UTS UNIVERSITY OF TECHNOLOGY SYDNEY

Faculty of Engineering & Information Technology

Intelligent Personalized Approaches for Semantic Search and Query Expansion

by

Omar Ghaleb Alshaweesh

A Thesis Submitted for the Degree of Doctor of Philosophy

February 22, 2019

CERTIFICATE

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as part of the collaborative doctoral degree and/or fully acknowledged within the text.

I also certify that the thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

This research is supported by the Australian Government Research Training Program.

Signature of Student:

Production Note:
Signature removed prior to publication.

Date:

22/2/2019

ACKNOWLEDGEMENTS

First and foremost, my sincere thanks go to Allah, who endowed me to complete this doctorate. In particular, I am grateful to Al-Hussein bin Talal University, Ma'an, Jordan for their financial support of this project.

Most of all, I wish to thank my supervisory Dr. Farookh Khadeer Hussain and Dr. Hai Yan for giving me the opportunity to perform this work and having guided and helped me throughout the research. Their assistance and advice have made this a rewarding experience.

I am extremely grateful to my mother, father, brothers and sisters for all of the sacrifices that you've made on my behalf. Your prayers for me have sustained me thus far. I will never be able to pay back the love and affection showered upon me by my family. I especially wish to thank my wife, Layla, who has been extremely supportive of me throughout this entire process and has made countless sacrifices to help me get to this point.

Finally, I would like to give my special thanks to my great friends. Their motivation and continuous support have helped make this research happen and a more than enjoyable experience. I am really very grateful for all you have done for me.

ABSTRACT

In today’s highly advanced technological world, the Internet has taken over all aspects of human life. Many services are advertised and provided to the users through online channels. The user looks for services and obtains them through different search engines. To obtain the best results that meet the needs and requirements of the users, researchers have extensively studied methods such as different personalization methods by which to improve the performance and efficiency of the retrieval process. A key part of the personalization process is the generation of user models. The most commonly used user models are still rather simplistic, representing the user as a vector of ratings or using a set of keywords. Recently, semantic techniques have had a significant importance in the field of personalized querying and personalized web search engines. This thesis focuses on both processes of personalized web search engines, first the reformulation of queries and second ranking query results.

The importance of personalized web search lies in its ability to identify users' interests based on their personal profiles. This work contributes to personalized web search services in three aspects. These contributions can be summarized as follows:

First, it creates user profiles based on a user’s browsing behaviour, as well as the semantic knowledge of a domain ontology, aiming to improve the quality of the search results. However, it is not easy to acquire personalized web search results, hence one of the problems that is encountered in this approach is how to get a precise representation of the user interests, as well as how to use it to find search results. The second contribution builds on the first contribution. A personalized web search approach is introduced by integrating user context history into the information retrieval process. This integration process aims to provide search results that meet the user’s needs. It also aims to create contextual profiles for the user based on several basic factors: user temporal behaviour during browsing, semantic knowledge of a specific domain ontology, as well as an algorithm based on re-ranking the search results.

The previous solutions were related to the re-ranking of the returned search results to match the user’s requirements. The third contribution includes a comparison of three-term weight methods in personalized query expansion. This model has been built to incorporate both latent semantics and weighting terms. Experiments conducted in the real world to evaluate the proposed personalized web search approach; show promising results in the quality of reformulation and re-ranking processes compared to Google engine techniques.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHM

LIST OF PUBLICATIONS

Journal Papers

Elshaweesh , O., Lu,H., Alwedyan ,M ., Chebil ,W, 'Context-Aware Personalized Web Search Using Navigation History'. International Journal on Semantic Web and Information Systems (IJSWIS), Submitted.

Conference Proceeding

ElShaweesh, O., Hussain, F.K., Lu, H., Al-Hassan, M. & Kharazmi, S. 2017, 'Personalized Web Search Based on Ontological User Profile in Transportation Domain', *International Conference on Neural Information Processing*, Springer, pp. 239-48.

CHAPTER 1

INTRODUCTION

1.1 Introduction

In this chapter, an overview of information retrieval is given in Section 1.2 and a discussion on a personalized search using semantic search and an ontology solution is presented in Section 1.3. The motivation for this thesis is presented in Section 1.4. Research objectives of the thesis are presented in Section 1.5. Then, Section 1.6 presents the contributions of the thesis. The scope of this thesis is defined in Section 1.7 after which the thesis plan is presented in Section 1.8. Finally, the chapter concludes in in Section 1.9.

1.2 Overview

A web information retrieval system is an intelligent software system that provides relevant web resources that match the users' queries (Wu, Feng & Luo 2009). Google offers a powerful web information retrieval system (www.google.com). For various reasons, such as too much information on the Internet, the difficulty or impossibility of processing information, lack of ability to understanding the semantics or meaning of the user query or available information on the WWW, multiple sources containing the same information etc.. , most search systems are unable to provide results that satisfy users as many returned web sources are often irrelevant or non-important to the query (Cristina Gonzalez 2011).

One of the most important factors when conducting a successful search using a search engine is to use queries which are clear and direct and not vague (Smyth et al, 2004). Vague search queries make the results ambiguous and therefore it is difficult to obtain relevant and accurate information to suit the user's requirements. Users of search engines use keywords which differ in terms of clarity. Some search using clear and short queries and obtain the information for which they are searching. Others who do not use appropriate query terms will not retrieve accurate information; rather, they will obtain general information. To solve these problems, search systems are evolving from a content-centric design (mainly keyword based) to a user-centric design whose main challenge is to analyze the intent of the users instead of answering single queries through using personalization methods to solve such problems of information retrieval.

1.3 Personalized Search Using a Semantic Search and Ontology Solution

Personalization is the process of presenting the right information to the right user at the right moment (Khare, Tewari & Dagdee 2014). The task of personalized search engines is to provide customized results to users. In order to learn about a user, systems must collect personal information, analyze it, and store the results of the analysis commonly known as a user profile. Profile information can be collected from users in two ways: explicitly, i.e. asking for information form users such as preferences or ratings; or implicitly, i.e. observing user behaviours such as the time spent reading an online document. The main challenges of personalization are 1) a limited understanding of the user's intent in the process of modelling the user's interests and preferences (Jiang & Tan 2009) and 2) how to model the user's needs to improve the search accuracy.

User profiles are used in information retrieval processes to retrieve information which accurately meets the user's needs. This study aims to reconcile search results with the user's preferences or interests. The provision of a user profile depends on the previously browsed pages (their browsing history) and the pages the user browses in each session can be added to the profile, including whether these are positive or negative preferences.

When a user searches on a subject, he sometimes may not retrieve information that meets his request as some documents may not have the terms of the query, even though they are semantically related to the query. To solve the problem of the deteriorating performance of a query in a personalized search, a semantic search is used to implement an ontology in the information retrieval process instead of a keyword-based search (Hourali & Montazer 2011). The semantic search is characterized by its reliance on the contextual handling of indexing and retrieval terms in documents. In this way, vague documents and keywords can be matched and disambiguated appropriately. These methods, such as sophisticated ranking techniques and query expansion techniques, can also help in the development of retrieval performance.

To solve this issue this thesis proposes an innovative solution as follows. Firstly, by building an ontological user profile approach for a personalized search to use it in the re-ranking process. The aim of this approach is to build user profiles based on user browsing behaviour and the semantic knowledge of a specific domain ontology to enhance the quality of the search results. The proposed approach utilizes a re-ranking algorithm to sort the results returned by the search engine to provide a search result that best relates to the user's query. Moreover, a contextual user profile based on navigation history is modelled in order to deliver the search results that best meet the user's needs. The key contributions are the technique of building contextual user profiles based on the user's browsing behaviour; the semantic knowledge of a specific domain ontology; an algorithm for re-ranking the original search results based on the user's contextual profile; and a new technique for personal query expansion. This approach finds similar texts for a particular user from a large number of documents. The results of applying this approach reduce the query search time and also minimize the effort of the users by extending the terms of the queries.

Personalized query expansion also plays an active role in improving the performance of the retrieval process, which represents the user's interests.

1.4 Motivation

Recently, there has been a dramatic increase in both the amount of information available on the Internet and its complexity due to the growth of this network (Lee 2002). It is obvious that this increase in the amount of information increases the burden on users when accessing information for which they are specifically looking as users finds many options thus, the process takes more time and effort.

One of the salient features of personalized and re-ranking systems is their ability to adapt the results to each user depending on his/her needs. This performance of personalized search engines is especially dependent on the knowledge of the behaviour and preferences of each user (Gauch et al. 2007). By doing so, customers receive what they need, which makes them feel satisfied which leads to greater profits for service providers (Cao et al. 2007).

Learning a user's preferences and acquiring them is one of the most important steps of personalization(Tao & Li 2009). For maximum user satisfaction, the users should be provided with the search results in the simplest and least time-consuming manner. To achieve this, a user profile must be built to determine their preferences and needs (Jannach et al. 2010).

The relationship between the user profile and personalization is theoretically positive. This means that the better or accurate the user profiles reflects the user's interests, the more efficient the personalization. In practice, acquiring information about users' interests and analyzing this involves information and knowledge.

In order to personalize a user's query, firstly their needs and preferences must be clearly and accurately defined. But, when sufficient information is not available, an incorrect retrieval problem appears (Gauch et al. 2007). For example, on a Web site such as Netflix, if the user is new, it will not be possible to find any of their ratings, which makes it difficult to know the user's needs and preferences (Chen et al. 2011). Thus, it becomes difficult to provide the user with a precise profile or knowledge of his preferences with high efficiency. Currently, there are several existing approaches to solve this problem. One example of these approaches is for the user to identify their needs by maintaining a user profile preference (Jannach et al. 2010) which includes user-specific personal information such as age, location, gender, and related items such as keywords, genres, categories, and descriptions of the product. Such information can have a positive impact on the performance of a personalized system (Gantner et al. 2010). However, it is not an easy task to find a radical solution to the incorrect retrieval problem.

Hence, it is necessary to find a way to properly purify the presented information to the users. Personalized query expansion systems are one of the most important methods of filtering and retrieving information. Such techniques help the user to obtain the information needed from a range of options (Bollen, Mao & Pepe 2010).

New methods of re-ranking retrieved documents or making recommendations can be developed to obtain new sources of information to assist profiling. Problems related to query expansion, such as information scalability, sparsity, and cold-start (Liang et al. 2010; Weng et al. 2008) can be solved.

The motives underpinning this thesis can be summarized as follows:

1) The environment including attributes such as the user's location, specific time of day can have an impact on the search results (Skillen et al. 2012) . Contextual awareness can also help improve user interaction by knowing their priorities and personal preferences and automatically adapt information to their circumstances (Chaari, Laforest & Celentano 2008).

2) The browsing web search history of the users, including, documents, cached web pages, emails etc. can be used to extract new keywords in order to re-rank the query results (Matthijs & Radlinski 2011).
3) To solve the problem of ambiguity in search queries, where most query terms contain few keywords, only two or three words. Therefore, the task of query expansion and re-ranking the results is to help the user add words to the search keywords in order to generate a more efficient and relevant query.
4) To incorporate the information related to pseudo-relevance feedback obtained from the user profile containing top-ranked documents in order to expand the initial query through the development of the expansion and re-ranking framework.

1.5 Research Objectives

Based on the issues related to a personalized web search, the objectives of this thesis are as follows:

1) To build an ontological user profile based on semantic fuzzy classification and user behaviour.
2) To build a concept-based user profile effectively for the purpose of search personalization based on context history.
3) To propose a new personalized query expansion approach using latent semantic technologies coupled with term weight approaches.
4) To validate the above developed methods by building a prototype of a personalized web search.

1.6 Research Contributions

Most current research effort aims at improving the retrieval effectiveness and robustness of personalized web search. In this section, we focus on three main contributions in this thesis, including a *personalized web search based on ontology*; *a context-aware personalized web search*; and a *latent semantic model for personalized query expansion*.

Contribution 1 - Personalized Web Search Based on Ontology

We propose a personalized search approach using an ontology-based user profile, taking into account the user's implicit browsing behaviour, semantic knowledge of concepts, and synonyms of term-based vectors extracted from the WordNet API. It also contains a comparison of three approaches in personalized web search based on an ontological user profile.

Contribution 2 - Context-Aware Personalized Web Search

The second contribution of this thesis is that, we present a personalized web search approach by incorporating the user context in the information retrieval process in order to deliver the search results that most meet the user's need.

Contribution 3 - A latent semantic personalized query expansion

The third contribution of this thesis is that, we propose a users' profile-based approach for query expansion. This approach is used for personalized query expansion and it also contains a comparison between three-term weight techniques in personalized query expansion.

The new query expansion model is based on term weight global techniques class level (TF-IDF-CF) and includes dimensionality reduction. It's performance has been s evaluated by comparing its effectiveness with two advanced models using the Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting techniques.

1.7 Scope of the Work

Most current research effort in this area aims at improving retrieval effectiveness and robustness, which can be done using automatic query expansion (AQE) and then re-ranks the results. In this section, we focus on three relatively well-established topics: personalized web search based on ontology, context-aware personalized web search, and a model for personal query expansion

In this section, we consider three main themes: first, we propose a personalized web search based on ontology, which represents the interests of the user in terms of user profiles. We build a personalized search approach based on an ontology-based user profile. The user behaviour is linked to the browsing process, knowledge of semantic concepts as well as synonyms obtained from the WordNet API. Second, we propose a context-aware personalized web search, which integrates user context in the information retrieval process in order to achieve the best results that accurately meet the needs and interests of the user. Third, we propose a personal query expansion Model which is able to deal with complex information, heterogeneity, and many other problems to accurately provide the user with the information he needs.

1.8 Plan of the Thesis

This study is concerned with a very important topic in information retrieval systems, personalized web search. To achieve the objectives of this study, the thesis is divided into eight chapters. We present a summary of each of the chapters following Chapter 1 as follows:

Chapter 2: Provides a review of the existing literature in areas that relate to personalized web search; such as information retrieval, web personalization, profile representation approach, user profile acquisition, query expansion and contextual web search. The research issues are identified based on the reviewed literature documented in this chapter.

Chapter 3: Describes the problem of the personalized web search technique and its terminologies. The research question that will be addressed in this thesis is defined. Furthermore, the research objectives and the research sub-questions are introduced in light of the mentioned issues.

Chapter 4: Provides an overview of the answers to the research questions presented in Chapter 3, each of which is discussed separately. The definitions of the personalized web Search based on Ontological User Profile is presented; along with the definitions of other terms used in this thesis such as Context-Aware Personalized Web Search using Navigation History and latent semantic QE.

Chapter 5: Proposes a Personalized Web Search Based on an Ontological User Profile in the transportation domain. This chapter proposes a personalized search approach using an ontology-based user profile. The aim of the proposed model is to build user profiles based on user browsing behaviour and the semantic knowledge of a specific domain ontology to enhance the quality of the search results. The proposed approach utilizes a re-ranking algorithm to sort the results returned by the search engine to provide a search result that best relates to the user query.

Chapter 6: Proposes a Context-Aware Personalized Web Search using Navigation History. The chapter presents a personalized web search approach by incorporating user context in the information retrieval process to deliver the search results that best meet the user's need. The key contributions of this chapter are the development of a technique for building contextual user profiles based on the user's browsing behaviour; the semantic knowledge of a specific domain ontology and an algorithm to re-rank the original search results based on the user's contextual profile.

Chapter 7: Addresses an innovative approach for user-profile based query expansion model. This contribution introduces a new approach for personalized QE, and it also incorporates a comparison of three-term weight techniques in personalized query expansion., the used local term weight is TF, where the global techniques document level (TF-IDF) and class level (TF-IDF-CF). Furthermore, the latent semantic method with class-level term weighting techniques is also implemented in this chapter.

Chapter 8: Concludes the thesis with the outcomes of the proposed methods, namely the Personalized Web Search Based model and personalized query expansion approaches. An outline of the future work is presented.

1.9 Conclusion

This chapter introduced this thesis which focuses on information retrieval, web personalization, semantic search and ontology solution. To develop the personalized results ranking and query expansion methodologies, three primary tasks are proposed in this thesis: a personalized web search based on ontology, a context-aware personalized web search and a latent semantic personalized query expansion.

The motivation for carrying out this research is discussed, followed by the scope of the research carried out in this thesis. The research contributions and research objectives are also outlined and discussed. Finally, an outline of each of the subsequent thesis chapters is presented.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

As discussed in chapter one, the task of personalized search engines is to provide the customized results to users. Researchers have intensively studied the possibility of improving the performance and efficiency of the retrieval process, focusing on the field of personalized querying and personalized web search engines. This chapter discusses the existing research in three dimensions of personalized information search as follows:

A. Web personalization and how to create a user profile and how to use this profile to personalize the search results using different re-ranking methods;
B. Contextual web search and how to provide services and products to users according to their preferences; and
C. Query expansion focusing on semantic query expansion to develop an efficient personalized search.

The existing literature and evaluation that relates to these research topics is described in Section 2.2 to Section 2.8.

Before the invention of the computer, people depended on archiving and storing information by writing it on paper. But this changed with the invention of the computer, where information is easily stored and retrieved digitally when needed. Hence, in this computer era, people can store information electronically and it can be retained for long periods and shared between various users. Furthermore, information can be retrieved easily. The main issues in information retrieval is ambiguity in the keyword or the query and the massive amount of available information within the World Wide Web (WWW) (Selvaretnam & Belkhatir 2012). For example, when a user searches with the keyword ‘flash’, he retrieves many suggestions, such as flash memory, flash card, flash (cleaning solution) or others, so the user uses one query, but he finds many unrelated searches for what he is looking for. So, there is a need for personalization which helps users to understand and identify what the user is looking for by depending on their profiles that are obtained from their preferences, which reflect the personal user interests that can be used in information retrieval. This chapter presents a literature review on areas that relate to the proposed study, the sections being organized as shown in Figure 2.1.

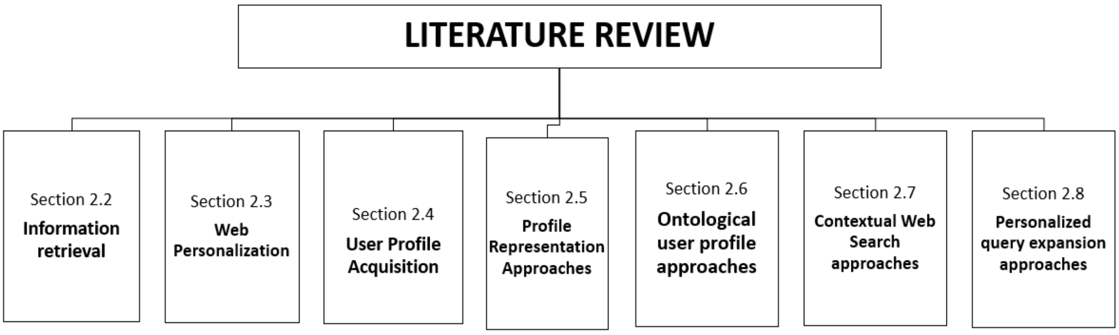

LITERATURE REVIEW
Section 2.2 Information retrieval
Section 2.3 Web Personalization
Section 2.4 User Profile Acquisition
Section 2.5 Profile Representation Approaches
Section 2.6 Ontological user profile approaches
Section 2.7 Contextual Web Search approaches
Section 2.8 Personalized query expansion approaches

Figure 2.1: The organization of the literature review in this thesis

2.2. Information Retrieval

Information retrieval can be defined as the process of finding information or documents that were previously stored and relate to the user's query history (Büttcher, Clarke & Cormack 2016). It is also the science of trying to find information within a document. This activity or process depends on using the keywords queries, but sometimes, if ambiguous queries or mismatching words are used to describe what the user want to search, this will lead to unrelated results (Xu & Croft 1996).

To help users find the information they need in the top results, an information retrieval system should analyze the query and provide the relevant web resources that match the user's needs. The following are the most common reasons for information overload (Klingberg 2009):

A. Large amount of information on the World Wide Web

B. Difficulty or impossibility of processing the retrieved information

C. Irrelevance or non-importance of most of the retrieved information

D. Multiple sources containing the same information.

Most of the time, users are unable to express precisely what they need or what they are searching for (Christopher, Prabhakar & Hinrich 2008). So, it is obvious that a query is not the best way for a user to describe his needs. To solve this problem, search systems are evolving from a content-centric design (mainly keyword based) to a user-centric design (Walther et al. 2011). Personalization methods have been developed to solve such problems of information retrieval.

2.3 Web Personalization

2.3.1 Overview

Personalizing the results of a web search has become an important area for researchers in the field of information retrieval. Despite this being the general target, researchers had their own specific goals when studying this topic. Figure 2.2 shows the summary of the literature review of web personalization.

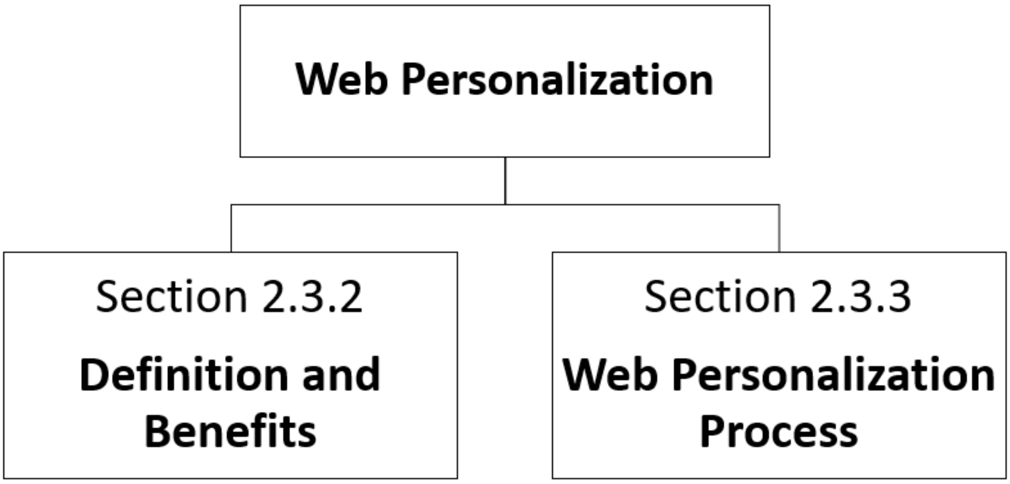


Figure 2.2: An overview of the literature review of web personalization.

2.3.2. Personalization Definition and Benefits

The word "personalize" refers to making or changing something to be suitable for a certain person and the term "personalization" is defined as "the ability to provide content and services tailored to individuals based on knowledge about their preferences and behaviour" (Liang et al. 2008). Web personalization is concerned with personalizing aspects of web resources - the content itself, links, web page structure and navigation. It is considered one of the most effective solutions to provide relevant information for which the user is searching, and this limits the problem of information overload on the Internet (Vuljanić, Rovan & Baranović 2010).

Researchers are working on developing many systems that will meet the needs of users in the field of information systems, and such systems that have been popular in the context of recent personalization studies (Lavie 2010). Therefore, personalization is considered one of the most important and effective tools for the user to mitigate information overload on the Internet (Isinkaye, Folajimi & Ojokoh 2015). Recently, personalizing a web search has received much attention from researchers because of its importance in web research activities such as the ability to reduce information overload and retrieve customized search results for the end user (Matthijs & Radlinski 2011).

Different researchers applied different personalized search techniques to improve Web personalization. In (Hong, Suh & Kim 2009), the authors present a literature review of personalized systems with different categories including the user's history, user interests, and concept and research application. Some researchers depend on a new personalization approach for web search; their approach can identify the interests of the users in the form of concepts according to their search results. Their approach uses two processes: re-ranking the search activities and profile updating (Leung, Lee & Lee 2010). In (Skillen et al. 2012), the research paid special attention to the dynamic components in the ontological models (contents, user profiles). They applied their methodology based on content-based sentiment rating and collaborative sentiment ranking. Another methodology is based on a weight computation method that builds an index of previously browsed web pages, where the weight is calculated based on the user's behaviour on that page. For example, downloading or just clicking where, based on the results of these weights they are re-ranked (Kumar, Park & Kang 2008). Another work by Skillen et al is focused on the ontological modelling of user profiles (dynamic and adaptive). A semantic web rule mark-up language and rule-based personalization were also considered (Skillen et al. 2014).

2.3.3 Web Personalization Process

The Web personalization process consists of collecting web data, processing the data and delivering the obtained outputs (Adomavicius & Tuzhilin 2005). Data collection is the main focus of this process, and this process collects data on user interactions and the providers of the required services. The importance of this phase is to obtain an information base by collecting comprehensive user information, which serves as a base for the next two phases of processing and presenting the output.

During the personalization process, two problems were presented, namely, having the correct knowledge of the user's interests and how to enhance the quality of the search in the network, where the personalized search attempt tries to overcome these problems and recover the search process (Fathy et al. 2014).

2.4 User Profile Acquisition

2.4.1 Overview

A user's needs and queries vary depending on their interests and the various topics at which they are looking. Fortunately, it is now possible to direct search using user-generated information and preferences automatically and precisely (Rekik et al. 2015). This information is not limited to the computer itself but has been extended in recent times to cellular phones as another source of information (Gupta & Chavhan 2014; Skillen et al. 2014). The main problem is that there is a huge amount of information that we do not need, but we are only looking at what matters to us, so we have to organize this information in a way that suits each user individually, depending on his interests and queries (Pelegrina, Martin-Bautista & Faber 2013). User profiles reflect personal user interests that can be used in information retrieval. The collection of user profiles depends on two techniques: the pre-processing technique (or manual technique), which is often dependent on a questionnaire, but it is impractical as it takes a long time. The second method depends on the interaction between the user and the system itself. It is an automatic learning technique which saves time and effort (Han, Chen & Li 2013). Figure 2.3 summarises the literature review on user profile acquisition.

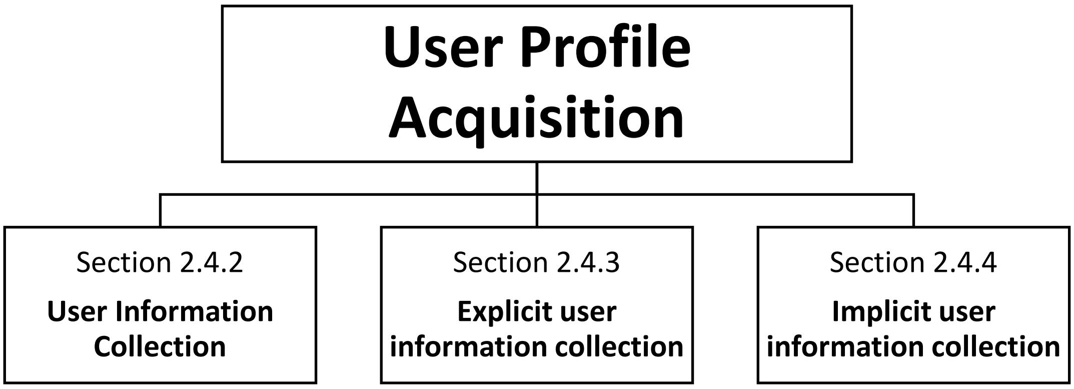

User Profile Acquisition
Section 2.4.2
User Information Collection
Section 2.4.3
Explicit user information collection
Section 2.4.4
Implicit user information collection

Figure 2.3: An overview of the literature review of the user profile acquisition.

2.4.2 User Information Collection

Collecting user information is the first stage of user profiling because of the importance of this information in determining user data, preferences, and needs (Cuomo, Nguyen & Singhal 2001). It is a key step in the process of delivering personalized recommendations or personalized search results. User preferences can be learnt by knowing users' interactions. These interactions consist of both types of information, whether explicit or implicit (also called implicit feedback and explicit feedback) (Jawaheer, Weller & Kostkova 2014). The following section details both types.

2.4.3 Explicit User Information Collection

The personal information of users is the primary source of information, and this personal information is obtained through questionnaires, registration forms, tracking users' queries or even when requesting an evaluation of an item (Gauch et al. 2007). In all of the above cases, the user enters his or her own information. For example, some websites provide the user with personal services and are asked to register (by providing their personal data) for profile work, in which case they can obtain user preferences and specify their needs in an explicit manner. Various sites support such services, such as eBay, where the user is asked to evaluate products and services and provide their views on them, and then the portal such as eBay develops the personal recommendations for the consumer based on the provided assessments and opinions.

Explicit information is divided into three different types. The first type is demographic information, such as age, occupation, gender, education and location. The second type is related to user preferences and interests such as brand, favourite products, topics of interest, as well as tastes. The third type is based on user feedback, such as feedback, reviews and comments (Gauch et al. 2007). Explicit information and ratings are very important in defining profile preferences for users in recommender systems (Jawaheer, Weller & Kostkova 2014) and there is a lot of literature body in applying and using explicit information for recommender systems.

The most important feature that characterizes explicit feedback is that they are easy to collect and are efficient with low noise. On the other hand, there are several disadvantages related to this type of information. Firstly there it involves collecting user information from the user and the user has to invest time and effort to express their preferences (Kathuria, Mittal & Chhabra 2017), which increases the burden on the user. Secondly, many users do not like sharing their personal information on websites, and sometimes they may use incorrect information, which negatively affects the performance and efficiency of the system. It is not easy to encourage users to provide their personal information in a precise way.

2.4.4 Implicit User Information Collection

Because of the difficulty in obtaining explicit information form the user or not being able to build a user profile, another source of information, called implicit information is used. This type of information depends on the behaviour or activities carried out by the user on the network, and is collected from several sources, including: click streams, logs, purchase records, browsing histories or even from visited pages. But the most important and common is browsing history (Gauch et al. 2007). (Yu, Liu & Zhao 2012) extracted contents for each page of browsing data determine user's real-time preference. One of the most important characteristics of this method of obtaining information is that it does not require the user’s time or effort to build profiles, in contrast to the previous method (Mnih 2011). Moreover, this method provides information easily and continuously, in addition to the possibility of updating it automatically. Implicit information makes it easy to obtain personalized recommendations. According to (Mnih 2011), such information helps in the development of performance by predicting recommendations, such as implicit ratings or item taxonomy. Some keywords or topics have attracted interest by using the concept-hierarchical method, which relies on browsing histories to know the user's interest (Kim & Chan 2008).

2.5 Profile Representation Approaches

2.5.1. Overview

The profiles of web users are represented using a variety of techniques. In this section, some of these approaches are presented with a brief overview of their characteristics. To provide this system with a set of weighted keywords, semantic networks, weighted concepts were used in addition to the traditional profiling approach. Here, we present an overview of the advantages and disadvantages of each of these techniques (Gauch et al. 2007). Figure 2.4 shows a summary of the literature review of the profile representation approaches.

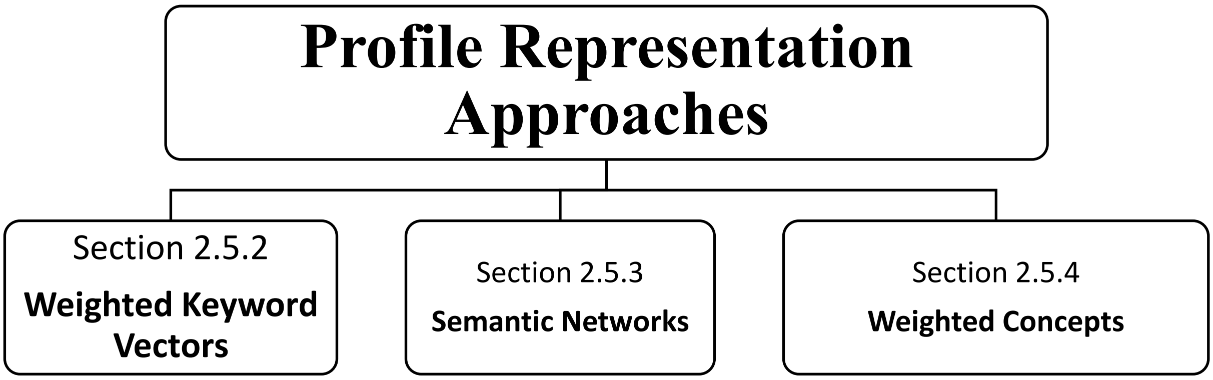


Figure 2.4: An overview of the literature review of the profile representation approaches.

2.5.2 Weighted Keyword Vectors

Weighted keyword vectors is one of the most commonly used techniques because of the ease of building keyword profiles, which are expressed in the words used in documents previously browsed by the user (Pennington, Socher & Manning 2014). This method has been used in several systems, such as the vector space model, which works effectively in comparing different profiles and items. One of the weaknesses of this method is the need for a huge volume of feedback to know the interests of the user. However, we cannot get keyword profiles for content that has not been browsed before. To a certain extent, this problem can be overcome by using modified versions of this approach.

2.5.3 Semantic Networks

Semantic networks is an effective solution to the problem of polysemy in the case of keyword profiles (Micarelli et al. 2007). Polysemy is defined as multiple meanings of a single word or term. In general, nodes make the semantic networks and express the concepts, which is the core component of semantic networks, and when two words are repeated in these concepts, arcs are produced. Then, the large data of a user must be discussed and analysed to resolve this problem.

2.5.4 Weighted Concepts

In this method, examples are treated as a training set of concepts, in which case concepts and vocabulary are linked (Trajkova & Gauch 2004). Where profiles based on weighted concepts are characterized by two basic advantages. Firstly, they do not require the user to generate a large amount of feedback, and secondly, they can bear the different changes and situations of terminology. A weakness of this system is that the requirements for operating the open directory are difficult to handle due to their complexity, despite the satisfactory results that can be accessed by using an external pre-classified database in the concept model.

2.6 Re-ranking Search Results

2.6.1 Overview

No doubt, the more accurate the response to the search results, the better the performance in the top rankings in the search engines, but practically, this is not an easy task, as achieving the results of high-precision queries requires deep and practical studies (Page et al. 1999). To resolve such barriers, with respect to the accuracy of top ranking results, many researchers proposed re-ranking models and re-ranking methods. This model rearranges documents that have been retrieved automatically, increasing the accuracy of the search results by top rankings (Takahashi & Kitagawa 2008).

Users of the information retrieval systems in general browse the returned search results with the top ranking (Wang et al. 2013). Re-ranking models are meant to help users to obtain the information they are looking for in the top retrieved documents.

In information retrieval systems, it is important to rank documents to find the degree of similarity between the queries and the documents obtained. Two methods can be used to rank search results. The first method is based on social information, which can be used by providing rankings with social relevancy. The second method is entirely based on personalized search results.

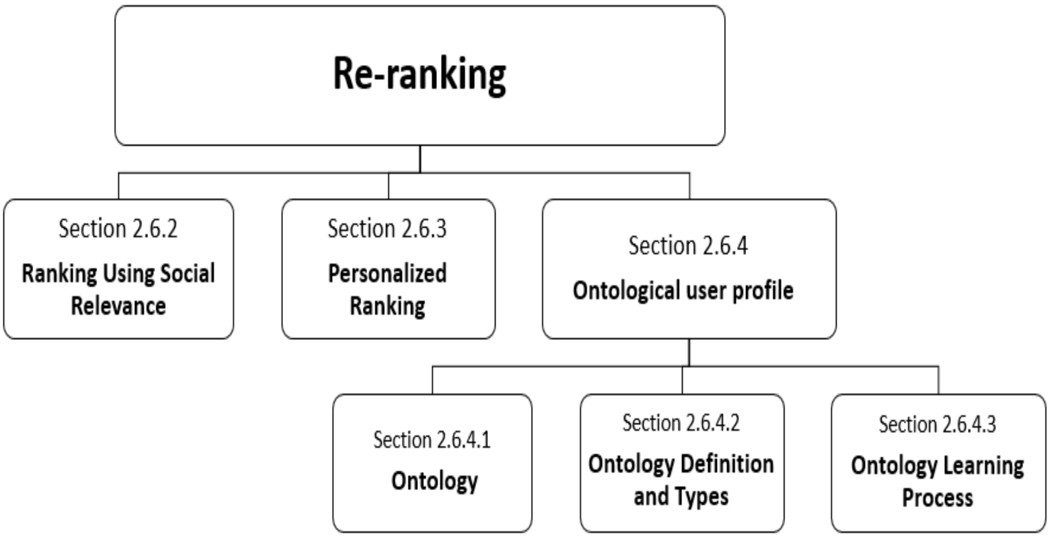


Figure 2.5: An overview of the literature review on re-ranking

2.6.2 Ranking Using Social Relevance

Many researchers studied how to develop rankings in documents using social relevancy, finding different approaches in this area. Social relevance can be defined as a collection of information obtained socially, i.e. the extent to which they are generally concerned and popular (Tseng 2015). The Social Page Rank is concerned with the importance and spread of documents to users, as well as the social annotations in a period of time. Many researchers are interested in ranking, including (Yanbe et al. 2007) , who proposed the SBRank approach, a system which finds the number of users who bookmarked a particular page on the Internet, helps to know the number of people interested in this page, so it was a strong indicator of the Web search.

2.6.3 Personalized Ranking for search results

Intuitively, each user has his or her own profile; and different interests as well. From this point, the interests and behaviours of all users cannot be generalized and provided with the same ranking of documents. Therefore, a personalized function for each user needs to be used, so that different rankings can be provided, which differs from one person to another so that customized search results can be obtained depending on the preferences of each user.

For achieve personalization Pelegrina et al ,(Pelegrina, Martin-Bautista & Faber 2013) proposed a mechanism that uses the user's search activities depending on a spreading activation technique accompanied by user profiles. An effective hybrid re-ranking search approach was used by modelling the interests of each user by means of a web search, after which the model was re-ranked, and the concepts used by the user were taken into account by the Open Directory Project [ODP] and classified as to whether it was a taxonomy document or a viewed document. These results were then rearranged according to their specific or public interests (Fathy et al. 2014). In some cases, more than one technique was combined to obtain better results. For example, the LSH-Minhash method with the SimHash dimension reduction algorithms were used to obtain the best results in terms of clustering. The TF-IDF algorithm and Slope algorithm were used to measure the importance of the keywords and their weights in relation to search and interest (He & Tang 2014).

Several studies deal with user profiles without users' feedback, with a rate of about 69%, with all concepts without reducing or shortening them. Several studies have also shown a marked improvement in the accuracy of research results in terms of ranking personalized academic information, using techniques like SPWS and SNP (Duong, Uddin & Nguyen 2013). The semantic identification of a user's interests helped develop a re-ranking process by giving relevant results (Fathy et al. 2014; Trajkova & Gauch 2004). The user interest model based on the hybrid algorithm proved its ability to resolve the "cold start" phenomenon, which refers to the inability to extract information that users need because of insufficient information (He & Tang 2014).

2.6.4 Re-ranking Based on an Ontological User Profile

2.6.4.1. Ontology

In (Gruber 2009) defined the ontology as formulated concepts in a particular area with a clarification of the relationship between them. An ontological profile is part of the concept of ontology which deals with each concept as a vector of terms with weights to find the relationship between these concepts (Solskinnsbakk & Gulla 2010). In this section, ontology definitions, ontology languages and examples and a description of the ontology development process are displayed in detail as a review of ontology. Next, measures are described for semantic similarity.

2.6.4.2 Ontology Definition and Types

In general, the term ontology is associated with philosophy, where ontology attempts to find answers to questions concerning the existence of things and their importance in the universe (Gruber 2009). However, the definitions of ontology differ based on different domains. For example, it is defined in computer science that a framework is built to systematically structure knowledge and information (Morbach, Wiesner & Marquardt 2009). It was also defined as the representation of the world in a formal way by a set of interrelated concepts. These primitives are formed for the representation of knowledge modelling, which is originally composed of certain properties, concepts, attributes, and relationships between concepts (Solskinnsbakk & Gulla 2010). Ontology concepts are shaped in a hierarchical way, consisting of basic concepts and properties in each domain (Gruber 1995).

There is a close relationship between concepts in ontologies and class in object-oriented programming where each concept consists of some properties related to features and this property holds a set and each group represents the limit of the values to be obtained. Concepts can also be classified into two types: either non-hierarchical relationships or hierarchical relationships in ontology. Ontologies can be divided into three types, depending on their degree and dependence level as follows (Conesa, Storey & Sugumaran 2010):

- Upper-level and independent ontologies: This type represents the semantics in the real world to be supportive of large applications. "Cyc project" is one of the examples that fall under this genre, and the real-world connotations are represented by a vast amount of knowledge.
- Domain ontologies: this type identifies relationships between concepts and rules of inference within the ranges in specific way, such as travel reservations, gourmet food and soccer.
- Application ontologies: This type is based on the description of concepts related in one area, for example, reasoning in a diagnosis of disease.

  Based on the classification of the previous three types of ontologies, this study is based on the third type of classifications, application ontologies, which is used in the field of e-Government services.

2.6.4.3 Ontology Learning Process

Ontological profiles remain the focus of attention for many researchers, as they use it to follow user concerns and favourites to give a clear image of the methods used to track the search areas of the users (Hawalah & Fasli 2015) For example personalized search results can be provided according to the ontology-based approach that includes preferences, personal information and interests (Rekik et al. 2015).

The main goal of a learning ontology is to integrate different disciplines to facilitate the construction process in ontologies and machine learning in particular (Cimiano et al. 2009). The process of ontology learning, is based on a semi-automatic approach, which means that it introduces some human interventions and lies in the adoption of balanced cooperative modelling (Gupta & Chavhan 2014). The framework consists of four stages in the cycle as shown in Figure 2.6.

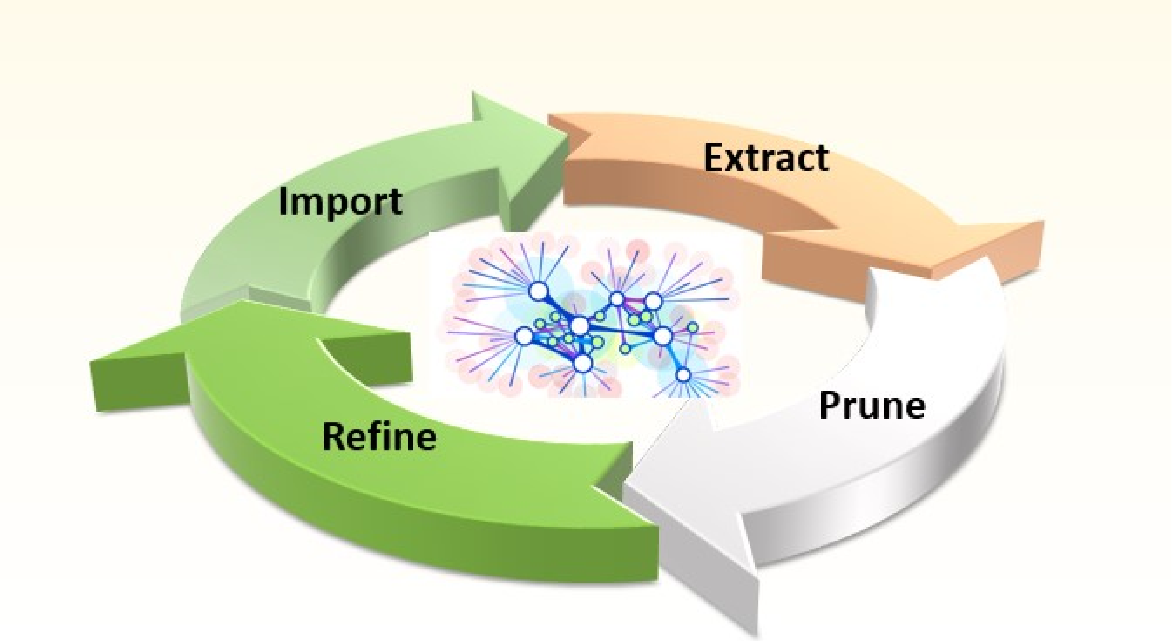


Figure 2.6: Ontology Learning Process Cycle (Maedche & Staab 2001).

The first step is to import existing ontologies and reuse them, and this process is done by integrating the existing structures or specifying the mapping rules between the ontology and its structures. For example, in the way ontological structures are used to facilitate the construction of domain-specific ontology in Cyc (Maedche & Staab 2001). The second step is to form the important parts in the ontology extraction phase with web documents. The third step is to take these ontologies and tailor them to the main objective. The fourth step is to purify the profits obtained from the domain ontology. The final step is to validate the results obtained in the ontology, based on their application criteria for the main target (Maedche & Staab 2001).

2.6.4.4 Personalized ontologies

For the purpose of personalization, the system develops a semantic profile for the users (Lekshmi & George 2016). Each concept is made up of two kinds of documents: viewed documents, which are utilized to find search interests on the user's web pages, and categorization documents based on the user's search history. In this process, more than one method is used to enhance the work of the search engine, most importantly: clustering, semantic user profiles and re-ranking (Gupta & Chavhan 2014). For researchers who relied on the ontology-based video recommender algorithm, which depends on the feedback that reflects users' needs and interests, their studies proved that this method is very effective in developing long-term user profiles (Hopfgartner & Jose 2014).

Different researchers have studied this topic from various perspectives, considering different aims and techniques. But overall, they aimed to capture the users' concerns based on their profiles and searches (Nanda, Omanwar & Deshpande 2014). Several researchers directed their thoughts to building their user profile accurately and relying on ontology concepts (Trajkova & Gauch 2004). Ontological profiles remained the focus of attention for several researchers, as they used it to follow user concerns and favourites to give a clear image of the methods used to track the search areas of the users (Hawalah & Fasli 2015).

In personalized ontologies, the activity of the user on the web was examined in terms of moving from page to page, documents, downloads or files that were opened, which were recorded on the history logs and used, thus obtaining relevant information (Calegari & Pasi 2010). Hence, user profiles reflect the user's interests, which contribute to the development and improvement of the search level (Calegari & Pasi 2010), Many techniques were utilized to embody their thoughts and achieve their goals. (Han, Chen & Li 2013) used Ontology-based User Profiles Acquisition (OUPA) to obtain ontology-based user profiles, which is based on assembling parts to build ontologies for each user automatically. This method combines two different methods to protect the user ontology: the branch and-bound search method and k-nearest-neighbour query algorithm. Several researchers used the general purpose ontology (YAGO) for the extraction process of knowledge and access to personal ontology automatically. This method identifies user’s interests by keywords that were used in the search process (Calegari & Pasi 2010).

In general, most of the technologies that used an ontology-based user framework were successful, as they helped users find the information they were looking for quickly and clearly from the vast amount of information on the Internet (Fernandez & Ponnusamy 2014; Han, Chen & Li 2013; Hawalah & Fasli 2015; Solskinnsbakk & Gulla 2010; Trajkova & Gauch 2004).

2.6.4.5 Re-ranking based on Ontological profile

Ontology was the main focus of most researchers in the personalized information retrieval field because it is of great importance when dealing with the interests of the user, as they defined the value of interest to each concept (Trajkova & Gauch 2004). Then, the results are re-ranked based on these values. For example, the Dynamic Category Interest Tree (DCIT) was created considering each category by adding a user interest score for each and was added through collaborative filtering and user history (Nanda, Omanwar & Deshpande 2014). The fuzzy classification approach was used to match the document of interest to the user and the appropriate category. Then there will be a percentage of relevancy for each category based on the matching for the keywords. In this case, according to the weighted tree, the user will get the re-ranked results based on his query. The approach proposed by (Sieg, Mobasher & Burke 2007b) has had a good effect on building an ontological user profile, since each concept has an initial value in the reference ontology repeatedly and implicitly. These values are updated by taking user behaviour into account. The concepts and pages visited are computed, and in this case, a score is assigned to a new interest based on a spreading activation technique. Furthermore, to compute the similarity between the retrieved results for the user query, the concept interest score and the similarity between the concept and the query are multiplied to calculate the final value of the retrieved documents. The re-ranked retrieved data has a great advantage in providing highly accurate results and documents.

(Daoud, Tamine-Lechani & Boughanem 2008) also used the domain ontology to model the context of the user. User concepts can be identified in each session of the search and the weights of these concepts are updated based on a linear combination formula. These weights are taken and assembled to determine the user's context. Based on the results of experiments conducted in this area, it was found that using the user context in re-ranking the results increases the retrieval precision. (Zidi et al. 2014) proposed a new framework based on ontology-based personalized retrieval. This approach integrates a domain ontology with the case base reasoning technique. In this case, user preferences are identified by their previous queries. When the user uses a new query (a new case), the system starts to format it based on the user's previous queries, and then obtains the required documents. After comparing this approach with the classic vector space model, this approach proved to have good efficiency.

A comparative analysis of the existing personalized web search approaches based on ontological user profile are summarized in Table 2.1.

Table 2.1: Comparative analysis of personalized web search based on ontological user profile

Existing research work	**Classification method**	**Do they consider browsing behaviour during weighting the ontology concepts**
(Sieg, Mobasher & Burke 2007b)	No fuzzy matching	No
(Hawalah & Fasli 2015)	No fuzzy matching	Yes
(Trajkova & Gauch 2004)	No fuzzy matching	No
(Nanda, Omanwar & Deshpande 2014)	Keyword-based fuzzy matching	No

Based on the thorough review of the existing literature on profile representation approaches, we identify the following shortcomings in this areas:

1. The existing work fails to consider the use of semantic fuzzy classification methods, which can be used to match the history of the user with the domain ontology.
2. In the process of ontology learning, not much attention has been given to taking into account the browsing behaviour of the user.

Due to these limitations in the previous studies, Chapter 5 proposes a personalized search using an ontology-based user profile. The aim of this approach is to build user profiles based on user browsing behavior. The semantic knowledge of a specific domain ontology is used to enhance the quality of the search results. The proposed approach utilizes a re-ranking algorithm to sort the results returned by the search engine to provide search results that best relates to the user query. Our algorithm evaluates the similarity between a user query, the retrieved search results and the ontological concepts. This similarity is computed by taking into account a user's explicit browsing behaviour, semantic knowledge of concepts, and synonyms of term-based vectors extracted from the WordNet API. A set of experiments using a case study from the transport service domain validates the effectiveness of the proposed approach and presents promising results.

2.7 Contextual Web Search

It is important to take into account user’s preferences during the search process so as to present more customized search results for the user. However, few researches have investigated this issue with context-aware computing in mind (Hong, Suh & Kim 2009). But recently, research has evolved in this area. The primary purpose of retrieval of a context is to know the activities of the user in order to make the connection between user preferences and the system as a whole (Limbu et al. 2009). Contextual data is especially important in knowing the behaviour and interests of users in different search engines, either by click (log) or by query auto-completion (QAC) log (Li et al. 2017). Thus, with the ever-increasing volume of information and user numbers, it is necessary to find techniques and systems that comply in the context effectively and efficiently (Melucci 2012).

Many researchers have studied this topic from their point of view and they have used different approaches and methods to prove their effectiveness and efficiency in improving the contextual web search. Several have investigated this topic using a new framework on context-aware computing to afford personalized services. Most of the previous systems relied on introducing user preferences manually, but what distinguishes this proposed system is that it connects user services and users' profiles automatically (Hong, Suh & Kim 2009). Finding accurate keywords in the search process is a difficult task. Some researchers have suggested a new approach that integrates semantic concepts and user context to get new and relevant keywords that give more accurate results and are in line with the user interests (Ahmadian, Nematbakhsh & Vahdat-Nejad 2011).

On the basis of several experiments, researchers found promising results in integrating ‘semantic concepts’ and ‘user context’ to get new and relevant keywords. Some of these experiments and models led to significant developments in the field of contextual web search when semantic relationships and context are linked. They also yielded promising results in relation to the accuracy of information retrieval (Ahmadian, Nematbakhsh & Vahdat-Nejad

2011). When combining term extraction, query expansion and document extraction, contextual search results were obtained as well as a realization of their usefulness in research in relation to educational resources on the Internet. Moreover, these techniques increase the speed of acquisition in context and accuracy (Roy et al. 2016). Search results for query suggestions yielded significant results in terms of efficiency and quality. The context-sensitive hybrid approach also played a major role in the development of retrieval quality of 2.92% taking into account ambiguous query words (Shokouhi et al. 2015).

Because it's hard for users to choose an appropriate query terms and manually use them during the search process, researchers have followed a method of automatically expanding the search based on the principle of analyzing the documents that have been retrieved before (Jiang et al. 2014). In order to establish relationships between document terms and query terms, some researchers relied on query logs in automatic query expansion, since this method approximates all kinds of queries (Mitra et al. 2014).

There have been many studies in the area of contextual personalized web search that deal with context-aware computing. Some of these studies are discussed in this section. Daoud et al. (Daoud, Tamine-Lechani & Boughanem 2008) focused on learning user interests in the long term and created a new model based on the user's contextual modelling. ODP ontology was also used in their approach with the aim of representing contexts. They also took into account the interests of the user. On the basis of a linear combination formula, concept weights for context were kept through sessions of related search. Their experiments show that the use of user context in the re-ranking of the results had a significantly positive effect on retrieval accuracy.

Other studies were concerned with the use of context information in ranking documents and their consequent problems. (Xiang et al. 2010) used real search logs in their experiments. Consequently, they created four different principles for context-aware ranking. Their experiments were subjected to implicit user click data as well as human judgments, achieving better performance in a commercial search engine when using a context-aware ranking approach.

(Mohammed, Duong & Jo 2010) examined the ontological method and found how it affected the formation of context for the user as well as the construction of high-quality personalization. By looking at the Open Directory Project (ODP), an ontological user profile was created. On the other hand, the history of the web search was built for the user implicitly, and then became an initial user profile. Experimental results indicate that their approach boosts the personalized search with re-ranking to efficiently adapt the search results based on the users' context. The recall and precision for their proposed approach compared to the non-personalized approaches for the top 10 results are: 0.2, 0.3 and 0.25, 0.35 respectively.

(Borisov et al. 2016) discussed the context bias effect between different users' reactions at different times. In addition, they proposed a context-aware time model. The principle of this model is to estimate the parameters concerning the distribution of probabilities within a time period between different user responses in a particular context. This model can also be used to calculate periods between user reactions in a given context as well as to predict those actions in a previously defined context. In addition, they pointed out that the ratio of query-result pairs (q, r) increased from 37% to 80%. According to some observations, it has been shown that the distributions of times between clicking on (r) and the second click of the same SERP are different in sessions, and obtain different results, where sometimes (r) is the first choice and sometimes not. At the same time, their observations proved that previous user actions affect the distribution of times in the overall search task. Start the search and represent the SERP first click, second start the search and represent the last SERP click, and finally, render the abandoned query, not clicking on the SERP and automatically giving the following query.

(Xie et al. 2016) dealt with three different topics for re-ranking the results based on contextual factors. First, a graph for a verbal context was constructed with a view to providing a description of search contexts in folksonomy. Second, set the identification of the elements of the context, only the important ones, and the secondary elements were ignored. Third, a variety of models of ranking in folksonomy were used order to facilitate a personalized search. To study these different topics, it was necessary to construct a verbal graph for the context by linking these elements to each other by taking semantic similarity into consideration. It is worth mentioning that the technique of iterative weight adjustment played an important role in converting the contextual graph to an unweighted one. But for essentiality properties and integrative properties, the researchers suggested using a new method that could achieve the dominating set while maintaining time complexity. The researchers conducted a number of experiments to ascertain the efficiency of their proposed technique. They tested it on MovieLens dataset, and this was done by comparing it with baselines in terms of performance in personalized search. Their studies and experiments proved its high efficiency which surpassed the other techniques in terms of performance in personalized search.

The abovementioned existing works show that using users' search context is promising for the enhancement of personalize web search. However, limiting the user's search context to the current search session would mean that the context being used is not rich enough. In our approach, we expanded from using a user's current search context to using a user's search contextual history , which was described as "a collection of past context and users' behaviour in relation to the past context" (Hong et al. 2009).

The existing personalized web search-based approaches based on contextual user profile are summarized and compared in Table 2.2.

Table 2.2. Comparative analysis of personalized web search based on contextual user profile

Existing research work	**Context factor**	**Context type**	**Area of research**
(Mohammed, Duong & Jo 2010)	Time	Current context	Personalized web search
(Daoud, Tamine-Lechani & Boughanem 2008)	Time	Current context	Personalized web search
(Leung, Lee & Lee 2010)	Location	Current context	Personalized web search
(Hawalah & Fasli 2014)	Time	Current context	Recommender system

The shortcomings of the existing literature on contextual web search are as follows:

1. The lack of knowledge of user context history is one of the main obstacles in the related work. Context history can be defined as a set of users' past behaviour and previous context. The main benefit of context history is its ability to develop a personalized web search that provides users with appropriate web navigation patterns.
2. The ontology-based user profile approach can be used for personalized web search. This approach is mainly based on the history of the context in addition to the identifying the semantic specific domain ontology. The suggested contextual profile for the user is utilised in order to re-rank the returned results of the search from the search engine, like Google. This is important to set the best search results.

To address the limitations in the previous work on contextual web search, in Chapter 6 we present a personalized web search approach by incorporating user context in the information retrieval process in order to deliver the search results that best meet the user's needs. The key contributions of our proposed approach are the method of building contextual user profiles based on the user's browsing behaviour, the semantic knowledge of a specific domain ontology and an algorithm for re-ranking the original search results based on the user's contextual profile. A prototype of a personalized web search system demonstrates the effectiveness of the proposed techniques with promising result of 35% precision improvement compared with the Google ranking scheme, which is the most commonly used search engine.

2.8 Query Expansion

2.8.1 Overview

Query expansion (QE) is the process of reformulating a given query to improve the search retrieval performance. This process includes a variety of techniques, such as finding synonyms and morphological forms of words, spelling and the grammatical correction of results (Leung et al. 2013). For an accurate personalized query expansion approach, it is necessary to pay attention to the methods of choosing query expansion terms. Hybrid query expansion methods (combining of existing query expansion methods) can be used to achieve better results (Singh & Sharan 2017).

Due to the importance of query expansion in several areas in computer science and other sciences, the researchers have studied this field from different perspectives based on diverse objectives. Some aimed to identify potential sources of plagiarism where a revised approach based on information retrieval was used (Nawab, Stevenson & Clough 2017). Some researchers were interested in studying the effect of query expansion in email search; and in their study, they relied on three different methods namely: 1) a global translation-based expansion model; 2) a personalized-based word embedding model; 3) the classical pseudo-relevance-feedback model (Kuzi et al. 2017). One of the studies included a number of techniques used in query expansion, detailing their weaknesses and comparing them against each other based on a number of criteria (Kathuria, Mittal & Chhabra 2017).

Some research implied that a simple-phrase query is better than individual query expansion. A simple-phrase query combines discriminative and popularity properties as well as robustness and shows its superiority over standard clarity scores (Thesprasith & Jaruskulchai 2016). In the field of engineering, their reliance on user's needs has had a significant impact on the development of retrieving engineering documents and has surpassed the keyword-based system and the existing query expansion system (Hahm et al. 2014).

Many QE methods and techniques have been introduced, each with specific results and characteristics that helped improve query development overall. Figure 2.7 shows a summary of the literature review on query expansion techniques.

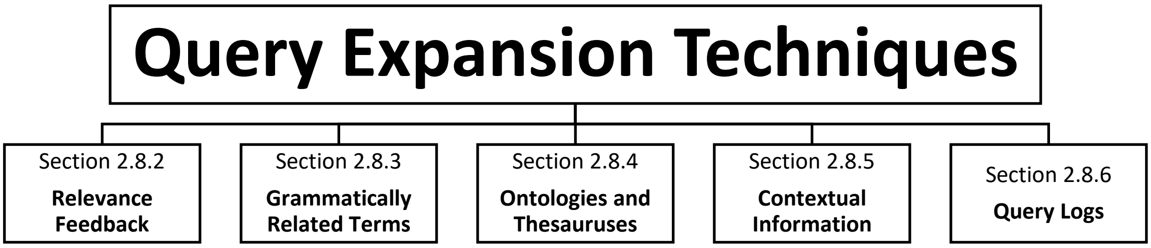

Query Expansion Techniques
Section 2.8.2 Relevance Feedback
Section 2.8.3 Grammatically Related Terms
Section 2.8.4 Ontologies and Thesauruses
Section 2.8.5 Contextual Information
Section 2.8.6 Query Logs

Figure 2.7: An overview of the literature review of query expansion

2.8.2. Query Expansion Using Relevance Feedback

This method depends on the original query for a set of high-ranking documents. This type of document contains other terms and is written in the original context of the query words (White, Ruthven & Jose 2002). As these highly ranked documents are used to build an expanded query, its importance lies in retrieving the documents that have been used to extract the answer. Sometimes, researchers rely on this approach because it gives them the freedom to choose and rank for some new search terms (Leuski 2000).

However, this is sometimes performed automatically in some systems, depending on the QE process, and is sometimes referred to as pseudo-relevance feedback, which is the system used in query expansion (Cao et al. 2008). In some other systems, statistically based term weighting can be used to define QE terms, as other heuristics can be used to perform this task. In QE-based systems, retrieved documents are used to extract key phrases from them, and similarities between the obtained key phrases and the query concept are analysed (Song et al. 2006).

Taking into account relevance feedback from the user the accuracy of query expansion can be improved significantly. Researchers have developed query expansion with poor words taking into account relevance feedback. It is worth noting that relevant documents constitute a fairly small sample; so it is possible to resort to non-relevant documents, and at the same time, sometimes relevant terms can attract non-relevant subjects. The number of terms used, the document-ranking functions, the document collection and the document queries are the basis that has an impact on relevance feedback in query expansion (Leuski 2000).

The importance of pseudo-relevancy feedback techniques helps to raise the overall efficiency, yet their results are not relevant specially when the set of relevant results is not homogenous and requires the end-user to put effort on top of the search itself. To address this problem in pseudo-relevance feedback methods, they had to be replaced in other ways to identify terms in the QE, such as the term co-occurrences measure. The co-occurrence measure quantifies the importance of words that are included in QE (Mitra, Singhal & Buckley 1998). In some documents, the importance of this method has been shown to reduce the amount of non-validated terms obtained from non-relevant documents, and QE is used to summarize the documents, which is the basis of this approach (Leuski 2000).

Thesaurus-based techniques, co-occurrence-based techniques and relevance feedback techniques were used to develop query keywords, and these algorithms were generally used to analyze the user's data (Chirita, Firan & Nejdl 2007).

2.8.3 Query Expansion Using Grammatically Related Terms

This is one of the methods used in QE to select a query based on grammatical terms which takes into account the roots of the words (stems) and synonyms of the terms. WordNet is the most commonly used application for finding synonyms. Stemming techniques are used in determining the stems or roots of words (Porter 1980). These algorithms are used to improve performance in general, such as the algorithms in (Porter 1980). Some of which showed clear improvement and others showed some degree of failure (Hull & Clyne 1996).

2.8.4 Query Expansion Using Ontologies and Thesauruses

A query can be expanded by using ontologies, where relevant query terms are specified. In general, there are several ways by which ontologies to assist with information retrieval can be built. Some approaches are based on online ontologies like WordNet and Mesh, where relevant terms are chosen for a query (Song et al. 2006). In a few domains, thesauruses are built manually, as in economic domains and environmental domains and can be built automatically depending on some related corpus, and this method automatically aggregates similar words into categories based on patterns in documents (Crouch & Yang 1992). Some researchers used a fuzzy logic-based method in a top-retrieved document to combine different weights for terms depending on the fuzzy rules. The goal of this process is to get an accurate weights to the query terms. In order to obtain a new vector, the original weights and the weights obtained are used, which is later used to retrieve the documents (Singh & Sharan 2017). Thus, fuzzy rules were also used in combination with the technique of word embedding which helps to find the semantic similarity between words (Leung, Lee & Lee 2010).

Another research study also sought to introduce a new method in query expansion, an unsupervised content-based similarity search for multimedia data (Houle et al. 2017). Because users are unaware of how to search for specific topics using the right query terms, researchers have followed a method of automatically expanding the search based on the principle of analyzing the documents that have been retrieved before (Khan, Khor & Chong 2004). Some studies based on global query semantics have made remarkable achievements in the field of recall and precision compared with other models based on word embeddings (Reyes et al, 2018). After testing the unsupervised content-based similarity approach, this method is found to be highly effective and highly efficient (Houle et al. 2017). When testing the efficiency of a query expansion in a similarity measure between email messages and queries, the results were impressive (Kuzi et al. 2017).

2.8.5 Query Expansion Using Contextual Information

Implicit and explicit data related to the user are valuable contextual information that improve the performance of the contextual search. Most of these systems rely on user behaviour, which reflects their personal interests. These behaviours are learnt and stored as part of the user profile, often based on search history. Some systems also consider relevancy feedback for contextual terms, where the system identifies and extracts them, allowing the user to choose between terms. This is called contextual relevance feedback (Conesa, Storey & Sugumaran 2010). In query expansion, domain-specific contextual information is used.

Contextual query expansion is an important area, as it improves the overall search performance. It uses a contextual search strategy, such as document extraction, in order to obtain more accurate results in the extraction of terms (Calvanese & Franconi 2018). A literature review of context-aware systems was undertaken by (Hong et al. 2009), based on the titles of the searched articles and the keyword index, accompanied with different layers including the user infrastructure layer, network layer, concept and research layer, application layer and middleware layer. Some researchers developed a system based on query suggestion by improving the performance in mining context-sensitive query reforms. The system's administrators tested their approach using a click-based scale (Moughrabi & Yamout 2016).

Some researchers aimed to prove the importance of using approximate string-matching techniques in query expansion. This is done to generate new terms in the query, In order to improve and develop web retrieval, web queries have to be expanded based on a collection of user's personal information repositories (Hahm et al. 2014; Zhou et al. 2017) depending on the short keyword queries used by the users. It was also used to develop the query expansion methods and find the relationships between the terms in a document (de Campos, Fernandez & Huete 1998).

2.8.6 Query Expansion Using Query-logs

In the past few decades the number of search engines have increased along with the number of users using the search engines. This has resulted in the accumulation in query logs which includes the search queries used by the users as well as the frequency of URL of Web pages resulting from the user's request for search results (Li et al. 2017). Query logs are seen as an indication of the user's interests and intentions based on the type of documents they are looking at, but we cannot be certain that they are exactly relevant between the Web pages and the query. When a user selects a web page by clicking it, it is a strong indication that the page he has chosen is related to the query he is using, even to a small extent. The basic idea in using query-logs in query expansion is to link documents to the same criteria that have been selected, since these documents contain terms related to query terms (Li et al. 2017). This relationship between query terms and document terms is based on query-logs, and this relationship may be used to define terms of expansion and to employ them in other queries (Xue et al. 2004). In the context of search engines, query expansion involves two main tasks: the first is to make some assessment of the inputs used by the users, and the second is to increase the investigation of the query, where both tasks aim to find more matching with more documents. To solve some of the problems related to query expansion, several techniques have been suggested including finding synonyms and different forms of the word and even autocorrecting it. Query expansion has been studied in different areas and for different purposes. In order to establish the relationships between document terms and query terms, some researchers relied on query logs in automatic query expansion, since this method approximates all kinds of queries (Cui et al. 2002). Their experiments gave promising results. Several researchers have studied this subject in order to increase the query with other elements of the same meaning, in which case the appropriate retrieval resources can be obtained with a variety of query expansion features (Shekarpour et al. 2013).Thus, some of the information in the logs of the user's query can achieve a high level of recommendations. Examples of this information are related to previous queries or documents that the user has accessed or used before. It is possible to rely on the query recommendation methods to classify queries according to their correlation with the query used (Dali, Fortuna, Tran, et al. 2012).In the QE method, proposed by (Liu et al. 2017), fuzzy rules were used, mainly aimed at enhancing the accuracy of classification in the documents ranking, in addition to a word embedding similarity calculation.

2.8.7 Evaluation of Query Expansion Techniques

Many studies have proposed various query expansion (QE) methods. The QE principle is based on several stages‹starting from computing the weight of the terms for the documents, then calculating the probabilistic relationships between the terms of the documents and the query, and finally, compiling the cohesion weight of the document terms by combining the entirely probabilistic terms for the query (Dali, Fortuna, Tran, et al. 2012). In relation to this issue, the QE method involves four techniques (Kathuria, Mittal & Chhabra 2017). The concept network-based query expansion is one of these techniques, which is based on conceptual semantic techniques (Hoeber, Yang & Yao 2005). Term weighting is another technique, and the addition of pseudo-relevancy feedback is a very important factor in raising the retrieval rate by about 7% (Hahm et al. 2014). The third technique is word sense disambiguation (WSD) (Hahm et al. 2014) and the fourth technique is the fuzzy logic-based approach (Singh & Sharan 2017). Several studies have used the semantic and linguistic features and measured their efficiency. These features were studied separately to determine which of these characteristics are best for query expansion. These studies were based on specific steps, namely: i) finding relevant words from the original keyword, taking into account semantic and linguistic characteristics; ii) using a weighting scheme to calculate the weight of the semantic features and linguistic features to calculate the efficiency of each feature and eliminate the ineffective ones; and (iii) identifying and calculating a relevance score for the relevant words, the importance of this step lying in creating a balance between recall and precision (Shekarpour et al. 2013).The QE method with the word embedding technique was used to check the similarity between the terms of the query in the ranking pipeline. The classification technique was used to boost the dataset's rank that matches the constraints of the query (Teodoro et al. 2017).

The results of the research in the area of query expansion show that the novel QE technique greatly surpasses the other traditional methods, such as BM25 in terms of document ranking performance (Nalisnick et al. 2016). Also, query expansion using the embedding method excels the traditional lexical methods, such as query expansion which is based on WordNet and QE-WordNet (Liu et al. 2017). Moreover, the results indicated that query expansion using fuzzy rules obtained better performance than the word embedding models that did not use fuzzy rules (Liu et al. 2017). It is clear that the outcome of a query which considers terms is more significant than the outcome of aa query which considers the whole document (Teodoro et al. 2017). The increasing use of a recommended query results in the increasing the value of precision (Dali, Fortuna, Tran, et al. 2012). It was also shown that linguistic and semantic features almost have the same effectiveness, however merging these boosts precision and recall (Shekarpour et al. 2013).

Some QE methods achieved very satisfactory results in terms of accuracy in querying and obtaining more relevant search results (Al-Khateeb, Al-Kubaisi & Al-Janabi 2017).

According to the researchers who used the ontology terms, their proposed system increased recall but was not good at precision; whereas better results were obtained when using a combination of query keywords and ontology terms (Azizan, Bakar & Noah 2016). The results showed that the auto-suggestion of the queries can considerably increase the performance of query completion and provide relevant and helpful suggestions (Moughrabi & Yamout 2016). (Rawashdeh et al. 2013) proved that it is worthwhile to build a model to predict user preferences based on their interests, but that like the idea is having a gap caused by the size of the data due to the large number of users, and to solve this gap they thought that reduced dimension could play a role in overcoming this challenge, and proceeding from this hypothesis the idea of the annotated project in.

Based on the review of the existing literature on query expansion, we identify the following shortcomings in this area:

- The traditional techniques for query expansion do not include the user's information, such as user's alternatives and choices when having many documents to extract the terms of expansion.
- Users are often unsure about the nature of the content they need or they find it difficult to describe the nature of the information they need in a concise and clear way.
- One of the reasons for the lack of effectiveness of information retrieval systems is vocabulary mismatch, for example, due to polysemy, where one word has different meanings, such as the word *orange*, which could be a fruit or a color and synonyms where one word has a similar meaning to another word , for instance, goal and objective.

To address the above mentioned limitations, Chapter 7 proposes a new technique for personal query expansion. In our proposed approach construction of the query occurs without any interventions from the user. This approach depends on finding similarities in text for a particular user from a large number of documents. As a result, applying this approach, the query search time and the user's effort is reduced by extending the terms of queries. Omar, we need critical tabular evaluation and comparison for each of the above three sections (Section 2.6, Section 2.7 and Section 2.8) to arrive at the relevant shortcomings.

2.9 Summary and Conclusion

This chapter deals with many topics raised in the literature review and discusses the various methods used by current information retrieval models. The literature discussed the best way to collect useful information to benefit the users which depends on the development of methods to gather information on the web.

This chapter also focused on several issues, most notably information retrieval, personalization, query expansion, query re-ranking and importance and finally user profiles. The related work referred to the importance of concept-based techniques, which are based on background knowledge and local instances that helps in providing valuable and relevant information. But with the great advances that have been made in this area, user profiles need to be developed more in terms of acquisition or representation. The literature review also discussed the necessary tasks and important processes in the data mining process. Because of the increasing importance of ontologies in general, especially in the areas of data mining and IR, we discussed several topics such as ontology-based techniques, semantic relationships and ontology learning. The literature review noted that personalization techniques in general have greatly assisted in searches, especially when a clear and unambiguous query was adopted. These clear queries help to easily access the required information, and the use of different methods of personalization raises the level of accuracy and recall in the query.

Therefore, the first major shortcoming in the existing literature is the lack of considering the semantic fuzzy classification methods, which match the history of the user together with the domain ontology. The second major shortcoming is the lack of knowledge of user context history. The third major shortcoming is not including user's information, such as user's alternatives and choices when having many documents from which to extract the query terms in addition to traditional information retrieval problems such as the difficulty in describing what the users' need to find in a few words and vocabulary mismatching.

CHAPTER 3

PROBLEM DEFINITION

3.1 Introduction

In this chapter, the main problem related to personalized web search that is addressed in this thesis is defined. The terminology which relates to the problem are detailed in Section 3.2, and the details of the problem are presented in Section 3.3. In this thesis, the research problem is divided into four research questions which are described in Section 3.4. The research aim and research objectives are presented in Section 3.5 and Section 3.6, respectively. In Section 3.7, the overview solution for solving the defined problem is proposed. Finally, the chapter concludes in Section 3.8.

In an information retrieval system, in general there is a gap between the user's need for information and the way in which the user query for this information. Often, users are unable to express their search query in a precise and concise way. So, in this case, a search query is not the best way for the user to describe his or her needs. To solve this problem, search systems are evolving from a content-centric design (mainly keyword based) to a user-centric design. User-Centered Design (UCD) is a process to design search engines around the people who will use them through the search personalization Personalization methods have been developed to solve such problems of information retrieval (Bai et al. 2017).

Web search personalization is a common technique that incorporates information retrieval systems to address the issue of the accuracy of the retrieved results to individuals (Lekshmi & George 2016). It is a promising solution to retrieving results based on individual needs. Personalization is achieved by modeling each individual by managing, building and representing users' preferences in the form of user profiles. The main challenge of personalization is how to model a user's needs to improve search accuracy. Numerous approaches have been developed in the web search domain to model user's needs by collecting user information explicitly through direct interaction between systems and users or implicitly based on browsing documents or past queries for personalization (Akhlaghian, Arzanian & Moradi 2010; Baazaoui-Zghal & Ghezala 2014; Daoud, Tamine-Lechani & Boughanem 2008; Ferreira-Satler et al. 2012; Nanda, Omanwar & Deshpande 2014).

The main challenges of modeling the user's interests and preferences is the limited understanding of the user's intent (Jiang & Tan 2009). The contents of the web are described in natural language which is often ambiguous due to the different meanings of words. One of the uses of the Semantic Web is to implement an ontology in the information retrieval process instead of a keyword-based search to overcome these limitations (Hourali & Montazer 2011). Therefore, many scholars applied ontology to the information retrieval system in two different ways: firstly, by building an ontological user profile which represents the user preferences and is subsequently used for the re-ranking of the query results; and secondly, for reformulating a user's query which results in increasing the probability of retrieving relevant documents. Ontological approaches are proven techniques to represent a user's interests and to generate more personalized results instead of traditional keyword-based approaches.

Ontology was the primary focus of the researches in carrying out personalized information retrieval. This is due to the applicability of ontologies in dealing and modeling with the interests of the user, as they defined the value of interest to each concept, So, in this thesis, generally we paid attention to building an ontological user profile approach for an information retrieval system to use it in the re-ranking process. In our work, we focus on modelling the contextual user profile of a user based on navigation history to deliver the search results that best meet the user's needs. Finally, we propose a new technique for personal query expansion, that takes into account the user's profile and derives relevant terms for QE. The results of applying this approach reduce the query search time and also reduce the effort of the users by extending the query terms.

3.2 Key Concepts

In this section we explain or define the key concepts that we use in this thesis.

3.2.1 Web Search Engines

A Web Search Engine is a service that enables Internet users to search for content via the World Wide Web (WWW), where the user enters his own keywords or key phrases into a search engine, then receives a Web content results list in the form of audios, images, videos, websites, etc. The content returned list obtained via a search engine is called a search engine results page (Schwartz 1998).

3.2.2 Personalized Web Search

A personalized search is defined as a process which is used to classify or customize the results for each user based on the context of the user, the search history and their preferences. These customised and relevant results are obtained by using filters (Esbitan & Barhoom 2012).

3.2.3 Search Engine Evaluation

To develop highly efficient and advanced search engines, an evaluation of these engines must be undertaken. This is to ensure the performance of search engines across different applications. An evaluation of these engines mainly depends on user interaction that can be represented by five evaluation criteria (Hofmann, Li & Radlinski 2016): relevance (effectiveness), efficiency, utility, user satisfaction, and connectivity, where two basic factors are used the efficiency of the search engine and the effectiveness. Effectiveness is determined by the ability of a search engine to ensure the user received the required search results, while efficiency is the measure of the search process in terms of speed in obtaining the correct results (Esbitan & Barhoom 2012).

3.2.4 Information Retrieval

Information retrieval (IR) can be defined as a science that meets the search requirements of the user by providing him with the information he needs by storing and organizing this information (Berger & Lafferty 1999).

3.2.5 Ontologies

In information technology sciences, an ontology can be defined as representational primitives set to model a domain of knowledge. Typically, the representational primitives are divided into classes (sets), attributes (properties), and relationships that reflect the relations among class members. The representational primitives' definitions involve information about their meaning and also the constraints on their logically consistent application. In terms of database systems, an ontology is considered as an abstraction level of data models, similar to relational and hierarchical models but destined for modelling knowledge concerning individuals (their attributes and their relationships to other individuals). Ontologies are mainly specified in the languages allowing abstraction regardless of strategies of implementation and data structures; in practice, ontology languages are much closer in expressive power to first-order logic than languages that are used to model databases. Therefore, ontologies can be considered as the semantic level, whilst database schema are considered as models of data at the physical or logical level. Because of their independence from lower level data models, ontologies are can be employed to integrate heterogeneous databases and specify interfaces to independent knowledge-based services (Gruber 2009).

3.2.6 Query Reformulation

Query reformulation is the process of converting a query from one form to another, i.e. from an initial query (q) to (q0). This conversion process takes two forms, either query expansion or query refinement. There is a clear difference between these two forms as follows (Bouadjenek, Sanner & Ferraro 2015) :

1. In query expansion, the objective is to remove the ambiguity of the initial query by adding new information to the query in its initial form;

2. In query refinement, transforming a query into a new query that reflect the user information need.

3.2.7 Query expansion

Query expansion is one of the common methods used in the user query reformulation process to reduce the number of irrelevant pages obtained from the information retrieval techniques (Zhou et al. 2017). Moreover, automatic query expansion is one of the essential fields of information retrieval systems on which researchers have been working to build new technologies and develop new systems (Roy et al. 2016).

3.2.8 Personalized Query Expansion

Information retrieval systems are limited by many factors reflecting the difficulty to satisfy user requirements expressed by short queries. Moreover, the user may express the conceptual content of the required information with query terms that do not match the terms appearing in the relevant documents. This vocabulary problem is more severe when the user queries are short or when the data sources to be searched are large, as in the case of Web-based retrieval. Also not all users are looking for the same items of interests and do not have the same preferences. Approach to alleviate these problems personalized query expansion methods are used. Personalized Query expansion aims to improve the user’s search results by adding new query terms to an existing query to provide each user with their required information according to their own preferences and to improve retrieval performance (Azad & Deepak 2017).

3.2.9 WordNet

WordNet is a lexical ontology which aims at highlighting the importance of the semantics of the lexicon with the relationships between various lexical concepts (Chaves 2001).

3.2.10 Semantics

Semantics refers to the meanings of words and sentences by associating them with each other. Using semantics overcomes the problem of polysemy in keyword profiles. Polysemy means multiple meanings for the same word or phrase (Cruse 2011).

3.2.11 Service Metadata

Metadata means data about data, which provides the user with additional information about the service, such as the service provider address, name, contact details or description (Chirita et al. 2005).

3.2.12 Fuzzy Logic

Fuzzy logic is a type of many-valued logic that deals with reasoning that is approximate instead of fixed and exact. Contrary to traditional logic theory, in which binary sets contain only two-valued logic (true or false), variables of fuzzy logic can have a truth value ranging between 0 and 1 degree. Fuzzy logic has been extended to be able to handle the concept of 'partial truth', where the truth value can range between completely true and completely false.

3.2.13. User context

The user context can regarded as the collection of all the attributes that capture user's intentions, perceptions or that of its surroundings. These factors cover various aspects such as physical, social, personal, professional, technical, task etc (Mylonas et al. 2008).

3.2.14. Context search

Context search is defined as the process of gathering information about a user's environment of interest and taking that into account during the search process (El Ghali & El Qadi 2017).

3.2.15. Category hierarchy

Each category in the user profile contains a query terms set, each of which has a weight that is calculated to represent the user's level of interest in a particular category (Liu, Yu & Meng 2004). Thereby, the user profile can be structured as a category hierarchy of Web pages.

3.3 Problem Overview and Problem Definition

With the increasing number of users and the information available on the Internet, there are new problems relating to searching on the Web. A personalized web search can be used to address this problem by taking into account the user's interests and preferences during the search and retrieval process. User profiles are used to inform the retrieval process where the user's search history is considered. There are numerous issues that must be considered in building a suitable model taking accurately models the preferences and interests of the user, including the vagueness of the intent of the user (Jiang & Tan 2009) and the vagueness of the different word meanings.

One of the most important problems relates to the search query as there are many query terms that can involve more than one meaning such as mouse, apple, python etc., thus different search results will be retrieved for the user, some related and others that are not related to their needs (Budzik & Hammond 2000). For example, when a user uses the search query “Orange”, they will receive more than one type of document. The first type of information retrieved relates to the colour orange, the second type relates to the orange fruit and the third type relates to the Orange cellular network. This is the main reason for retrieving documents that are unrelated to the search because of ambiguity in the query words (Mielke et al. 2015). To address this problem, more than one word can be used in the query. For example, to solve the problem of ambiguity in the previous search query "Orange", we might add another word that explains the subject the user is looking at, such as "Orange color" or "Orange fruit" or “Orange telecommunications Company”. In this case, the search results will be more precise than the previous scenario where general terms are used. Semantic technologies play an active role in this area, helping the user to obtain relevant information by linking the meanings of words or sentences within a text or a document using the user’s interest (Allan et al. 2003). These technologies accurately provide the user with the services the user needs.

Users can dispense with the automatic methods of linking categories such as automatic query-topic detection (Stamou & Ntoulas 2009) and relying on traditional manual methods. For example, before a user submits a query, he can browse categories hierarchically and then choose the desired category. When using these categories in a query, the user obtains more effective and related results. But the disadvantage of this method is that the user will get a very large hierarchy category, so it is sometimes difficult for the unassuming user to choose the required paths and select the correct categories. This also involves extra time to reach the desired results before submitting his queries. However, these categories do not fully correspond to the users' needs and intentions in their search (Rose & Levinson 2004), thus, the approach of personalized query expansion can be used to meet the needs of all users regardless of their different personal needs.

In Section 2.8, in which we reviewed personalized query expansion methods, we introduced the importance of ontology in some methods, which is used in the expansion of terms used in the query. The ontology is a modelling methodology that is used to represent information semantically. In this way, the query is modified by adding the new terms to improve the probability of related retrieved knowledge.

There are two types of ontologies, a general ontology, such as WordNet, and a domain-specific ontology (introduced in section 2.6.4). A general ontology is used in the query expansion process of information retrieval systems. In this case, synonyms are added to the original query terms. In the case of a domain-specific ontology, the querying process is improved if there are domain specific terms in the relevant document.

Gaining access to the information required to personalize a query is a challenge. The user content can be expanded to include factors that describe the user’s interests. It is difficult to create a user context-aware information retrieval system that takes into account all relevant contextual factors during information retrieval and search. To overcome these obstacles in the field of information retrieval, this thesis identifies several key pressing issues and different methods to solve them, which are summed up in the following three strategies:

a) In Building user profiles, which is a strategy that relies on semantic knowledge and user behaviour in the process of browsing according to the specific domain ontology;

b) proposing a new approach to personalize the web search which is a process that emphasizes the role of the user context in the process of information retrieval so as to meet the needs of the user based on his/her interests; and

c) suggesting a different approach for personalized QE considering the cooperation between latent semantic net and the weight of global term.

The literature review, problem definition, the questions of the research, the aim and the objectives of the research are discussed in section 3.4, section 3.5 and section 3.6, respectively.

3.4 Research Questions

Based on the gaps identified in Chapter 2 and further outlined in Section 3.3, we define our primary research question as below:

How can intelligent methods be developed to provide personalized web search results for users based on their profile?

This question involves four sub-questions, which are listed as follows:

Research Sub-question 1:

How can user's search be personalized taking into account the user's search history?

Research Sub-question 2:

How can a context-aware personalized web search be undertaken taking into account the user' search history?

Research Sub-question 3:

How can personalized query expansion be undertaken?

Research Sub-question 4:

How can we validate and compare the developed approaches (for research sub-question 1 to research sub-question 3) in terms of accuracy?

3.5 Aim of the research

To develop an intelligent and personalized approach for a semantic search and query expansion.

3.6 Research Objectives

These objectives of this thesis are as follows:

Research Objective 1:

To build an ontological user profile based on semantic fuzzy classification and user behaviour for results re-ranking to generate personalized search results

Research Objective 2:

To build a context-aware user profile effectively based on context history for results re-ranking to generate personalized search results

Research Objective 3:

To develop a user profile as a weighted keyword vector using latent semantic technologies coupled with term weight approaches for personalized query expansion.

Research Objective 4:

To validate the aforementioned methods by building a prototype of a personalized web search.

3.7 Research Approach to Problem Solving:

Taking into account the aforementioned objectives, the primary objective of this thesis is to develop intelligent and accurate mechanisms for a personalized web search. In the next section, various scientific research methodologies are outlined and then details of our research methodology is provided.

3.7.1 Existing Research Methods

There are two approaches to research into information systems: the social science approach where the survey and interview are the basic components of knowledge based on a systematic plan (Bryman & Bell 2015); and the science and engineering-based approach in which empirical or measurable information is the source of knowledge (Peffers et al. 2007).

The science and engineering-based research approach solves problems using different types of technologies, devices, and methodologies.(Galliers 1992) point out the three hierarchical levels: the perceptual level which helps to design and implement new techniques, the conceptual level that aims to analyze and create new concepts or ideas; and the practical level. These levels test and validates the proposed approaches using both laboratory tests and real-world tests (Galliers 1992).

The social science-based research approach aims at explaining social phenomena and researching them to have a better understanding of social evidence. This approach involves the use of two techniques, quantitative, which uses measurable evidence and can be implemented with the help of statistical models to analyse the initial data in its raw form; and subjective evidence from interview techniques and personal observation for clear and deep knowledge.

3.7.2 Choice of Science and Engineering-based Research Method

The science and engineering research-based approach is employed for developing a new methodology for personalized web search. Figure 3.1 shows the working of this method.

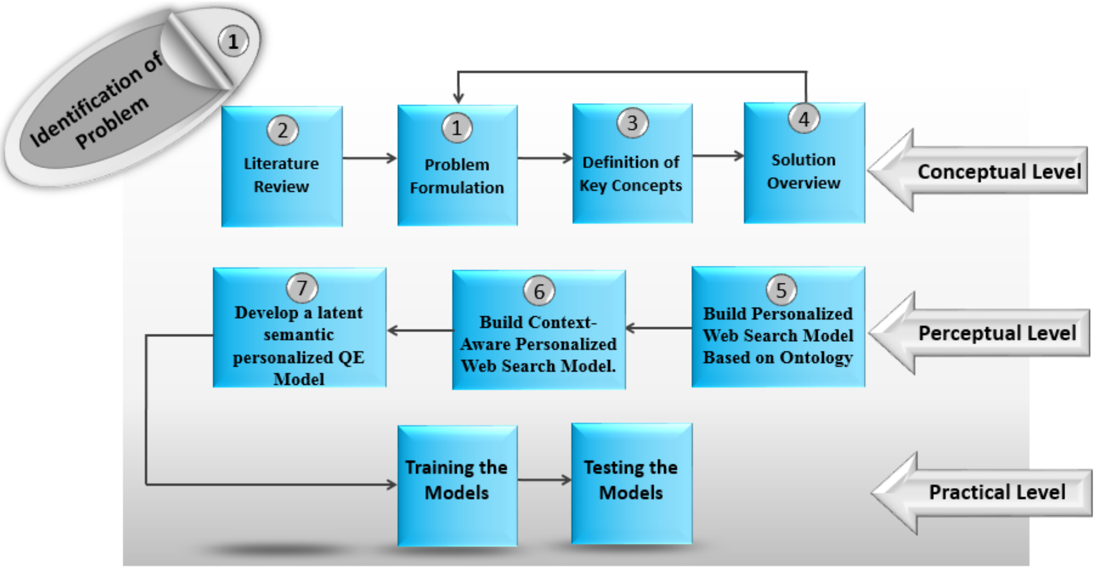

1
Identification of Problem
2
Literature Review
1
Problem Formulation
3
Definition of Key Concepts
4
Solution Overview
Conceptual Level
7
Develop a latent semantic personalized QE Model
6
Build Context-Aware Personalized Web Search Model.
5
Build Personalized Web Search Model Based on Ontology
Perceptual Level
Training the Models
Testing the Models
Practical Level

Figure 3.1 A science and engineering-based research method

The proposed methodology consists of three research levels: the conceptual level, perceptual level and practical level, based on Gallier et al.'s (Galliers 1992) research. In the conceptual level, the research problems in the area of personalized web search are identified followed by applying sub-processes in the same level for specifying concepts and various ideas of a new personalized web search methodology. Then, in the perceptual level, other sub-processes are applied to develop and implement the proposed methodology. Finally, in the practical level, the proposed methodology is tested. These three levels are illustrated as follows:

Conceptual Level

This research level involves four processes: literature review, formulation of problem, definition of key concepts and conceptual solution. The description of these four processes that aim to create new concepts and ideas for a personalized web search methodology are as follows:

1. Literature review

The literature in the personalized web search area was reviewed to find the gaps in the research or the problematic issues in relation to the identified research problems. By focusing on the sub-problems in this thesis, the literature can be classified into three groups: *ontological user profile for re-ranking the results*, *contextual web search for web personalization*; and *personalized query expansion*.

2. Problem formulation

This process applies the research issues identified by reviewing the current literature in order to formulate the research problem. Also, the research questions and the objectives of the research are defined and presented in Section 3.4, Section 3.5 and Section 3.6. In light of the literature review, the problem is divided into three sub-problems: ontological user profile for reranking the results, contextual web search for web personalization; and personalized query expansion.

Definition of key concepts

After applying the process of problem formulation, a set of key concepts are defined in Section 3.2 such as: web search engines, personalized web search, search engine evaluation, information retrieval, ontologies, service querying, query reformulation, query expansion, personalized query expansion, semantic, service metadata, user context, context search etc…. The concepts are used in defining the conceptual solution.

Conceptual solution

This process aims to produce various solutions conceptually for the research. Depending on the process of problem formulation, the solution is divided into three parts that are described in Chapter 4. The techniques of soft computing are applied in this thesis to tackle the defined research problems.

Perceptual Level

The perceptual level is concerned with the development and implementation of the solution. This level comprises the development of methodology addressing the identified research issues in this thesis and the development of prototype to validate the solution developed.

Methodology development:

Based on the conceptual solution in the conceptual level, the methodology for personalized web search was developed. Based on the identified the proposed methodology in this thesis is sectioned into three subparts: ontological user profile for re-ranking the results, contextual web search for web personalization, and personalized query expansion. The working of each of these subparts are explained in detail in Chapter 5, Chapter 6 and Chapter 7 respectively.

Development of prototype:

After introducing the proposed methodology, the ontological user profile prototypes, contextual user profile and personalized query expansion are developed. To validate the methodology, a case study is presented. The personalized web search methodology is implemented using the Java programming language. The details of the prototype and its working is presented and explained in detail in Chapter 8.

Practical Level

This level is concerned with the testing and validating of the proposed methodology. For the testing process, the transport domain ontology and corpus dataset are used to test the proposed methodology. For the validating process, the performance measures of information retrieval such as precision, recall and f-measure are employed to evaluate the proposed methodology. In some cases, in which the experiments were inadmissible, fine-tuning is needed to evaluate the performance of the proposed methodology.

3.8 Conclusion

In this chapter, the research problem is identified in the area of a personalized web search. This issue is addressed by decomposing it into a subset of topics, namely personalized web search based on ontological user profile, context-aware personalized web search, and semantic personal query expansion. Based on the literature review in those research areas, four research objectives were identified. We subsequently proposed solution for each of the research issues, thereby resulting in research solution. The most important concepts were presented in this chapter. The science and engineering-based research technique is used in the retrieval of user services to improve the personalized web search-based methodology. In addition, we identified the reasons why we chose this technique, with a detailed explanation of the processes undertaken.

CHAPTER 4

SOLUTION OVERVIEW

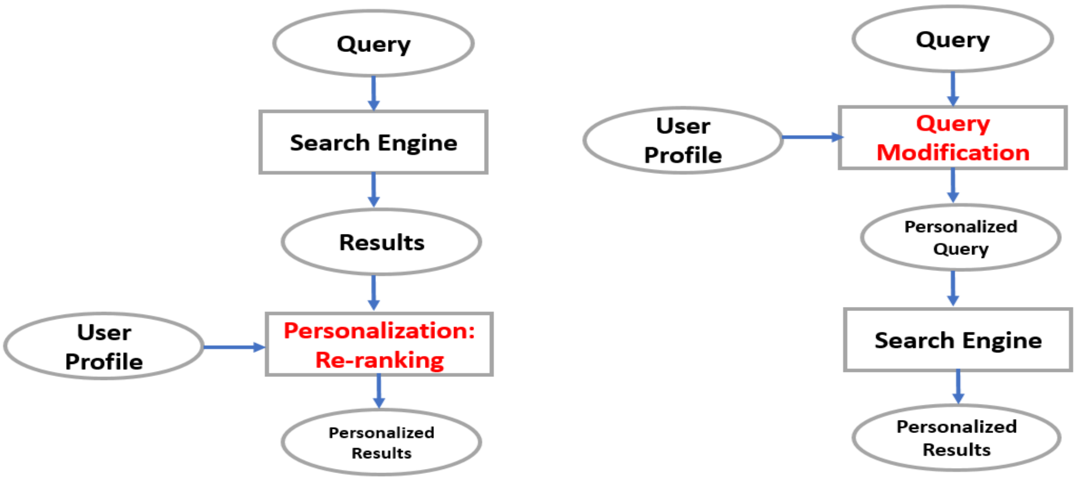
4.1 Introduction

As discussed in the second and third chapters, in this thesis we identify the issues related to personalized results ranking and query expansion into the following three main categories:

a) Lack of personalized web search approaches based on ontological user profile.
b) Lack of personalized and re-ranking of Web search using context history.
c) Lack of semantic approaches for query expansion.

This chapter presents an overview of the solutions for each of the research questions identified in Chapter 3. Also, based on the identified science and engineering-based research methodology, the research problem is defined, and then the conceptual solution is presented to clarify the issues related to a semantic personalized web search.

An ontological user profile is important in the process of information retrieval because it contains information about the user's interests which can be used to re-rank the search results as shown in Fig 4.1.



A) Re-ranking activity B) Reformulating the user query (query expansion)

Figure 4.1 The phases where the user's profile impacts on the personalization process during the retrieval process (Micarelli et al. 2007).

4.2 Overview of the Solution for Personalized Web Search

In this section, the approaches used to address the objectives of this study are presented. Four research questions are posed in this thesis according to its objectives. We proposed a semantic personalized web solution that consists of three main parts:

a. A personalized web search based on an ontological user profile that aims to build user profiles based on the users' behaviour and the web pages that they visited,
b. A context-aware personalized and re-ranking web search using navigation history and
c. A personalized query reformulation.

Finally, to evaluate the proposed methods, we used three measures widely used in information retrieval: precision, F-measure, and recall.

4.3 Overview of the Solution for Personalized Web Search Based on Ontological User Profile

This section presents a step-wise overview of the solution for carrying out personalized web search based on an ontological user profile. The overview of the solution is as follows:

Step 1: Constructing a vector of weighted terms

A weighted vector is constructed for the user based on the web pages visited. These pages are represented by pre-processing metadata from the visited web page. Subsequently, the stemmed and stop words are eliminated. These vectors of terms include some terms obtained from the page itself and a set of keywords.

On the other hand, a reference ontology concept can be represented as a vector of terms. Synonyms may play an important role in adding useful knowledge which leads to retrieving content which is more accurate. On this basis, the web page vector or the concept vector are both based on the WordNet API, and each term has some added synonyms for the corresponding vector.

Step 2: Finding the term weight values in the vectors.

During this step, the term values in the defined vectors of the web page and the concepts are computed.

Step 3: Calculating the similarity between a page and concept.

The similarity between the visited page vector (P) and the concept vector (C) is calculated by computing the cosine between two constructed vectors, and then building the results obtained in the similarity matrix and saving them as outputs, which is called page-concept.

Step 4: Finding a user's interest score for a specific concept

The matrix (i.e. IS [P × C]) expresses the degree of user interest by the concepts used, and in this matrix, each concept used by the user expresses the degree of interest. The user may browse some URLs for a very short period of time, or a web page may be opened by mistake. Such pages cannot be compared to pages that have been opened for a long time, saved or printed. So not all pages have the same degree of interest. A particular mechanism is used to measure the interest level of each user based on his browsing history to obtain a more accurate allocation process. This mechanism tracks the pages the user opened and measures the weight value of each page. These values are used to determine the degree of user interest in the concepts in domain ontology. We used Chromium browser automation software which records the users' browsing behaviour.

Three values can be calculated to determine the weight of each page opened: search result, browsing and favourites. When a user opens a web page, the browser status can be obtained, and the weight status is set to 0.5. But, when the user browses a web page that was used, by the retrieved URL, and based on a query, the search result status is obtained, and the weight status is set to 0.75. Finally, the weight status of a favourite page is calculated based on the pages that have been printed or stored and the weight status is set to 1.

Step 5: Re-ranking the results returned by the search engine

The results related to a personalized search are based on the results obtained from the search engines after rearranging them according to the ontological user profile. This profile and the retrieved search results are the inputs for the query that the user is looking for. Thus, as explained in the second step, each retrieved URL has a similarity concept, taking into account the similarity matrix [P × C].

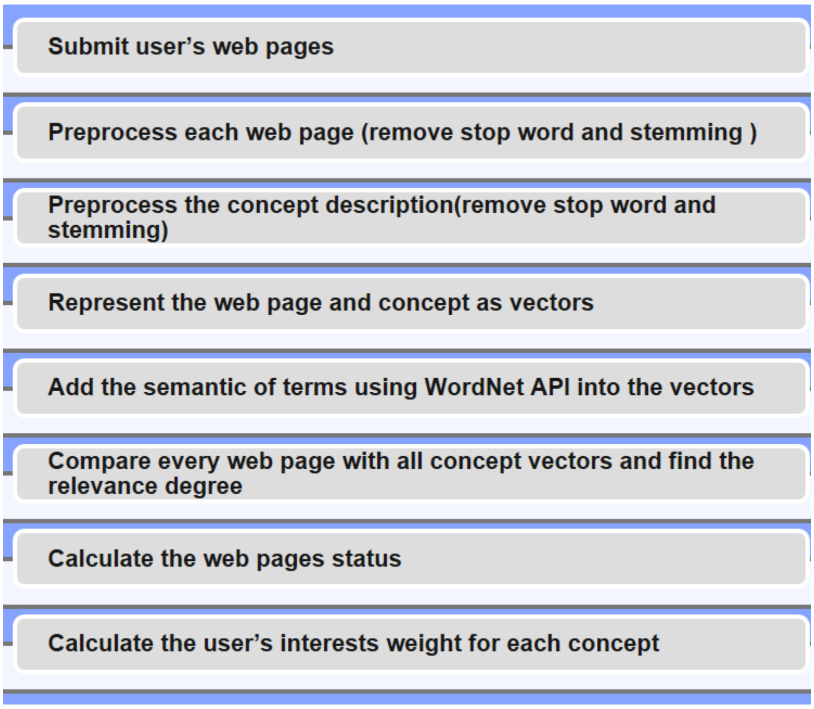

Submit user’s web pages
Preprocess each web page (remove stop word and stemming)
Preprocess the concept description(remove stop word and stemming)
Represent the web page and concept as vectors
Add the semantic of terms using WordNet API into the vectors
Compare every web page with all concept vectors and find the relevance degree
Calculate the web pages status
Calculate the user’s interests weight for each concept

Figure 4.2 Stepwise process of building an ontological user profile

4.4 Overview of the Solution for Context-Aware Personalized Web Search Using Navigation History

In this section, we present the solution overview of the approach used for the context-aware personalized retrieval methodology.

The context history used in our proposed approach consists of two basic elements, the users' behaviour and the past context. To obtain the user’s context history, the web navigational patterns of the users must be obtained to improve the performance of personalized web search. For example, if the user searches using the query "travel", he will obtain different results depending on the time that the search query is carried out. For example, if the user uses a query “travel”, the search results can be different depending on when the query was issued or carried out. For example, the results returned on a weekend may be different from those on weekdays, as the user may take a plane to travel on weekends, but a train, taxi or buses on weekdays (based on his personal preferences). This temporal differential search results can be employed using a contextual user profile as illustrated in the previous example to determine the user’s needs taking into account the user's semantic knowledge and context history. A contextual profile is used to rearrange the results obtained by different search engines to get the closest results to the user’s query.

The main idea of the proposed solution is the reranking of the obtained results from the search engines by determining the degree of similarity between the contextual profile of the user and the retrieved results. Our proposed approach comprises three steps as below.

(a) The first step is to get a user query and using the WordNet API to expand it.
(b) The second step is to retrieve the search results after completing the above step by a keyword-based search engine.
(c) The third and last step is to re-rank the obtained results, depending on the contextual profile of the user, and then the user received the relevant results.

The work in this proposed model involves several procedures: building a contextual profile for the user based on navigation history, which includes the user's interests and preferences, by referring to the user's web logs or search histories. URLs can be extracted for the pages the user visited, which can be obtained from the user's search history. These URLs are then subjected to analysis in order to extract some site-specific data, such as descriptions of webpages and keywords.

Depending on the user's temporal context, four-time values are defined for each concept, divided as follows: weekday day-time (WDD), weekend day-time (WED), weekday night-time (WDN) and weekend night-time (WEN). Based on these, the pages visited are divided based on the browsing history and a numerical value for each concept is determined.

The overview of the solution is as follows:

Step 1: Obtaining the weights of the concepts and pages that the user visited implicitly.
To represent four-time concepts and pages, an extended vector space model (EVSM) is used. To build a vector space for both concepts and visited web pages, it relies on the description and keywords on each web page to find the synonyms for each term depending on the WordNet API. These synonyms provide deeper knowledge to determine the content more precisely, and then the weights of the web pages and concepts are determined in the expanded vector space.
Step 2: Finding the similarity degree between the visited pages and the concepts.

In this step, we compare the visited pages (P) and concepts (C) to find the degree of similarity by computing the cosine of the angle between the two constructed vectors and store the results as outputs in the similarity matrix.

Step 3: Generating a user-concept interest score matrix.

We generate a user-concept interest score matrix [P × C] by finding a degree of interest for the different concepts and the results are represented by the matrix. In this case, the user's interest in this page is represented by the "weight status" based on the search history. There are three different ways to express weight status, either to be searching, favoring, or browsing, and this is determined by the actions the user takes when opening a page. We define the first situation of "searching" as the process of entering a page by getting a link by the results obtained from the query. We define the second situation "being a favourite" as either storing or printing a page. We define the situation of "browsing" as when the user is looking for a page. We used Chromium browser automation software which records the users' browsing behaviour. Based on these three cases, the user's interest in a page is calculated.

Step 4: Re- arranging the search engines results.

Based on the query, we re-rank the results obtained from the search engines and obtain the personalized search results based on the contextual profile of the user. The contextual profile and the original search results are the inputs that are used in the procedure of re-ranking the search results.

4.5 Overview of the Solution for Personalized Query Expansion

The goal of query expansion is to associate some terms that are most relevant to the original query (Kathuria, Mittal & Chhabra 2017). For this purpose, query expansion solutions have been integrated with the service querying solution.

We present a new query expansion model based on user's profiles by considering the term weight global techniques class level (TF-IDF-CF) and dimensionality reduction. The proposed approach is evaluated by comparing its effectiveness with two advanced models using the Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting techniques, using the benchmarks of recall, precision and F-score. The evaluation results show that our proposed model is approximately 3% better for the short terms and a 15% improvement for the long terms across all the measures compared with the two other models.

The overview of the steps in our proposed approach are as follows:

Step 1: Data collection

To validate our approach, we gather the documents manually that belong to a group of 19 users and each user has implied 50 queries. The information collected contained a total of 950 documents, and the number of words in these documents totalled 2767. Each document contains a meta-description of the pages that are compiled and obtained through the "BuzzStream" tool from the link (http: /tools.buzzstream.com/meta-tag-extractor) (Shreves 2012).

Step 2: Feature term extraction

Our proposed approach for feature extraction is based on three different methods. These methods are used to represent the data collected from the original documents in a statistical way. These include: Term Frequency-Inverse Document Frequency Class Frequency (TF-IDF-CF) (De Silva & Haddela 2013), Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) (Xu & Croft 1996). These feature terms are formed like a matrix and called the mother matrix.

The main importance of representation techniques is in measuring the weight of the terms in the text (Ren & Sohrab 2013). Local weighting (LW) represents the number of times the term is repeated in a single document, indicating the importance of that document, and global weighting (GW) represents the number queries in the text based on the use of a logarithm (Xia & Chai 2011).

Step 3: Matrix dimensionality reduction

After completing the previous stage, a matrix can be built of column headers and row headers, where the labels for the texts are column headers (T1, T2 ... .tm) and the words for the corpus (collection of documents) are row headers. This matrix contains several elements, and based on one of the different representation methods mentioned above, these elements represent different statistical values. This will result in a matrix with high dimensionality with a number of words and text files, which is called the mother matrix (origin matrix). Because of the large size of this matrix, it is not possible to work with it, but it is needed to reduce dimension to improve its performance in terms of accuracy and time. Dimensionality is reduced by using the singular value decomposition (SVD) method or the feature selection method. The importance of this approach is to improve the performance of the system in general, in order to find the most appropriate features. As a result of applying SVD, the original matrix (mother matrix), and three sub-matrices are derived. These matrices reduce the complexity of the calculations, eliminating the features that have been repeated minimizing the dimensions and reducing the features that are used in the personalized query expansion model.

Step 4: Query expansion

After obtaining the mother matrix, the degree of similarity between documents can be calculated based on the cosine similarity for each user. Then the highest terms values in the user profile is used with an initial query in order to expand and reformulate the query. Note that the documents are divided into two parts: testing and training. Approximately 70% of these documents are used for training and 30% for the testing.

4.6 Overview of the Solution for Validation of the proposed Methodology for Semantic-based information Retrieval

This part of the thesis concerns the solution that helps to test the effectiveness of the proposed methods. In the information retrieval field, different types of measures are used to measure the level of performance in semantic-based approaches (Baeza-Yates & Ribeiro-Neto 1999). These measures include F-measure, precision and recall and are used for testing the accuracy of the proposed ontology-based personalized and re-ranking approaches. The personalized and re-ranking web search using navigation history and the semantic query reformulation are illustrated as follows:

Precision is one of the important measures in retrieval systems that summarizes and compares the search results in order to maintain the accuracy level by measuring how well the search engine is able to rejecting non-relevant documents (Sieg, Mobasher & Burke 2007a):

$$precision = \frac{\#R_{rel}}{\#R} \qquad (4.1)$$

Where,

#Rrel: represents the number of relevant results retrieved.

#R: represents the number of results retrieved.

Recall is also an important measure in retrieval systems that is used to summarize and compare the search results to maintain the effectiveness by measuring how well the search engine is able to obtain all the relevant documents for a given query (Sieg, Mobasher & Burke 2007a):

$$Recall = \frac{\#R_{rel}}{\#D_{rel}} \tag{4.2}$$

Where,

#Rrel: represents the number of relevant results retrieved.

#D: represents the number of relevant documents.

F-measure: this measures the performance of precision and of recall. This work represents the F-measure in the following way:

$$F - score = 2 * \frac{precision * recall}{precision + recall} \tag{4.3}$$

Benchmarking of the proposed approaches

In our work, we have benchmarked the performance of our proposed approaches developed in response to Research Question 1 against Google and that presented by Samen et al (Samen, Ezin & Onana 2017). For benchmarking of these three approaches, we used the *Precision* measure. Furthermore, the performance of the solution developed in response to Research Question 2 was compared with the performance of the solution developed in Research Question 1. Finally, the proposed approach developed in response to Research Question 3 was benchmarked against two advanced models using the Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) weighting techniques.

4.7 Conclusion

In this chapter, an overview of the proposed solution of the semantic personalized web search is given. Based on the research questions and objectives identified in Chapter 3, the proposed semantic personalized web solution was divided into three main parts: the personalized web search based on an ontological user profile, the personalized and re-ranking web search using navigation history and a new model of personalized query expansion. The outline and working of each of the three approaches was explained. An overview of the approach for experimentation and benchmarking that is used in this thesis was discussed.

In the next chapter, the proposed personalized web search based on an ontological user profile is explained in detail.

CHAPTER 5

PERSONALIZED WEB SEARCH BASED ON ONTOLOGICAL USER PROFILE

5.1. Introduction

In recent years, the World Wide Web has become widely used for various activities and in various sectors such as Health, Transportation etc. The Internet provides users with an enormous amount of information, which answers their various problems or concerns. This information can be searched for and retrieved using search engines. However, the information provided to all users for a specific query is similar. As users' interests and preferences differ, it is important to help users find relevant information through personalized systems (Sieg, Mobasher & Burke 2007b). Personalization can be achieved by modelling each individual by managing, building and representing user preferences in the form of user profiles (Gauch et al. 2007). The main challenge of personalized web search is to model the user's needs and interests with the aim of enhancing search quality. In order to model the needs of the user, many systems have been developed in the web search domain. This process involves collecting information related to the user directly and explicitly or implicitly and tailoring the search results of the user the collected information about the user (Akhlaghian, Arzanian & Moradi 2010; Baazaoui-Zghal & Ghezala 2014; Daoud, Tamine-Lechani & Boughanem 2008; Ferreira-Satler et al. 2012; Nanda, Omanwar & Deshpande 2014).

In the process of modelling user's preferences and interests, there are a number of issues that need to be addressed, the most important of which are: (i) ignorance of user intent; and (ii) the use of natural language in the contents of the web, where this language contains different meanings of words, so it is not clear language. To overcome these limitations, a number of research studies have attempted to incorporate ontology within the information retrieval process (Hourali & Montazer 2011). Employing ontology to personalized web search systems in this manner is performed in two different ways. The first approach is by building an ontological user profile, which represents user preferences, and using it for the re-ranking of the query results. The second approach is to reformulate user queries in such a way that the probability of retrieving relevant documents (based on user's interests) is significantly increased. The ontology-based personalization approaches have proven to be highly effective, as they have outperformed traditional keyword-based approaches and helped to create more personalized results (Hopfgartner & Jose 2014). This topic is still the focus of man researchers and is still under consideration.

In this thesis, the proposed web search personalization approach aims to build an ontology-based user profile based on an implicit new weighting technique. This method attempts to assign interest scores to concepts in the targeted domain ontology using a semantic fuzzy classification technique that takes into account the browsing behaviour of users while scoring the concepts. The effectiveness of the proposed personalized search approach has been validated in a case study in a transport service domain. Based on the obtained results, the proposed approach has proven to be very effective in terms of quality research results.

5.2 Existing Personalized Web Search Approaches

Ontology is one of the most successful solutions to address the issues of cold start and ambiguity in words. Moreover, it is an important and reliable representation in personalized retrieval systems in terms of users' interests (Calegari & Pasi 2010; Duong, Uddin & Nguyen 2013; Gauch et al. 2007). In addition, ontology has been used for query expansion in a personalized web search engine. A new generation of ontologies based on fuzzy theory with two degrees of uncertainty was proposed in (Hourali & Montazer 2011). The concepts and relations between concepts were assigned by an expert in the related domain, and the proposed ontology was used for the query expansion process. The results show that the fuzzy ontology approach increased the precision results compared to the crisp ontology approach. An online information retrieval system named SIRO was proposed in (Baazaoui et al. 2008). The authors combined a domain ontology with a service ontology to assist user query reformulation. In their approach, the relationship between every concept in the domain ontology and the tasks in the service ontology is established. When a user makes a query, the query is expanded by the addition of new terms in the service ontology which are related to the matching concepts. A comparison of the proposed approach with other systems showed a considerable improvement in the results. An individual fuzzy ontology method for integration into the query reformulating process was proposed by (Baazaoui-Zghal & Ghezala 2014). Three main components were introduced: automatic individual fuzzy ontology building, query reformulation based on fuzzy ontology, and document classification. The individual fuzzy ontology is built automatically by assigning an initial membership value to the ontology concepts and relations and the values are then updated continuously based on the user's previous queries. In the query expansion process, negative terms are eliminated from the original query terms that reflect the negative preferences of the user. This method has had a very positive effect on improving the search results' quality.

Some researchers have relied on ontology studies as the most important elements that express the interests of the user, where ontology determines the value for each concept. The importance of these values lies in the re-ranking process of the results related to the user's query. For example,(Nanda, Omanwar & Deshpande 2014) created a dynamic category interest tree (DCIT) in which a user's interest score is added implicitly to each category in the tree through the browsing history of the user and collaborative filtering. To match the user's document of interest with the corresponding category, a fuzzy classification approach is used. The final results obtained by Google for the user's query will be re-ranked based on the weighted tree. The authors in (Duong, Uddin & Nguyen 2013) proposed a semantic framework for a personalized search using a semantic web technique. Three issues are addressed in this work: a way to detect the interests of the user, modelling the user profile and expanding the query of the user. This approach proved to be superior to other approaches such as SPWS and SOnP in terms of efficiency in performance and accuracy.

The following part of the thesis presents the basic flowchart of the semantic personalized web search methodology. In the other sections, each step of the methodology is detailed.

5.3 Personalized Web Search Based on Ontological User Profile Methodology

In our proposed approach the initial results obtained from the search engines are rearranged. This process rearranges the results based on the degree of match between the retrieved results and the ontological profile of the user. This process helps to obtain user interests by providing personalized search results. In this section, the workflow of the semantic personalized web search methodology is outlined in Figure 5.1. Given a query Q, the personalized web search approach is able to re-rank the search engine results that are relevant to Q depending on the ontological user interest, where the proposed approach effectively enhances the overall quality of the search result by offering users the most suitable result with the least search effort. The personalized web search methodology involves the following stages:

1- **Expanding Query**: the user's query is received and expanded using a WordNet API.

2- **Passing of expanded query:** the expanded query is passed to the keyword-based search engine and the related results are retrieved.

3- **Re-ranking Query**: the retrieved results are re-ranked based on the ontological user profile, and the personalized results are presented to the user.

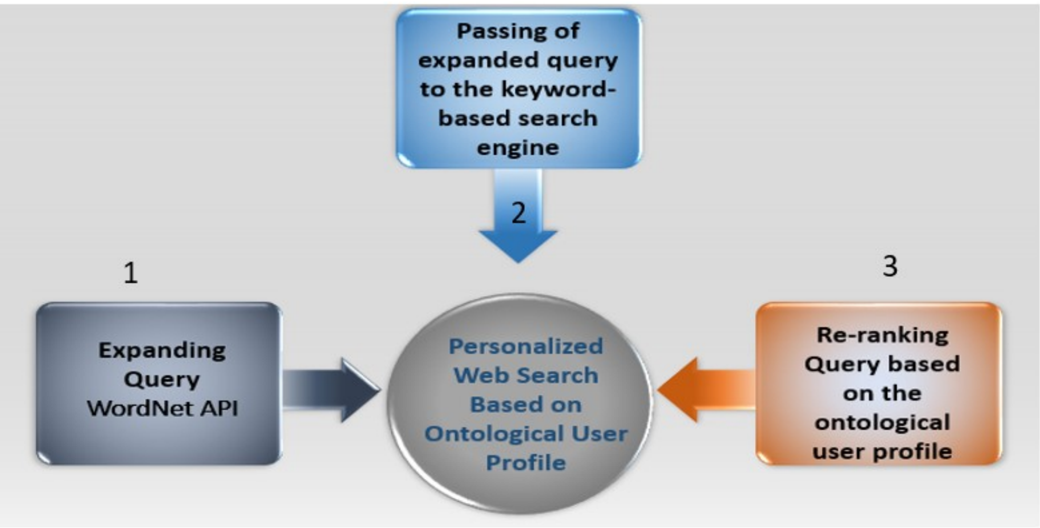

Passing of expanded query to the keyword-based search engine
2
1
Expanding Query WordNet API
Personalized Web Search Based on Ontological User Profile
3
Re-ranking Query based on the ontological user profile

Figure 5.1: Steps of the Personalized Web Search Based on Ontological User Profile Methodology.

Our proposed approach involves two features: the use of ontology, which is the basis of this approach and the ontological user profile. It offers a structured and unambiguous representation of knowledge in a well-defined format using ontology modelling languages such as the Ontology Web Language (OWL). Semantic knowledge embedded in the ontology can be exploited to support the semantic analysis of heterogeneous content and to return useful information for better decision making (Al-Hassan, Lu & Lu 2015; Sánchez et al. 2012). The idea of using ontology in our personalized web search methodology is based on the success of semantic approaches on the Internet as well as the different needs of the users.

Depending on the user's search history and the existing reference domain ontology, an ontological user profile is built and developed. This profile can be used for two primary objectives. The first objective is to extract concepts from user profiles that reflect the user's interests and the second objective is that the ontological user profile can be used to re-rank the search results.

In this part of the thesis, the reference ontology which is demonstrated is the transport service domain ontology that was developed by (Dong, Hussain & Chang 2011). The transport service ontology is based on metadata extracted from transport service websites. The major mission of transport service metadata is to extract meaningful information regarding transport services from downloaded web pages. The proposed approach comprises two components: building the ontology-based user profile and the re-ranking algorithm. Each is detailed in the remainder of the thesis.

5.4 Building an Ontology-Based User Profile

This is where a user profile is created as a record of structured data. The user profile is intended to preserve user data in terms of interests or preferences. Users' preferences and interests are identified in light of their search history (i.e. web log). The search history of users is used to extract the URLs of web pages that have been visited at the given website, and the URLs are analyzed to fetch the metadata of those web pages, i.e. keywords and descriptions of web pages based on the meta tag in the HTML. Two types of data are maintained in the user file. The first type instances, which includes transport service items, which present the important pages for an active user. The second type is concerned with concepts, which work to classify service items using ontology. Each concept is annotated with a weight reflecting a user's degree of interest. It is important to emphasize that one of the main contributions of this approach is that a user's interest score values of concepts are computed semantically based on a semantic fuzzy-based classification approach. This approach considers the overlapping nature of concepts; thus, it calculates the relevance degree between each web page visited and every concept, which in turn is used to find the user's interest scores for the concepts. The calculation process in which the interest scores of concepts is calculated, based on four steps, is as follows:

Step 1: Building user-weighted vectors for visited pages and concepts

During this step, each user gets his own weighted vector, which is created based on the concepts and pages he browses. These concepts and pages are then represented as the vector of terms. To represent each visited web page in this way, the description of its metadata is pre-processed to remove stop and stemmed words using the Porter stemmer (Porter 1980). On this basis, the vector of terms is created from a web page and consists of keywords and some terms obtained from the description of the page.

On the other hand, each concept expresses a vector of those terms. These terms include all the terms extracted from the metadata of keywords and descriptions of all the web pages that are classified under a specific concept. Intuitively, the synonyms of terms can add meaningful knowledge that may help to determine more accurate content. Consequently, the WordNet API is used in the expansion process of the web page vector or the concept vector, which adds synonyms for each term to the corresponding vector. The vector for a specific webpage $\overrightarrow{p_j}$ can be defined as follows:

$$\overrightarrow{p_j} = \left(w_{t_1,j}, w_{t_2,j}, \dots, w_{t_n,j}, w_{s_1,j}, w_{s_2,j}, \dots, w_{s_m,j}\right) \tag{5.1}$$

where $w_{t_1,j}$represents the weight of the first term in the vector space of a webpage p_j, and $w_{s_l,j}$represents the synonym of a specific term in the vector space of a webpagep_j.Formally, the vector for a specific concept $c_i$is defined as follows:

$$\overrightarrow{c_i} = \left(w_{t_1,i}, w_{t_2,i}, \dots, w_{t_n,i}, w_{s_1,i}, w_{s_2,i}, \dots, w_{s_m,i}\right) \tag{5.2}$$

where $w_{t_1,i}$represents the weight of the first term in the vector space of a concept c_i, and $w_{s_l,i}$represents the synonym of a specific term in the vector space of a concept c_i.

Step 2: Discovering the weight values on the level of concept and web page for all vector terms.

This phase involves calculating a weight value for all terms in the vectors, whether web page vectors or concept vectors. Formally, the weight values of a specific term and its synonym are computed using equations 3 and 4 as follows:

$$w(t) = occ(t) + occ(t)\sqrt{1/_{occ(syn_t)}} \tag{5.3}$$

$$w(syn_t) = occ(syn_t) + occ(t)\sqrt{1/_{occ(syn_t)}} \tag{5.4}$$

where $w(t)$ is the weight of term *t*, $occ(t)$ is the number of occurrences of term *t*, $occ(syn_t)$ is the number of words with the same meaning as term *t* (i.e. synonyms), $w(syn_t)$ is the weight of the synonym of term *t*.

Step 3: Finding the degree of similarity between the concepts and all the URLs that have been visited, and then the results obtained in the similarity matrix are stored and used as output later.

This phase involves calculating the degree of similarity between the concepts and all the URLs visited based on cosine similarity. The defined weighted vectors of a specific web page and concept, as detailed in step 1, are considered to find their relevance (i.e. similarity) degree. The computed similarity results are stored as output in a similarity matrix named page-concept, i.e. $SIM[P \times C]$ matrix. This equation represents the calculation process in which the similarity $sim(p_j, c_i)$ is computed between the c_i and p_j:

$$sim(p_j, c_i) = cos(\overrightarrow{p_j}, \overrightarrow{c_i}) = \frac{\overrightarrow{p_j}.\overrightarrow{c_i}}{\|\overrightarrow{p_j}\|_2 \times \|\overrightarrow{c_i}\|_2} = \frac{\sum_{t_k \in T_{p_j,c_i}} w_{p_j,t_k} \times w_{c_i,t_k}}{\sqrt{\sum_{t_k \in T_{p_j,c_i}} w^2_{p_j,t_k}} \sqrt{\sum_{t_k \in T_{p_j,c_i}} w^2_{c_i,t_k}}} \qquad (5.5)$$

where $p_j. c_i$ denotes the dot-product between the two vectors $\overrightarrow{p_j}$ and $\overrightarrow{c_i}$, T_{p_j,c_i} is the set of terms available in both vectors $\overrightarrow{p_j}$ and $\overrightarrow{c_i}$, $w_{p_j,t_k}$and w_{c_i,t_k} is the weight value of term t_k in the vector of p_j and the vector of concept c_i, respectively. The similarity between a page and concept is calculated by computing the cosine of the angle between two vectors. The results of cosine similarity range between 0 and 1. On this basis, the relationship between the matching value of a web page and a concept is positive, and the higher the matching value, the higher the similarity between them.

Step 4: Finding a user's interest score for a specific concept

This step involves implementing the following equation to calculate the interest score of a concept c_i for the user:

$$IS(c_i) = \frac{\sum_{\forall p \in P, c = c_i} sim(p_j, c_i)}{\sum_{\forall c \in C} \sum_{\forall p \in P} sim(p_j, c_i)} \qquad (5.6)$$

where P is the set of all the visited URLs in a user profile, C is the set of all concepts in the reference ontology, $\sum_{\forall p \in P, c = c_i} sim(p_j, c_i)$ is the sum of all the relevance degrees of web pages for a concept c_i, and $\sum_{\forall c \in C} \sum_{\forall p \in P} sim(p_j, c_i)$is the sum of all relevance degrees of web pages for all concepts. A user-concept interest score matrix (i.e. IS [U × C]) is output for this phase, where the user's interest score for a concept c_i can be expressed by that entry.

Intuitively, not all search history contents from URLs have the same importance to the user. It is possible that the user has opened one of these links by mistake or the user opened it for a few seconds, in which case, these pages cannot be compared to the pages that are stored or printed. For a more personalized process, a mechanism is introduced to determine the level of user interest in the URLs in the search history. This mechanism collects a user's behaviour while browsing web pages and allocates a weight status value to every URL. The status value is then considered to calculate the user's interest score of concepts available in the reference domain ontology.

Browse, search result, and favourites represent different weight states for the visited web pages. The first states, browsing, occurs when a certain web page is browsed. The search result status occurs when a user browses a retrieved URL based on a query. The favourite status represents the occasions on which a user saves or prints a web page. A weight status of (1, 0.75 and 0.5) is assigned to the favourites, search result and browse status, respectively. Table 5.1 details the weighting process.

Table 5.1: Concept weights based on user interests

User's web pages/Concepts	C1	C2	C3	………	Cn	Page state
P1	.8	.5	.3		.2	Status weight
P2	.5	.7	.3		.1	Status weight
P3	.3	.8	.5		.7	Status weight
P4	.7	.6	.4		.4	Status weight
…..	..	..	..		..	..
….	..	..	..		..	..
Pm	..	..	..		..	..
Sum	X=∑C1	Y=∑C2	Z=∑C3		∑w(ci)	

The status of the page's weight is inspected in the proposed approach and detailed in the experiment section. To demonstrate this, each entry in the similarity matrix $SIM[P, C]$, as described in step 3, is weighted by multiplying the similarity of p_j and c_i by the corresponding weight status value of p_j. The user's interest scores for the concepts are then computed using equation (5.6), based on the obtained weighted similarity. The experiments show that using the page's weight status helps to improve the quality of the results. This will be described in detail in the experiment evaluation section.

Algorithm 5.1: Re-ranking of the retrieved search results based on ontological user profile

Input: ontological user profile for a target user (u), set of retrieved search results for a query (Q)

Output: re-ranked search results, i.e. personalized results for the target user (u)

Process:

```
    Set $C = \{c_1, c_2, \dots, c_m\}$
    Set $URL = \{URL_1, URL_2, \dots, URL_n\}$
For each $URL_j \in URL$
$c$= find the highest matching concept to $URL_j$
        compute $sim(Q, c)$

        compute $sim(Q, URL_j)$
            $IS(URL_j) = sim(Q, URL_j) \times sim(Q, c) \times IS(c)$
End For
Sort URLs based on interest score $IS$
```

5.5 Search Personalization – Re-ranking Algorithm

The results of the personalized search are expanded by re-arranging the results that are returned by the search engine for a specific query considering the ontological user profile. Algorithm 5.1 shows the steps of the results retrieval re-ranking process. As can be seen in Algorithm 5.1, the input is the ontological user profile and the retrieved search results for a specific query. Each retrieved result (i.e. URL) is first represented in a weighted vector and expanded as described in step 1 and the weight values of the terms computed according to equations (5.3) and (5.4). Secondly, the highest similarity concept is obtained for each retrieved URL, based on the similarity matrix$[P \times C]$, as presented in step 2. Finally, the interest score of the user is calculated for each URL in the retrieved search result by applying the following equation:

$$IS(URL_j) = sim(Q, URL_j) \times sim(Q, c) \times IS(c) \qquad (5.7)$$

where $sim(Q, URL_j)$ is the similarity between the query Q and a retrieved search result URL_j, $sim(Q, c)$ is the similarity between the query Q and the concept c with the highest match with URL_j, and $IS(c)$ is the interest score of the concept c which has the highest match with URL_j.The similarity in equation (5.7) is calculated using cosine similarity, as presented in equation (5.5).

5.6 Experiment Summary

The aim of the experiment evaluation is to measure the performance of this approach. The proposed approach is compared with the Google ranking scheme, to measure the improvement in precision in the variant results of top-n.

5.6.1 Description of the Data Set Used For Experiments

Based on the work conducted by (Dong, Hussain & Chang 2011), this part of the thesis uses a transport service domain ontology. This ontology was built based on a survey of websites of transport service companies. The concepts relating to the transport service ontology are divided into two groups: abstract concepts and actual concepts. The abstract concepts refer to the abstract domain and the sub-domain of service concepts, whereas the actual concepts relate to real services (e.g. transport web pages); that is, the actual service concepts can link to the metadata of a service description entity (i.e. SDE). Service description and service name are the main properties for all service concepts.

The proposed approach was evaluated against Google, which is the most commonly used search engine, and ten participants were invited to engage in our experiments. These users were asked to search various websites using some transportation terms. The users were required to write down the web pages they visited that were of interest. To capture user behaviour on the web pages and to help build the user profile, we used the Chromium browser automation software, which records all users' browsing behaviour.

After this, we asked the participants to use six different queries from the selected set into the Google search engine. Every user selected the results of interest from the first 30 results returned by Google for each query result. We called this collection of pages *"the interesting* *pages data set"*. The users also used our model to retrieve the data set of interest for the same queries. The top 5, 10, 15 and 20 links were then compared. The precision for the top 5, 10, 15, and 20 links is calculated to evaluate the performance of the proposed approach.

5.6.2 Average Precision of top-n Results

We calculated the degree of improvement in the precision obtained by the proposed approach and the Google search results at different top-n. On the whole, the results show that our personalized system based on the proposed approach achieved high precision at all top-n, especially at top 5 and 10. Figure 5.2 shows the average precision for the proposed approach without considering user behaviour (PWB), the proposed approach with user behaviour considered (PB), and the Google search engine. As shown in Figure 5.2 the proposed approach with PB shows an improvement in precision of 13%, 12.5%, 10.5% and 5% for the top 5, 10, 15 and 20 links, respectively over the Google search engine. The proposed PWB approach demonstrates improvements in precision of 9.5%, 10%, 8.5% and 4.5% for the top 5, 10, 15 and 20 links respectively, over the Google search engine. Nevertheless, the results demonstrate that the behaviour of the user has an encouraging impact on the results of the search.

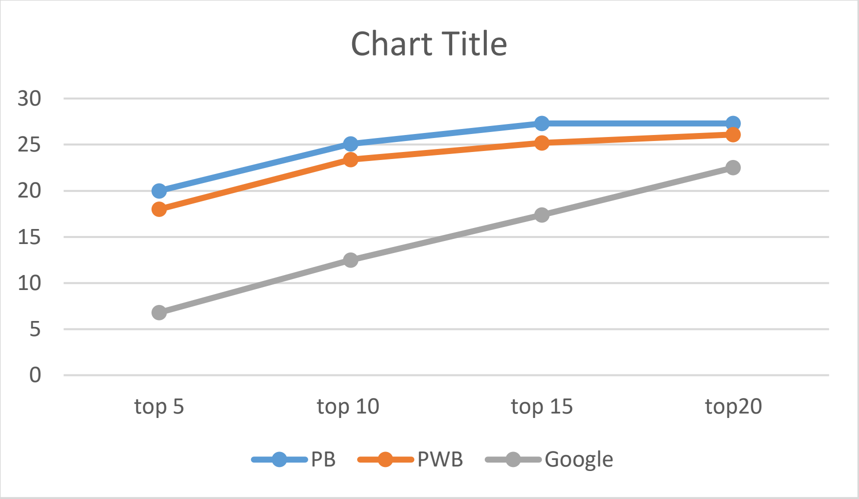


Figure 5.2: Average precision for top-n results

5.6.3 The Impact of User Behaviour on Precision

In our experiment, an improvement in precision is noticed when the behaviour of the user is taken into account when building the user profiles and exploited to rank the retrieved results. Figure 5.3 presents the average precision for the results of the personalized search (either with user behaviour or without it). The results show that the proposed personalized approach based on user behaviour achieves slightly better precision at all top-n. There is an improvement in precision of 1.5%, 2.5%, 2% and 0.05% for the top 5, 10, 15 and 20 links, respectively.

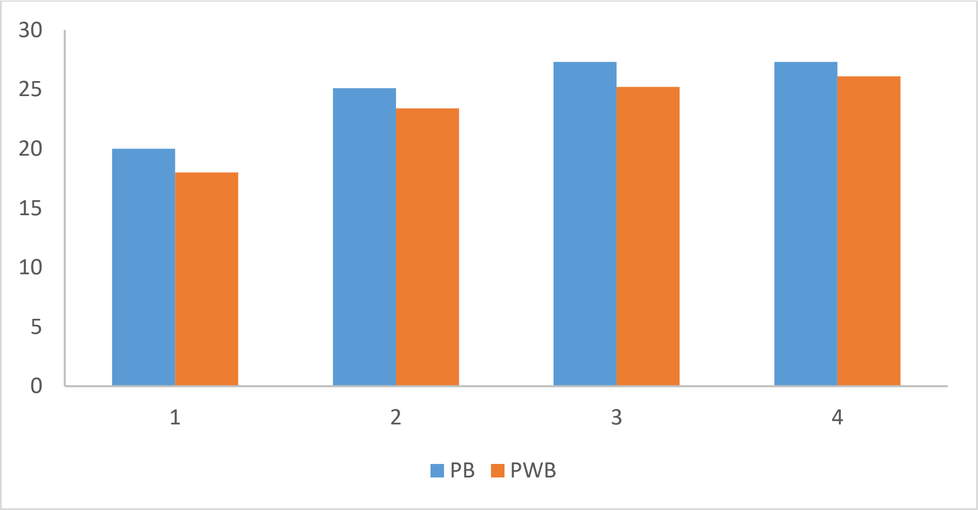


Figure 5.3: The effect of user behaviour on precision.

5.7 Experiment Evaluation and Benchmarking

The experiment evaluation based on comparisons measures the performance of our approach compared to other similar approaches presented in this field. This section details the results of the comparison of the proposed approach with two commonly used approaches, namely: the model proposed by (Samen, Ezin & Onana 2017) and the Google ranking approach. These comparisons are based on the improvement in precision in the variant results of top-n. Also, a benchmark dataset called the Open Directory Project (ODP) was used.

The rest of the chapter is as follows. In the next section (Section 5.7.1), we give an overview of the ODP project. In Section 5.7.2, we provide an accurate description of the proposed approach by Samen after which in Section 5.7.3 we benchmark the three approaches.

5.7.1 The Open Directory Project (ODP)

As it is difficult to obtain topics directly from a query, we infer the topics from the clicked webpage which reflects the query intent. In this thesis, we leverage a sophisticated content-based classifier that categorizes webpages into ODP categories (Bennett & Nguyen 2009).

In web page taxonomy, the "DMOZ" Open Directory Project 1 (ODP) is the largest, most comprehensive human-edited directory of the web. This high quality and free web taxonomy resource has been used in a number of prior research studies (Adar et al. 2002; Chirita et al. 2005; Ma, Pant & Sheng 2007; Qiu & Cho 2006). Some of these studies follow a similar idea to ours. They use the ODP categories as topics to calculate a set of topic-biased PageRanks, which are used in a personalized search. Following their work, we also used the depth three of the ODP ontology as the dimensions of the topic space.

ODP releases all the data in RDF format. In the RDF file, each of the web pages included in ODP attaches a short description. All the descriptions of the web pages under a category can be merged to create a term vector of the corresponding category. Then the topic vector of a web page can be calculated using the cosine similarity of the category's term vector and the social annotations of the web page. Similarly, the interest vector of a user can be calculated using the cosine similarity of the category's term vector and the social annotations made by the user (Xu et al. 2008).

5.7.2 Samen's Approach

The volume of data on the web has grown significantly in recent years. Hence, it has become increasingly difficult for a user to access the right information in a short time. However, several works have proposed algorithms to re-rank the user's search results on the web by taking into account their profile. (Samen, Ezin & Onana 2017) propose an approach to re-rank users' search results based on a dynamic and hybrid modelling of user profiles. Their approach takes into account the user's interests identified during his browsing session and the history of his search on the web. They use a multi-agent system to collect both explicit and implicit user data and process this data to detect the users' interests represented as ontological concepts. Experiments conducted with their model show that it is able to re-rank user search results with better accuracy than that of the Google search engine.

They present a dynamic and hybrid user profile that is able to learn and adapt user's interests based on data firstly obtained explicitly from the user and by observing user behaviour and mapped to a reference ontology. The user profile consists of four layers: session-based layer, explicit profile layer, short-term layer and long-term layer. Each layer consists of one or more agents that are responsible for a set of tasks. The use of a multi-agent system resides in its ability to address complex problems by dividing it into sub-problems which can be handled by agents.

Their model needs to track user behaviour, add, update and delete user interests and dynamically process explicit user interests. Generally, user interests tend to change. Indeed, a user may lose an interest in an item or a concept in which he was interested in the past. Hence, it is important to detect this change in behaviour and to adapt the user profile to improve the user's information search.

5.7.3 Comparison Results and Evaluation of the Accuracy of Re-ranking Approaches

Ten participants were invited to engage in our experiments. These users were asked to search websites using a set of terms. Users were required to write down the web pages they visited that were of interest.

After this, we asked the participants to place six different queries into the Google search engine and every user selects the results of interest from the first 20 results returned by Google for each query result. We called this collection of pages *the interesting pages data set*. The users also used our model and Samen's model to retrieve the data set of interest for the same queries. The top 5, 10, 15 and 20 links were then compared. The precision for the top 5, 10, 15, and 20 links were calculated to evaluate the performance of all approaches.

We compared the degree of improvement in precision, between the proposed approach, Samen's approach and the Google search results at different top-n results. On the whole, it is shown that out approach, i.e., the personalized system based on the proposed approach achieved high precision at all top-n, especially at top 15 and 20. Tables 5.2, 5.3 and 5.4 show the average results for the proposed approach considering user behaviour (PB), Samen's approach, and the Google search engine. As can be seen from Table 5.2 the proposed approach shows an improvement in results of 0.65, 0.83, 0.9 and 0.91 for the top 5, 10, 15 and 20 links, respectively over the Google search engine and Samen's approach. The Google search engine approach achieves the following results of 0.46, 0.66, 0.73 and 0.77 for the top 5, 10, 15 and 20 links, respectively. Samen's approach achieves the following results of 0.63, 0.8, 0.86 and 0.87 for the top 5, 10, 15 and 20 links, respectively. This demonstrates that the behaviour of the user has an encouraging impact on the search results.

Table 5.2: Results of the proposed approach

Our approach	top 5	top10	top 15	top 20
User 1	**0.7**	**0.833333**	**0.933333**	**0.933333**
User 2	**0.633333**	**0.85**	**0.9**	**0.883333**
User 3	**0.666667**	**0.783333**	**0.916667**	**0.9**
User 4	**0.666667**	**0.816667**	**0.883333**	**0.9**
User 5	**0.7**	**0.866667**	**0.866667**	**0.9**
User 6	**0.633333**	**0.816667**	**0.9**	**0.833333**
User 7	**0.633333**	**0.816667**	**0.916667**	**0.933333**
User 8	**0.666667**	**0.816667**	**0.916667**	**0.9**
User 9	**0.633333**	**0.85**	**0.9**	**0.933333**
User 10	**0.633333**	**0.833333**	**0.866667**	**0.933333**
Average Precision of our approach	0.657	0.83	0.9	0.91

Table 5.3: Results of Samen's approach

Samens' approach	Top 5	Top 10	Top 15	Top 20
User 1	**0.666667**	**0.833333**	**0.883333**	**0.883333**
User 2	**0.666667**	**0.8**	**0.866667**	**0.933333**
User 3	**0.6**	**0.833333**	**0.866667**	**0.833333**
User 4	**0.633333**	**0.783333**	**0.833333**	**0.883333**
User 5	**0.666667**	**0.75**	**0.866667**	**0.9**
User 6	**0.6**	**0.8**	**0.833333**	**0.833333**
User 7	**0.6**	**0.75**	**0.866667**	**0.833333**
User 8	**0.633333**	**0.85**	**0.866667**	**0.85**
User 9	0.6	0.8	0.85	0.9
User 10	**0.6**	**0.8**	**0.866667**	**0.883333**
Average Precision of Samen's approach	0.63	0.8	0.86	0.87

Table 5.4: Results of Google's approach

Google	top 5	top10	top 15	top 20
User 1	0.45	0.7	0.75	0.75
User 2	0.466667	0.633333	0.7	0.7
User 3	0.466667	0.6	0.8	0.716667
User 4	0.5	0.666667	0.75	0.75
User 5	0.466667	0.683333	0.716667	0.75
User 6	0.466667	0.6	0.766667	0.783333
User 7	0.433333	0.633333	0.7	0.783333
User 8	0.5	0.683333	0.683333	0.833333
User 9	0.4	0.666667	0.683333	0.8
User 10	0.4	0.683333	0.716667	0.783333
Avg Precision using Google	0.46	0.66	0.73	0.77

In our experiment, the improvement in precision is noticed when the behaviour of the user is taken into consideration in building the user profiles and is exploited to rank the retrieved results. Figure 5.4 compares the average precision for the results of the personalized search with the other approaches. The results show that our proposed personalized approach based on user behaviour achieves slightly better precision at all top-n. There is an improvement in precision of 0.656666667, 0.828333333, 0.9 and 0.905 for the top 5, 10, 15 and 20 links, respectively whereas the results of the Google search engine and Saman's model were as follows: (0.62667, 0.8, 0.86, 0.87333) and (0.455, 0.655, 0.72667, 0.765) respectively for the same number of users.

Table 5.5 : Average precision for top-n results for the three approaches

Top pages	Our approach	Samens' approach	Google
Top 5	0.656666667	0.62667	0.455
Top 10	0.828333333	0.8	0.655
Top 15	0.9	0.86	0.72667
Top 20	0.905	0.87333	0.765

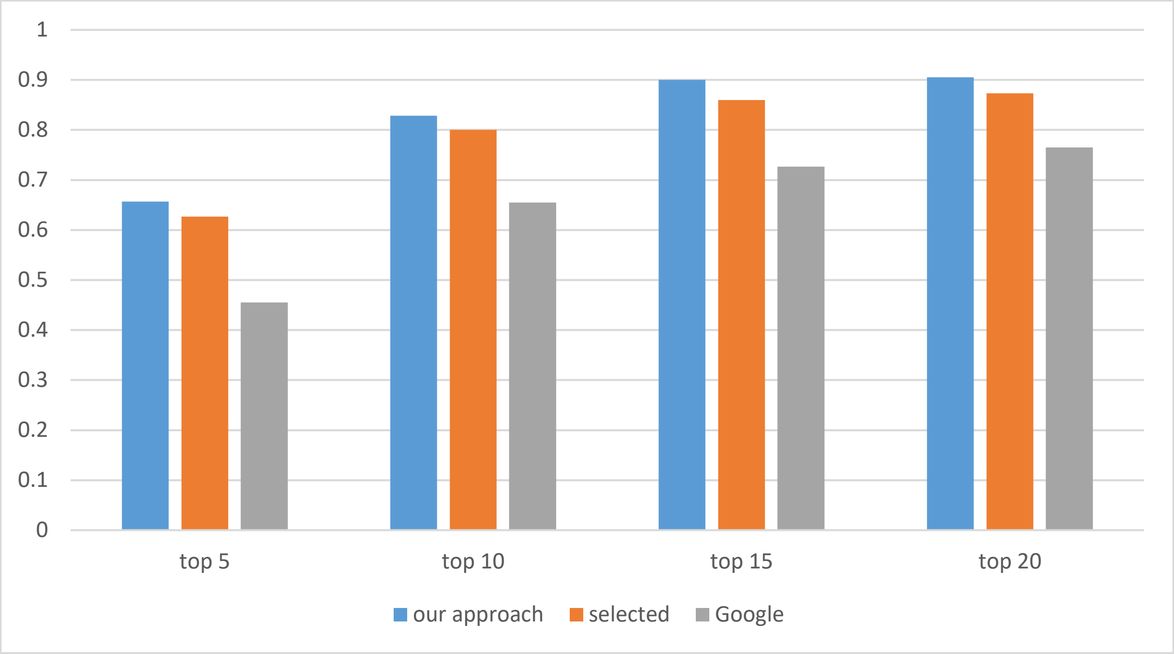


Figure 5.4: Comparison of average precision for top-n results for the three approaches

5.8 Conclusion

Traditional search engines provide the same results for all users if they all enter the same query. Recently, personalized search engines, to take into account users' interests and preferences. Personalized search engines rely on the semantic meanings of words to provide more relevant information efficiently that meets the needs of users with different requirements. The main feature of building a personalized web search is to represent user interests in terms of user profiles. The major shortcomings of the current personalization search approaches as mentioned in chapter two are two-fold as follows(i) they do not use semantic fuzzy classification methods in personalization search, which matches the history of the user with the domain ontolog; (ii) Not much attention has been given to taking into account the browsing behaviour of the user in the process of ontology learning used in search personalization.

Our solution proposes a personalized search approach using an ontology-based user profile. The aim of this approach is to build user profiles based on user browsing behaviour and the semantic knowledge of a specific domain ontology to enhance the quality of the search results. Our proposed approach utilizes a re-ranking algorithm to sort the results returned by the search engine to provide search results that best relate to the user’s query. This algorithm evaluates the similarity between a user’s query, the retrieved search results and the ontological concepts. The similarity is computed by taking into account a user's explicit browsing behaviour, semantic knowledge of concepts, and synonyms of term-based vectors extracted from the WordNet API. The results of the experiments, which are conducted using a case study from the transport service domain, showed that the proposed approach proved its effectiveness and demonstrated promising results. Hence, we found that the use of semantic fuzzy classification is an efficient approach to assign an interest weight to each concept in order to build ontological user profiles. Furthermore, benchmarking experiments with the current best performing personalization approach (Samen, Ezin & Onana 2017) and Google show that our proposed approach achieves higher precision than both the approaches. Taking into consideration user behaviour results in an improvement in search precision.

CHAPTER 6

CONTEXT-AWARE PERSONALIZED WEB SEARCH

6.1. Introduction

Context plays an important role in a wide variety of areas, such as automatic image analysis, computational linguistics and information retrieval (IR) (Vallet et al. 2007). Context is any information that can be used to describe the situation of an entity, such as location, time, or the preferences of an entity, where an entity is a person, place, or object which is relevant to the relationship between a user and a system (Yau & Karim 2004). A context-aware system uses context to provide relevant information and/or services to the user, where relevancy depends on the user's tasks (Hwang et al. 2007). (Allan et al. 2003) defined contextual IR as follows: "search technologies and knowledge about a query and a user's context are combined into a single framework to provide the most appropriate answer to meet a user's information need". In general, context-aware systems are adaptive to the desires of the different users based on their profiles, needs and requirements (Adomavicius & Tuzhilin 2011).

Common web search systems adapt the idea of matching user queries and documents (Schwartz 1998). These systems ignore the large volume of information on the actual users and search contexts. So, when users enter the same query, they retrieve the same results. (Budzik & Hammond 2000) show that the main reason for this is that user context has not been taken into account in the information retrieval process. For example, if a user working in computer science formulates a query about "Apple", documents related to fruit will be incorrectly retrieved. To improve the accuracy of information retrieval systems, an optimal retrieval system needs to effectively model the users' preferences and exploit the contextual information related to their web searches. This chapter presents a personalized web search approach by incorporating user context in the information retrieval process to deliver the search results that most meet the user's need. The most important contributions are the methods of building contextual profiles for users according to their browsing behaviour, such as context history, the semantic knowledge of a specific domain ontology and the technique of re-ranking the original search results according to the contextual profile of the user.

6.2. Contextual personalized web search approaches

Modelling user profiles is important for different types of personalization. A user profile is defined as a digital representation of a user that indicates the preferences and interests of that user. Data on users are collected either explicitly by obtaining feedback from users or implicitly by monitoring the user's behaviour when they are browsing the web. One of the most important representations of a user's interests in a personalized retrieval system is the use of ontology, which is a promising solution for solving word ambiguity and the cold start problem (Baazaoui et al. 2008; Calegari & Pasi 2010). Therefore, many studies have been conducted on context-aware computing in different fields such as mobile applications, recommender systems and information retrieval. In this study, we concentrate on a way to represent the context of the user for enabling personalized web search.

A large body of research has studied context-aware computing in relation to contextual personalized web searches and the different approaches are presented in this section. (Daoud, Tamine-Lechani & Boughanem 2008) described a new approach for learning long-term user interests based on modelling concept-based user contexts identified in related search sessions. They used the depth three of the ODP ontology to represent user contexts. Unlike most of the existing related works, they also focus on learning the user interests as the user context with each user's interest being represented as a set of semantically related concepts. Maintaining the concept weights of the user context is achieved across related search sessions based on a linear combination formula. The results of different experiments show that the re-ranking process of the search results using user context improves retrieval precision.

One of the basic directions for the provision of personalized services in context-aware computing and the utilization of context history was suggested by (Hong et al. 2009). The researchers indicated the need for and the usability of context history which had not been considered in previous research. This research forms the basic direction of design and the development guidelines for context-aware computing systems. Several researchers have studied the problem of using context information in ranking documents in a Web search. (Xiang et al. 2010) carried out an experimental study on real search logs and developed several principles for context-aware ranking. The evaluation process uses human judgment and the user's implicit click data. The experiment outcomes show that the context-aware ranking approach improves the ranking of a commercial search engine.

Furthermore, in (Mohammed, Duong & Jo 2010), the authors investigated an ontological technique to model the user and context for effective personalization. An ontological user profile is built with reference to a widely organized domain concept hierarchy called the Open Directory Project (ODP). The web search history for a specific user is implicitly created and modelled with the reference ontology to build a user profile. A lack of consideration of the users' context history is one of the constraints of the existing approaches. Context history comprises the past context and the users' behaviour in relation to the past context. A personalized web search can be improved by extracting useful web navigation patterns based on context history. For example, if a user submits a query on a weekday related to "travel", the search results will be different to those if he had submitted the same query on a weekend. Furthermore, the results of a query submitted during the day will be also different to the results retrieved from a query which is submitted at night, as it is possible for the user to travel by plane on weekends, whereas he uses trains, taxies and buses on weekdays.

The abovementioned existing works show that using users' search context is promising for the enhancement of personalize web search. However, limiting the user's search context to the current search session, would mean that the context being used is not rich enough. In our approach, we expanded from using a user's current search context to using a user's search contextual history, which was described as "a collection of past context and users' behaviour in relation to the past context" (Hong et al. 2009). The key contribution of our proposed technique is building user profiles taking into account the search contextual histories of the users and the semantic domain knowledge represented in an ontology.

The proposed user contextual profile is used to re-rank the search results returned by the search engine (e.g. Google) to provide the search result that best relate to the user's context. In the following sub-section, the flowchart for the contextual personalized web search methodology is presented and in subsequent sections, the methodology and the steps involved are detailed.

6.3. Contextual personalized web search methodology

The aim of the proposed approach is to provide personalized search results which reflect the context of the user, and to do so, the search results that are returned by a search engine are rearranged based on the degree of similarity between the retrieved results and the contextual profile of the user. This approach consists of three steps.

(a) Firstly, the user's query is received and expanded using the WordNet API.
(b) Secondly, the expanded query is passed to the keyword-based search engine and the search results are retrieved.
(c) Lastly, the retrieved results are re-ranked based on the user's contextual profile and the personalized results are returned to the user.

The proposed approach effectively enhances the overall quality of the search results by offering users the most suitable results with the least search effort.

A fundamental problem is finding a way to construct an effective contextual user profile. The idea of developing the proposed personalized web search approach using an ontology comes from the success of semantic-enhanced approaches to web searches and the variety of user's needs. An ontological user profile is created based on an existing reference domain ontology and a user's context, where the users' interests are closely associated with the relevant concepts in the reference ontology. The contextual user profile is beneficial to the proposed approach in two ways:

(i) the concepts that are extracted from the user profiles may have interest scores, and these scores point out the weight of these concepts to users for that context; and
(ii) the results of the search can be re-ranked by applying a contextual user profile. Figure 6.1 shows the flow diagram of this approach.

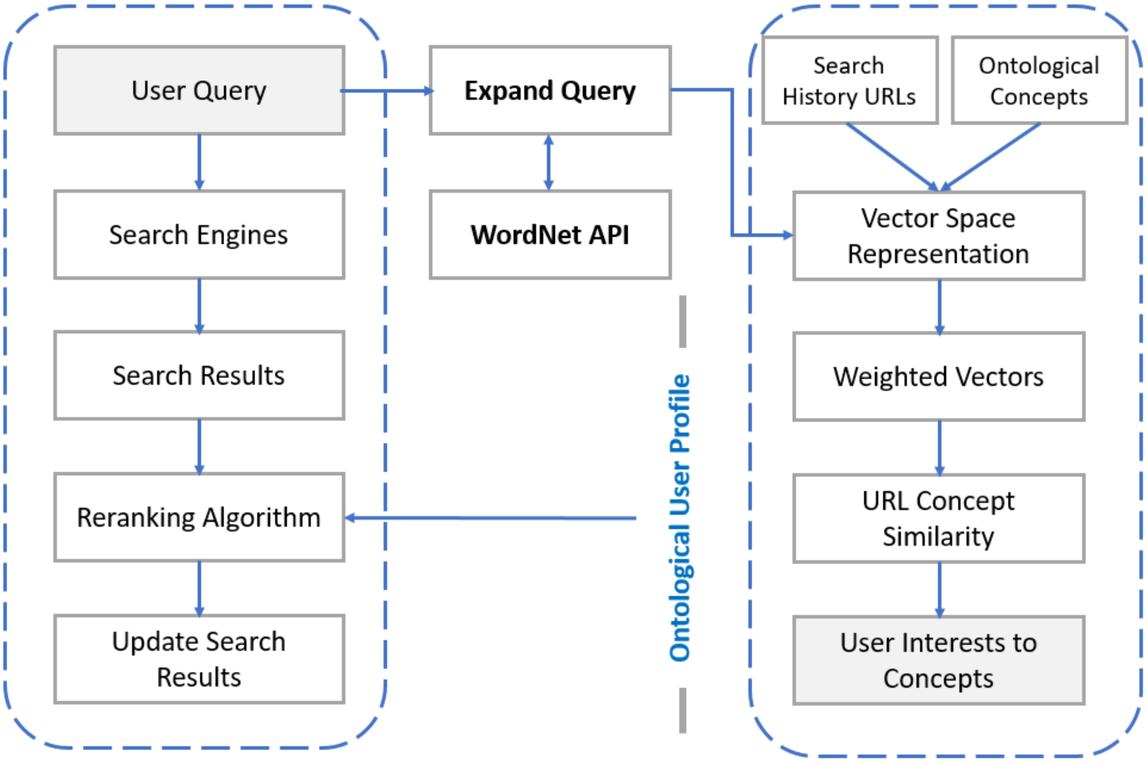


Figure 6.1: Flowchart of procedures for the proposed approach.

6.4 Building an Ontology-Based User Contextual profile

The steps in this model involve constructing a contextual profile, like a structured data log, where the users' related data include their interests and preferences. The preferences and interests of the users are obtained from their search histories or web logs. The primary purpose of using search history is to extract the URLs from the web pages that were visited by users at different times. It is also important to examine the URLs as this step helps to retrieve the metadata, such as the keywords and web page descriptions of the web sites referred by the HTML meta-tag. A personalized web search gives the search results in the same order regardless of the time the query is submitted to the search engine due to the consistency of the ontological domain and its concepts based on the browsing history of the user. In this research, we propose the design of a dynamic ontological user profile to retrieve more precise and personalized results in light of the context of the user, taking into account time as a time property, thus, there is more than one value that represents the weight of the user's interests in a certain context for each concept of the ontological user profile. Figure 6.2 shows an example to simplify the existing concept in the ontological user profile with regard to the interest weights of different users.

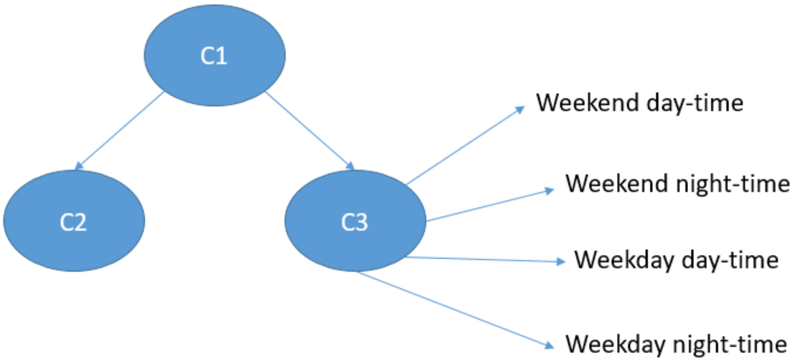


Figure 6.2: Concept values based on user context history

To calculate the interest score values of the concepts of the user, a semantic fuzzy-based classification method is applied. This method adopts the idea of concepts that are overlapping in nature. Hence, the relevance degree between each visited web page and every concept is computed by this technique, then, the results are used to determine the interest scores of the concepts for the user. Figure 6.2 depicts how a user's interest score for concepts is calculated and represented. We assign four values for every concept to represent and model the temporal context of the user, namely weekend day-time (WED), weekend night-time (WEN), weekday day-time (WDD), and weekday night-time (WDN). We extract the temporal context from the browsing history of the user by dividing the search history into these four groups. Based on the search history time, we assign a value to each concept. There are three steps in this process as follows:

Step 1: Intelligently and implicitly find the user's weight for the visited web pages and concepts

A. *Apply an Extended Vector Space Model (EVSM) to demonstrate concepts and web sites.*

When constructing a vector space, the metadata of every web page needs to be used, both keywords and descriptions. The description metadata and the raw text are pre-treated, the stop words are removed and the Porter stemmer is used to stop the word so as to remove the distinct terms. Finally, every web page is shown as a term vector that contains both the distinct terms and the keywords that were removed from the description. Also, every concept found in the reference ontology is shown as a vector of the distinct terms. These terms consist of all the terms, both keywords and the metadata descriptions of all the web pages, that are classified in a specific concept. As previously discussed, the synonyms of terms can help deepen the knowledge which, in turn, helps to define more accurate content. Hence, it is necessary to use WordNet API to expand the vector space of the web pages and concepts to find the synonyms for each term. The synonyms of the terms in the original vectors of the page and concepts to be included are also added. The expanded vector space for a particular web page $\mathrm{p_j}$ is explained as follows:

$$\vec{\mathrm{p_j}} = \left(\mathrm{w_{t_1,j}, w_{t_2,j}, \dots, w_{t_n,j}, ws_{1,j}, ws_{2,j}, \dots, ws_{m,j}}\right) \tag{6.1}$$

where $\mathrm{w_{t_1,j}}$ symbolizes the weight of the initial term located in the web page$\mathrm{p_j}$ vector space, while $\mathrm{ws_{l,j}}$ symbolizes the synonym of a particular term located in the web page$\mathrm{p_j}$ vector space. The expanded vector space for a particular concept $\mathrm{c_i}$ is explained as follows:

$$\vec{\mathrm{c_i}} = \left(\mathrm{w_{t_1,i}, w_{t_2,i}, \dots, w_{t_n,i}, ws_{1,i}, ws_{2,i}, \dots, ws_{n,i}}\right) \tag{6.2}$$

where $\mathrm{w_{t_1,i}}$symbolizes the weight of the initial term in a concept $\mathrm{c_i}$ vector space, while $\mathrm{w_{s_1,i}}$ symbolises the synonym of a particular term in a concept $\mathrm{c_i}$ vector space.

B. Determine the weight values for all the terms in the expanded vector space for both the concept and the web page.

In order to determine the weight of a particular term and its synonyms, the following two equations are used:

$$\mathrm{w(t) = occ(t) + occ(t)\sqrt{1/_{(t)}}} \tag{6.3}$$

$$\mathrm{w(syn_t) = occc(syn_t) + occ(t)}\sqrt{1/_{\mathrm{occ(t)}}} \tag{6.4}$$

where $\mathrm{w(t)}$ symbolises the weight of term t, $\mathrm{occ(t)}$ represents the number of occurrences of term t, $\mathrm{occ(syn_t)}$ represents the number of words that have a similar meaning to term t (i.e. synonyms), $\mathrm{w(syn_t)}$ represents the weight of term t’s synonyms, and $\mathrm{occ(syn_t)}$ represents the number of occurrences of term t’s synonyms.

Step 2: Find the similarity between each visited URL and concepts and then store the result in a similarity matrix as the output.

In this step, cosine similarity is calculated to compute the similarity between the concepts and the visited web sites, which can be extracted from the user profile. The degree of similarity of the expanded weighted vector of a particular web page and vector, is computed, then, the calculated similarity outcomes are saved in a similarity matrix called page- concept or $[\mathrm{P} \times \mathrm{C}]$ matrix., as the output. The following equation is used to calculate the similarity $\mathrm{sim(p_j, c_i)}$ between the two coefficients $\mathrm{p_j}$ and $\mathrm{c_i}$.

$$\mathrm{sim}(\mathrm{p_j}, \mathrm{c_i}) = \cos(\overrightarrow{\mathrm{p_j}}, \overrightarrow{\mathrm{c_i}}) = \frac{\overrightarrow{\mathrm{p_j}}.\overrightarrow{\mathrm{c_i}}}{\|\overrightarrow{\mathrm{p_j}}\|_2 \times \|\overrightarrow{\mathrm{c_i}}\|_2} = \frac{\sum_{\mathrm{t_k} \in \mathrm{T_{p_j,c_i}}} \mathrm{w_{p_j,t_k}} \times \mathrm{w_{c_i,t_k}}}{\sqrt{\sum_{\mathrm{t_k} \in \mathrm{T_{p_j,c_i}}} \mathrm{w^2_{p_j,t_k}}} \sqrt{\sum_{\mathrm{t_k} \in \mathrm{T_{p_j,c_i}}} \mathrm{w^2_{c_i,t_k}}}} \tag{6.5}$$

where $\overrightarrow{\mathrm{p_j}}.\overrightarrow{\mathrm{c_i}}$represents the dot-product between the vector $\overrightarrow{\mathrm{p_j}}$ and the vector$\overrightarrow{\mathrm{c_i}}$, $\mathrm{T_{p_j,c_i}}$ represents the group of terms which are found in the two vectors $\overrightarrow{\mathrm{p_j}}$ and$\overrightarrow{\mathrm{c_i}}$, while $\mathrm{w_{p_j,t_k}}$ is the weight of $\mathrm{t_k}$ term in $\mathrm{p_j}$ vector, and $\mathrm{w_{c_i,t_k}}$ is the weight of $\mathrm{t_k}$ term in the $\mathrm{c_i}$ concept. To calculate the degree of similarity between a concept and a visited web page, the angle between the two vectors has to be found first, and to compute the degree of similarity the angle's cosine is used. The results of the cosine similarity range from 0 to 1. Thus, a high matching value between a web page and a concept indicates high similarity.

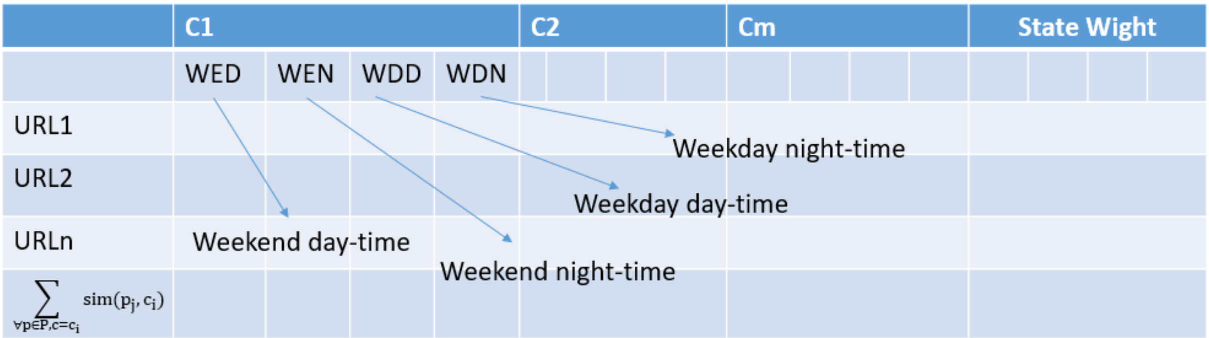
Step 3: Find a user’s interest score for a specific concept

In this phase, the user’s interest score is calculated considering concept $\mathrm{c_i}$ using the following equation:

$$\mathrm{Context\_Weight(c_i)} = \sum\nolimits_{\forall \mathrm{p} \in \mathrm{P}, \mathrm{c}=\mathrm{c_i}, T=ti} \mathrm{sim(p_j, c_i)} / \sum\nolimits_{\forall \mathrm{c} \in \mathrm{C}} \sum\nolimits_{\forall \mathrm{p} \in \mathrm{P}} \mathrm{sim(p_j, c_i)} \tag{6.6}$$

where P represents the set of all URLs that were visited by a user in the user profile, C represents the set of all concepts found in the reference ontology, $\sum_{\forall \mathrm{p} \in \mathrm{P}, \mathrm{c}=\mathrm{c_i}, T=ti} \mathrm{sim}(\mathrm{p_j}, \mathrm{c_i})$ indicates the sum of all the relevance degrees of pages in relation to a particular concept $\mathrm{c_i}$ at time t_i, $\sum_{\forall \mathrm{p} \in \mathrm{P}, \mathrm{c}=\mathrm{c_i}, T=ti} \mathrm{sim}(\mathrm{p_j}, \mathrm{c_i})$ is the sum of all the relevance degrees of pages in relation to all the concepts at time t_i.

The user-concept interest score matrix [U×C] represents the results that are obtained from this step, as illustrated in Figure 6.3, where each entry in the matrix represents the user's interest score for a concept $\mathrm{c_i}$ at a specified time.



	C1				C2				Cm				State Wight			
	WED	WEN	WDD	WDN												
URL1																
URL2																
URLn																
$\sum_{\forall \mathrm{p} \in \mathrm{P}, \mathrm{c}=\mathrm{c_i}} \mathrm{sim}(\mathrm{p_j}, \mathrm{c_i})$																

Weekday night-time
Weekday day-time
Weekend day-time
Weekend night-time

Figure 6.3: Matrix of user-concept interest score

It is clear that different users do not have the same degree of interest in all URLs that are in their search history. For example, they may open a page by mistake and leave it quickly, or they may browse pages for a very short time without any real interest in them. Hence, we introduce the term “weight status” for each URL in the search history of a user to capture the extent of this user’s interest in the web page. This weight status can be represented as one of the following three status options: being a *favourite*, *browsing* and *searching* for results. The weight status is determined based on the user’s actions when they are on the web page. The “favourite” status refers to the action of either “printing” or “saving” a web page; the browsing status refers to the action of “browsing” a web page; the searching for results status refers to the action of entering this web page through a link from the query results. To make use of this weight status, it is essential to compute the interest score of a user based on the concepts found in the web page. So, a value is pre-defined for each status option as follows: a numerical value of 1 is assigned for favourite; 0.75 for searching and 0.5 for browsing.

By multiplying the similarity degree of c_i and p_j with the corresponding weight status value of p_j, the weight of each entry of the similarity matrix [P×C] is computed, as explained earlier in step 2. Then, equation (6.6) is used to calculate the interest scores of the concepts of the user on the basis of the weighted similarity. The novelty of this approach stems from its use of context history in determining the user's concept interest scores as the previous works on personalized web searches do not address context history when designing the ontological user profile.

6.5 Personalized Search results using a Re-ranking algorithm

By re-arranging the results returned by the search engine for a given query, the personalized search results, which are referred to as the original search results, are retrieved in light of the contextual profile of the user. The re-ranking procedure is described in Algorithm 6.1. For a particular query, the input for the re-ranking procedure includes the user’s contextual profile and the original search results for that query. The first step is the representation of each URL, which is the returned result, in a weighted vector which is then extended as shown in step 1. After this, equations 6.3 and 6.4 are used to calculate the weight values of the terms. In the second step, the similarity matrix $[\mathrm{P} \times \mathrm{C}]$ is used to determine the highest similarity concept for every returned URL. The last step calculates the user's interest score for every URL found in the returned search results, using the following equation:

$$\mathrm{IS}(\mathrm{URL_j}) = \mathrm{sim}(\mathrm{Q}, \mathrm{URL_j}) \times \mathrm{sim}(\mathrm{Q}, \mathrm{c}) \times \mathrm{Context\_Wight}(\mathrm{c}) \qquad (6.7)$$

where $\mathrm{sim}(\mathrm{Q}, \mathrm{URL_j})$ represents the similarity between $\mathrm{URL_j}$a web page in the original search result and Q the query, $\mathrm{sim}(\mathrm{Q}, \mathrm{c})$ represents the similarity between concept C and Q the query with the highest match with $\mathrm{URL_j}$, and$\mathrm{Context_weight}(\mathrm{c})$ represents the interest score of concept c in a particular context that has the highest match with $\mathrm{URL_j}$. Cosine similarity as explained in equation 6.5 is used to compute the similarity in equation 6.7.

Algorithm 6.1: Re-ranking of the retrieved search results based on contextual profile

Input: ontological user profile for a target user (u), set of retrieved search results for a query (Q)

Output: re-ranked search results, i.e. personalized results for the target user (u)

```
Process:

        Set $C = \{c_1, c_2, \ldots, c_m\}$
        Set$URL = \{URL_1, URL_2, \ldots, URL_n\}$
For each$URL_j \in URL$
c= find the highest matching concept to $URL_j$
                compute $sim(Q, c)$

                compute $sim(Q, URL_j)$
                  $IS(URL_j) = sim(Q, URL_j) \times sim(Q, c) \times Context\_Wight(c)$
        End For
 Sort URL based on interest score IS
```

6.6 Experiment Description

The proposed approach and the newly developed techniques are implemented in a prototype for the transport service domain in Australia. The prototype results are compared with the results from the ranking scheme for Google, being the most commonly used search engine. The search results quality is measured according to the precision of different top-n results.

Reference domain ontology

The web pages of various transportation companies were examined to obtain the reference domain ontology that was used in the prototype, as well as to utilize the domain ontology on the basis of the work done in (Dong, Hussain & Chang 2011). In addition, Yellow Pages business web sites were studied to obtain the metadata of the service description entity (SDE). Two sets of concepts are related to the service ontology, grouped according to the basis of the metadata of SDE: actual concepts and abstract concepts. The actual concepts are related to the actual services, such as the transportation web sites, which means that the concepts of a real service may connect to the SDE metadata, while the abstract concepts represent both the abstract domain of the service concepts in addition to their sub-domain. These concepts have two significant features: service name and service description.

Data Acquisition

Ten participants took part in the experiment. These participants (users) were asked to search the Internet and browse web pages in the transportation domain using a given set of terms for eight weeks. Then, the users were asked to record the web sites they visited and in which they were interested. To capture the user behaviour on the web pages and to help build the user profile, we used Chromium browser automation software, which records all the browsing behaviour of the users. The users were also asked to submit different queries from the set they selected before submitting them to the Google search engine. It is important to capture the querying time to decide when to re-rank the results and which context value to use. For each query result, every user choses the results of interest from the first 30 web pages in the original search results returned by Google. These selected web pages form the data set of interesting pages. These participants are also asked to utilize the prototype to retrieve the data set of interesting pages for the same queries. The top 5, 10, 15 and 20 links for these two sets of interesting pages are then compared in terms of precision.

i. Experiment Cases

To evaluate the effectiveness of the proposed approach, three experimental cases were examined. Case 1 is a base case in which the users use the Google search engine to run the queries. The results are the original search results. Case 2 is the first test case in which the original search results are re-ranked using the prototype with the option of "Personalized Without Context (PWC)". The second test case is case 3, where the prototype with the option of "Personalized with Context (PC)" is used to re-rank the original search results.

It is worth mentioning that the prototype was implemented and presented based on our previous work (documented in Chapter 5) which has proved that the techniques of capturing a user behaviour and building a user profile using such a user behaviour are superior over the related competing work. Therefore, this work starts with a superior personalised web search system (Case 1) and expanded it by adding the factor of users' search contextual histories in building the user ontological profile (Case 3) with the aim of evaluating the effectiveness of using a user's search contextual history.

ii. Overall Performance Evaluation

Two sets of comparisons were undertaken to evaluate the overall performance of the proposed approach and the related methods. Comparison 1 evaluates the degree of improvement in the average precision of the top-n results obtained by the Google search engine at different top-n and our proposed approach. In general, the results show that the personalized system based on the proposed approach achieved high precision at all top-n, and particularly at top 5 and 10. Figure 6.4 presents the average precision of the proposed approach without taking user context into consideration (PWC), the average precision of the proposed approach with user context (PC), and the average precision of the Google search engine. It can be seen from Figure 6.4 that our proposed approach with (PC) achieves an improvement in precision of 35%, 29%, 22% and 17% for the top 5, 10, 15 and 20 links, respectively over the Google search engine. The proposed (PWC) approach demonstrates an improvement in precision of 22%, 19%, 18% and 14% for the top 5, 10, 15 and 20 links respectively over the Google search engine.

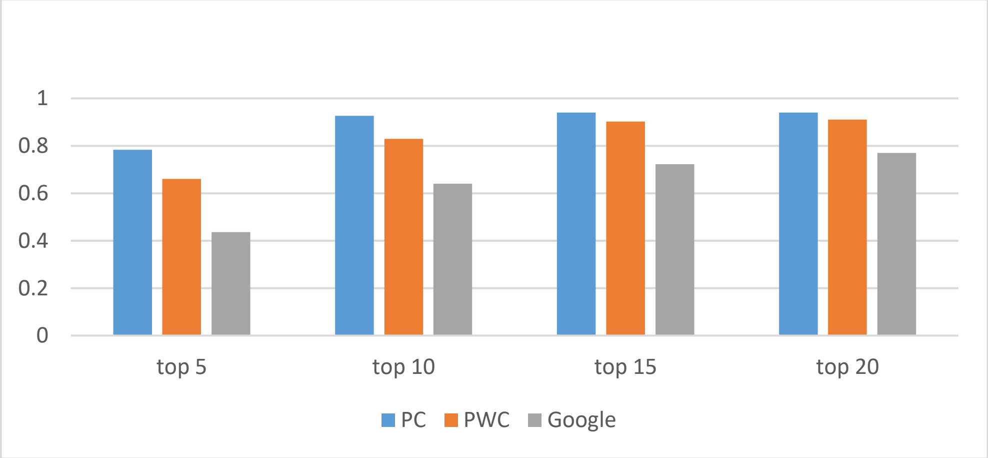


Figure 6.4: Average precision for top-n results

Comparison 2 evaluates the difference in precision in the process of building the contextual user profile with and without considering user context history. There is a noticeable improvement in precision when the users' context is included in the construction of the user profiles and is used to rank the retrieved results. Figure 6.5 shows the average precision for the personalized search results with and without considering user context. The results show that the proposed personalized approach based on user context achieves slightly better precision at all top-n. There is an improvement in precision of 12%, 9%, 0.04% and 0.03% for the top 5, 10, 15 and 20 links, respectively as shown in Figure 6.5

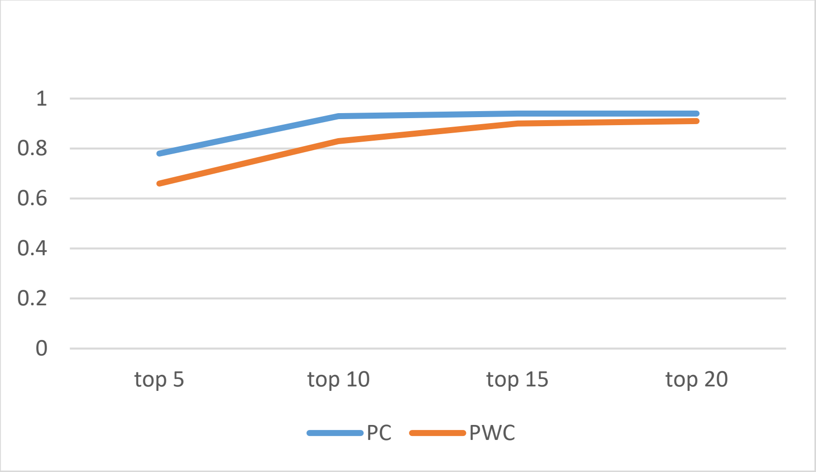


Figure 6.5: The effect of user context on precision

6.7 Conclusions

Most of the commonly used web search engines use the one-size-fits-all approach to the delivery of search results and only consider the input query. Hence the same results are provided to all users for any given query. This approach cannot provide the most desirable information to the users. Personalized search results based on the users' interests and preferences are highly effective in meeting the users' needs. One of the key problems in achieving personalised web search results is ensuring the accurate representation of user context and making use of it in retrieving the search results.

The solution presented in this thesis for effective personalised web search approach incorporates user context in the information retrieval process to provide the user with the most highly related results. The key contributions are the technique of building the contextual user profiles based on the user's browsing behaviour and semantic knowledge of the specific domain ontology and the algorithm to re-rank the original search results based on the user's contextual profile.

In this chapter, a personalized search results approach is proposed using a developed context-aware ranking model based on contextual user profiles. The user profile model provides an extensible model that utilizes context history for a well-defined user search pattern. The search history of a user is collected implicitly as the user's temporal context, and a dynamic user profile is generated by modelling the profile based on the reference ontology. We also consider user's behaviours in browsing web pages, and use the web page's weight status value when constructing user profiles. The results of the experiments in which a prototype is used in the domain of transportation services shows that the proposed approach and the new techniques are highly effective, achieving a maximum improvement in precision of 35%.

CHAPTER 7

PERSONALIZED QUERY EXPANSION

7.1. Introduction

The Internet has had significant impact on everyday lives, not only because of its wide usage but also due to the activities that it enables, such as interaction, business, marketing etc (Colledge & Barnes 2011). One of the significant advantages of the Internet is the availability of a huge amount of information on all fields, from education and medicine to sport and space exploration. This massive amount of information has been an issue of consideration by researchers in terms of how to make the process of information retrieval faster, easier and more accurate (Larson 2010). Internet users seek to obtain precise information related to their search queries such as searching for particular documents or information. However, if someone wants to retrieve information from the Internet, they need to know the exact web address which may not always be possible.

With the increasing volume of data published on the Internet, it is difficult for search engines to return accurate results that meet the quality of service requirements of the users. This in turn has prompted researchers to use specialized techniques to improve the search process. Query Expansion (QE) is one such technique. Query Expansion is a technique to reformulate a seed query and enhance it by adding suitable additional terms to enhance the performance of information retrieval (Gan & Hong 2015). Researchers have developed query recommendation methods that suggest the most related query to the initial query to get better and precise search results (Vidinli & Ozcan 2016).

Another issue associated with QE is vocabulary, which varies from author to author. Another problem is that some user queries do not include the terms used in the documents for which they are searching. Furthermore, some web users search for information from search engines using a very short query of one or two words (Wen, Nie & Zhang 2001). The accuracy of user queries depends on the number of words in the query, as queries of only a few words have a negative impact on the query results (Blanco, Ottaviano & Meij 2015). Hence, QE has been investigated to overcome these problems and to enhance information retrieval in terms of time and accuracy (Curé et al. 2015) Figure 7.1 shows the general principle of query expansion.

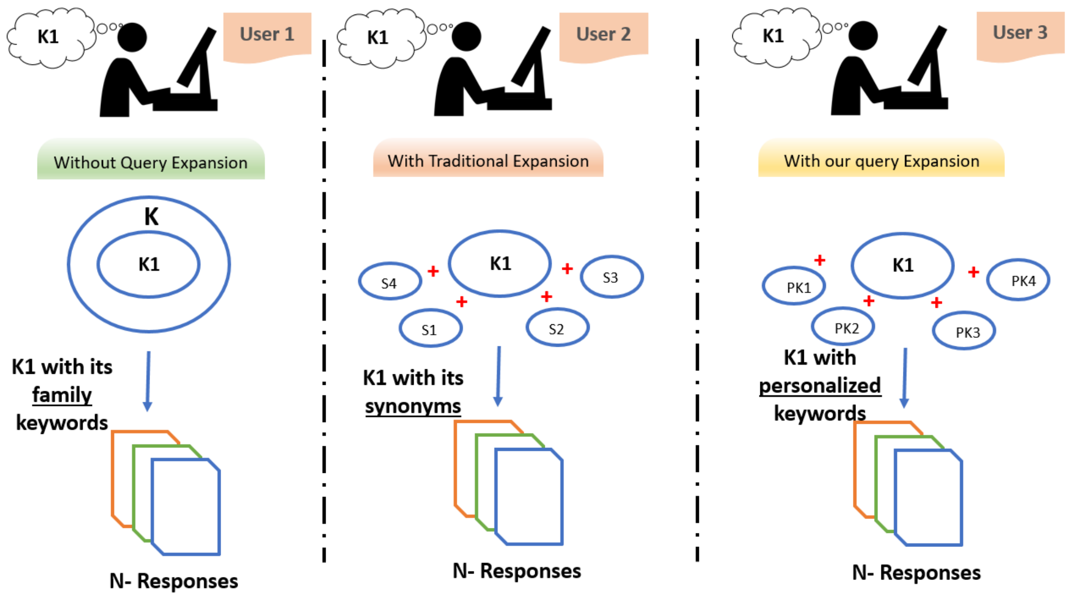


Figure 7.1: Basics of query expansion

According to (Shekarpour et al. 2013) , there are many query expansion techniques, some of which are as follows:

- Find and search for synonyms. For example, the words *little, few and small* are synonyms so if any of these are used in a query, the retrieved outcomes should be very similar.
- Find all the different morphological forms of words, where the word and its derivatives will be retrieved as one word (these derivatives are generated by adding prefixes and suffixes to the stem word). For example, when the query contains any of the words *teach, teaching or teacher*, the same result will be retrieved.
- Fix spelling errors by providing suggestions to correct them, known as auto-correction. For example, if we incorrectly write *Australea*, in the query it is automatically corrected to *Australia*.
- Reformulate the original user queries in light of the users' habitual behaviour in searching, based on web logs.

7.2. Challenges of query expansion

The traditional approach to query expansion is to expand the query with additional terms that are extracted from a variety of information sources, such as a thesaurus-like WordNet, or from top-ranking documents. However, these traditional techniques have a limitation in that they do not include the user's information, such as user's alternatives and choices when having many documents to extract the terms of expansion (Dalton, Dietz & Allan 2014).

To obtain the best query results using information retrieval systems, exact query terms are required to return a high-quality results (Curé et al. 2015). Consequently, a user who tends to select very general keywords when querying will receive hundreds of thousands of irrelevant documents. This kind of user behaviour is an obstacle as these users may be unsure about the nature of the content they need or find it difficult to describe the nature of the information they are looking for in a few concise words.

One common reason that contributes to the lack of effectiveness of information retrieval systems is vocabulary mismatch, which is a fundamental problem in information retrieval (Elkateb 2014). Moreover, Natural language classified vocabulary mismatch as a common phenomenon as different words (polysemes or synonyms) are often used for similar concepts or items (Elkateb 2014). The term *polysemy* refers to the same word with different meanings, such as the word *orange*, which could be a fruit or a colour. *Synonyms* are words which have similar or nearly the same meaning as another word or other words, for instance, the words *goal* and *objective*.

7.3. Personalized Query Expansion techniques

In the context of search engines, QE involves evaluating the input of users and increasing the result of the querying process to find a match with relevant documents. There are many methods which are used for query expansion to overcome various obstacles, such as finding synonyms, finding different forms of words and auto correction. Many researchers have studied query expansion for different reasons and objectives. Some aimed to expand the user's original query with other query elements of similar meaning in order to increase the opportunity of retrieving appropriate resources, based on semantic and linguistic inference over linked open data (Shekarpour et al. 2013). Furthermore, it is important to provide high-level recommendations by using specific information from the query logs, such as the users' previous queries and documents they have clicked on before. This approach uses a query recommendation algorithm to categorize the queries according to their degree of relevance to the input query (Dali, Fortuna, Tran, et al. 2012). A novel QE technique proposed by (Liu et al. 2017) utilized fuzzy rules and a word embedding similarity calculation to model the user's interests, based on context-related words instead of synonyms. Using fuzzy rules reinforces the accuracy of document ranking. A ranking pipeline to enhance the search of biomedical datasets was also introduced in (Teodoro et al. 2017) using a variety of novel techniques and methods. This method comprises the following steps: calculating the weight of the document terms, computing the probabilistic relationships between the document terms and query terms, and calculating the document term's cohesion weight through merging the probabilistic of entirely terms of query (Dali, Fortuna, Tran, et al. 2012). Four QE techniques are used (Kathuria, Mittal & Chhabra 2017). The first is network-based query expansion, which is based on conceptual semantic theories (Hoeber, Yang & Yao 2005). The second is term weighting. When applying pseudo-relevance feedback in this technique, the retrieval performance of the MAP criterion is increased by 7% (Hahm et al. 2014). Word sense disambiguation (WSD) is the third query expansion technique (Hahm et al. 2014) and the fourth is the fuzzy logic-based approach (Singh & Sharan 2017). Some measured the effectiveness of linguistic and semantic features separately to determine which of these features are the most effective in query expansion. To achieve the proposed approach, several steps were followed: (i) extract all the related words from the original keyword based on linguistic and semantic features; (ii) compute the weight of the linguistic and semantic features by utilizing a weighting scheme in order to detect the effectiveness of each feature and to remove ineffective features; and (iii) compute and assign a relevance score to the related words to balance between precision and recall (Shekarpour et al. 2013). For the ranking pipeline to boost the search of the biomedical datasets, a QE model using a word embedding algorithm was applied to measure the similarity of query terms and a classification method was used to enhance the rank of datasets matching the query constraints (Teodoro et al. 2017).

Several interesting results have emanated from the existing research showing that the novel QE technique greatly surpasses other traditional methods, such as BM25, in terms of document ranking performance. Also, query expansion using the embedding method outperforms the traditional lexical methods, such as query expansion-based WordNet and QE-WordNet. Moreover, the results indicate that query expansion using fuzzy rules achieved better performance than the word embedding models that do not use fuzzy rules (Liu et al. 2017). It is clear that the results of a query which consider terms are more significant than the results of a query which consider the whole document (Teodoro et al. 2017). The increasing use of recommended query results in increasing for the short queries precision (Dali, Fortuna, Duc, et al. 2012).

Therefore, based on the previous work, we find that query expansion should be based on a large number of documents using more ways to extract features from these documents and to reduce the dimensionality that occurs because of the existence of a large number of documents and features. To address this gap, we propose a new technique for personal query expansion. The construction of a query occurs intelligently and implicitly without any intervention from the user. This approach involves finding similar texts for a particular user from a large number of documents.

7.4. Personalized Query Expansion Methodology

Over the past number of years, researchers have studied the query logs saved in web servers with web mining methods in order to improve the usability and effectiveness of search engines (Rieh 2006). These kinds of studies aim to enhance the performance of search engines by, for example, providing recommendations, categorizing queries, targeted advertising, ranking and many other improvements. Logs contain a large number of queries. In the following, the related document to one query should be considered (Cramer et al. 2013). In the proposed model, several experiments are conducted based on a subset of the whole databases. If a couple of queries share similar relevant documents, this refers to the relationship between these couple of queries in some manner, and the linked queries terms could be considered as participant terms for query expansion, in addition to the relevant documents terms for related queries could be used in order to expand the related query.

This chapter proposes a new technique for personal query expansion. The construction of a query has been occurred implicitly without any interventions from the user. This approach finds similar texts for a particular user from a large number of documents. This approach reduces the query search time and reduces the effort required from the users by extending the query terms. To build this model, there is a need to collect some documents manually. Then, we are able to investigate the performance of the approach. The proposed method comprises the following five steps: gathering the corpus, pre-processing, term weighting, dimensionality reduction and expansion.

Step 1: Data collection

The proposed method uses a manually collected corpus, consisting of 950 documents and each document contains the meta description for the URLs that were opened by users. This corpus relates to 19 users, with 50 queries for each user and the number of words in the corpus is 2767. To obtain the meta description for the users, we used the BuzzStream tool from the link (http://tools.buzzstream.com/meta-tag-extractor). The description of the data is shown in Table 7.1.

Table 7.1: Our dataset description

Component	Number
Attributes	2767
Instances	950
Distinct label	19
Text file/ distinct label	50
Sum of weights	950
Labels	Art, Biology, Radiology, Graphic Design, History, Optics, Journalism, Computer Science, Physics, Music, Biomedical Physics, Geography, Geology, Medicine, Drawing, English, Math, Tourism Sport.

Step 2: Pre-processing

This section details the steps for the proposed query building technique:

1. **Tokenization** (split the dataset into terms using white space, tab key or enter key) these terms are called tokens.
2. **Digits removal**: This step removes the numbers.
3. **Punctuation removal**: In this step, all punctuation marks (? ,: ,. , ; etc.) and symbols ((,*< , #, etc.) are removed.
4. **Stop words removal**: In this step, a set of words that is not meaningful is extracted from each query. We used the stop word list that was collected by Gerard Salton and Chris Buckley for the empirical SMART Information Retrieval system (SMART stop word list), which is available at the following link: http://www.lextek.com/manuals/onix/stopwords2.html.

5. **Stemming**: A morphological analysis step to remove prefixes and suffixes and the remaining part is called the stem. Figure 7.2 shows an example of text pre-processing.

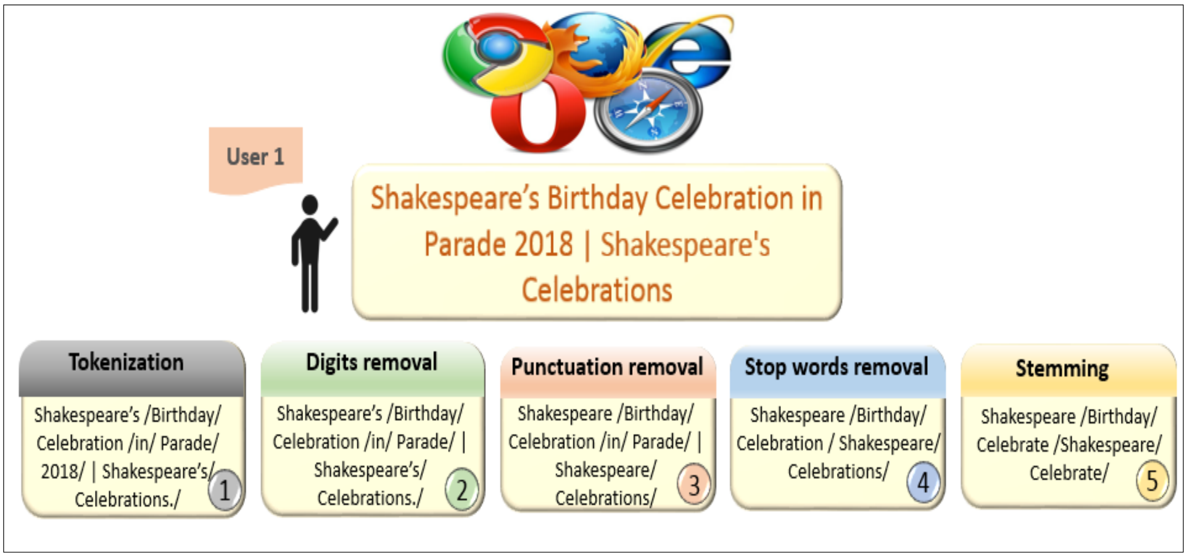

User 1
Shakespeare's Birthday Celebration in Parade 2018 | Shakespeare's Celebrations
Tokenization
Shakespeare's /Birthday/ Celebration /in/ Parade/ 2018/ | Shakespeare's/ Celebrations./
1
Digits removal
Shakespeare's /Birthday/ Celebration /in/ Parade/ | Shakespeare's/ Celebrations./
2
Punctuation removal
Shakespeare /Birthday/ Celebration /in/ Parade/ | Shakespeare/ Celebrations/
3
Stop words removal
Shakespeare /Birthday/ Celebration / Shakespeare/ Celebrations/
4
Stemming
Shakespeare /Birthday/ Celebrate /Shakespeare/ Celebrate/
5

Figure 7.2: Example of a user query with pre-processing steps.

Step 3: Feature Extraction

The proposed method uses the following three algorithms for the feature extraction phase from the original documents: the weights' term frequency (TF), term frequency-inverse document frequency (TF-IDF) (Xu & Croft 1996) and term frequency-inverse document frequency class frequency (TF-IDF-CF) (De Silva & Haddela 2013) to represent the data in statistical form. The results are represented as a matrix, called the mother matrix, as shown in Figure 7.3.

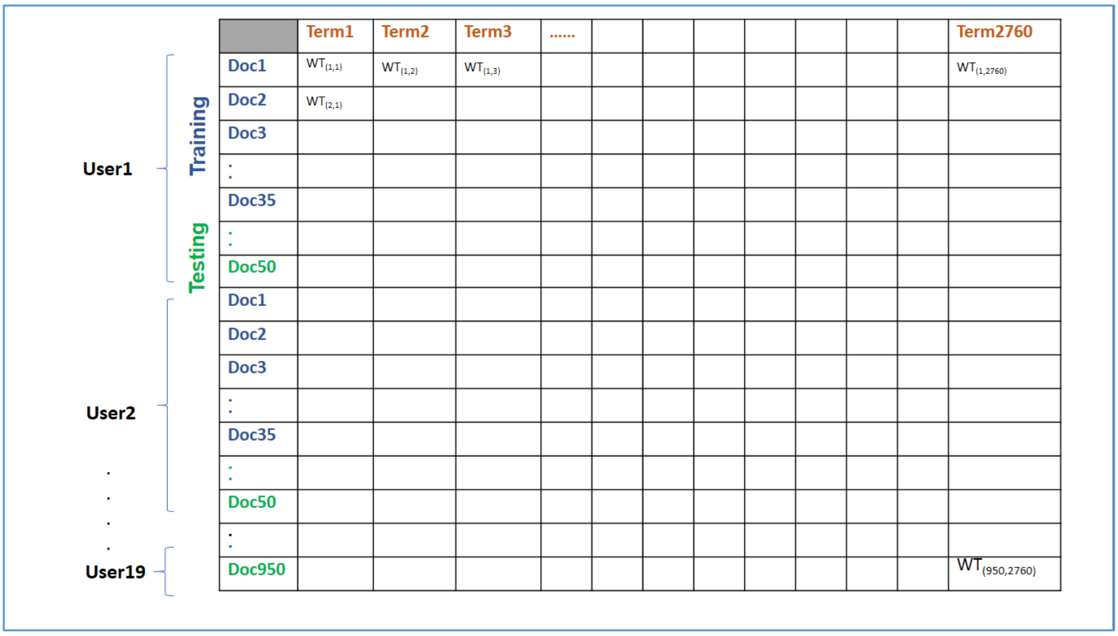
We create the matrix based on the intersection between documents and terms, where the terms are the columns (terms = 2760) of the matrix and the documents are the rows of the matrix, the total number of rows being 950. The term weight is the value of the intersection between each document and row. This term weight can be computed using different methods, so, we have a vector for each document. Each document is represented in a document vector $v_i = \{v_1, v_2, \dots v_{950}\}$, where 950 is the number of documents. The terms are represented as the vectors of terms $t_k= \{t_1, t_2, \dots t_{2760}\}$. Algorithm 7.1 shows the steps in constructing the mother matrix.

Algorithm 7.1: Feature extraction

Input: texts file

Output: mother matrix.

1. Create the matrix (documents and terms)
2. Let the # of columns or terms = 2760
3. Let the # of rows or documents = 950
4. Each document = document vector vi = {v1, v2, … v950},
5. The terms = vector of terms tk= {t1, t2, … t2760}.
6. The term weight = The intersection between each document and columns.


User1
Training
Testing
User2
.
.
.
.
User19

	Term1	Term2	Term3	......								Term2760
Doc1	$WT_{(1,1)}$	$WT_{(1,2)}$	$WT_{(1,3)}$									$WT_{(1,2760)}$
Doc2	$WT_{(2,1)}$											
Doc3												
:												
Doc35												
:												
Doc50												
Doc1												
Doc2												
Doc3												
:												
Doc35												
:												
Doc50												
:												
Doc950												$WT_{(950,2760)}$

Figure 7.3: The mother matrix for documents and terms

Term Weighting Techniques: The main function of the term weighting approach is to measure the term’s weight in a document. This work uses three term weighting techniques: TF, TF-IDF, and TF-IDF-CF (Ren & Sohrab 2013), that are constructed using two methods. The first method is local weighting (LW). LW can be defined as the number of repeated terms in a document and it specifies the term's degree of importance to a document. The second method is global weighting (GW). This method applies a logarithm to represent the number of queries shown in the text (Xia & Chai 2011). We explain the methods below.

A. **Local term weighting** (term frequency)**:** This method calculates the actual weighting of terms locally within a specific text regardless of its weight in the global range (Sabbah et al. 2017). The expansion of a query file relies on the frequency of terms, where TF (fr, t) refers to the frequency of term (*fr*) in text (t). The weight of the frequency can be computed by distributing the term frequency through the all-out term frequency in a text file. Equation 1 shows the process of computing term frequency:

$$TF(fr,t) = \frac{fr,t}{Max(fr,t)} \tag{7.1}$$

where *f* indicates the number of frequency and *t* represents the text at a specific point.

B. **Global term weighting:** This method calculates the weighting of terms globally, taking into account all the existing terms (Cuellar, Diaz & Ponce-de-Leon-Senti 2015).The importance of this scheme lies in the fact that it gives an accurate weight of the term because it is looking for finding the ability to repeat a certain term in a specific text or category, taking into account the lack of existence of the same term in other texts. Global term weighting has two levels: the document level which is called inverse document frequency (IDF), and the class level which is called term frequency-inverse document frequency class frequency (TF - IDF - CF).

C. **Document-level frequency: Inverse Document Frequency (IDF):** This refers to a case when a term appears in one text but does not appear in every text *TFr,t* is the number of occurrences of term *r* in each text separately, whereas *TF-IDF* computes the percent of occurrences in a certain text and the percent of occurrences in all other texts, and *IDF* indicates the non-occurrence of the term in the rest of the texts (Singhal, Buckley & Mitra 1996). Equations 7.2 and 7.3 calculate IDF and TF-IDF respectively:

$$IDF\ =\ log_{10}(\tfrac{N}{n}) \tag{7.2}$$

According to equations 1 and 2, TF-IDF can be computed as in equation 3

$$TF - IDF = TF(t,q) * log\frac{N}{n} \tag{7.3}$$

Document-level methods for term extraction suffer from several weaknesses, as this method requires long documents (Kathuria, Mittal & Chhabra 2017). Moreover, it is not necessary that the term of high *TF-IDF* is associated with the query of the user, and it doesn't have the ability to catch the semantics, nor can it take the context of the query into consideration. Thus, we implement the class frequency level as described in the next subsection.

D. Class Frequency Level: Term Frequency-Inverse Document Frequency Class Frequency (TF-IDF-CF):

We use the TF-IDF-CF method utilizing class (personal) frequency. This method is calculated by Equation 7.4:

$$TFIDF - CF = \log\left(tf_{t,d} + 1\right) * \log\left(\frac{N+1}{n_t}\right) * \left(\frac{n_{pt,d}}{N_{pt}}\right) \qquad \text{Equation 7.4}$$

Where $n_{pt,d}$ denotes the number of texts where query d appears related to person p. N_{pt} denotes the number of texts related to a query. This technique gives the term weight relative to a certain class compared to other classes. The terms with higher frequencies for each class (person user) are selected for the expansion stage.

Step 4: Dimensionality Reduction

After the implementation of the previous step, we produce a matrix whose column headers are the texts labels (T1, T2 …. Tm), and the row headers are the words that make up the corpus. Each element in this matrix represents a statistical value (feature) produced by one of the methods of representing the term weight of the previous three methods as a result of the previous step, hence we have a high dimensionality matrix with size of number of words x a number of text files.

The obtained matrix is called the mother matrix or the origin matrix, which is difficult to deal with as it is, so it must be re-represented in a different manner with a lower dimension, because the large size negatively affects the performance of the system in terms of time and accuracy. To solve this problem, we use singular value decomposition (SVD) as one of the dimension reduction techniques or a feature selection algorithm, which selects a subset of the proper features to improve system performance. Thus, the main goal of these algorithms is to re-represent the mother matrix with three child matrices to reduce the size of the dimensions, exclude repetitive or non-system features and reduce the number of features required to learn the system, hence reducing the complexity of the calculations required for the system.

SVD is a powerful mathematical concept that deals with dimensionality reduction, based on arithmetic operations in matrices. To have discriminative features, it is preferable to use SVD over the original matrix in calculations (Xu, Zhang & Hu 2007).

SVD is used for several reasons. Firstly, SVD deals with dimensionality reduction of matrix A. Secondly, SVD can be a number of matrices that assigns data with numerically different representations while preserving the semantic meaning. Thirdly, SVD offers multiple solutions to problems related to least-squares and it also has the ability to handle certain cases related to singular or near-singular matrices.

Also, SVD displays the geometry of the matrix, which is a significant factor in the process of calculating the matrix. It is a way to switch from one vector to another in a different space. The extent of the change between the basic geometry of these vector spaces can be determined depending on the component elements of the SVD matrix. Any factor present in the model can determine the single value of matrix A with size r × t. Equation 7.5 represents the use of SVD technique.

$$A_{r\times t} = U_{r\times m}\times\sum\nolimits_{m\times m}\times V^{*}{}_{m\times t} \tag{7.5}$$

where A denotes the term (r) -text (t) matrix, U denotes the orthogonal matrix r × m, $\Sigma_{m\times m}$ denotes the vector matrix m × m, and V^* denotes the text matrix m × t. Figure 7.4 illustrates a term-text matrix based on SVD.

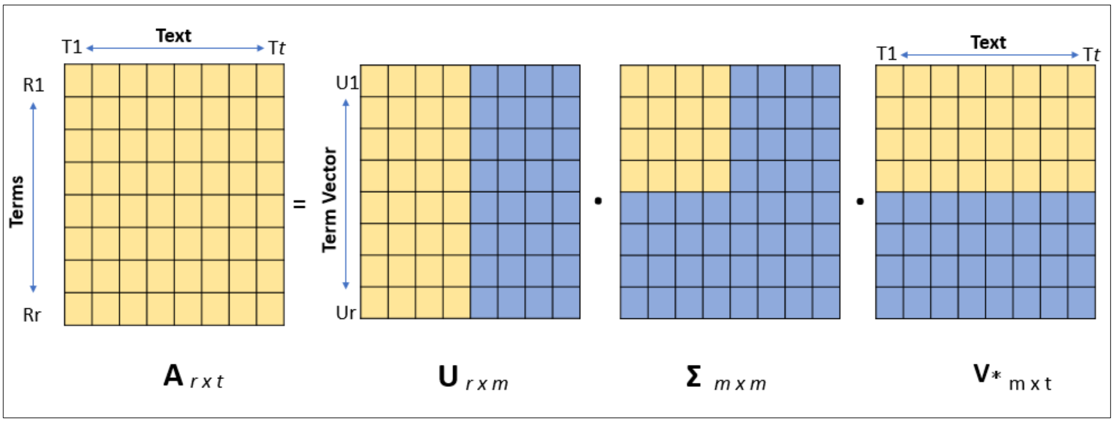


Figure 7.4: A term-text matrix based on SVD

$U_{r\times m}$ is a column matrix of the term vectors, $\Sigma_{m\times m}$ is a diagonal matrix consisting of the square roots of the eigenvalues of U or V. $V^{*}_{n\times t}$ is a row matrix of the text vectors.

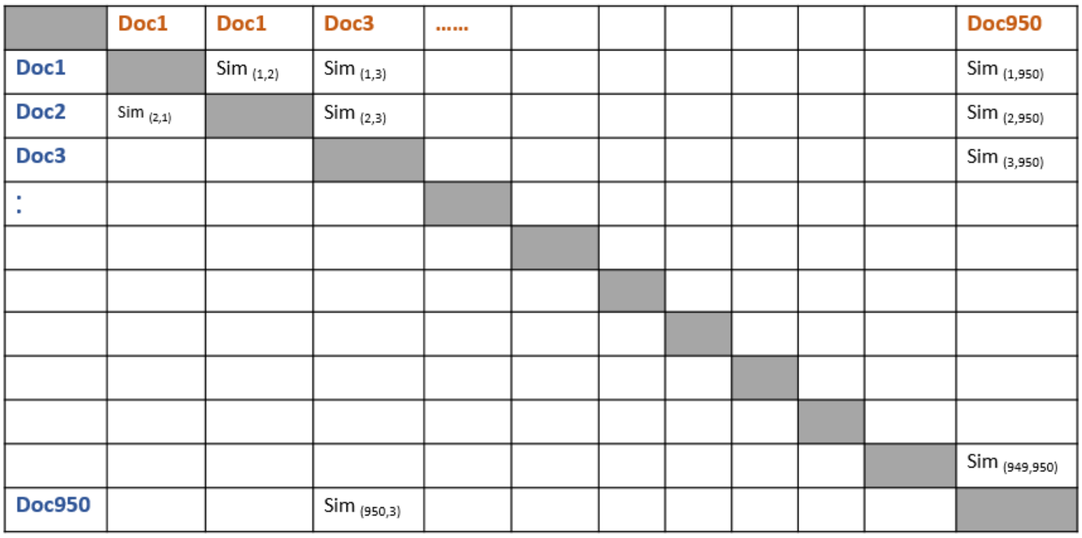
Step 5: Query expansion

After the mother matrix is formulated, we compute the similarity between documents at the term level for each user using cosine similarity. After this, we take the five largest similarity values for each user. These values are then considered with the initial query to reformulate the query with expansion. The documents are divided into two parts, training and testing. To train the proposed model, we used 70% of the documents as training documents, and the remaining 30% were used as the testing dataset with the five top-ranked terms to apply query expansion.

Cosine Similarity: This measures vector similarities which represents document vectors and searches for queries containing several related shared documents, and then computes the relationship degree between these terms and the relevant documents. Cosine similarity is one of the most commonly used similarity algorithms. It is used to calculate the similarity between a pair of vectors as shown in equation 7.6:

$$CosSim(di, dj) = \cos(di, dj) = \frac{di \times dj}{|di| \times |dj|} \tag{7.6}$$

Where di and d_j are the vectors that represent two documents. Figure 7.5 summarizes the proposed method. Calculating the similarity between vector terms using a cosine similarity algorithm is an essential step in query expansion. This is done by taking each pair of document vectors as a separate element and detecting the matrix by calculating the degree of similarity between each pair of document vectors based on cosine similarity, as shown in Figure 7.5:



	Doc1	Doc1	Doc3	……							Doc950
Doc1		Sim $_{(1,2)}$	Sim $_{(1,3)}$								Sim $_{(1,950)}$
Doc2	Sim $_{(2,1)}$		Sim $_{(2,3)}$								Sim $_{(2,950)}$
Doc3											Sim $_{(3,950)}$
:											
											Sim $_{(949,950)}$
Doc950			Sim $_{(950,3)}$								

Figure 7.5: The similarity between each pair of document vectors

The aim of the similarity calculation step is to retrieve the terms that are most similar to each other in the data collection that best represented a specific user. Each user is given a set of terms, which are the optimal set of terms to generate a new and expanded query. Figure 7.6 summarizes the proposed method.

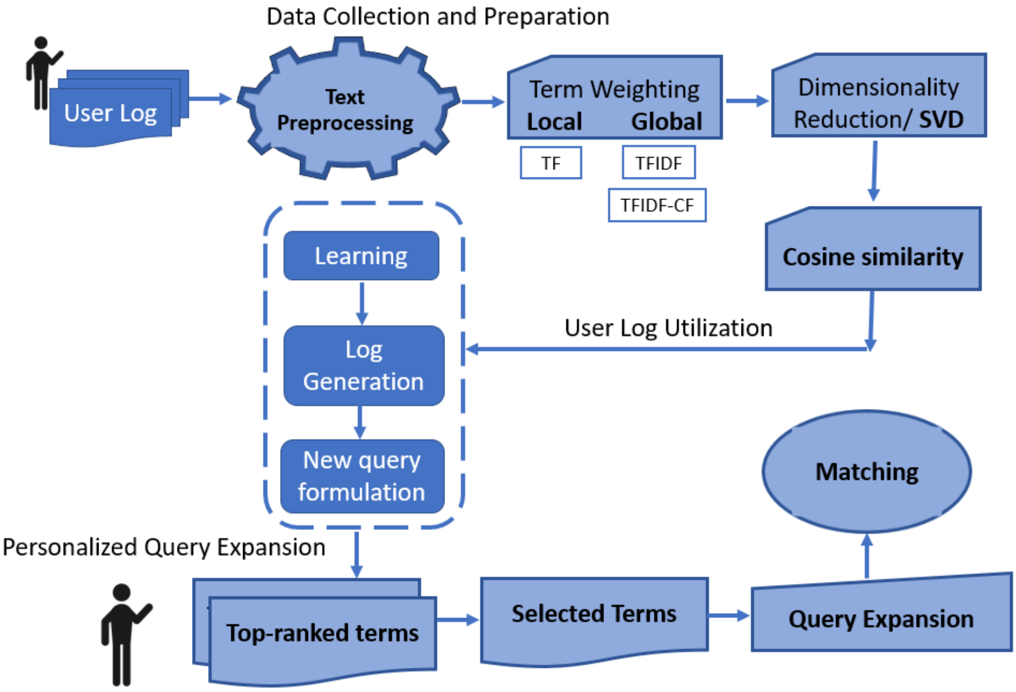


Figure 7.6: A simplified flowchart of our proposed model of the query expansion process.

The created query vectors are utilized as input for the query expansion algorithm. Regarding the QE algorithm, representation methods are used in addition to similarity techniques, such as cosine similarity. After obtaining the cosine similarity results, we generate a log for each user and train the system based on the users' words. These steps are repeated by re-forming new queries, selecting the most frequently used words and specifying them as top-ranked terms with logs to select the most frequently used words to expand the query and retrieve relevant results.

7.5. Experiments

This sub-section describes the set-up of the prototype and the experiments carried out. Subsequently, we explain the evaluation measures used to assess the proposed model. Then, the experiment results are detailed with a justification and analysis.

Case 1: Experiments without Query Expansion

In this subsection, we describe the experiments using our proposed model based on the dataset collected. These experiments were divided into two categories, long and short, based on the length of the query terms, each category comprising 25 documents. A query was deemed to be short if it contained one or two words, and a query was deemed to be long if it contained more than three words. In both cases, three methods were used to retrieve the term, TF, TF-IDF and TF-IDF-CF. Table 7.2 to Table 7.4 depict the results without query expansion, while Table 7.5 to Table 7.7 depict the results after query expansion, based on the top-ranked term. 70% of the data were used for sample training and 30% for sample testing to ensure that both groups are a random sample of the same distribution. Table 7.2 shows the results of using TF without query expansion. The results show that the short texts (S-Term) are superior to the long texts in all subjects, except medicine and tourism. The average of the results are as follows: recall = 0.824, precision = 0.878 and F-Score = 0.836.

Table 7.2: Results of experiment without query expansion using TF

Topic	Recall		Precision		F1-Score	
	L-Term	S- Term	L-Term	S-Term	L-Term	S-Term
Art	0.714	0.857	1	1	0.833	0.923
Biology	0.714	1	1	0.875	0.833	0.933
Radiology	1	1	0.538	0.777	0.7	0.875
Graphic design	1	1	1	1	1	1
History	0.571	1	0.666	0.583	0.615	0.736
Optics	1	1	0.777	1	0.875	1
Journalism	1	1	0.875	0.7	0.933	0.823
CS	1	1	0.777	1	0.875	1
Physics	0.428	0.833	1	1	0.6	0.909
Music	0.571	1	0.8	0.875	0.666	0.933
Bio medical Physics	1	1	0.777	1	0.875	1
Geography	1	1	0.583	1	0.736	1
Geology	0.571	0.714	0.363	1	0.444	0.833
Medicine	0.5	0.375	0.75	0.375	0.545	0.428
Drawing	1	1	0.875	0.875	0.933	0.933
English	0.428	0.428	1	1	0.6	0.6
Math	0.142	0.166	1	1	0.25	0.583
Tourism	0.857	0.571	0.6	1	0.705	0.727
Sport	0.428	0.714	0.6	0.625	0.5	0.666
Average	0.732	0.824	0.788	0.878	0.711	0.836

Table 7.3 shows the results of using TF-IDF without query expansion. Similar to the previous experiment, the results indicate that the short texts (S- Term) are superior to the long texts in all subjects, except medicine and tourism. The average of the results are as follows: recall = 0.777, precision =0.776 and F-Score = 0.753.

Table 7.3: Results of experiment without query expansion using TF-IDF

Topic	Recall		Precision		F1-Score	
	L-Term	S- Term	L-Term	S-Term	L-Term	S-Term
Art	1	0.625	1	1	1	0.769
Biology	0.75	0.875	0.857	0.7	0.799	0.777
Radiology	1	1	0.615	0.727	0.761	0.842
Graphic design	0.875	1	1	0.888	0.933	0.941
History	0.5	0.875	0.8	0.7	0.615	0.777
Optics	1	1	0.615	0.888	0.761	0.941
Journalism	0.875	0.625	0.777	0.625	0.823	0.625
CS	0.75	1	0.75	0.533	0.75	0.695
Physics	0.5	0.714	0.8	1	0.615	0.833
Music	0.5	1	1	0.571	0.666	0.727
Bio medical Physics	0.75	1	0.857	0.888	0.799	0.941
Geography	0.875	1	0.583	1	0.7	1
Geology	0.375	0.625	0.3	1	0.333	0.769
Medicine	0.5	0.285	0.5	0.4	0.5	0.333
Drawing	0.875	1	0.7	0.888	0.777	0.941
English	0.5	0.5	1	1	0.666	0.666
Math	0.125	0.142	0.5	0.333	0.2	0.2
Tourism	0.875	0.625	0.777	0.833	0.823	0.714
Sport	0.375	0.875	0.375	0.777	0.375	0.823
Average	0.684	0.777	0.726	0.776	0.678	0.753

Table 7.4 shows the results of using TF-IDF-CF without query expansion. The results show that the short texts (S-Term) are superior to the long texts (L-Term) across all topics without exception. The average of the results are as follows: recall=0.924, precision=0.938 and F1-Score =0.923. It should be noted that the results of this experiment compared to the previous two experiments demonstrate the high effectiveness of our proposed TF-IDF-CF method.

Table 7.4: Results of experiment without query expansion using TF-IDF-CF

Topic	Recall		Precision		F1-Score	
	L-Term	S- Term	L-Term	S-Term	L-Term	S-Term
Art	0.75	0.875	1	1	0.857	0.933
Biology	0.875	1	1	1	0.933	1
Radiology	1	1	0.727	0.727	0.842	0.842
Graphic design	1	1	0.888	1	0.941	1
History	0.625	1	0.416	0.8	0.5	0.888
Optics	0.625	1	1	1	0.769	1
Journalism	0.75	1	1	0.888	0.857	0.941
CS	1	1	0.8	1	0.888	1
Physics	0.5	0.857	0.8	1	0.615	0.923
Music	0.75	1	0.75	1	0.75	1
Bio medical Physics	1	1	0.8	1	0.888	1
Geography	1	1	1	1	1	1
Geology	0.625	0.625	1	1	0.769	0.769
Medicine	0.5	0.714	1	0.833	0.666	0.769
Drawing	1	1	1	0.888	1	0.941
English	0.5	0.625	1	1	0.666	0.769
Math	0.5	1	0.571	1	0.533	1
Tourism	0.875	1	0.538	1	0.666	1
Sport	0.875	0.875	0.5	0.7	0.636	0.777
Average	0.776	0.924	0.831	0.938	0.777	0.923

Case 2: Experiments with query expansion

Experiments were then conducted using query expansion based on the top-ranked term. The results indicate the superiority of the short texts (S-Term) over the long texts (L-Term) for most topics using the three methods to extract the term. TF-IDF-CF achieved the best average results, with an F-score of 0.94, whereas TF-IDF = 0.921 and TF = 0.916.

Table 7.5: Results of experiment with query expansion using TF

Topic	Recall		Precision		F1-Score	
	L-Term	S-Term	L-Term	S-Term	L-Term	S-Term
Art	0.857	1	0.857	1	0.857	1
Biology	0.857	1	1	1	0.923	1
Radiology	1	1	0.777	1	0.875	1
Graphic design	1	1	0.777	1	0.875	1
History	0.714	1	0.714	0.777	0.714	0.875
Optics	1	1	1	1	1	1
Journalism	0.857	1	0.75	0.7	0.799	0.823
CS	1	1	0.875	1	0.933	1
Physics	0.571	0.833	0.8	1	0.666	0.909
Music	0.714	1	1	1	0.833	1
Bio medical Physics	1	1	0.777	1	0.875	1
Geography	1	1	1	1	1	1
Geology	0.714	1	0.714	1	0.714	1
Medicine	0.571	0.5	0.8	0.75	0.666	0.6
Drawing	1	1	1	1	1	1
English	0.714	0.428	1	1	0.833	0.6
Math	0.714	1	1	1	0.833	1
Tourism	1	1	0.777	0.875	0.875	0.933
Sport	0.714	0.714	0.625	0.625	0.666	0.666
Average	0.842	0.919	0.855	0.933	0.839	0.916

Table 7.6: Results of experiment with query expansion using TF-IDF

Topic	Recall		Precision		F1-Score	
	L-Term	S- Term	L-Term	S-Term	L-Term	S-Term
Art	0.75	1	1	1	0.857	1
Biology	0.875	1	0.875	1	0.875	1
Radiology	0.875	1	0.636	0.888	0.736	0.941
Graphic design	0.625	1	0.833	1	0.714	1
History	0.75	1	0.75	0.8	0.75	0.88
Optics	1	1	0.888	1	0.941	1
Journalism	0.75	1	0.75	0.727	0.75	0.842
CS	0.75	1	0.75	1	0.75	1
Physics	0.625	0.857	0.714	1	0.666	0.923
Music	0.75	1	1	1	0.857	1
Bio medical Physics	1	1	0.666	1	0.8	1
Geography	1	1	0.533	1	0.695	1
Geology	0.5	0.875	0.5	1	0.5	0.933
Medicine	0.375	0.571	0.6	.8	0.461	0.666
Drawing	1	1	0.8	1	0.888	1
English	0.75	.5	1	1	0.857	0.666
Math	0.375	1	0.75	1	0.5	1
Tourism	0.875	1	0.777	0.888	0.823	0.941
Sport	0.375	0.75	0.5	0.66	0.428	0.705
Average	0.736	0.923	0.753	0.935	0.729	0.921

Table 7.7: Results of experiment with query expansion using TF-IDF-CF

Topic	Recall		Precision		F1-Score	
	L-Term	S- Term	L-Term	S-Term	L-Term	S-Term
Art	1	1	1	1	1	1
Biology	0.875	1	1	1	0.933	1
Radiology	1	1	0.888	1	0.941	1
Graphic design	1	1	1	1	1	1
History	0.875	1	0.875	0.888	0.875	0.941
Optics	1	1	1	1	1	1
Journalism	0.875	1	1	0.888	0.933	0.941
CS	1	1	1	1	1	1
Physics	0.	0.857	0.	1	0.	0.923
Music	0.875	1	1	1	0.933	1
Bio medical Physics	1	1	0.8	1	0.888	1
Geographic	1	1	1	1	1	1
Geology	0.625	1	1	1	0.769	1
Medicine	0.5	0.714	0.5	0.833	0.5	0.769
Drawing	1	1	1	1	1	1
English	0.625	0.5	1	1	0.769	0.666
Math	1	1	0.727	1	0.842	1
Tourism	1	1	0.666	0.888	0.8	0.941
Sport	0.875	1	0.777	0.727	0.823	0.842
Average	0.881	0.951	0.898	0.959	0.880	0.948

Figures 7.7 and 7.8 compare the three methods with and without query expansion for the short and long texts, the results demonstrating the significant superiority of query expansion, especially when using TF-IDF-CF.

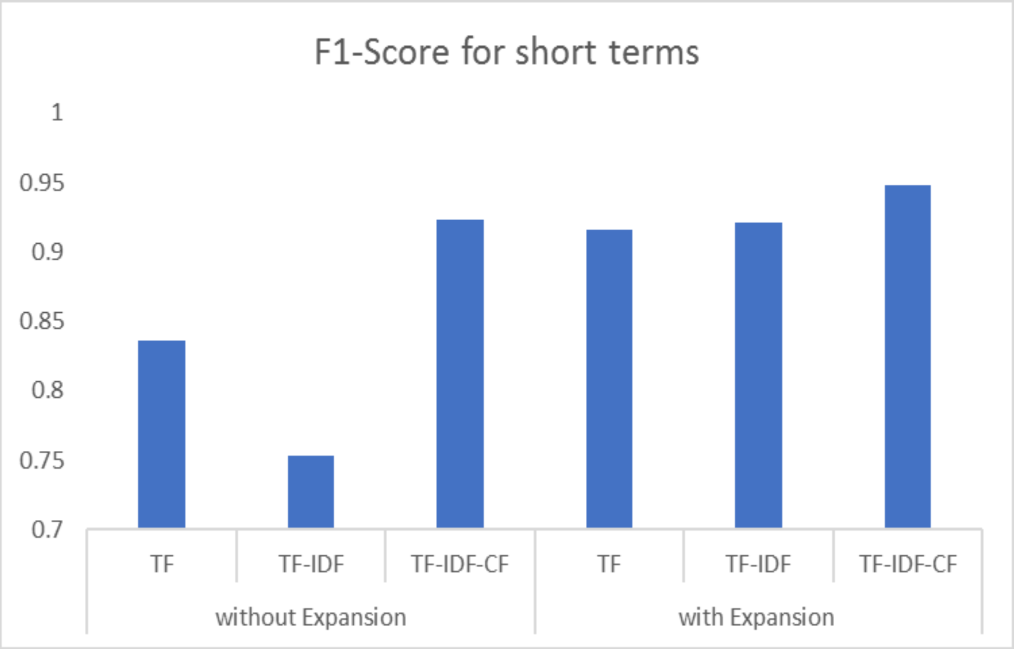


Figure 7.7: F1-score for short texts.

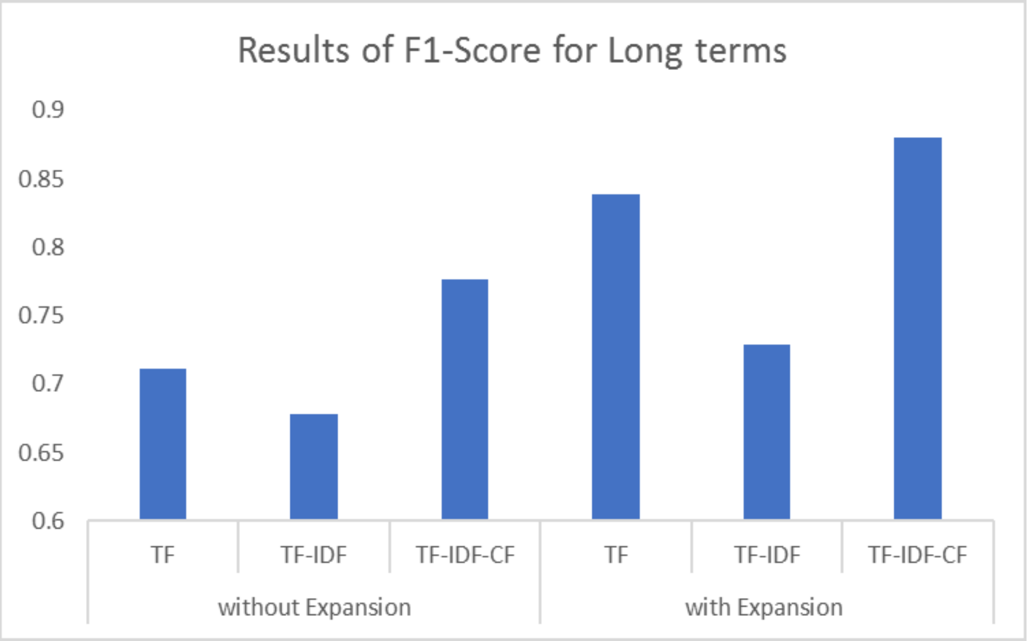


Figure 7.8: F1-score for long.

In view of the results presented in the two above Figure7.7 and Figure 7.8,. it is clear that the comparison of the three methods with and without the expansion of the query showed a big difference for the long text and the short text. The results show great superiority in query expansion for our method, especially when using (TF-IDF-CF). The F-score results at the short text level were slightly better after expansion. However, the (TF-IDF) achieved a clear improvement after the expansion was applied and the (TF-IDF-CF) also achieved some improvement after expansion.

As for the long text, the three methods significantly improved the result of the F-score. The three techniques (TF-IDF-CF; TF-IDF; and TF) achieved the following results before expansion (0.77, 0.67, 0.70) respectively, while these techniques achieved the following results after expansion (0.88, 0.73, 0.8). This reflects a clear improvement in the effectiveness of query expansion using our proposed expansion approaches.

7.6 Conclusions

The huge volume of information available on the Internet makes web browsing a challenging issue for users, hence there is an urgent need to assist users to find relevant documents that are most similar to their search queries. Thus, researchers have developed several query expansion (QE) methods to overcome the challenge of information retrieval and to provide users with useful data relating to their search queries. High dimensionality is the main challenge in the information retrieval area, especially in relation to query expansion. To solve this problem, many dimensionality reduction techniques have been developed, such as SVD. This research developed a query expansion model based on user profiles. This chapter introduced the new approach for personalized QE, and compares three term weighting techniques in personalized query expansion: the local term weighting technique, TF, and the global term weighting technique at the document level (TF-IDF) and the class level (TF-IDF-CF). The experiment results show that the implementation of the SVD algorithm with the class-level term weighting techniques achieves a significant improvement of (approximately 3% and 15 % for the short term query and long term query respectively. Furthermore, it successfully retrieves personalized relevant search results which match the user’s interests.

CHAPTER 8

CONCLUSION

8.1 Introduction

A web information retrieval system is an intelligent software system that provides relevant web resources that match the users' queries. To solve the existing issues in information retrieval systems, such as vague search queries and not using appropriate queries, there is a need to analyze the intent of the users instead of answering single queries through a personalized search using a semantic search and an ontology solution. As is evident from the state-of-the-art survey discussed in Chapter 2, several proposals have been made by various researchers to personalize searches using semantic approaches. However, the major shortcomings of these proposals are: 1) they do not take into account the user context history for a personalized search; 2) they do not consider the use of an ontology-based user profile for personalized search retrieval; and 3) they do not take into account the user's information during the process of query expansion.

The contributions of this thesis can be summarized as follows: first, it creates user profiles based on a user's browsing behaviour, as well as the semantic knowledge of the domain ontology, aiming to improve the quality of the search results; second, it proposes intelligent methods that integrate user context history into the information retrieval process for a personalised web search; and third, it proposes a three-term weighting method for personalized query expansion.

In the next section, we outline the problems addressed in this thesis. In Section 8.3, we outline the contributions of this thesis. Section 8.4 concludes the thesis and sets the stage for future work.

8.2 Problems Addressed in this thesis

In this thesis, we addressed three major issues associated with a personalized search using a semantic search and an ontology solution. This thesis addresses the issues by:

1. building an ontological user profile based on semantic fuzzy classification and user behaviour.
2. building an effective concept-based user profile for the purpose of search personalization based on context history.
3. proposing a new personalized query expansion approach using latent semantic technologies coupled with term weight approaches.
4. validating the aforementioned developed methods by building a prototype of a personalized web search.

8.3 Contributions of this thesis

The major contribution of this thesis to the existing literature is that it proposes three methodologies to improve the performance of a personalized search by taking into account the ontological user profile, contextual user profile and personal query expansion.

The three methodologies encompassed by this thesis are as follows:

A. We propose a personalized search approach using an ontology-based user profile, taking into account the user's implicit browsing behaviour, semantic knowledge of concepts, and synonyms of term-based vectors extracted from the WordNet API.

B. We present a personalized web search approach by incorporating user context in the information retrieval process in order to deliver the search results that most meet the user’s need.

C. We propose a users’ profile-based approach for query expansion using TD-IDF-CF weighing techniques. This approach is used for personalized query expansion and it also compares three-term weight techniques in personalized query expansion.

8.3.1 Contribution 1: Personalized search using ontological user profile

The first major contribution of this thesis is building an ontological user profile approach for a personalized search to use it in the re-ranking process. The aim of this approach is to build user profiles based on user browsing behaviour and the semantic knowledge of the specific domain ontology to enhance the quality of the search results. The proposed approach utilizes a re-ranking algorithm to sort the results returned by the search engine to provide a search result that best relates to the user's query.

This algorithm evaluates the similarity between a user query, the retrieved search results and the ontological concepts. This similarity is computed by taking into account a user's explicit browsing behaviour, semantic knowledge of concepts, and synonyms of term-based vectors extracted from the WordNet API. A set of experiments using a case study from a transport service domain validates the effectiveness of the proposed approach and demonstrates promising results.

The experiment evaluation based on comparisons measures the performance of this approach compared to other similar approaches presented in this field. This section details the results of the comparison of the proposed approach with two commonly used approaches, namely: the model proposed by (Samen, Ezin & Onana 2017) and the Google ranking approach. These comparisons are based on the improvement in precision in the variant results of top-n. Also, a benchmark dataset called the Open Directory Project (ODP) was used.

8.3.2 Contribution 2: Personalized search taking into account contextual user profile

The second major contribution of this thesis is building a contextual user profile based on navigation history which is modeled in order to deliver the search results that best meet the user's needs. The key contributions are the technique of building contextual user profiles based on the user's browsing behaviour; the semantic knowledge of a specific domain ontology; and an algorithm for re-ranking the original search results based on the user's contextual profile.

The steps in this model involve constructing a contextual profile, like a structured data log, where the users' related data include their interests and preferences. The preferences and interests of the users are obtained from their search histories or web logs. The primary purpose of using search history is to extract the URLs from the web pages that were visited by users at different times. It is also important to examine the URLs, as this step helps to retrieve the metadata, such as the keywords and web page descriptions of the web sites referred by the HTML meta tag. A personalized web search gives the search results in the same order regardless of the time the query is submitted to the search engine due to the consistency of the ontological domain and its concepts based on the browsing history of the user.

8.3.3 Contribution 3: Personalized query expansion using TF-IDF-CF

The third major contribution of this thesis is proposing a new technique for personal query expansion. The construction of a query occurs without any intervention from the user. This approach finds similar texts for a particular user from a large number of documents. The results of applying this approach reduce the query search time and also minimize the effort of the users by extending the terms of the queries. Personalized query expansion also plays an active role in improving the performance of the retrieval process, which represents the user's interests.

In order to build this approach, there is a need to collect some documents manually. Then, we investigate the performance of the approach. The proposed method comprises the following five steps: gathering the corpus, pre-processing, term weighting, dimensionality reduction, and expansion. It introduces a new approach for personalized QE and it also compares three-term weight techniques in personalized query expansion. The used local term weight is TF, where the global techniques document level (TF-IDF) and class level (TF-IDF-CF).

8.4 Conclusion and Future Work

In this thesis, a user profile is built to enhance the role of Web search personalization and search engines. At the same time, an ontology is created which relies on the user's search history, as it helps to obtain more efficient and accurate user profiles through a document hierarchy. For the best search results, the most frequent queries are identified by an interest score where a personalized ontology is built in order to obtain a more accurate profile of local and global directories. The importance of the ontological approach lies in the precise calculation of the most appropriate documents which are the most relevant to the page content and search profiles of the user. The user profile shows the user's interest in a specific web page during the search process. It is worth mentioning that these profiles are updated consistently to ensure the accuracy of the interest scores. Ontological profiles assist in creating Web search personalization and user profile reformulation.

There is an urgent need to use the clicked documents to improve the quality of the search and to facilitate the retrieval methods and information. To address this issue, an ontological user profile is proposed, based on the browsing history of the user's search. Based on the experimentation and validation, this profile has proven its effectiveness in determining the weight of each concept in the domain ontology based on user interest, with the aim of creating ontological user profiles in the transportation domain. The proposed approach proved to be highly effective, achieving better results than Google. Our experiments show that user behaviour results are of great importance in improving the overall accuracy of the search results.

The proposed approach with user behaviour considered (PB) shows an improvement in precision of 13%, 12.5%, 10.5% and 5% for the top 5, 10, 15 and 20 links, respectively over the Google search engine. The proposed without considering user behaviour (PWB) approach demonstrates improvements in precision of 9.5%, 10%, 8.5% and 4.5% for the top 5, 10, 15 and 20 links respectively, over the Google search engine. However, user behaviour has demonstrated a positive impact on the search result. Also, the results show that the proposed personalized approach based on user behaviour and semantic fuzzy classification achieves slightly better precision at all *top-n*. There is an improvement in precision of 1.5%, 2.5%, 2% and 0.05% for the top 5, 10, 15 and 20 links respectively.

The key contributions of the second approach are the development of the technique of building the contextual user profiles based on the user's browsing behaviour and the semantic knowledge of the specific domain ontology and the development of the algorithm to re-rank the original search results based on the user's contextual profile. The search history of a user is collected implicitly as the user's temporal context, and a dynamic user profile is generated by modelling the profile based on the reference ontology. We also consider a user's behaviour in browsing web pages and use the web page's weight status value when constructing user profiles. We noted the importance of modelling user behaviours during the browsing process and used a temporal factor to create user profiles with high efficiency and accuracy in the personalization search engine. In relation to the experiment results, the proposed approach with user context (PC) achieves an improvement in precision of 35%, 29%, 22% and 17% for the top 5, 10, 15 and 20 links, respectively over the Google search engine. The proposed without taking user context into consideration (PWC) approach demonstrates an improvement in precision of 22%, 19%, 18% and 14% for the top 5, 10, 15 and 20 links respectively over the Google search engine. Also, the results show that the proposed personalized approach based on user context achieves slightly better precision at all top-n. There is an improvement in precision of 12%, 9%, 0.04% and 0.03% for the top 5, 10, 15 and 20 links, respectively.

Several query expansion (QE) methods and techniques provide a solution to the challenge of information retrieval issues and provide users with useful data related to their search queries. To solve this problem, many dimensionality reduction techniques are used, such SVD. The proposed solution addresses a new study, based on users' profiles, to develop the query expansion model. It introduced a new approach for personalized QE and it also compares three-term weight techniques in personalized query expansion., the used local term weight is TF, where the global techniques document level (TF-IDF) and class level (TF-IDF-CF), The experiment results show that the implementation of the SVD algorithm with class-level term-weighting techniques yield a significant improvement, and successfully personalized relevant search results according to the user's interests. The results were evaluated based on F- score with values of (0.959) with short terms and (0.898) on long terms.

8.5 FUTURE WORK

In this research, a semantic technique and an ontology user profile are used to build an effective user profile to ensure the best query results in the re-ranking process. Furthermore, a term weighting and a latent semantic method are used for effective personalized query expansion to achieve the best query reformulation. Based on the work carried out and documented in this thesis, we present the following directions for future work:

- It is achievable to develop a personalized web search model, taking into account factors related to both time and location.
- In the future, we aim to develop intelligent ways to capture and improve the accuracy of the user profiles.
- In the future, we plan to improve the already used personal query expansion by applying some methods in the classification process and taking into account semantic information retrieval.
- Finally, in the retrieval of information, we will aim to develop intelligent methods to capture users' intent for formulating search queries and query expansion.

References:

Adar, E., Breuel, T.M., Cass, T.A., Pitkow, J.E. & Schuetze, H. 2002, 'System and method for searching and recommending documents in a collection using share bookmarks', Google Patents.

Adomavicius, G. & Tuzhilin, A. 2005, 'Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions', *IEEE Transactions on Knowledge & Data Engineering*, no. 6, pp. 734-49.

Adomavicius, G. & Tuzhilin, A. 2011, 'Context-aware recommender systems', *Recommender systems handbook*, Springer, pp. 217-53.

Ahmadian, N., Nematbakhsh, M.A. & Vahdat-Nejad, H. 2011, 'A context aware approach to semantic query expansion', *Innovations in Information Technology (IIT), 2011 International Conference on*, IEEE, pp. 57-60.

Akhlaghian, F., Arzanian, B. & Moradi, P. 2010, 'A personalized search engine using ontology-based fuzzy concept networks', *Data Storage and Data Engineering (DSDE), 2010 International Conference on*, IEEE, pp. 137-41.

Al-Hassan, M., Lu, H. & Lu, J. 2015, 'A semantic enhanced hybrid recommendation approach: A case study of e-Government tourism service recommendation system', *Decision Support Systems*, vol. 72, pp. 97-109.

Al-Khateeb, B., Al-Kubaisi, A.J. & Al-Janabi, S.T. 2017, 'Query reformulation using WordNet and genetic algorithm', *New Trends in Information & Communications Technology Applications (NTICT), 2017 Annual Conference on*, IEEE, pp. 91-6.

Allan, J., Aslam, J., Belkin, N., Buckley, C., Callan, J., Croft, B., Dumais, S., Fuhr, N., Harman, D. & Harper, D.J. 2003, 'Challenges in information retrieval and language modeling: report of a workshop held at the center for intelligent information retrieval, University of Massachusetts Amherst, September 2002', *ACM SIGIR Forum*, vol. 37, ACM, pp. 31-47.

Azad, H.K. & Deepak, A. 2017, 'Query Expansion Techniques for Information Retrieval: a Survey', *arXiv preprint arXiv:1708.00247*.

Azizan, A., Bakar, Z.A. & Noah, S.A. 2016, 'Query reformulation using ontology and keyword for durian web search', *Information Retrieval and Knowledge Management (CAMP), 2016 Third International Conference on*, IEEE, pp. 94-100.

Baazaoui-Zghal, H. & Ghezala, H.B. 2014, 'A fuzzy-ontology-driven method for a personalized query reformulation', *Fuzzy Systems (FUZZ-IEEE), 2014 IEEE International Conference on*, IEEE, pp. 1640-7.

Baazaoui, H., Aufaure, M.-A., Soussi, R., Laboratoy, R.-G. & de la Manouba, E.C.U. 2008, 'Towards an on-line semantic information retrieval system based on fuzzy ontologies', *Journal of digital information management*, vol. 6, no. 5, p. 375.

Baeza-Yates, R. & Ribeiro-Neto, B. 1999, *Modern information retrieval*, vol. 463, ACM press New York.

Bai, X., Cambazoglu, B.B., Gullo, F., Mantrach, A. & Silvestri, F. 2017, 'Exploiting search history of users for news personalization', *Information Sciences*, vol. 385, pp. 125-37.

Bennett, P.N. & Nguyen, N. 2009, 'Refined experts: improving classification in large taxonomies', *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 11-8.

Berger, A. & Lafferty, J. 1999, 'Information retrieval as statistical translation', *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 222-9.

Blanco, R., Ottaviano, G. & Meij, E. 2015, 'Fast and space-efficient entity linking for queries', *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, ACM, pp. 179-88.

Bollen, J., Mao, H. & Pepe, A. 2010, 'Determining the Public Mood State by Analysis of Microblogging Posts', *ALIFE*, pp. 667-8.

Borisov, A., Markov, I., de Rijke, M. & Serdyukov, P. 2016, 'A context-aware time model for web search', *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, ACM, pp. 205-14.

Bouadjenek, M.R., Sanner, S. & Ferraro, G. 2015, 'A study of query reformulation for patent prior art search with partial patent applications', *Proceedings of the 15th International Conference on Artificial Intelligence and Law*, ACM, pp. 23-32.

Bryman, A. & Bell, E. 2015, *Business research methods*, Oxford University Press, USA.

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

Büttcher, S., Clarke, C.L. & Cormack, G.V. 2016, *Information retrieval: Implementing and evaluating search engines*, Mit Press.

Calegari, S. & Pasi, G. 2010, 'Ontology-based information behaviour to improve web search', *Future Internet*, vol. 2, no. 4, pp. 533-58.

Calvanese, D. & Franconi, E. 2018, 'First-Order Ontology Mediated Database Querying via Query Reformulation', *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, Springer, pp. 169-85.

Cao, G., Nie, J.-Y., Gao, J. & Robertson, S. 2008, 'Selecting good expansion terms for pseudo-relevance feedback', *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 243-50.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F. & Li, H. 2007, 'Learning to rank: from pairwise approach to listwise approach', *Proceedings of the 24th international conference on Machine learning*, ACM, pp. 129-36.

Chaari, T., Laforest, F. & Celentano, A. 2008, 'Adaptation in context-aware pervasive information systems: the SECAS project', *International Journal of Pervasive Computing and Communications*, vol. 3, no. 4, pp. 400-25.

Chaves, R.P. 2001, 'WordNet and Automated Text Summarization', *NLPRS*, pp. 109-16.

Chen, Y., Wu, C., Xie, M. & Guo, X. 2011, 'Solving the sparsity problem in recommender systems using association retrieval', *Journal of computers*, vol. 6, no. 9, pp. 1896-902.

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

Chirita, P.A., Nejdl, W., Paiu, R. & Kohlschütter, C. 2005, 'Using ODP metadata to personalize search', *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 178-85.

Christopher, D.M., Prabhakar, R. & Hinrich, S. 2008, 'Introduction to information retrieval', *An Introduction To Information Retrieval*, vol. 151, no. 177, p. 5.

Cimiano, P., Mädche, A., Staab, S. & Völker, J. 2009, 'Ontology learning', *Handbook on ontologies*, Springer, pp. 245-67.

Colledge, M. & Barnes, J. 2011, 'Internet searching using semantic disambiguation and expansion', Google Patents.

Conesa, J., Storey, V.C. & Sugumaran, V. 2010, 'Usability of upper level ontologies: The case of ResearchCyc', *Data & Knowledge Engineering*, vol. 69, no. 4, pp. 343-56.

Cramer, M., Zhai, C.X., Shen, X. & Tan, B. 2013, 'Real time implicit user modeling for personalized search', Google Patents.

Crouch, C.J. & Yang, B. 1992, 'Experiments in automatic statistical thesaurus construction', *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 77-88.

Cruse, A. 2011, 'Meaning in language: An introduction to semantics and pragmatics'.

Cuellar, D., Diaz, E. & Ponce-de-Leon-Senti, E. 2015, 'Improving Information Retrieval Through a Global Term Weighting Scheme', *Mexican Conference on Pattern Recognition*, Springer, pp. 246-57.

Cui, H., Wen, J.-R., Nie, J.-Y. & Ma, W.-Y. 2002, 'Probabilistic query expansion using query logs', *Proceedings of the 11th international conference on World Wide Web*, ACM, pp. 325-32.

Cuomo, G.A., Nguyen, B.Q. & Singhal, S.K. 2001, 'Method and system for collecting user profile information over the world-wide web in the presence of dynamic content using document comparators', Google Patents.

Curé, O.C., Maurer, H., Shah, N.H. & Le Pendu, P. 2015, 'A formal concept analysis and semantic query expansion cooperation to refine health outcomes of interest', *BMC medical informatics and decision making*, vol. 15, no. 1, p. S8.

Dali, L., Fortuna, B., Duc, T.T. & Mladenić, D. 2012, 'Query-independent learning to rank for rdf entity search', *Extended Semantic Web Conference*, Springer, pp. 484-98.

Dali, L., Fortuna, B., Tran, D.T. & Mladenic, D. 2012, ' Query-independent learning to rank for rdf entity search', *ESWC*, pp. 484–98.

Dalton, J., Dietz, L. & Allan, J. 2014, 'Entity query feature expansion using knowledge base links', *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, pp. 365-74.

Daoud, M., Tamine-Lechani, L. & Boughanem, M. 2008, 'Using a concept-based user context for search personalization', *Proc. of the 2008 Internat. Conf. of Data Mining and Knowledge Engineering*.

de Campos, L.M., Fernandez, J.M. & Huete, J.F. 1998, 'Query expansion in information retrieval systems using a Bayesian network-based thesaurus', *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, Morgan Kaufmann Publishers Inc., pp. 53-60.

De Silva, J. & Haddela, P.S. 2013, 'A term weighting method for identifying emotions from text content', *Industrial and Information Systems (ICIIS), 2013 8th IEEE International Conference on*, IEEE, pp. 381-6.

Dong, H., Hussain, F.K. & Chang, E. 2011, 'A service search engine for the industrial digital ecosystems', *IEEE Transactions on Industrial Electronics*, vol. 58, no. 6, pp. 2183-96.

Duong, T.H., Uddin, M.N. & Nguyen, C.D. 2013, 'Personalized semantic search using ODP: a study case in academic domain', *International Conference on Computational Science and Its Applications*, Springer, pp. 607-19.

El Ghali, B. & El Qadi, A. 2017, 'Context-aware query expansion method using Language Models and Latent Semantic Analyses', *Knowledge and Information Systems*, vol. 50, no. 3, pp. 751-62.

Elkateb, S. 2014, 'Words, Concepts and Meaning Representation', *Zawia University Bulletin– ISSUE No. 16*, vol. 4.

Esbitan, S.M. & Barhoom, D.T.S. 2012, 'A Personalized Context-Dependent Web Search Engine Using Word Net (Sama Search Engine)', Islamic University of Gaza.

Fathy, N., Gharib, T.F., Badr, N.L., Mashat, A.S. & Abraham, A. 2014, 'A Personalized Approach for Re-ranking Search Results Using User Preferences', *J. UCS*, vol. 20, no. 9, pp. 1232-58.

Fernandez, F.M.H. & Ponnusamy, R. 2014, 'User behavior framework for personalized library ontology', *Proceedings of the National Conference on Man Machine Interaction (NCMMI 2014), India*, pp. 62-8.

Ferreira-Satler, M., Romero, F.P., Menendez-Dominguez, V.H., Zapata, A. & Prieto, M.E. 2012, 'Fuzzy ontologies-based user profiles applied to enhance e-learning activities', *Soft Computing*, vol. 16, no. 7, pp. 1129-41.

Galliers, R. 1992, *Information systems research: Issues, methods and practical guidelines*, Blackwell Scientific.

Gan, L. & Hong, H. 2015, 'Improving query expansion for information retrieval using wikipedia', *International Journal of Database Theory and Application*, vol. 8, no. 3, pp. 27-40.

Gantner, Z., Drumond, L., Freudenthaler, C., Rendle, S. & Schmidt-Thieme, L. 2010, 'Learning attribute-to-feature mappings for cold-start recommendations', *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, IEEE, pp. 176-85.

Gauch, S., Speretta, M., Chandramouli, A. & Micarelli, A. 2007, 'User profiles for personalized information access', *The adaptive web*, Springer, pp. 54-89.

Gruber, T. 2009, 'Ontology', *Encyclopedia of database systems*, pp. 1963-5.

Gruber, T.R. 1995, 'Toward principles for the design of ontologies used for knowledge sharing?', *International journal of human-computer studies*, vol. 43, no. 5-6, pp. 907-28.

Gupta, D. & Chavhan, N. 2014, 'Ontological user profiling for adaptive re-ranking in mobile web search', *Convergence of Technology (I2CT), 2014 International Conference for*, IEEE, pp. 1-5.

Hahm, G.J., Yi, M.Y., Lee, J.H. & Suh, H.W. 2014, 'A personalized query expansion approach for engineering document retrieval', *Advanced Engineering Informatics*, vol. 28, no. 4, pp. 344-59.

Han, L., Chen, G. & Li, M. 2013, 'A method for the acquisition of ontology-based user profiles', *Advances in Engineering Software*, vol. 65, pp. 132-7.

Hawalah, A. & Fasli, M. 2014, 'Utilizing contextual ontological user profiles for personalized recommendations', *Expert Systems with Applications*, vol. 41, no. 10, pp. 4777-97.

Hawalah, A. & Fasli, M. 2015, 'Dynamic user profiles for web personalisation', *Expert Systems with Applications*, vol. 42, no. 5, pp. 2547-69.

He, Y. & Tang, Y. 2014, 'Research of User Profile Model in Personalized Search', *Applied Mechanics & Materials*.

Hoeber, O., Yang, X.-D. & Yao, Y. 2005, 'Conceptual query expansion', *International Atlantic Web Intelligence Conference*, Springer, pp. 190-6.

Hofmann, K., Li, L. & Radlinski, F. 2016, 'Online evaluation for information retrieval', *Foundations and Trends® in Information Retrieval*, vol. 10, no. 1, pp. 1-117.

Hong, J.-y., Suh, E.-h. & Kim, S.-J. 2009, 'Context-aware systems: A literature review and classification', *Expert Systems with applications*, vol. 36, no. 4, pp. 8509-22.

Hong, J., Suh, E.-H., Kim, J. & Kim, S. 2009, 'Context-aware system for proactive personalized service based on context history', *Expert Systems with Applications*, vol. 36, no. 4, pp. 7448-57.

Hopfgartner, F. & Jose, J.M. 2014, 'An experimental evaluation of ontology-based user profiles', *Multimedia tools and applications*, vol. 73, no. 2, pp. 1029-51.

Houle, M.E., Ma, X., Oria, V. & Sun, J. 2017, 'Query expansion for content-based similarity search using local and global features', *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 3, p. 25.

Hourali, M. & Montazer, G.A. 2011, 'An intelligent information retrieval approach based on two degrees of uncertainty fuzzy ontology', *Advances in Fuzzy Systems*, vol. 2011, p. 7.

Hull, D. & Clyne, T.W. 1996, *An introduction to composite materials*, Cambridge university press.

Hwang, H.-s., Shin, S.-h., Kim, K.-u., Lee, S.-C. & Kim, C.-s. 2007, 'A Context-aware System Architecture using Personal Information based on Ontology', *Software Engineering Research, Management & Applications, 2007. SERA 2007. 5th ACIS International Conference on*, IEEE, pp. 610-5.

Isinkaye, F., Folajimi, Y. & Ojokoh, B. 2015, 'Recommendation systems: Principles, methods and evaluation', *Egyptian Informatics Journal*, vol. 16, no. 3, pp. 261-73.

Jannach, D., Zanker, M., Felfernig, A. & Friedrich, G. 2010, *Recommender systems: an introduction*, Cambridge University Press.

Jawaheer, G., Weller, P. & Kostkova, P. 2014, 'Modeling user preferences in recommender systems: A classification framework for explicit and implicit user feedback', *ACM Transactions on Interactive Intelligent Systems (TiiS)*, vol. 4, no. 2, p. 8.

Jiang, J.-Y., Ke, Y.-Y., Chien, P.-Y. & Cheng, P.-J. 2014, 'Learning user reformulation behavior for query auto-completion', *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, pp. 445-54.

Jiang, X. & Tan, A.-H. 2009, 'Learning and inferencing in user ontology for personalized Semantic Web search', *Information sciences*, vol. 179, no. 16, pp. 2794-808.

Kathuria, N., Mittal, K. & Chhabra, A. 2017, 'A Comprehensive Survey on Query Expansion Techniques, their Issues and Challenges', *International Journal of Computer Applications*, vol. 168, no. 12.

Khan, M.S., Khor, S. & Chong, A. 2004, 'Fuzzy cognitive maps with genetic algorithm for goal-oriented decision support', *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 12, no. supp02, pp. 31-42.

Khare, P., Tewari, V. & Dagdee, N. 2014, 'User profile mining and personalization of web services', *International Journal of Computer Applications*, vol. 105, no. 13.

Kim, H.-R. & Chan, P.K. 2008, 'Learning implicit user interest hierarchy for context in personalization', *Applied Intelligence*, vol. 28, no. 2, pp. 153-66.

Klingberg, T. 2009, *The overflowing brain: Information overload and the limits of working memory*, Oxford University Press.

Kumar, H., Park, S. & Kang, S. 2008, 'A personalized url re-ranking methodology using user's browsing behavior', *KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications*, Springer, pp. 212-21.

Kuzi, S., Carmel, D., Libov, A. & Raviv, A. 2017, 'Query Expansion for Email Search', *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 849-52.

Larson, R.R. 2010, 'Introduction to information retrieval', *Journal of the American Society for Information Science and Technology*, vol. 61, no. 4, pp. 852-3.

Lavie, A. 2010, 'Evaluating the output of machine translation systems', *AMTA Tutorial*, p. 86.

Lee, K.R. 2002, 'Impacts of Information Technology on Society in the new Century', *Konsbruck Robert Lee. Route de Chavannes C*, vol. 27.

Lekshmi, S.B. & George, A. 2016, 'Survey on Profile Creation and Privacy in Personalized Web Search', *International Journal of Engineering Research*, vol. 5, no. 6, pp. 482-3.

Leung, C.H., Li, Y., Milani, A. & Franzoni, V. 2013, 'Collective evolutionary concept distance based query expansion for effective web document retrieval', *International Conference on Computational Science and Its Applications*, Springer, pp. 657-72.

Leung, K.W.-T., Lee, D.L. & Lee, W.-C. 2010, 'Personalized web search with location preferences', *Data Engineering (ICDE), 2010 IEEE 26th International Conference on*, IEEE, pp. 701-12.

Leuski, A. 2000, 'Relevance and reinforcement in interactive browsing', *Proceedings of the ninth international conference on Information and knowledge management*, ACM, pp. 119-26.

Li, L., Deng, H., Dong, A., Chang, Y., Baeza-Yates, R. & Zha, H. 2017, 'Exploring Query Auto-Completion and Click Logs for Contextual-Aware Web Search and Query Suggestion', *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 539-48.

Liang, T.-P., Yang, Y.-F., Chen, D.-N. & Ku, Y.-C. 2008, 'A semantic-expansion approach to personalized knowledge recommendation', *Decision Support Systems*, vol. 45, no. 3, pp. 401-12.

Liang, Y., Siddaramu, T., Yesuf, J. & Sarkany, N. 2010, 'Fermentable sugar release from Jatropha seed cakes following lime pretreatment and enzymatic hydrolysis', *Bioresource technology*, vol. 101, no. 16, pp. 6417-24.

Limbu, D.K., Connor, A.M., Pears, R. & MacDonell, S.G. 2009, 'Improving web search using contextual retrieval', *2009 Sixth International Conference on Information Technology: New Generations*, IEEE, pp. 1329-34.

Liu, F., Yu, C. & Meng, W. 2004, 'Personalized web search for improving retrieval effectiveness', *IEEE Transactions on knowledge and data engineering*, vol. 16, no. 1, pp. 28-40.

Liu, Q., Huang, H., Lut, J., Gao, Y. & Zhang, G. 2017, 'Enhanced word embedding similarity measures using fuzzy rules for query expansion', *Fuzzy Systems (FUZZ-IEEE), 2017 IEEE International Conference on*, IEEE, pp. 1-6.

Ma, Z., Pant, G. & Sheng, O.R.L. 2007, 'Interest-based personalized search', *ACM Transactions on Information Systems (TOIS)*, vol. 25, no. 1, p. 5.

Maedche, A. & Staab, S. 2001, 'Ontology learning for the semantic web', *IEEE Intelligent systems*, vol. 16, no. 2, pp. 72-9.

Matthijs, N. & Radlinski, F. 2011, 'Personalizing web search using long term browsing history', *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, pp. 25-34.

Melucci, M. 2012, 'Contextual search: A computational framework', *Foundations and Trends® in Information Retrieval*, vol. 6, no. 4–5, pp. 257-405.

Micarelli, A., Gasparetti, F., Sciarrone, F. & Gauch, S. 2007, 'Personalized search on the world wide web', *The adaptive web*, Springer, pp. 195-230.

Mielke, S., Pelke, M., Pospiech, S. & Mertens, R. 2015, 'Flexible semantic query expansion for process exploration', *Semantic Computing (ICSC), 2015 IEEE International Conference on*, IEEE, pp. 440-3.

Mitra, B., Shokouhi, M., Radlinski, F. & Hofmann, K. 2014, 'On user interactions with query auto-completion', *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*, ACM, pp. 1055-8.

Mitra, M., Singhal, A. & Buckley, C. 1998, 'Improving automatic query expansion', *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 206-14.

Mnih, A. 2011, 'Taxonomy-informed latent factor models for implicit feedback', *Proceedings of the 2011 International Conference on KDD Cup 2011-Volume 18*, JMLR. org, pp. 169-81.

Mohammed, N.U., Duong, T.H. & Jo, G.S. 2010, 'Contextual information search based on ontological user profile', *International Conference on Computational Collective Intelligence*, Springer, pp. 490-500.

Morbach, J., Wiesner, A. & Marquardt, W. 2009, 'OntoCAPE—A (re) usable ontology for computer-aided process engineering', *Computers & Chemical Engineering*, vol. 33, no. 10, pp. 1546-56.

Moughrabi, I.A. & Yamout, F.A. 2016, 'A weight dissemination look at relevance feedback and query reformulation', *SAI Computing Conference (SAI), 2016*, IEEE, pp. 458-63.

Mylonas, P., Vallet, D., Castells, P., Fernández, M. & Avrithis, Y. 2008, 'Personalized information retrieval based on context and ontological knowledge', *The Knowledge Engineering Review*, vol. 23, no. 1, pp. 73-100.

Nalisnick, E., Mitra, B., Craswell, N. & Caruana, R. 2016, 'Improving document ranking with dual word embeddings', *Proceedings of the 25th International Conference Companion on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 83-4.

Nanda, A., Omanwar, R. & Deshpande, B. 2014, 'Implicitly learning a user interest profile for personalization of web search using collaborative filtering', *Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)-Volume 02*, IEEE Computer Society, pp. 54-62.

Nawab, R.M.A., Stevenson, M. & Clough, P. 2017, 'An IR-Based approach utilizing query expansion for plagiarism detection in MEDLINE', *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, vol. 14, no. 4, pp. 796-804.

Page, L., Brin, S., Motwani, R. & Winograd, T. 1999, *The PageRank citation ranking: Bringing order to the web*, Stanford InfoLab.

Peffers, K., Tuunanen, T., Rothenberger, M.A. & Chatterjee, S. 2007, 'A design science research methodology for information systems research', *Journal of management information systems*, vol. 24, no. 3, pp. 45-77.

Pelegrina, A.B., Martin-Bautista, M.J. & Faber, P. 2013, 'Contextualization and personalization of queries to knowledge bases using spreading activation', *International Conference on Flexible Query Answering Systems*, Springer, pp. 671-82.

Pennington, J., Socher, R. & Manning, C. 2014, 'Glove: Global vectors for word representation', *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532-43.

Porter, M.F. 1980, 'An algorithm for suffix stripping', *Program*, vol. 14, no. 3, pp. 130-7.

Qiu, F. & Cho, J. 2006, 'Automatic identification of user interest for personalized search', *Proceedings of the 15th international conference on World Wide Web*, ACM, pp. 727-36.

Rawashdeh, M., Kim, H.-N., Alja'am, J.M. & El Saddik, A. 2013, 'Folksonomy link prediction based on a tripartite graph for tag recommendation', *Journal of Intelligent Information Systems*, vol. 40, no. 2, pp. 307-25.

Rekik, W., Le Hégarat-Mascle, S., Reynaud, R., Kallel, A. & Hamida, A.B. 2015, 'Dynamic estimation of the discernment frame in belief function theory: Application to object detection', *Information Sciences*, vol. 306, pp. 132-49.

Ren, F. & Sohrab, M.G. 2013, 'Class-indexing-based term weighting for automatic text classification', *Information Sciences*, vol. 236, pp. 109-25.

Rieh, S.Y. 2006, 'Analysis of multiple query reformulations on the web: The interactive information retrieval context', *Information Processing & Management*, vol. 42, no. 3, pp. 751-68.

Rose, D.E. & Levinson, D. 2004, 'Understanding user goals in web search', *Proceedings of the 13th international conference on World Wide Web*, ACM, pp. 13-9.

Roy, D., Paul, D., Mitra, M. & Garain, U. 2016, 'Using word embeddings for automatic query expansion', *arXiv preprint arXiv:1606.07608*.

Sabbah, T., Selamat, A., Selamat, M.H., Al-Anzi, F.S., Viedma, E.H., Krejcar, O. & Fujita, H. 2017, 'Modified frequency-based term weighting schemes for text classification', *Applied Soft Computing*, vol. 58, pp. 193-206.

Samen, Y.U.T., Ezin, E.C. & Onana, C.A. 2017, 'An Approach of Re-Ranking Search Results based on a Dynamic and Hybrid Modeling of User Profile', *International Journal of Computer Applications*, vol. 158, no. 4.

Sánchez, D., Batet, M., Isern, D. & Valls, A. 2012, 'Ontology-based semantic similarity: A new feature-based approach', *Expert systems with applications*, vol. 39, no. 9, pp. 7718-28.

Schwartz, C. 1998, 'Web search engines', *Journal of the American Society for Information Science*, vol. 49, no. 11, pp. 973-82.

Selvaretnam, B. & Belkhatir, M. 2012, 'Natural language technology and query expansion: issues, state-of-the-art and perspectives', *Journal of Intelligent Information Systems*, vol. 38, no. 3, pp. 709-40.

Shekarpour, S., Hoffner, K., Lehmann, J. & Auer, S. 2013, 'Keyword query expansion on linked data using linguistic and semantic features', *2013 IEEE Seventh International Conference on Semantic Computing*, IEEE, pp. 191-7.

Shokouhi, M., Sloan, M., Bennett, P.N., Collins-Thompson, K. & Sarkizova, S. 2015, 'Query suggestion and data fusion in contextual disambiguation', *Proceedings of the 24th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, pp. 971-80.

Shreves, R. 2012, *Drupal search engine optimization*, Packt Publishing Ltd.

Sieg, A., Mobasher, B. & Burke, R. 2007a, 'Web search personalization with ontological user profiles', *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM, pp. 525-34.

Sieg, A., Mobasher, B. & Burke, R.D. 2007b, 'Learning ontology-based user profiles: A semantic approach to personalized web search', *IEEE Intelligent Informatics Bulletin*, vol. 8, no. 1, pp. 7-18.

Singh, J. & Sharan, A. 2017, 'A new fuzzy logic-based query expansion model for efficient information retrieval using relevance feedback approach', *Neural Computing and Applications*, vol. 28, no. 9, pp. 2557-80.

Singhal, A., Buckley, C. & Mitra, M. 1996, 'Pivoted document length normalization', *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 21-9.

Skillen, K.-L., Chen, L., Nugent, C.D., Donnelly, M.P., Burns, W. & Solheim, I. 2012, 'Ontological user profile modeling for context-aware application personalization', *International Conference on Ubiquitous Computing and Ambient Intelligence*, Springer, pp. 261-8.

Skillen, K.-L., Chen, L., Nugent, C.D., Donnelly, M.P., Burns, W. & Solheim, I. 2014, 'Ontological user modelling and semantic rule-based reasoning for personalisation of Help-On-Demand services in pervasive environments', *Future Generation Computer Systems*, vol. 34, pp. 97-109.

Solskinnsbakk, G. & Gulla, J.A. 2010, 'Combining ontological profiles with context in information retrieval', *Data & Knowledge Engineering*, vol. 69, no. 3, pp. 251-60.

Song, M., Song, I.Y., Allen, R.B. & Obradovic, Z. 2006, 'Keyphrase extraction-based query expansion in digital libraries', *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM, pp. 202-9.

Stamou, S. & Ntoulas, A. 2009, 'Search personalization through query and page topical analysis', *User modeling and user-adapted interaction*, vol. 19, no. 1-2, pp. 5-33.

Takahashi, T. & Kitagawa, H. 2008, 'S-bits: social-bookmarking induced topic search', *Web-Age Information Management, 2008. WAIM'08. The Ninth International Conference on*, IEEE, pp. 25-30.

Tao, X. & Li, Y. 2009, 'A user profiles acquiring approach using pseudo-relevance feedback', *International Conference on Rough Sets and Knowledge Technology*, Springer, pp. 658-65.

Teodoro, D., Mottin, L., Gobeill, J., Gaudinat, A., Vachon, T. & Ruch, P. 2017, 'Improving average ranking precision in user searches for biomedical research datasets', *Database*, vol. 2017.

Thesprasith, O. & Jaruskulchai, C. 2016, 'Simple-phrase score for selective query expansion in health Information Retrieval', *Computer Science and Engineering Conference (ICSEC), 2016 International*, IEEE, pp. 1-6.

Trajkova, J. & Gauch, S. 2004, 'Improving ontology-based user profiles', *Coupling approaches, coupling media and coupling languages for information retrieval*, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 380-90.

Tseng, E. 2015, 'Ranking objects by social relevance', Google Patents.

Vallet, D., Castells, P., Fernández, M., Mylonas, P. & Avrithis, Y. 2007, 'Personalized content retrieval in context using ontological knowledge', *IEEE Transactions on circuits and systems for video technology*, vol. 17, no. 3, pp. 336-46.

Vidinli, I.B. & Ozcan, R. 2016, 'New query suggestion framework and algorithms: A case study for an educational search engine', *Information Processing & Management*, vol. 52, no. 5, pp. 733-52.

Vuljanić, D., Rovan, L. & Baranović, M. 2010, 'Semantically enhanced web personalization approaches and techniques', *Information Technology Interfaces (ITI), 2010 32nd International Conference on*, IEEE, pp. 217-22.

Walther, E., Lu, Q., Ku, D., Lee, K., Tam, C.-M. & Diab, A. 2011, 'Search systems and methods with integration of user annotations', Google Patents.

Wang, H., He, X., Chang, M.-W., Song, Y., White, R.W. & Chu, W. 2013, 'Personalized ranking model adaptation for web search', *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 323-32.

Wen, J.-R., Nie, J.-Y. & Zhang, H.-J. 2001, 'Clustering user queries of a search engine', *Proceedings of the 10th international conference on World Wide Web*, acm, pp. 162-8.

Weng, J., Li, Y., Xu, W., Shi, L., Zhang, Q., Zhu, D., Hu, Y., Zhou, Z., Yan, X. & Tian, H. 2008, 'Effect of intensive insulin therapy on β-cell function and glycaemic control in patients with newly diagnosed type 2 diabetes: a multicentre randomised parallel-group trial', *The Lancet*, vol. 371, no. 9626, pp. 1753-60.

White, R.W., Ruthven, I. & Jose, J.M. 2002, 'Finding relevant documents using top ranking sentences: an evaluation of two alternative schemes', *Proceedings of the 25th annual* *international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 57-64.

Wu, L., Feng, J. & Luo, Y. 2009, 'A personalized intelligent web retrieval system based on the knowledge-base concept and latent semantic indexing model', *Software Engineering Research, Management and Applications, 2009. SERA'09. 7th ACIS International Conference on*, IEEE, pp. 45-50.

Xia, T. & Chai, Y. 2011, 'An improvement to TF-IDF: Term Distribution based Term Weight Algorithm', *JSW*, vol. 6, no. 3, pp. 413-20.

Xiang, B., Jiang, D., Pei, J., Sun, X., Chen, E. & Li, H. 2010, 'Context-aware ranking in web search', *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 451-8.

Xie, H., Li, X., Wang, T., Chen, L., Li, K., Wang, F.L., Cai, Y., Li, Q. & Min, H. 2016, 'Personalized search for social media via dominating verbal context', *Neurocomputing*, vol. 172, pp. 27-37.

Xu, J. & Croft, W.B. 1996, 'Query expansion using local and global document analysis', *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 4-11.

Xu, S., Bao, S., Fei, B., Su, Z. & Yu, Y. 2008, 'Exploring folksonomy for personalized search', *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM, pp. 155-62.

Xu, X., Zhang, X. & Hu, X. 2007, 'Using two-stage concept-based singular value decomposition technique as a query expansion strategy', *Advanced Information Networking and Applications Workshops, 2007, AINAW'07. 21st International Conference on*, vol. 1, IEEE, pp. 295-300.

Xue, G.-R., Zeng, H.-J., Chen, Z., Yu, Y., Ma, W.-Y., Xi, W. & Fan, W. 2004, 'Optimizing web search using web click-through data', *Proceedings of the thirteenth ACM international conference on Information and knowledge management*, ACM, pp. 118-26.

Yanbe, Y., Jatowt, A., Nakamura, S. & Tanaka, K. 2007, 'Towards improving web search by utilizing social bookmarks', *International Conference on Web Engineering*, Springer, pp. 343-57.

Yau, S.S. & Karim, F. 2004, 'An adaptive middleware for context-sensitive communications for real-time applications in ubiquitous computing environments', *Real-Time Systems*, vol. 26, no. 1, pp. 29-61.

Yu, J., Liu, F. & Zhao, H. 2012, 'Building user profile based on concept and relation for web personalized services', *International Conference on Innovation and Information Management*, Citeseer.

Zhou, Z., Wang, Y., Wu, Q.J., Yang, C.-N. & Sun, X. 2017, 'Effective and efficient global context verification for image copy detection', *IEEE Transactions on Information Forensics and Security*, vol. 12, no. 1, pp. 48-63.

Zidi, A., Bouhana, A., Abed, M. & Fekih, A. 2014, 'An ontology-based personalized retrieval model using case base reasoning', *Procedia Computer Science*, vol. 35, pp. 213-22.