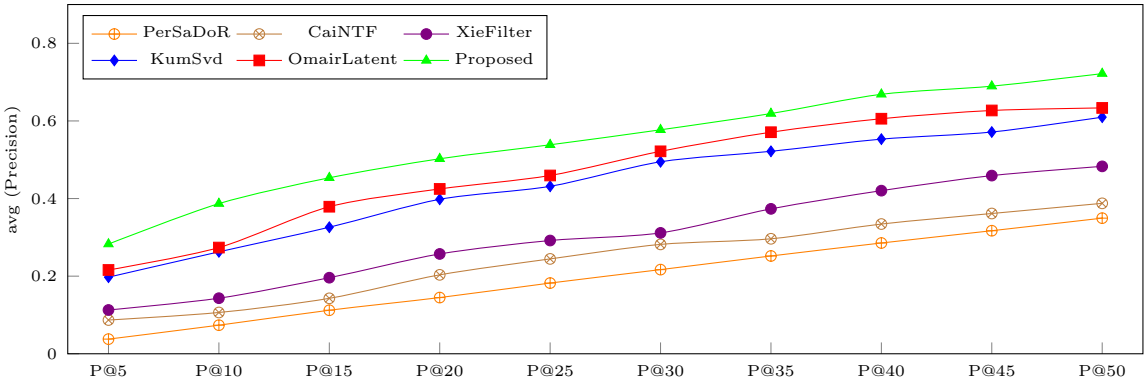
As many as five baseline methodologies for web search personalization were selected and used in the experiments to validate the effectiveness of the proposed personalization model.

The first methodology, *i.e.*, *PerSaDoR* was given by Bouadjenek et al. [31], where the UTF-IUF approach was used to model a user profile and a personalized representation of the web resources used to model resource profile. The second methodology, *i.e.*, *CaiNTF*, was a personalization methodology for web search [34], where the weights of tags both in a user profile and resource profile were determined through NTF values. The third methodology, *i.e.*, *KumSvd*, was presented by Kumar et al. [30] for personalizing the user's web search as per user preferences where both user's own tags and its augmentation were used to model user profile. Further, the profile was used for query disambiguation and matching document was searched and ranked using the Vector Space Model. TF-IDF values and clusters of semantically similar tags were used in *KumSvd* to measure tag weights in order to give modSvdCUIP. The fourth methodology, *i.e.*, *OmairLatent* was proposed by Shafiq et al. [7], where not only the tags used by a user himself, but also the tags recommended by latent friendship circle of a user were used to model user preferences which were then matched with web search results for reordering. The last methodology, *i.e.*, *XieFilter* was proposed by Xie et al. [129], in which tags of user and resource profile were weighted using NTF values. But before constructing resource profile by *XieFilter* social filtering of tags was performed. The baselines selected to have a comparison resembled the current mainstream approaches in order to personalize the user's web search in a social collaborative tagging system. However, the major differences existed in the selection of various strategies for UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score computation process. Moreover, either single or no strategy was used by baselines for user profile enrichment to deal with sparsity problem.

5.4 Results and Discussion

The experimental results of the proposed personalization model have been compared with those of various other baseline methodologies. The metrics described in Table 3.2 have been used to evaluate and quantify the comparison. Line-plots and bar-plots have been used to highlight the difference in the results. Line-plots are used in the case of multi-valued parameters for a personalized ranking; and for a single-valued parameters bar-plots have been used. A uniform color coding has been used in each experiment where

a designated color code is assigned to the proposed model and baseline methodologies in order to clearly visualize even a small change in the performance of personalization methodology in all the experiments.



(a)

Figure 5.3: A comparative analysis of the proposed model and baseline methodologies on the basis of average precision metric

Fig. 5.3 illustrates the difference in the performance of proposed model and other baseline methodologies based on precision, *i.e.*, $P@K$ evaluation metric. It is an important and most commonly used metric for measuring the accuracy of obtained personalized ranking hierarchy of web pages. Basically, it highlights the implications of keeping more percentage of web pages relevant to both the query and the user issuing that query at higher positions in the list of web pages returned as a result of user's query. The experiment was repeated with different values of K (X-axis) in each subsequent iteration. The results obtained from each iteration were averaged over all the users in a dataset to give an average precision (Y-axis) as it was not feasible to represent the results for every user. Greater the value of K, higher would be the precision. It was achieved by the proposed model and baseline methodologies under consideration after verification from the behaviour of line-plots corresponding to respective personalization methodologies exhibited in Fig. 5.3.

It was observed that the proposed model achieved a value of 0.7219 when K was equal to 50, *i.e.*, $P@50$. It indicated that out of top 50 web pages returned as a result of user's query by the proposed methodology, approximately 36 of them were relevant to both query and the query issuer. While comparing these results with those of other baseline methodologies, *i.e.*, OmairLatent and PerSaDor, it was found that the number of relevant web pages among top 50 were approximately 31 and 17 with the respective percentage of 63.35 and 34.94. This count was lesser than that of the proposed model. Even at $K = 5$,

a precision of 0.28272 was achieved by the proposed model which was 30% higher than the dominant baseline. Here, the results of precision were calculated based on the average of all the users and their respective queries in the dataset. As is evident from each value of K in Fig. 5.3, the proposed model has mostly outperformed each baseline methodology by a considerable margin. In other words, an analysis of the results provides that the users can obtain most of the relevant information by just browsing top 50 web pages returned by the proposed model. Here, the value of K is restricted to 50 as the searching behavior of users indicates that they rarely visit web pages beyond this. Moreover, the growth rate also decreases with an increase in the value of K after a specific point as can be visualized from the slopes of line-plot in Fig. 5.3. Therefore, on the basis of precision results, it can be said that the ranking hierarchy of web pages constructed by the proposed model is better than that of other baseline methodologies.

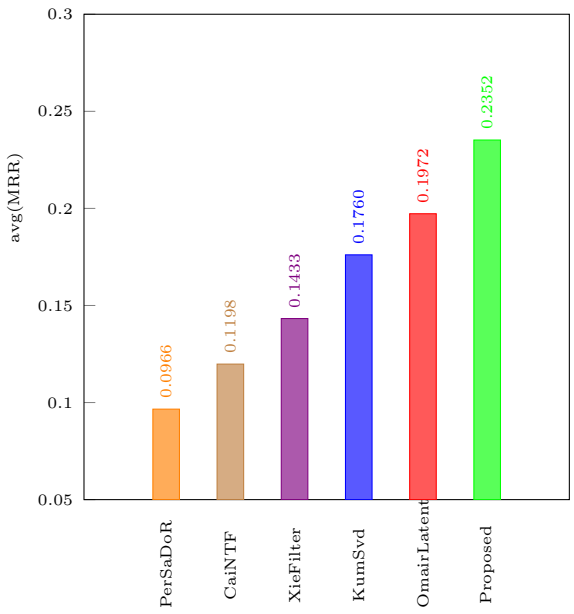


Figure 5.4: Comparative analysis of the proposed model and baseline methodologies on the basis of average MRR metric

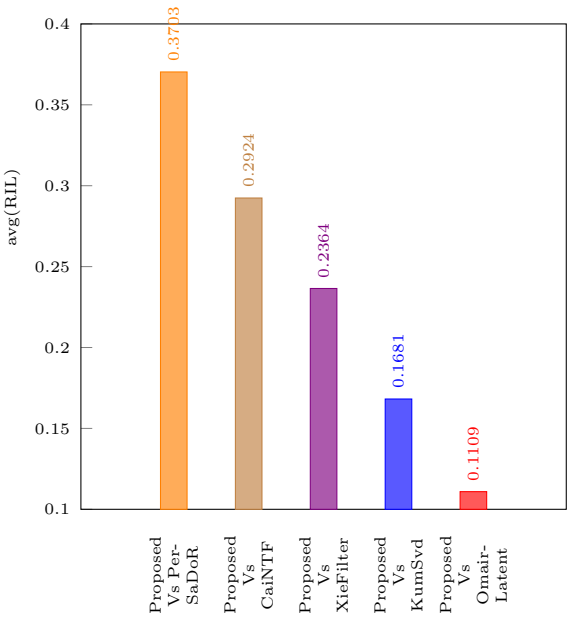


Figure 5.5: Comparative analysis of the proposed model and baseline methodologies on the basis of average RIL metric

Fig. 5.4 presents the comparative analysis results of the proposed model and baseline methodology obtained through MRR metric showing their web page ranking. The evaluation metric MRR is the most important performance measure to quantify the web page ranking methodology used by a search engine in the field of information retrieval. Basically, MRR indicates the presence of most relevant web page closer to the head of web page ranking hierarchy. Higher the value of MRR, more would be the chance of top-

ranked web pages being relevant to the user. In other words, it can be also said that higher the value of MRR, lesser would be the user's effort to find the desired information. The figure highlights that the proposed model has attained the highest MRR value of 0.2352 among all the baselines. It has outperformed the dominant baseline, *i.e.*, OmairLatent by 19.26%, and the least performing baseline by considerably high margin. Similarly, in the case of $P@K$ metric, for the sake of simplicity, MRR has also been averaged over the total number of users in the dataset. The results shown in Fig. 5.4 depicts that the chances of user relevant information to be present at higher ranks are greater for ranking hierarchy obtained by proposed model in comparison to the baseline methodologies.

A comparison of personalized web page ranking hierarchies obtained by the proposed model and respective baseline methodologies based on RIL evaluation metric is shown in Fig. 5.5. Basically, RIL quantifies the relative improvement in the rank of user relevant web pages assigned by two different personalized ranking methodologies. It concludes that improvement and deterioration in a rank of a relevant web page, assigned by a ranking methodology, is always w.r.t. some bases; and a change in bases causes a change in the value of RIL, while ranking methodology remains the same. However, in the case of absolute measure, there is nothing like a bases. Therefore, to compute RIL, it is essential to compare the ranks assigned to the same web page by two different personalized ranking methodologies out of which the second one will act as a bases and for the first one, improvement in web page rank is computed. Improvement can be both positive and negative; where positive means upward movement of a relevant web page in the ranking hierarchy, while a negative value of RIL means degradation of rank, *i.e.*, downward movement in the ranking hierarchy. More the positive value of RIL, greater would be the superiority of personalized ranking methodology over the one which acts as a bases in RIL computation. The results exhibited in Fig. 5.5 show a relative improvement of minimum 11% in the ranking of relevant web pages by the proposed model w.r.t. every baseline methodology. The maximum and minimum RIL for PerSaDor and OmairLatent have been observed as 37.03 % and 11.09% respectively. Therefore, as per results, the proposed model has outperformed every baseline taken into consideration in terms of qualitative web page ranking hierarchy. Similarly to the experiments for evaluation metric $P@k$ and MRR, for RIL also the results have been averaged over all the users in the dataset. The relative improvement in the ranking of web pages by various baseline methodologies in

comparison to the proposed one and other baselines were summarized in Table 5.1.

Table 5.1: Performance of web search personalization methodologies based on RIL metric

Methodology	PerSaDoR	CaiNTF	XieFilter	KumSvd	OmairLatent	Proposed
PerSaDoR	0.00	-7.68	-10.27	-14.81	-19.41	-23.11
CaiNTF	11.17	0.00	-9.46	-12.36	-16.68	-20.78
XieFilter	15.83	12.55	0.00	-10.23	-11.80	-15.88
KumSvd	22.35	19.44	16.34	0.00	-6.53	-11.27
OmairLatent	29.10	26.19	18.85	9.77	0.00	-7.52
Proposed	37.03	29.24	23.64	16.81	11.09	0.00

On the basis of trends as shown in Figs. 5.3 to 5.5 and results of relative improvement in ranking of web pages summarized in Table 5.1, certain observations can be made to deduce some facts. The First one is inline with the facts of UIP modeling in Chapter 4, *i.e.* baselines PerSaDoR, CaiNTF, and XieFilter, which do not employ any UIP enrichment strategy to model a user's UIP, are always the least performers. As they are not able to either enlist every interest of the user or give proper weightage to the enlisted interest. The other remaining baselines, *i.e.*, KumSvd and OmairLatent equipped with some enrichment strategy are able to perform better than PerSaDoR, CaiNTF, and XieFilter, but their performance is not as good as that of the proposed model. Thus, it justifies the claim of the present work that it is essential to employ a UIP enrichment strategy in UIP modeling.

Secondly, the usage of Intelligent Collaborative Filter (ICF) for outlier detection while modeling an RIP in the proposed model greatly influenced its performance which can be clearly noticed from the evaluation results. Even the baseline XieFilter, which had also used social filtering of resource profiles where topic communities are discovered using Latent Dirichlet Allocation method is able to perform better than baselines PerSaDoR and CaiNTF in which no outlier detection was involved. But the method used to discover topic communities for both users and resources in the proposed model can produce more real-world approximations than the method used by XieFilter. This observation towards the construction of an efficient web page personalization methodology also establishes the claim of this study that outlier detection is essential in RIP modeling, otherwise, it can degrade the relevancy of user relevant web pages.

Last but not least observation is regarding the adoption of a suitable mapping strategy to map Query-RIP and UIP-RIP to measure query relevancy and user interest relevancy

score respectively. A web resource, being only relevant to a query, has no place in a personalization methodology, while a web resource being only relevant to user's UIP is also irrelevant as query represents the user's current interest and UIP represents user potential interest. Thus, the degree of relevancy of a web resource for query and user must be measured to finally compute the post-relevancy score. These mappings contributed significantly towards the performance of web page ranking hierarchy obtained from the proposed model and helped to outperform all other methodologies in terms of qualitative ranking. The baseline CaiNTF, XieFilter, and OmairLatent had also benefited from these mappings to surpass their counterparts, but not good as the proposed model. Based on experimental results of various evaluation metrics, the overall performance rating of personalized web ranking under the proposed model is the greatest among all baselines.

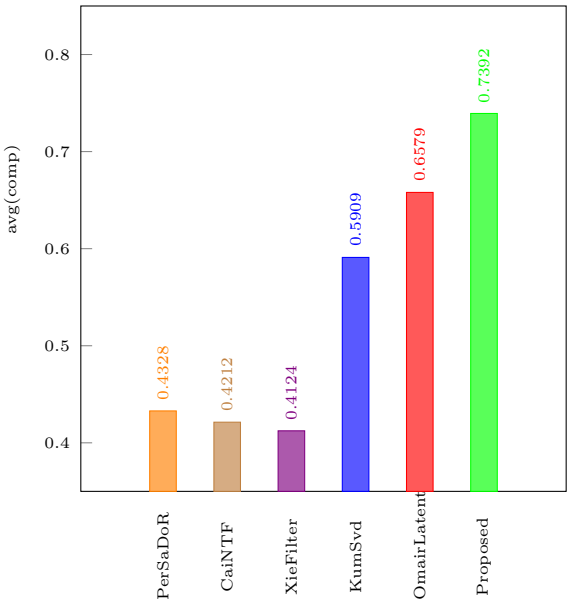


Figure 5.6: Comparative analysis of the proposed model and baseline methodologies on the basis of average completeness metric

Fig. 5.6 provides a comparative analysis between a list of web pages, relevant to a user, constructed by the proposed model and baseline methodologies respectively on the basis of *comp* metric. Basically, *comp* is a measure of user relevant web pages that belong to both the testing set and hierarchy of web pages retrieved as a response to user's query by a personalization methodology. The location of a web page inside the web page ranking hierarchy doesn't affect the value of *comp*. Fig. 5.6 clearly shows that the proposed model has outperformed all other baselines by successfully incorporating 73.92% user relevant web pages into the final hierarchy of web pages. Similarly to $P@K$, MRR, and RIL,

for the sake of simplicity, *comp* is also averaged over all the users in the dataset. The results exhibited in Fig. 5.6 are by and large in accordance with the results of baselines KumSvd and OmairLatent as shown in Figs. 5.3 to 5.5. However, the trends are not the same for baselines PerSaDoR, CaiNTF and XieFilter as *comp* value for PerSaDoR is greater than those of CaiNTF and XieFilter. On basis of this change in trend, it can be said that resource profile enrichment strategy used by PerSaDoR led to increase in percentage of relevant web pages obtained by it more significantly than CaiNTF and XieFilter. Nevertheless of it, the ranking assigned to these relevant web pages is inferior than those assigned by CaiNTF and XieFilter as is evident from the results shown in Figs. 5.3 to 5.5 and Table 5.1. Therefore, with a large percentage of the relevant web pages in the final ranking hierarchy, ranks assigned to these web pages also become important. As without an efficient web page ranking, the problem remains the same as rarely a user browses the web pages ranked at the lower positions in the hierarchy and will start the query reformulation in a hope to find the desired information. The combined effect of all the supporting modules of a web search personalization model had helped the proposed model to outperform all other methodologies in every aspect.

5.5 Summary

In this chapter, firstly a methodology has been for modeling the RIP of a web resource. Basically the responsibility of RIP is to enlist the topics or items to which the content present in a web resource is directly or indirectly related. Along with topic or item list, RIP also provide degree of affinity with which a topic or an item can illustrate the content of a web resource. Here in every profile either UIP or RIP, tags are used to represent viewpoint, topic, item etc. After modeling the RIP of a web resource, the major problem that lies in personalization model is the suitable mapping of user preferences,*i.e.*, UIP and web resource topics, *i.e.*, RIP. So as to provide a list of web resource arranged in the order preferred by a query issuing user. However, considering only the user's UIP for customization or re-ordering of web resource list is also not feasible as UIP represents the items or things which a user usually prefer to view. But the present query issued by a user represents the user's current requirements, therefore, mapping of user's query with RIP of a web resource is also necessary. In fact these two mappings,*i.e.*, UIP-RIP and

Query-RIP are two phases of a same coin which can not be separated from each other. So to take the benefits of these mappings for an user satisfactory ranking of web resources and high performing personalization model, these mappings have been adopted in the work presented in this thesis. However, there are some inbuilt deficiencies in traditional approaches of mapping, so we have proposed a novel methodology to eradicate these deficiencies and provide more real world approximation of the mapping problem. The two relevancy scores are generated by the proposed methodology of mapping. Which will further help to generate the post-relevancy score of a web resource and based on this score final personalized ranking of web resources is performed. An extensive set of experiments are performed to analyze the performance of proposed personalization model. The results of experiments have quantified by using various evaluation metrics and provides the answer to some key research questions raised in Literature Review Chapter of this thesis.

RQ3 *How to construct a RIP of a web resource in collaborative tagging system ?*

To answer this question, RIP corresponding to every web resource in the dataset have been constructed by utilizing the tags used by various users to annotate that web resource. The degree of affinity with which a tag can illustrate a web resource is computed using NTF approach. The RIP obtained by utilizing tags of every annotating user for that web resource is collective RIP, but this RIP has presence of some outliers which can mislead the ranking algorithm by misplacing a web resource in a ranking hierarchy. So to deal with these outliers in the proposed methodology of RIP modeling, an Intelligent Collaborative Filter (ICF) is designed which can automatically identify the users who's tags are resulting an outlier problem in RIP. The ICF will also identify user's who are not eligible to tag a particular web resource according to user's UIP. The filter is capable of removing these type of relations from C3TG graph so to give refined C3TG which further generate a refined RIP.

RQ4 *How can we perform query relevancy and user interest relevancy mapping for a web resource ?*

To answer this question, $Q_{map}R(\overrightarrow{Q}_i, \overrightarrow{R}_k)$ and $U_{map}R(\overrightarrow{U}_i, \overrightarrow{R}_k)$ functions have been designed under the proposed methodology for mapping operations. The $Q_{map}R$ and $U_{map}R$ will help to measure the relevancy of Query-RIP and UIP-RIP in the form of query relevancy score $S_k^{q_i}$ and user interest relevancy score $S_k^{u_i}$ respectively. In the proposed

methodology for mapping special attention has given to the issue that exist in the traditional approaches. Almost every work in literature has either ignored these mappings or just have used the keyword matching that too without considering the affinity of tags. But in today's time, when there are lots of terms to refer a thing, terms which are not matched between query-RIP or UIP-RIP cannot be neglected as these terms are strongly related in the real-world. So the proposed methodology has also considered the influence of these unmatched terms between the function parameters. Moreover, the mapping problem is modeled as a fuzzy satisfaction problem by the proposed methodology with significant attention to boundary conditions.

RQ5 *How to compute a suitable trade-off parameter value for every instance in a personalization system?*

To answer this question, $PRScore(\overrightarrow{U_i}, \overrightarrow{Q_i}, \overrightarrow{R_k})$ function has been designed in the proposed personalization model for computation of post-relevancy score of a web resource, which is then further used to perform the personalized ranking of web resources. The $PRScore$ function provides an answer to the argument that *How to trade-off the influence of query relevance score and user interest relevance score towards the post-relevance score ?*. In addition to answer this argument, the $PRScore$ function in proposed model has also raised one more argument that *Can a trade-off parameter value is independent of satire that which user has issued which query ?*. In general terms this argument mean that *Can single value of trade-off parameter is suitable in every personalized search instance ?*. The proposed model of personalization has handled this argument by calculating trade-off parameter value β as alignment of user's UIP and Query.

How can we design a personalization model for improving the user satisfaction towards the web search ?

To answer this generic research question raised in the Section 2.5 of this thesis, the task of designing a personalization model has been divided into different subtasks. Every supporting modules of personalization discussed in this thesis corresponds to a dedicated subtask. The task of predicting the user preferences along with degree of interest in each preference is handled by UIP modeling module of proposed personalization model in User Interest Profile (UIP) chapter, which provides a user's UIP as output. The task of constructing the resource illustration profile is done by RIP modeling module of proposed

model in current chapter, to give RIP as output. The mapping task of Query-RIP and UIP-RIP is taken care by query and user interest relevancy mapping modules of proposed model in current chapter. While procedure to compute final personalized ranking based on user and query relevance score is done by post-relevance score module of proposed personalization model also covered in this chapter. The combined interrelated contribution of above mentioned supporting modules in form of solutions to various issues in personalization helps to design a high performing personalization module. The efficiency of proposed personalization model in terms of user satisfaction can be analyzed and verified from results pattern in Figs. 5.3 to 5.6 and Table 5.1. The results corresponding to various evaluation metrics presents a different aspect of user satisfaction. The next chapter, concludes the research work done for this thesis along with few key findings and future scope of the research.

Chapter 6

Conclusion and Scope for Future Research

From past the few years, web search personalization in the field of information retrieval has been the most widely studied and discussed research area in the symposiums, conferences, etc. on either world wide web services or internet applications. Today, search personalization is the primary requirement of any user using the services of any commercial web search platform especially search engines as it helps the users to automatically find an information according to their preferences and needs. Basically, a personalization system is a kind of expert and intelligent system which can automatically learn about the preferences of a user in order to provide the search results as per their relevance to a user. A web search personalization model is not a single standalone entity, but an interrelated contribution of multiple supporting modules. The key modules to support personalization are UIP, RIP modeling, UIP-RIP, Query-RIP mapping, and post-relevancy score calculation. On the basis of computed post-relevancy scores, final personalized ranking of web resources is performed. The performance of a personalization model towards producing a qualitative personalized ranking of web resources highly depends on the performance of constituent supporting modules. While the composition, *i.e.*, approaches used to perform the tasks corresponding to a supporting module, is the major governing factor of module performance.

The present work starts with a discussion on justifying the need for personalization of web search in accordance with user preferences, followed by various issues that obstruct the performance of a personalization model. The requirement of a qualitative personalized ranking system by the information retrieval community has been a motivation to frame the major research question: "How can we design a personalization model for improving the user satisfaction towards the web search ?". In reponse to this question, some sub-questions have also been formulated as sub-problems in order to provide a solution to the main question. Each sub-question or sub-problem will correspond to a task of a dedicated

supporting module of personalization.

In this thesis, web search personalization is formulated as a web resource re-ranking problem, where ranking of web resources has been made according to their relevance for a user. The composition of supporting modules as discussed above, and adopted in the proposed model for personalization of web search, has made the current work distinct from the extant personalization methodologies. The dataset and evaluation metrics used to conduct the experiments and performance evaluation have also been discussed in the respective chapter. Further, the sorting algorithm, *i.e.*, CBIS used to perform sorting task in different supporting modules has also been discussed as the temporary or final data generated during the working of various algorithms in a supporting module is mostly in a partial sorted order. Therefore, following a traditional approach of sorting irrespective of probable data patterns will just contribute to high number of computation cycles and model complexity.

The task of user interest prediction along with various other issues involved in qualitative prediction of user profile has been taken care of by User Interest Profile (UIP) modeling module. Moreover, it also justifies importance of UIP for improving the performance of a personalization model. The UIP must be strong enough to cover maximum number of tags in which the user is or may be interested. Apart from it, proper weightage to tags must be assigned according to the user's level of interest in that tag. The construction of a UIP only on the basis of tags used by a user himself is not sufficient for a strong UIP as confirmed by the evaluation results. Thus, some tag enrichment strategy must be employed. To achieve this objective, instead of utilizing only society or tag relations, both society relationship network and real-world tag relationships have to be utilized. The employment of a single or no strategy at all for UIP enrichment cannot lead to a strong and efficient UIP as confirmed by the results of experimental evaluations. So, keeping all these factors in mind, the UIP has been constructed using the proposed multi-level UIP modeling methodology. UIP modeling is the backbone module of any personalization model; and any claim of performing personalization without UIP is baseless.

The fulfillment of remaining supporting modules, *i.e.*, RIP modeling, UIP-RIP, Query-RIP mapping and post-relevancy score calculation of personalization have also been performed by the proposed personalization model. Resource Illustration Profile (RIP) module

focuses on modeling of collaborative filtered RIP rather than a collective RIP as the presence of intentional or unintentional outlier tags in collective RIP can mislead the ranking mechanism of the proposed model. A novel Intelligent Collaborative Filter (ICF) has been designed to perform the task of tag filtering based on the mapping of User-Resource TOI communities. The real-world associations between User-TOI and Resource-TOI were computed using Word-vectors generated through Word2vec model. It has made the proposed methodology for RIP modeling unique in itself. After modeling RIP, the model focuses on the computation of UIP-RIP and Query-RIP mapping relevancy scores which have been further used to compute post-relevancy score of a web resource. Based on this score of a web resource, its position in ranking hierarchy is decided by the proposed personalization model.

The concluding chapter provides the key findings of work performed in this thesis and highlighted in Section 6.1. The directions for further research in the area under consideration appear in Section 6.2.

6.1 Key Findings

The present research work aims to personalize the web search of a user in order to increase his satisfaction towards web search in terms of qualitative personalized ranking of web resources. The personalized ranking means arranging the web resources in a hierarchical manner as per their relevance to a user.

This research work proposes a novel methodology for modeling a strong UIP module. Todays, a single strategy used to construct a UIP does not fully serve the purpose; and multiple strategies need to be followed. Hence, in this work, three different strategies have been employed to construct a three-level UIP, *i.e.*, FBUIP, CRUIP, and FRUIP with a dedicated set of protocols working at each level to mine the user information from selected strategies in a well-defined manner. User's own tags have been utilized to construct a FBUIP. The basis of CRUIP is real-world tag relationships, whereas FRUIP is based on a user's real society relationship network. Both tag relationships and society relationship network are the UIP enrichment strategies. Two major distinctions exist in the proposed and extant methodologies. The first one is the usage of Word2vec model in order to generate real-world word embeddings or word vectors. To measure the semantic

relatedness between two tags, a similarity measure is applied on word vectors corresponding to the tags under consideration. More the semantic relatedness between two tags, greater are their chances to be in the same cluster. The second one is the utilization of user real society network explicitly defined by the user. The weightage assigned to different members of the society varies because some of them belong to the Inner Private Circle (IPC), while others fall outside this circle.

Extensive experiments on a dataset of del.icio.us have been conducted to evaluate the performance of the proposed methodology for UIP modeling on the basis of four evaluation metrics, *i.e.*, MRR, *imp*, *completeness* and $P@k$. To study the contribution of both the UIP enrichment strategies for improving the performance of final UIP constructed by the proposed methodology, two intermediate UIPs have also been constructed. The impact of different parameters, similarity measures, and number of clusters in a tag cluster set on the performance of intermediate UIP corresponding to CRUIP strategy have also been analyzed.

Key Findings of UIP modeling: Experimental results shows that the selection of different parameters, similarity measures, and number of clusters have affected the performance of intermediate UIP which lead to further affect the performance of final UIP. The jaccard similarity measure at number of cluster $N_{clus} = 34$ is an ideal choice to obtain maximum performance as per evaluation results. The contribution of CRUIP and FRUIP strategies towards the performance of final UIP varies under different conditions. The obtained results confirm that the proposed methodology has outperformed the state of the art methodologies in terms of a strong and efficient UIP construction. In addition to this, results of MRR, *imp*, *completeness* and $P@k$ evaluation metrics also confirm the claim of this work that enrichment of a UIP is necessary; and a single strategy does not work efficiently in the construction of a strong UIP.

Although a UIP has always been the backbone of a personalization model and most extensively studied research topic for web search personalization, but other supporting modules also have a significant impact on the performance a personalization model. One such module is RIP modeling. The RIP of a web resource enlists the topics or items about which information is provided in a web resource. However, traditional RIP modeling methodologies suffer from the problem of outliers.

The methodology used to construct an RIP in the proposed personalization model utilizes the concept of collaborative filtering to deal with the outlier problem. A novel ICF filter has been designed to work on the principle of topic community modeling to generate a filtered C3TG graph, from which RIP has been modeled using NTF approach.

Key Findings of RIP modeling: Outliers present in the original C3TG graph must be filtered out as their presence can mislead the ranking mechanism to list the relevant web resource lower in the ranking hierarchy. As the experimental results confirms that methodology having no mechanism of outlier detection has performed less than the one equipped with suitable filters.

The selection of suitable mapping functions for the identification of the relationship between a Query-RIP pair and UIP-RIP pair has also contributed a lot towards the performance of the proposed model. Both these relationships, in turn, have helped to measure the query relevancy and user interest relevancy score for a web resource, *i.e.*, the extent to which a web resource is relevant to the query and user respectively. A web resource being only relevant to the query has no place in a personalization methodology. Similarly, a web resource being only relevant to a user's UIP is also irrelevant as the query represents the user's current interest, whereas UIP represents the user's potential interest. Thus, it is essential to perform these mappings in order to finally compute the post-relevancy score for a web resource. Generally, these mappings are neglected by the researchers; and a few who have done these mappings neglected the influence of terms present in a query or user's UIP, but not in RIP of a web resource inspite of being directly related to the remaining terms of RIP. The failure to take these terms into consideration can deprive a user of a relevant web resource. However, the mapping functions in the proposed model have considered these terms using the concept of semantic relatedness and modeled the mapping functions $Q_{map}R(\overrightarrow{Q}_i, \overrightarrow{R}_k)$ and $U_{map}R(\overrightarrow{U}_i, \overrightarrow{R}_k)$ as a fuzzy satisfaction problem.

Key Findings of UIP-RIP and Query-RIP mapping: The final personalized ranking of web resources is highly dependent on UIP-RIP and Query-RIP mapping relevancy scores as UIP-RIP mapping helps to list web resources according to the preferred interest of a user. Contrary to this, Query-RIP mapping identifies the list of web resources which are relevant to query issued by a user, and its impact in personalization cannot be neglected as a query represents the current interest of a user, while a UIP represents

the thing which a user prefers to view. Moreover, the unmatched terms between the corresponding vectors may be significantly related to each other in the real-world.

The trade-off parameter selection which is last but not the least reason responsible for distinctness of the proposed model has been taken care by post-relevancy score calculation module. Based on this parameter, the influence of query relevancy and user interest relevancy scores is adjusted to compute the post-relevancy score of a web resource. Almost every researcher had selected a single value between [0,1] for trade-off parameter, and was fixed for a personalized system irrespective of the query and the user. It is argued that selection of a common value for the trade-off parameter is undesirable as this approach is not as per the objectives of a personalized system. Thus, this work proposes a novel approach to calculate trade-off parameter values which vary from one user-query pair to another.

Key Findings of post-relevancy score calculation: Single value for trade-off parameter is not inline with the principles of a personalization system. Extensive experiments have been conducted to evaluate the effectiveness of the proposed model as a whole, which is based on web search personalization using the del.icio.us dataset. According to the results of various evaluation metrics, the web page ranking hierarchy obtained by the proposed model has outperformed all other baselines. It also lends support to the claim of this study that it is essential to perform UIP enrichment, RIP filtering, and mapping of UIP-RIP and Query-RIP for the construction of an efficient web search personalization model.

6.2 Scope for Future Research

As the current research work examines only the specific objectives formulated for the study, there is sufficient scope for further improvements in the field of web search personalization. Every possible improvement discussed here represents the goals of our future research work, which are as follows:

- Mostly, every approach studied in the literature including the proposed one has reflected a long-term user interest without any consideration to short-term interest of the user. For example, a user who had previously preferred spicy food may now be

showing interest in boiled vegetables. This issue is known as Information Requirement (IR) drift. The system responsible for constructing a user's UIP must also take care of dynamic drift or temporal drift in user interest. However, incorporating temporal element in user's UIP has posed many challenges like whether change in interest is permanent or seasonal.

- Similarly to UIP, this IR drift must be taken care of in the case of RIP modeling too as some web resources like restaurants or small traders generally offer different foods or products respectively on seasonal basis.
- In addition to separate vector for short-term interest of a user, UIP modeling methodology must also incorporate tag transference to mark the movement of tag representing a user interest from one level to another, if UIP is constructed in multi-level fashion *i.e.*, highly preferable, averagely preferable, occasionally preferable, etc.
- Most of the research studies on UIP modeling have assumed that all the annotations made by a user represent the user's favourite things; and utilizing each one of them to create a single UIP vector is not reasonable. As annotations made by a user not only include the user's preferable item, but also many otherthings which make a person annoyed. Therefore, a separate vector of annoying tags should be constructed as the performance of a personalized recommender system is highly affected by it.
- The plan is to extend the current work further by adopting sentiments aspect of the tags for cluster formation and topical community in modeling of UIP and RIP respectively in order to improve their performance.

References

[1] M. de Kunder, "The size of the World Wide Web (The Internet)," http://www.worldwidewebsize.com/, accessed: 2019-01-6.

[2] Y. Press, "Total number of videos on youtube," https://www.youtube.com/yt/about/press/, accessed: 2019-01-6.

[3] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer networks and ISDN systems*, vol. 30, no. 1-7, pp. 107–117, 1998.

[4] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[5] R. Lempel and S. Moran, "The stochastic approach for link-structure analysis (salsa) and the tkc effect," *Computer Networks*, vol. 33, no. 1-6, pp. 387–401, 2000.

[6] S. Goel, R. Kumar, M. Kumar, and V. Chopra, "An efficient page ranking approach based on vector norms using snorm (p) algorithm," *Information Processing & Management*, vol. 56, no. 3, pp. 1053–1066, 2019.

[7] O. Shafiq, R. Alhajj, and J. G. Rokne, "On personalizing web search using social network analysis," *Information Sciences*, vol. 314, pp. 55–76, 2015.

[8] T. Vu, "Dynamic user profiling for search personalisation," Ph.D. dissertation, The Open University, 2017.

[9] F. Liu, C. Yu, and W. Meng, "Personalized web search for improving retrieval effectiveness," *IEEE Transactions on knowledge and data engineering*, vol. 16, no. 1, pp. 28–40, 2004.

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

[11] B. Tan, X. Shen, and C. Zhai, "Mining long-term search history to improve search accuracy," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2006, pp. 718–723.

[12] K. Makvana, P. Shah, and P. Shah, "A novel approach to personalize web search through user profiling and query reformulation," in *Proceedings of the IEEE International Conference on Data Mining and Intelligent Computing (ICDMIC).* IEEE, 2014, pp. 1–10.

[13] P.-A. Chirita, C. S. Firan, and W. Nejdl, "Summarizing local context to personalize global web search," in *Proceedings of the 15th ACM international conference on Information and knowledge management.* ACM, 2006, pp. 287–296.

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

[15] S. Chawla, "A novel approach of cluster based optimal ranking of clicked URLs using genetic algorithm for effective personalized web search," *Applied Soft Computing*, vol. 46, pp. 90–103, 2016.

[16] X. Shen, B. Tan, and C. Zhai, "Context-sensitive information retrieval using implicit feedback," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2005, pp. 43–50.

[17] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub, "Laicos: An open source platform for personalized social web search," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2013, pp. 1446–1449.

[18] Y. A. Kim and G. W. Park, "Topic-driven SocialRank: Personalized search result ranking by identifying similar, credible users in a social network," *Knowledge-Based Systems*, vol. 54, pp. 230–242, 2013.

[19] M. R. Morris, J. Teevan, and K. Panovich, "What do people ask their social networks, and why?: a survey study of status message q&a behavior," in *Proceedings*

of the SIGCHI conference on Human factors in computing systems. ACM, 2010, pp. 1739–1748.

[20] P. A. Chirita, W. Nejdl, R. Paiu, and C. Kohlschütter, "Using odp metadata to personalize search," in *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2005, pp. 178–185.

[21] S. Gauch, M. Speretta, A. Chandramouli, and A. Micarelli, "User profiles for personalized information access," in *The adaptive web.* Springer, 2007, pp. 54–89.

[22] L. Yang, Q. Guo, Y. Song, S. Meng, M. Shokouhi, K. McDonald, and W. B. Croft, "Modeling user interests for zero-query ranking," in *Proceedings of European Conference on Information Retrieval.* Springer, 2016, pp. 171–184.

[23] Q. Du, H. Xie, Y. Cai, H.-f. Leung, Q. Li, H. Min, and F. L. Wang, "Folksonomy-based personalized search by hybrid user profiles in multiple levels," *Neurocomputing*, vol. 204, pp. 142–152, 2016.

[24] M. Gupta, R. Li, Z. Yin, and J. Han, "Survey on social tagging techniques," *ACM Sigkdd Explorations Newsletter*, vol. 12, no. 1, pp. 58–72, 2010.

[25] P. Heymann, G. Koutrika, and H. Garcia-Molina, "Can social bookmarking improve web search?" in *Proceedings of the 2008 International Conference on Web Search and Data Mining.* ACM, 2008, pp. 195–206.

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

[27] D. Carmel, H. Roitman, and E. Yom-Tov, "Social bookmark weighting for search and recommendation," *The VLDB journal*, vol. 19, no. 6, pp. 761–775, 2010.

[28] S. A. Yahia, M. Benedikt, L. V. Lakshmanan, and J. Stoyanovich, "Efficient network aware search in collaborative tagging sites," *Proceedings of the VLDB Endowment*, vol. 1, no. 1, pp. 710–721, 2008.

[29] J. Teevan, S. T. Dumais, and E. Horvitz, "Potential for personalization," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 17, no. 1, pp. 1–35, 2010.

[30] H. Kumar, S. Lee, and H.-G. Kim, "Exploiting social bookmarking services to build clustered user interest profile for personalized search," *Information Sciences*, vol. 281, pp. 399–417, 2014.

[31] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and A. Vakali, "Persador: Personalized social document representation for improving web search," *Information Sciences*, vol. 369, pp. 614–633, 2016.

[32] H. Xie, X. Li, T. Wang, R. Y. Lau, T.-L. Wong, L. Chen, F. L. Wang, and Q. Li, "Incorporating sentiment into tag-based user profiles and resource profiles for personalized search in folksonomy," *Information Processing & Management*, vol. 52, no. 1, pp. 61–72, 2016.

[33] H. Xie, Q. Li, X. Mao, X. Li, Y. Cai, and Y. Rao, "Community-aware user profile enrichment in folksonomy," *Neural Networks*, vol. 58, pp. 111–121, 2014.

[34] Y. Cai, Q. Li, H. Xie, and H. Min, "Exploring personalized searches using tag-based user profiles and resource profiles in folksonomy," *Neural Networks*, vol. 58, pp. 98–110, 2014.

[35] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang, "Enhancing personalized search by mining and modeling task behavior," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1411–1420.

[36] H. Wang, Y. Song, M.-W. Chang, X. He, R. W. White, and W. Chu, "Learning to extract cross-session search tasks," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1353–1364.

[37] D. Shen, J.-T. Sun, Q. Yang, and Z. Chen, "Building bridges for web query classification," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 131–138.

[38] Y. Cai, Q. Li, H. Xie, and L. Yu, "Personalized resource search by tag-based user profile and resource profile," in *Proceedings of International Conference on Web*

Information Systems Engineering. Springer, 2010, pp. 510–523.

[39] A. Hassan and R. W. White, "Personalized models of search satisfaction," in *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*. ACM, 2013, pp. 2009–2018.

[40] M. Shokouhi, R. W. White, P. Bennett, and F. Radlinski, "Fighting search engine amnesia: Reranking repeated results," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 273–282.

[41] J. Yan, W. Chu, and R. W. White, "Cohort modeling for enhanced personalized search," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 505–514.

[42] Y. Ustinovskiy, G. Gusev, and P. Serdyukov, "An optimization framework for weighting implicit relevance labels for personalized web search," in *Proceedings of the 24th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2015, pp. 1144–1154.

[43] S. Salehi, J. T. Du, and H. Ashman, "Examining personalization in academic web search," in *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 2015, pp. 103–111.

[44] P. Lofgren, S. Banerjee, and A. Goel, "Personalized pagerank estimation and search: A bidirectional approach," in *Proceedings of the 9th ACM International Conference on Web Search and Data Mining*. ACM, 2016, pp. 163–172.

[45] S. Bao, G. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su, "Optimizing web search using social annotations," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 501–510.

[46] H. Kumar and H.-G. Kim, "Using folksonomies for building user interest profile," in *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Springer, 2011, pp. 438–441.

[47] F. Liu and H. J. Lee, "Use of social network information to enhance collaborative filtering performance," *Expert systems with applications*, vol. 37, no. 7, pp. 4772–

4778, 2010.

[48] J. Hannon, M. Bennett, and B. Smyth, "Recommending twitter users to follow using content and collaborative filtering approaches," in *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 2010, pp. 199–206.

[49] X. Luo, Y. Ouyang, and Z. Xiong, "Improving neighborhood based collaborative filtering via integrated folksonomy information," *Pattern Recognition Letters*, vol. 33, no. 3, pp. 263–270, 2012.

[50] X. Yang, Y. Guo, Y. Liu, and H. Steck, "A survey of collaborative filtering based social recommender systems," *Computer Communications*, vol. 41, pp. 1–10, 2014.

[51] M. Speretta and S. Gauch, "Personalized search based on user search histories," in *Proceedings of the 2005 IEEE/WIC/ACM international conference on web intelligence*. IEEE, 2005, pp. 622–628.

[52] A. Sieg, B. Mobasher, and R. Burke, "Web search personalization with ontological user profiles," in *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*. ACM, 2007, pp. 525–534.

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

[54] P. N. Bennett, R. W. White, W. Chu, S. T. Dumais, P. Bailey, F. Borisyuk, and X. Cui, "Modeling the impact of short-and long-term behavior on search personalization," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 185–194.

[55] F. Cai, S. Wang, and M. de Rijke, "Behavior-based personalization in web search," *Journal of the Association for Information Science and Technology*, vol. 68, no. 4, pp. 855–868, 2017.

[56] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Transactions on Information Systems (TOIS)*, vol. 23, no. 2, pp. 147–168, 2005.

[57] C. Liu, R. W. White, and S. Dumais, "Understanding web browsing behaviors through Weibull analysis of dwell time," in *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2010, pp. 379–386.

[58] H.-C. Chen, T.-Y. Chen, S.-S. Tseng, K.-M. Chang, F. Chang, and M.-H. Jiang, "The hot security topics analysis for the announcements in icann website by using web crawler and association rule technology," in *Proceedings of International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Springer, 2019, pp. 562–571.

[59] S. Chawla, "Personalised web search using ACO with information scent," *International Journal of Knowledge and Web Intelligence*, vol. 4, no. 2-3, pp. 238–259, 2013.

[60] S. Chawla and P. Bedi, "Personalized web search using information scent," in *Proceedings of Innovations and Advanced Techniques in Systems, Computing Sciences and Software Engineering*. Springer, 2008, pp. 483–488.

[61] Q. Guo and E. Agichtein, "Beyond dwell time: estimating document relevance from cursor movements and other post-click searcher behavior," in *Proceedings of the 21st international conference on World Wide Web*. ACM, 2012, pp. 569–578.

[62] J. Huang, R. W. White, G. Buscher, and K. Wang, "Improving searcher models using mouse cursor activity," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 195–204.

[63] G. Buscher, L. Van Elst, and A. Dengel, "Segment-level display time as implicit feedback: a comparison to eye tracking," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 67–74.

[64] G. Buscher, A. Dengel, and L. Van Elst, "Query expansion using gaze-based feedback on the subdocument level," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 387–394.

[65] E. Vicente-López, L. M. de Campos, J. M. Fernández-Luna, and J. F. Huete, "Use of textual and conceptual profiles for personalized retrieval of political documents," *Knowledge-Based Systems*, vol. 112, pp. 127–141, 2016.

[66] Z. Dou, R. Song, J.-R. Wen, and X. Yuan, "Evaluating the effectiveness of personalized web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 8, pp. 1178–1190, 2008.

[67] V. Balakrishnan and X. Zhang, "Implicit user behaviours to improve post-retrieval document relevancy," *Computers in Human Behavior*, vol. 33, pp. 104–112, 2014.

[68] M. R. Ghorab, D. Zhou, A. O'connor, and V. Wade, "Personalised information retrieval: survey and classification," *User Modeling and User-Adapted Interaction*, vol. 23, no. 4, pp. 381–443, 2013.

[69] Y. Li and N. Zhong, "Mining ontology for automatically acquiring web user information needs," *IEEE transactions on Knowledge and Data Engineering*, vol. 18, no. 4, pp. 554–568, 2006.

[70] A. Micarelli and F. Sciarrone, "Anatomy and empirical evaluation of an adaptive web-based information filtering system," *User Modeling and User-Adapted Interaction*, vol. 14, no. 2-3, pp. 159–200, 2004.

[71] J. Budzik and K. J. Hammond, "User interactions with everyday applications as context for just-in-time information access," in *Proceedings of the 5th international conference on intelligent user interfaces*. ACM, 2000, pp. 44–51.

[72] A. Hannak, P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson, "Measuring personalization of web search," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 527–538.

[73] K. W.-T. Leung, W. Ng, and D. L. Lee, "Personalized concept-based clustering of search engine queries," *IEEE transactions on knowledge and data engineering*, vol. 20, no. 11, pp. 1505–1518, 2008.

[74] J. Teevan, M. R. Morris, and S. Bush, "Discovering and using groups to improve personalized search," in *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining*. ACM, 2009, pp. 15–24.

[75] I. Feddaoui, F. Felhi, and J. Akaichi, "Multidimensional user profile construction for web services selection: social networks case study," *Social Network Analysis and Mining*, vol. 8, no. 1, pp. 1–9, 2018.

[76] P. Vashisth, P. Khurana, P. Bedi, and S. K. Agarwal, "Capturing user preferences through interactive visualization to improve recommendations," in *Proceedings of International Conference on Application of Computing and Communication Technologies*. Springer, 2018, pp. 65–76.

[77] H.-C. Chen, Q.-H. Ruan, P.-C. Yeh, and Z.-M. Lin, "Intelligent management model based on cbr-sda approach–an example of smart life recommendation system for choosing clothes and accessories," in *Proceedings of 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS)*. IEEE, 2016, pp. 446–451.

[78] K. Bischoff, C. S. Firan, W. Nejdl, and R. Paiu, "Can all tags be used for search?" in *Proceedings of the 17th ACM conference on Information and knowledge management*. ACM, 2008, pp. 193–202.

[79] S. Maniu and B. Cautis, "Taagle: Efficient, personalized search in collaborative tagging networks," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 661–664.

[80] S. A. Golder and B. A. Huberman, "Usage patterns of collaborative tagging systems," *Journal of information science*, vol. 32, no. 2, pp. 198–208, 2006.

[81] A. Yeung, C. Man, N. Gibbins, and N. Shadbolt, "A study of user profile generation from folksonomies," in *Proceedings of the workshop on Workshop on Social Web and Knowledge Management, world wide web conference (WWW2008)*. Web & Internet Science, 2008, pp. 1–8.

[82] J. Gemmell, A. Shepitsen, B. Mobasher, and R. Burke, "Personalization in folksonomies based on tag clustering," *Intelligent techniques for web personalization & recommender systems*, vol. 12, pp. 37–48, 2008.

[83] E. Michlmayr and S. Cayzer, "Learning user profiles from tagging data and leveraging them for personal (ized) information access," in *Proceedings of the workshop on*

tagging and metadata for social information organization, 16th international world wide web conference (WWW2007). Eigenverlag, 2007, pp. 1–7.

[84] M. G. Noll and C. Meinel, "Web search personalization via social bookmarking and tagging," in *The semantic web*. Springer, 2007, pp. 367–380.

[85] D. Vallet, I. Cantador, and J. M. Jose, "Personalizing web search with folksonomy-based user and document profiles," in *Proceedings of European conference on information retrieval*. Springer, 2010, pp. 420–431.

[86] D. Zhou, S. Lawless, and V. Wade, "Improving search via personalized query expansion using social media," *Information retrieval*, vol. 15, no. 3-4, pp. 218–242, 2012.

[87] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," in *Proceedings of the 2008 ACM conference on Recommender systems*. ACM, 2008, pp. 259–266.

[88] A. Kacem, M. Boughanem, and R. Faiz, "Time-sensitive user profile for optimizing search personlization," in *Proceedings of International conference on user modeling, adaptation, and personalization*. Springer, 2014, pp. 111–121.

[89] A. Younus, C. O'Riordan, and G. Pasi, "A language modeling approach to personalized search based on users' microblog behavior," in *Proceedings of European Conference on Information Retrieval*. Springer, 2014, pp. 727–732.

[90] D. Horowitz, D. Contreras, and M. Salamó, "Eventaware: A mobile recommender system for events," *Pattern Recognition Letters*, vol. 105, pp. 121–134, 2018.

[91] Z. Xu, O. Tifrea-Marciuska, T. Lukasiewicz, M. V. Martinez, G. I. Simari, and C. Chen, "Lightweight tag-aware personalized recommendation on the social web using ontological similarity," *IEEE Access*, vol. 6, pp. 35 590–35 610, 2018.

[92] A. Hawalah and M. Fasli, "Dynamic user profiles for web personalisation," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2547–2569, 2015.

[93] X. Han, Z. Shen, C. Miao, and X. Luo, "Folksonomy-based ontological user interest profile modeling and its application in personalized search," in *Proceedings of International Conference on Active Media Technology*. Springer, 2010, pp. 34–46.

[94] J. Sang, C. Xu, and J. Liu, "User-aware image tag refinement via ternary semantic analysis," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 883–895, 2012.

[95] A. Harpale, Y. Yang, S. Gopal, D. He, and Z. Yue, "Citedata: a new multi-faceted dataset for evaluating personalized search performance," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 549–558.

[96] H. Xie, A. Liu, F. L. Wang, T.-L. Wong, X. Liu, and Y. Rao, "Revisit tag-based profiles in the folksonomy: How many tags are sufficient for profiling?" in *Proceedings of IEEE International Conference on Big Data and Smart Computing (BigComp)*. IEEE, 2017, pp. 274–277.

[97] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, "An efficient non-negative matrix-factorization-based approach to collaborative filtering for recommender systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273–1284, 2014.

[98] M. Belkin, I. Matveeva, and P. Niyogi, "Tikhonov regularization and semi-supervised learning on large graphs," in *Proceedings of 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3. IEEE, 2004, pp. iii–1000.

[99] X. Luo, M. Zhou, S. Li, Z. You, Y. Xia, and Q. Zhu, "A nonnegative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method," *IEEE transactions on neural networks and learning systems*, vol. 27, no. 3, pp. 579–592, 2016.

[100] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and G. Stumme, "Evaluating similarity measures for emergent semantics of social tagging," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 641–650.

[101] D. Zhou, X. Wu, W. Zhao, S. Lawless, and J. Liu, "Query expansion with enriched user profiles for personalized search utilizing folksonomy data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 7, pp. 1536–1548, 2017.

[102] P. Mulhem, N. O. Amer, and M. Géry, "Axiomatic term-based personalized query expansion using bookmarking system," in *Proceedings of International Conference*

on Database and Expert Systems Applications. Springer, 2016, pp. 235–243.

[103] D. Varshney, S. Kumar, and V. Gupta, "Predicting information diffusion probabilities in social networks: A bayesian networks based approach," *Knowledge-Based Systems*, vol. 133, pp. 66–76, 2017.

[104] D. Valcarce, J. Parapar, and Á. Barreiro, "Finding and analysing good neighbourhoods to improve collaborative filtering," *Knowledge-Based Systems*, vol. 159, pp. 193–202, 2018.

[105] J. Hunag, X. Yuan, N. Zhong, and Y. Yao, "Modeling tag-aware recommendations based on user preferences," *International Journal of Information Technology & Decision Making*, vol. 14, no. 05, pp. 947–970, 2015.

[106] S. Kanoje, D. Mukhopadhyay, and S. Girase, "User profiling for university recommender system using automatic information retrieval," *Procedia Computer Science*, vol. 78, pp. 5–12, 2016.

[107] D. Wu, K. Yang, T. Wang, W. Luo, H. Min, and Y. Cai, "Integrating opinion leader and user preference for recommendation," in *Proceedings of International Conference on Database Systems for Advanced Applications*. Springer, 2015, pp. 17–28.

[108] T. Heath, E. Motta, and M. Petre, "Computing word-of-mouth trust relationships in social networks from semantic web and web 2.0 data sources," in *Proceedings of the 4th European Semantic Web Conference (ESWC 2007)*, 2007, pp. 44–56.

[109] C.-L. Huang, P.-H. Yeh, C.-W. Lin, and D.-C. Wu, "Utilizing user tag-based interests in recommender systems for social resource sharing websites," *Knowledge-Based Systems*, vol. 56, pp. 86–96, 2014.

[110] Y. Zeng, X. Ren, Y. Qin, N. Zhong, Z. Huang, and Y. Wang, "Social relation based scalable semantic search refinement," in *Proceedings of the 1st Asian Workshop on Scalable Semantic Data Processing (AS2DP 2009), co-located with the 2009 Asian Semantic Web Conference (ASWC 2009)*. Citeseer, 2009, pp. 1–10.

[111] J. Zhang, Y. Yang, Q. Tian, L. Zhuo, and X. Liu, "Personalized social image recommendation method based on user-image-tag model," *IEEE Transactions on Multi-*

media, vol. 19, no. 11, pp. 2439–2449, 2017.

[112] Z. Yang, H. Wang, and S. Li, "A social network recommendation algorithm based on information aging," in *Proceedings of the 10th EAI International Conference on Mobile Multimedia Communications*. ICST (Institute for Computer Sciences and Social-Informatics, 2017, pp. 62–68.

[113] X. Luo, J. Sun, Z. Wang, S. Li, and M. Shang, "Symmetric and nonnegative latent factor models for undirected, high-dimensional, and sparse networks in industrial applications," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 6, pp. 3098–3107, 2017.

[114] M. Rawashdeh, M. F. Alhamid, J. M. Alja'am, A. Alnusair, and A. El Saddik, "Tag-based personalized recommendation in social media services," *Multimedia Tools and Applications*, vol. 75, no. 21, pp. 13 299–13 315, 2016.

[115] M. Y. H. Al-Shamri, "User profiling approaches for demographic recommender systems," *Knowledge-Based Systems*, vol. 100, pp. 175–187, 2016.

[116] Z. Saoud and S. Kechid, "Integrating social profile to improve the source selection and the result merging process in distributed information retrieval," *Information Sciences*, vol. 336, pp. 115–128, 2016.

[117] C. Boulis and M. Ostendorf, "Text classification by augmenting the bag-of-words representation with redundancy-compensated bigrams," in *Proceedings of the International Workshop in Feature Selection for Data Mining*. Citeseer, 2005, pp. 9–16.

[118] D. Sontag, K. Collins-Thompson, P. N. Bennett, R. W. White, S. Dumais, and B. Billerbeck, "Probabilistic models for personalizing web search," in *Proceedings of the 5th ACM international conference on Web search and data mining*. ACM, 2012, pp. 433–442.

[119] P. Bedi and Richa, "Parallel proactive cross domain context aware recommender system," *Journal of Intelligent & Fuzzy Systems*, vol. 34, no. 3, pp. 1521–1533, 2018.

[120] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," in *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI2013)*. Cornell University, 2013, pp. 1–10.

[121] T. Hofmann, "Probabilistic latent semantic analysis," in *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*. Morgan Kaufmann Publishers Inc., 1999, pp. 289–296.

[122] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[123] Y. Sun, J. Han, J. Gao, and Y. Yu, "Itopicmodel: Information network-integrated topic modeling," in *Proceedings of 9th IEEE International Conference on Data Mining*. IEEE, 2009, pp. 493–502.

[124] A. P. García-Plaza, V. Fresno, R. M. Unanue, and A. Zubiaga, "Using fuzzy logic to leverage html markup for web page representation," *IEEE Transactions on Fuzzy Systems*, vol. 25, no. 4, pp. 919–933, 2016.

[125] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 101–110.

[126] A. Gruber, M. Rosen-Zvi, and Y. Weiss, "Latent topic models for hypertext," in *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI2008)*. Cornell University, 2012, pp. 1–10.

[127] X. Cheng, X. Yan, Y. Lan, and J. Guo, "Btm: Topic modeling over short texts," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 12, pp. 2928–2941, 2014.

[128] M. R. Bouadjenek, H. Hacid, and M. Bouzeghoub, "Social networks and information retrieval, how are they converging? a survey, a taxonomy and an analysis of social information retrieval approaches and platforms," *Information Systems*, vol. 56, pp. 1–18, 2016.

[129] H.-R. Xie, Q. Li, and Y. Cai, "Community-aware resource profiling for personalized search in folksonomy," *Journal of computer science and technology*, vol. 27, no. 3,

pp. 599–610, 2012.

[130] R. W. White, P. N. Bennett, and S. T. Dumais, "Predicting short-term interests using activity-based search context," in *Proceedings of the 19th ACM international conference on Information and knowledge management*. ACM, 2010, pp. 1009–1018.

[131] H. K. Azad and A. Deepak, "A new approach for query expansion using wikipedia and wordnet," *Information Sciences*, vol. 492, pp. 147–163, 2019.

[132] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar, "Rank aggregation methods for the web," in *Proceedings of the 10th international conference on World Wide Web*. ACM, 2001, pp. 613–622.

[133] T. T. Vu, D. Song, A. Willis, S. N. Tran, and J. Li, "Improving search personalisation with dynamic group formation," in *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014, pp. 951–954.

[134] S. Shrivastav, S. Kumar, and K. Kumar, "Towards an ontology based framework for searching multimedia contents on the web," *Multimedia Tools and Applications*, vol. 76, no. 18, pp. 18 657–18 686, 2017.

[135] S. Gauch, J. Chaffee, and A. Pretschner, "Ontology-based personalized search and browsing," *Web Intelligence and Agent Systems: An international Journal*, vol. 1, no. 3, 4, pp. 219–234, 2003.

[136] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, 1953.

[137] G. Kumaran and V. R. Carvalho, "Reducing long queries using query quality predictors," in *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2009, pp. 564–571.

[138] E. N. Efthimiadis, "Interactive query expansion: A user-based evaluation in a relevance feedback environment," *Journal of the American Society for Information Science*, vol. 51, no. 11, pp. 989–1003, 2000.

[139] P.-A. Chirita, C. S. Firan, and W. Nejdl, "Personalized query expansion for the web," in *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2007, pp. 7–14.

[140] Z. Yin, M. Shokouhi, and N. Craswell, "Query expansion using external evidence," in *Proceedings of European Conference on Information Retrieval*. Springer, 2009, pp. 362–374.

[141] V. P. Singh, R. Srivastava, Y. Pathak, S. Tiwari, and K. Kaur, "Content-based image retrieval based on supervised learning and statistical-based moments," *Modern Physics Letters B*, p. 1950213, 2019.

[142] I. A. Adeyanju, D. Song, M. Albakour, U. Kruschwitz, A. De Roeck, and M. Fasli, "Adaptation of the concept hierarchy model with search logs for query recommendation on intranets," in *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2012, pp. 5–14.

[143] F. Cai, M. De Rijke *et al.*, "A survey of query auto completion in information retrieval," *Foundations and Trends® in Information Retrieval*, vol. 10, no. 4, pp. 273–363, 2016.

[144] C. Carpineto and G. Romano, "A survey of automatic query expansion in information retrieval," *ACM Computing Surveys (CSUR)*, vol. 44, no. 1, pp. 1–50, 2012.

[145] M. P. Kato, T. Sakai, and K. Tanaka, "When do people use query suggestion? a query suggestion log analysis," *Information retrieval*, vol. 16, no. 6, pp. 725–746, 2013.

[146] F. Cai, S. Liang, and M. De Rijke, "Time-sensitive personalized query auto-completion," in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*. ACM, 2014, pp. 1599–1608.

[147] M. Shokouhi, "Learning to personalize query auto-completion," in *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2013, pp. 103–112.

[148] C. Lioma, R. Blanco, and M.-F. Moens, "A logical inference approach to query expansion with social tags," in *Proceedings of Conference on the Theory of Infor-*

mation Retrieval. Springer, 2009, pp. 358–361.

[149] S. Jin, H. Lin, and S. Su, "Query expansion based on folksonomy tag co-occurrence analysis," in *Proceedings of IEEE International Conference on Granular Computing*. IEEE, 2009, pp. 300–305.

[150] Y. Lin, H. Lin, S. Jin, and Z. Ye, "Social annotation in query expansion: a machine learning approach," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 405–414.

[151] M. Bender, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, R. Schenkel, and G. Weikum, "Exploiting social relations for query expansion and result ranking," in *Proceedings of IEEE 24th International Conference on Data Engineering Workshop*. IEEE, 2008, pp. 501–506.

[152] R. Schenkel, T. Crecelius, M. Kacimi, S. Michel, T. Neumann, J. X. Parreira, and G. Weikum, "Efficient top-k querying over social-tagging networks," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2008, pp. 523–530.

[153] M. Bertier, R. Guerraoui, V. Leroy, and A.-M. Kermarrec, "Toward personalized query expansion," in *Proceedings of the 2nd ACM EuroSys Workshop on Social Network Systems*. ACM, 2009, pp. 7–12.

[154] C. Biancalana, A. Micarelli, and C. Squarcella, "Nereau: a social approach to query expansion," in *Proceedings of the 10th ACM workshop on Web information and data management*. ACM, 2008, pp. 95–102.

[155] M. R. Bouadjenek, H. Hacid, M. Bouzeghoub, and J. Daigremont, "Personalized social query expansion using social bookmarking systems," in *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM, 2011, pp. 1113–1114.

[156] M. F. Porter *et al.*, "An algorithm for suffix stripping." *Program*, vol. 14, no. 3, pp. 130–137, 1980.

[157] F. Wikimedia. (2017) The data dumps of wikipedia. [Online]. Available: https://meta.wikimedia.org/wiki/Data_dumps

[158] S. Goel and R. Kumar, "Brownian motus and clustered binary insertion sort methods: An efficient progress over traditional methods," *Future Generation Computer Systems*, vol. 86, pp. 266–280, 2018.

[159] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. ACM press New York, 1999.

[160] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 746–751.

[161] DMOZ, "Open directory project," https://www.http://dmoz-odp.org/, accessed: 2018-11-6.

[162] S. Tapia-Fernández, E. Rodríguez, J. Velázquez, F. Seco, and A. R. Jiménez, "Location aware web: concept, protocol and system," in *Proceedings of the IEEE International Conference on Industrial Technology (ICIT)*. IEEE, 2015, pp. 3424–3429.

[163] Y. Zhou, X. Xie, C. Wang, Y. Gong, and W.-Y. Ma, "Hybrid index structures for location-based web search," in *Proceedings of the 14th ACM international conference on Information and knowledge management*. ACM, 2005, pp. 155–162.