Sheffield Hallam University

Providing personalised information based on individual interests and preferences.

AL SHARJI, Safiya.

Available from Sheffield Hallam University Research Archive (SHURA) at:

http://shura.shu.ac.uk/19228/

This document is the author deposited version. You are advised to consult the publisher's version if you wish to cite from it.

Published version

AL SHARJI, Safiya. (2016). Providing personalised information based on individual interests and preferences. Doctoral, Sheffield Hallam University (United Kingdom)..

Copyright and re-use policy

See http://shura.shu.ac.uk/information.html

Sheffield Hallam University Research Archive
http://shura.shu.ac.uk

Learning and Information Services
Adsetts Centre, City Campus
Sheffield S1 1WD

102 112 788 4

REFERENCE

ProQuest Number: 10694108

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

ProQuest 10694108

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.
This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Providing Personalised Information Based on Individual Interests and Preferences

Safiya Al Sharji

A thesis submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Doctor of Philosophy

April, 2016

Table of Contents

Dedication

This thesis is dedicated to the Al Harthy family. Without the love and support of my husband Hamed and my children Fatma, Mohammed and Sultan, this thesis would not have been possible.

ABSTRACT

The main aim of personalised Information Retrieval (IR) is to provide an effective IR system whereby relevant information can be presented according to individual users' interests and preferences. In response to their queries, all Web users expect to obtain the search results in a rank order with the most relevant items at the lowest ranks. Effective IR systems rank the less relevant documents below the relevant documents. However, a commonly stated problem of Web browsers is to match the users' queries to the information base. The key challenge is to return a list of search results containing a low level of non-relevant documents while not missing out the relevant documents.

To address this problem, keyword-based search of Vector Space Model is employed as an IR technique to model the Web users and build their interest profiles. Semantic-based search through Ontology is further employed to represent documents matching the users' needs without being directly contained in the users' specified keywords. The users' log files are one of the most important sources from which implicit feedback is detected through their profiles. These provide valuable information based on which alternative learning approaches (i.e. dwell-based search) can be incorporated into the IR standard measures (i.e. *tf-idf*) allowing a further improvement of personalisation of Web document search, thus increasing the performance of IR systems.

To incorporate such a non-textual data type (i.e. dwell) into the hybridisation of the keyword-based and semantic-based searches entails a complex interaction of information attributes in the index structure. A dwell-based filter called dwell-tf-Idf that allows a standard tokeniser to be converted into a keyword tokeniser is thus proposed. The proposed filter uses an efficient hybrid indexing technique to bring textual and non-textual data types under one umbrella, thus making a move beyond simple keyword matching to improve future retrieval applications for web browsers. Adopting precision and recall, the most common evaluation measure, the superiority of the hybridisation of these approaches lies in pushing significantly relevant documents to the top of the ranked lists, as compared to any traditional search system. The results were empirically confirmed through human subjects who conducted several real-life Web searches.

ACKNOWLEDGMENT

Following my utmost gratitude to Almighty Allah, the most Merciful and the most Compassionate for blessing me with good health, endurance and the strength to complete this study, I would like to thank Martin and Elizabeth. Martin initiated the whole process and attentively assured it was moving in the right direction. Elizabeth complemented him in always getting the product right. Our discussions were related to the topic, although they would start with some family or international students' matters. Thank you very much both of you. Your patience in teaching me the A, B, C of research, your tireless encouragement and support throughout this entire mission will never be forgotten. Part of your contributions includes an initial compulsion which leveraged my academic skills from zero to hero (PhD).

Thanks to both the administrative staff of C_3RI for their tireless advises and the Learning Centre for their continuous guidance as well as my colleagues and friends (research students) in Science Park, at Unit 12, for those discussions which were a source of constant improvement. All of this contributed in making the whole experience easier and more enjoyable.

Thanks also to Tante Khadija Salman, Sheilah Augustin, Shaikh Nasser Al Ruquaishy and Shaikh Amor Al Toqui for encouraging me to accept the offer from Sheffield Hallam University to immerse myself in research. This has been one of the most invaluable experiences of my life.

My very special thanks go to my family, including my niece Maya Al Azri and all my sisters in general, and Hamida in particular - who, while I was away, looked after *Nasra Al Azri* - an exceptional Mum - an Angel Grand-Mum of my Kids. Without their support and sacrifices, I would never have been able to make it this far. These deeds in caring for our Angel are indeed worth more than a written paragraph.

Finally, my employer (the Omani Ministry of Manpower) gave me a scholarship without which, I could not have survived. Thank you for the sponsorship, the aftermath of a PhD is, according to some, a bit weird, especially for workaholics.

Publications Based on this Work:

Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth (2013). "Enhancing the Degree of Personalisation through Vector Space Model and Profile Ontology", Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), IEEE RIVF 2013 International Conference, pp. 248-252.

Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth (2015). "A Dwell Time-Based Technique for Personalised Ranking Model", In: Database and Expert Systems Applications, Springer International Publishing, pp. 205-214.

Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth (2016). "A Relevance-Focused Search Application for Personalised Ranking Model", In: Database and Expert Systems Applications, Springer International Publishing, pp. 244-253.

LIST OF FIGURES

LIST OF TABLES

Glossary of Terms

Word	Description
Degree of Relevancy	The degree of relevancy is the extent to which a document is relevant to the query and based on how much the user's information needs are satisfied. It is determined by performing both the keyword-based similarity comparison and the concept-based similarity comparison.
Index Term Database (ITD)	The ITD is a database of index terms which is built based on all terms extracted. It consists of all index terms representing the document's features for the search the user is involved with. It also consists of a representation of the distribution of document indices for the document collections related to the user's query.
Information Retrieval (IR)	An IR system is a program designed for analysing, processing and storing sources of information in order to retrieve those that match a particular user's requirements. Baeza-Yates & Ribeiro-Neto (1999) define it as a program that studies the retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language.
Interest score	The interest score provides the relevance of the document with respect to the relevance feedback. It is based on the similarity comparison between the document representation and the keyword query representation and the user-profile's representation, taking thus into consideration the particular user's interests.
Keywords' query	Keywords' query is the set of words issued by the user while querying for information from the Web. It is used interchangeably with the word query separately or together.
Language Model (LM)	The LM is a branch of the probabilistic modelling and one of the most popular IR models. It is employed in natural language processing applications as a probability distribution over sequences of words. At search time, the top ranked document is the one which' language model assigns the highest probability to the query.

Word	Description
Normalised Discounted Cumulative Gain (NDCG)	The NDCG is the measure of ranking quality which is derived from both the Cumulative Gain (CG) and a Discounted Cumulative Gain (DCG) ranking measures. These are IR measure employed to quantify the gain (i.e. usefulness) of a document.
Personalisation degree	The degree of personalisation is an adjustable parameter which is based on the degree of relevancy. The higher the degree of relevancy, the higher the personalisation degree. This might increase or decrease depending on how much personalisation leads to relevant or non-relevant document retrieval for each individual user.
Personalised Information Retrieval (PIR)	PIR is the retrieval approach that is based on PSE and it is treated as synonymous of both PRM and PSE.
Personalised Ranking Model (PRM)	PRM is a main component of the three models of the project: keyword-based, semantic-based and dwell-based. Each model can be applied without the functionality of the other respective models. PRM is the combined Web search model that is based on the similarity merge of two or different models of the system.
Personalised Search Engine (PSE)	PSE is the retrieval system that implements the PRM to produce the ranked list for individual user.
Profile Ontology (PO) model	PO model is the name of the proposed information retrieval model which is based on the semantic-based approach. The two terms (PO and semantic-based) are treated as synonymous of each other and aim at determining the relevancy of documents expanded with Ontology terms by employing the term-weighting function approach of the Vector Space Model.
Relevance score	The relevance score provides the relevance of the document with respect to the keyword-based similarity comparison between the document representation and the keyword query representation. It is based on the binary search without taking into consideration particular interests of the user.

Word	Description
Search Engine (SE)	Merriam-Webster defines a SE as a computer program that is used to look for information on the Internet. It is thus a program that is employed particularly for finding particular sites on the World Wide Web (e.g. Google) in order to search for and identify items in a database corresponding to keywords or characters specified by the user.
Vector Space Modeling (VSM)	VSM is the name of the proposed information retrieval model which is based on the keyword-based approach. The two are treated as synonymous of each other and aim at determining the document relevancy by employing the term-weighting function approach of the Vector Space Model.

Chapter 1 Introduction

This thesis describes a framework of a novel approach to provide search results that are personalised to the needs of individual users. Instead of testing hypotheses, the thesis focuses on investigating, developing, and validating personalisation tools and their measurement methods. The research analysis is based on the assumption that, not only does the number of documents matching the searchers' input keywords differ in their degree of relevancy, but also it far surpasses the number of documents the searchers conceivably filter through; therefore, ranking the documents returned in a list of search results by their order of relevancy is crucial for search engines (Manning, Raghavan and Schütze 2008) to provide personalised information.

The key concept of this thesis is that, on account of information overload, if IR systems (i.e. search engines) can adopt personalised search to accurately filter the searchers' current context and personalise the search results, their performance will be improved, and the potential value perceived by the users will increase. Personalised search is often defined as an effort to provide to the users, individualised collections of documents based on some form of IR models which represent their individual needs as well as the context of their interaction activities (Micarelli, et al. 2007); whereby search results are tailored based on the users' preferences expressed in the form of queries.

1.1 Personalised Information Retrieval

This section provides the problem definition and the problem features upon which the current approach is built. The problems of personalised information - referred to as individual search results - are outlined to highlight the key challenges in personalising the information systems and users' information needs, prior to introducing possible solutions for addressing them.

1.1.1 Problem Definition

Information Retrieval (IR) is generally perceived as a subject field that covers both the representation and retrieval parts of information (Jones and Willet Peter 1997). The representation part is the processing (i.e. abstracting, indexing) and management (i.e. categorisation) of information, while the retrieval part is the extraction of information. The retrieval sub-field can further be divided into information access (i.e. getting or obtaining information), information seeking (i.e. user interaction with the system) and information searching (i.e. looking for information). In terms of information searching, both information access and information seeking are interim components of the process[1] (Manning, Raghavan and Schütze 2008), which form the basis of this work that occurs in a digital environment[2].

The main aim of Personalised Information Retrieval (PIR) is to provide an effective IR system whereby relevant information can be presented according to the users' interests and preferences. Effective IR systems (i.e. those with good predictor functions) provide ranked lists of search results with the less relevant documents below the relevant documents, thereby allowing searchers to browse them without having to surf through many extraneous documents. Such systems have two main characteristics: (1) Coverage: they display most of the relevant documents at the lowest ranks, and (2) Accuracy: they display a low level of non-relevant documents at the lowest ranks.

To evaluate the achievement of these fundamental characteristics or performance of an IR system in general, the two most common measures used are known as recall and precision (Cleverdon, Mills and Keen 1966, Manning, Raghavan and Schütze 2008). Recall is simply the proportion of the relevant documents displayed in the results list; and precision is the proportion of all relevant documents retrieved by the system (deferred till Section 2.4.1).

An effective IR system always seeks to maximise both recall and precision. Recall can be maximised if all relevant documents are displayed in the list of results, in which case the system will have achieved 100% recall; similarly,

---

[1] The sub-field 'information searching' will be synonymously used to mean information access and information seeking together or separately, depending on the IR environment under discussion.

[2] Since textual IR is the most popular form of information for IR models in such an environment, the terms 'information retrieval' and 'textual information retrieval' will also be treated as synonymous in this work as in most IR systems.

precision can be maximised if there is no non-relevant document displayed in the list of results, in which case.the system will have achieved 100% precision. As shown in Figure 1-1, these two measures are often inversely proportional (Voorhees and Harman 2000) and the key challenge is to strike a balance between them (Cleverdon, Mills and Keen 1966); requiring that users be thus trained into becoming experienced in managing to trade them off for their particular information needs (Marcus 1991); this study aims at achieving a balance between them while minimising users' involvement.

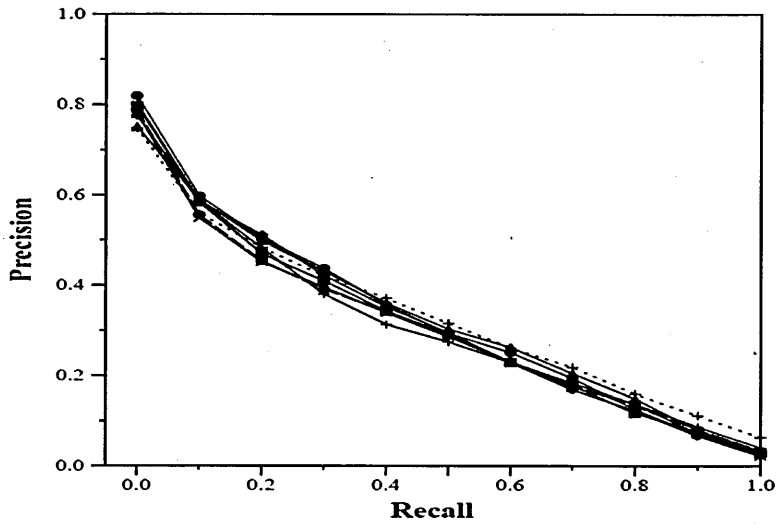

Figure 1-1 Graph showing the trade-off between recall and precision for an IR system (adapted from Voorhees and Harman (2000)). When all relevant documents are retrieved - the recall reaches 1 - the precision is at its lowest. The recall plummets, when the precision reaches 1 - the Ranked list contains only relevant documents.

Most of the relevant documents for retrieval are the users' search intents (Teevan, Dumais and Horvitz 2005b), which reside in the users' minds (i.e. searchers' information needs). These documents (i.e. information objects) share some vocabulary in common (i.e. have similar terms) with the query keywords. They represent the searchers' information needs which they seek from the Web by expressing them in the form of keywords. In terms of users' information needs, they are referred to as interest-based information.

With regard to information searching, there are several types of information access (or retrieval) for information needs, but in this framework, the focus is on the most common approach of information access used in many IR models known as the keyword-based search technique (Baeza-Yates and Ribeiro-Neto 1999, Haav and Lubi 2001, Castells, Fernandez and Vallet 2007). However, retrieving relevant information for different individuals characterised by different

information needs and different information seeking strategies based on these search techniques is the main challenge in PIR; and it is the focus of this thesis: developing interest-based search techniques to provide personalised search results characterised by both coverage and accuracy.

In terms of users' information seeking, the author holds the hypothesis that the acknowledgement of personalising Web search results showed the way for a great improvement in the search experience (Matthijs and Radlinski 2011). The problem of information needs can thus be tackled using a keyword-based search technique to match the keyword query with both the previously clicked documents and the information base residing on the Web based on individual interests. Personalised information is thus built upon this information access and information seeking process. In other words, it is an IR tool that allows individual users to interact with the system to retrieve *immediately* (i.e. achieving coverage) and *exactly* (i.e. achieving accuracy) what they need when they type in their keyword queries in the form of a natural language statement.

The following scenario presents the problem of personalised information in a more realistic manner in discerning individual search goals (Teevan, Dumais and Horvitz 2005b). To provide *relevant documents*, a personalised search for the input keyword *IR* returns at the top of the list, results such as the SIGIR homepage for the IR researcher, and Infrared Light for the chemist, while for the financial analyst, it returns the stock quotes for Ingersoll-Rand. Figure 1-2 exemplifies the scenario of a typical set of personalised search results for the keyword *IR* as googled by an IR researcher.

In this research, both the input keywords and the contents of the previous web pages visited by the Web users (represented in a form of users' profiles) are employed to be matched with the information objects. They actually express the items representing users' interests and preferences. They can be used to estimate the relevance of the documents in the targeted collection by using the term weighting (see definition 2.1 - deferred till Section 2.2) score to provide a corresponding term weighting for each document. Using such a score to rank the documents in decreasing order is referred to as a Personalised Ranking Model (PRM), the documents that are potentially most interesting to the user

are pushed at the lowest ranks (i.e. the top ranks) in the results list (Salton and McGill 1983, Salton 1989) to provide the PIR.

1.1.2 Problem Features

This section features the problem of PIR with respect to document collection, user patterns, and research methodology in order to provide a sensible approach to the problem currently being addressed.

From the data point of view, personalised information can be perceived as a collection of documents each of which is characterised by a unique metadata (e.g. keywords/terms or document content). For instance, each document presented by the system to the user consists of numerous terms (i.e. important terms, less important terms etc.), thus, the term position, its frequency in that document (identified by document number) are major features in determining the document's importance with respect to the input keyword (Salton and McGill 1983, Manning, Raghavan and Schütze 2008); whereas the keyword (i.e. input query-number), the IP address and the dwell time are features that obviously characterise individual users' patterns during their interaction with the system. Given that these documents are sourced from the Web, while the user holds the search intent (i.e. interests), this characterisation is even more valid.

Since a search engine (SE) interacts with several individuals as stated earlier, the system must adapt to different behavioural patterns conceived by different individuals. Such a system forms a Personalised Search Engine (PSE) and allows the individual user's information access and information seeking to be taken into account in order to provide personalised search results.

At this point, it is important to note the assumption that during the information seeking and searching process, the users' queries consist of similar terms available in both documents previously visited and the information base of documents. Throughout this work, the design of the proposed PRM to provide the PIR model is based on this assumption.

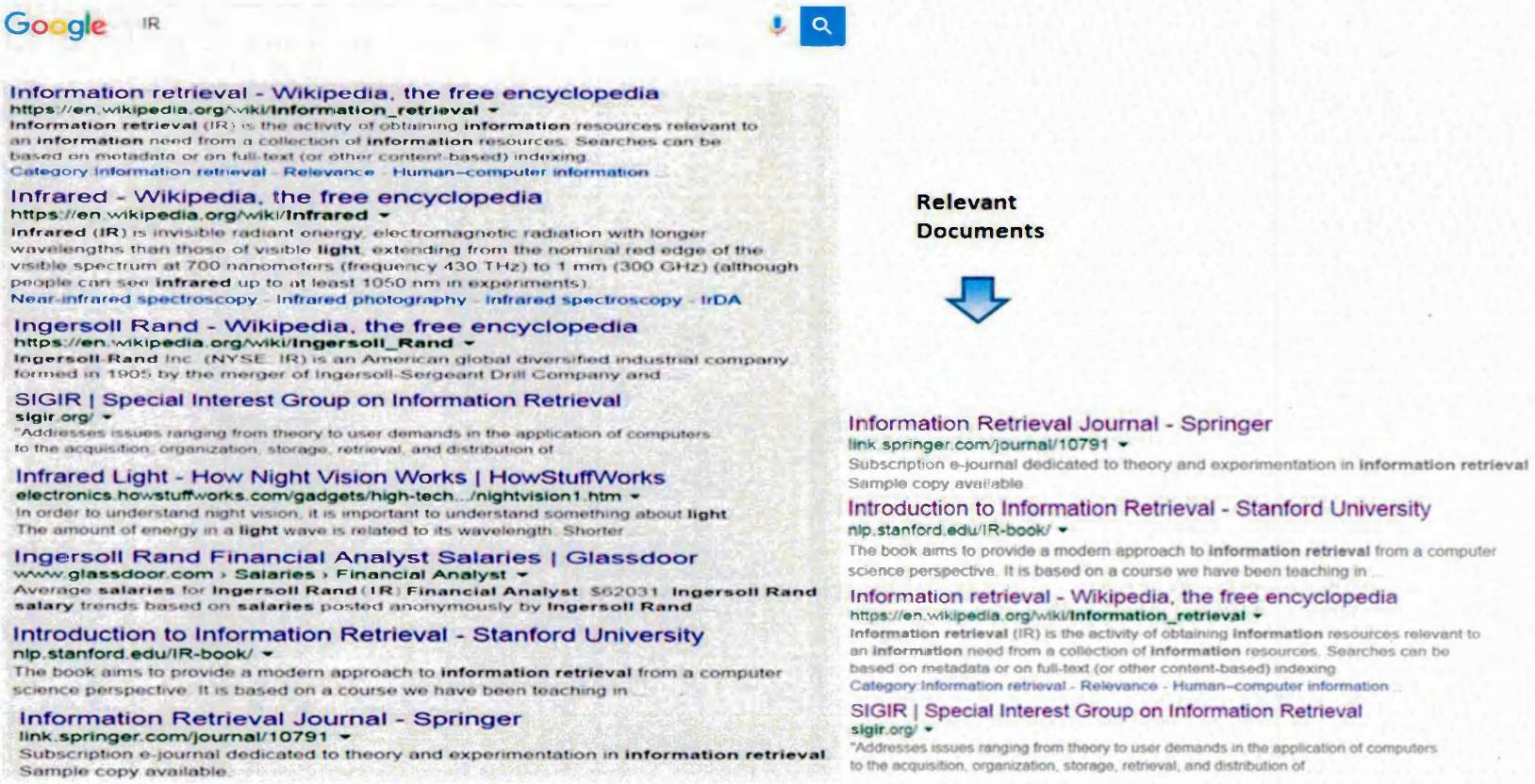

Figure 1-2 An example of personalised information retrieval: searching for relevant documents from the Web. A ranked list of returned results (ieft) for a keyword-based search shows results matching with the query in many different documents, while the relevant documents (right) are for a user with particular interests.

1.2 Personalised Information

This section introduces the current approach to the problem. First, the information seeking (i.e. information needs) process in PIR is described to provide the basis for the proposed approach, before introducing the current retrieval model in more detail.

1.2.1 Model of Information Seeking

To provide an effective model of information seeking for personalised search results, it is important to propose a hypothetical model of the problem and derive specific approaches from it. Figure 1-3 shows a model of information seeking for search results ranked according to documents which might be useful to searchers (i.e. their interests). Here, the same scenario is presented with the chemist, financial analyst and researcher labelled as user X, user Y and user Z respectively (due to limited space), basing their information needs on individual interests.

From the retrieval dimension, the users' information needs ought somehow to be formulated into realistic queries which allow the retrieval process to be initiated. The information seeking model should support queries which are both keyword-based (i.e. the user's input keyword) and interest-based (i.e. the user's useful documents). This means that even though users have individual interests, they should find documents relevant to their information needs. The proposed retrieval system uses the users' keyword queries as inputs to produce the results in the form of a ranked list depending on relevant documents based on users' individual interests. This overall information retrieval process - from the input keyword query to the display of the ranked list - forms a framework on which the current approach is established.

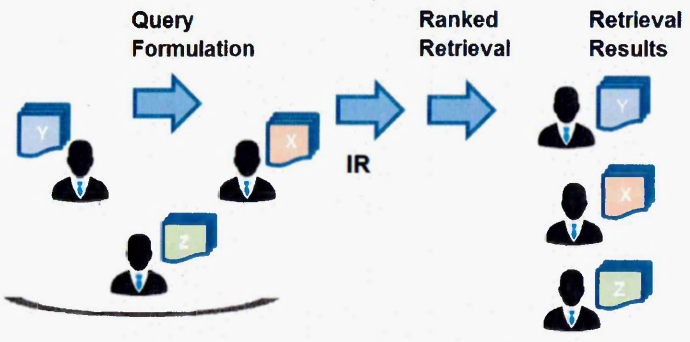

Figure 1-3 Information Seeking Model for Personalised Information

In summary, the proposed IR framework aims to find relevant results based on an individual user's interests which in turn are based on the importance of the keywords (i.e. terms) which are available in the documents with respect to input keywords. The evaluation framework is based on a model of the users' information seeking in general and users' interests in particular, with the goal of achieving both a high coverage and a high accuracy.

1.2.2 Retrieval Framework

Based on the description of the model of information seeking provided above, a vector space model (i.e. keyword-based search) and profile ontology (i.e. semantic-based search) are proposed as the primary retrieval models. More specifically, the keyword-based search allows users to find those documents (i.e. digital collections) matching their keyword queries; while a semantic-based search allows a set of related terms (i.e. concepts which function as expanded input queries) to be integrated into those documents for matching terms related to the input keywords which are not directly expressed in the query.

These retrieval models are based on standard weighting schemes known as $tf \bullet idf$ functions (deferred till Section 2.2) which are often evaluated by the effects of the term frequency (see Section 1.1.2) involved in a document. While the keyword-based search utilises the term-weighting scheme to determine the estimation of document weights to provide relevant documents on search space, the semantic-based search applies the same mechanism to detect the relevant documents which include concepts not expressed in the original input keywords. By developing a means of using the associations of users' interests between documents, the chain of associations can similarly surrogate a form of information retrieval (or access). A comparison of the two retrieval models is shown in Table 1-1 indicating that the two models are based on different assumptions around user's interests, although they are still both based on the statistical co-occurrences of terms, they require different types of input to provide relevant documents linked to the input query. They are both general retrieval techniques which use similarity measures to match and weigh terms occurring in users' queries with terms occurring in documents in Web information base in order to meet their information needs.

Table 1-1 Comparison between keyword-based and semantic-based searches

	Keyword-based Search	Semantic-based search
User's Interests	Relevant documents	Related Relevant document(s)
Input	Keyword	Keyword + Related-Terms

The proposed information searching and browsing models are thus combined into a single scheme in a natural way to provide the textual retrieval model. For instance, the proposed approach not only minimises the effort required from users, since both approaches (i.e. keyword-based and semantic-based searches) do not require their involvement in predicting useful documents, but it also automatically builds associations between relevant documents to achieve maximum recall and precision without involving the user. Furthermore, it also incorporates a more precise good indicator than term recurrence, according to empirical studies by Agichtein, Brill and Dumais (2006), Hassan and White (2013) and Jiang, Pei and Li (2013), namely the dwell, as a *learning* feature to achieve a more robust IR model. Such model pushes relevant documents to the top of the list returned and provides search results with both high coverage and accuracy.

1.2.3 Research Question

A major challenge in performing a personalised search is to learn about individual users' interests and preferences. A number of research studies which relied on the original $tf \bullet idf$ functions to employ the hybridisation of the Vector Space Model (VSM) and the Profile Ontology (PO) proved effective in the ranking process to provide personalised search results (Gauch, Chaffee and Pretschner 2003, Vallet, Fernández and Castells 2005, Castells, Fernandez and Vallet 2007). While the combination process is somehow recursive in nature, accuracy is a critical criterion for the success of a personalisation solution. The current work thus seeks to answer the following question: '*How can the term-weighting function approach of the Vector Space Model (VSM) improve the personalisation of Web document searches?*'

1.3 Objectives

The following objectives are set with the main goal of investigating an answer to the above stated research question:

1. to use a set of parameters comprising of keyword-entered (i.e. search query) and document-clicked (i.e. URL-visited) based on a real life collection of search logs in detecting the searchers' individual interests;
2. to identify salient features describing the search queries issued and the document contents using the term-weighting schemes of the VSM approaches in order to exploit an implicit mapping between query-terms and document-terms to identify relevant documents based on individual interests;
3. to combine (1) and (2) and apply IR ranking strategies in addition to incorporating other features to re-rank the search results of a non-personalised search engine; and
4. to identify some shortcomings of the existing term-weighting functions to reduce the gap between relevance probability and retrieval probability to improve the accuracy of future retrieval applications for web browsers.

1.4 Contributions

This research resulted in the following contributions:

1. A suite of approaches (VSM-based and ontology-based) which exploits an implicit mapping between two fields (i.e. both query-terms and document-terms) to provide an effective PSE. Compared with the Google Search Engine (GSE), the scheme devised in the proposed approach achieved a 14% improvement in relevance at the 10 top ranks.
2. A multi-strategy novel approach to IR for the re-ranking of search results returned by a non-personalised search engine based on a real life collection of search logs to provide a reliable PRM of information retrieval.
3. An indexing mechanism which is based upon a filter pattern called dwell-tf-idf; makes a move beyond simple keyword matching to improve the accuracy of future retrieval applications for web browsers.

1.5 Summary

A brief contextualisation of this work and its main contributions as outlined in the preceding sections are detailed in a total of six chapters organised as follows:

Chapter 2 - provides related or similar work on IR standard approaches and some of the typical systems adopted in the proposed model. Various models of IR are presented. A thorough discussion of term weighting approaches and the term weighting framework used in the current work are outlined.

Chapter 3 - provides related or similar work on IR approaches to personalised search involving models of user context which expand the user's query through ontology terms. A number of techniques to derive interest scores implicitly and the similarity matching procedures are reviewed.

Chapter 4 - points out the gaps identified in the above chapters and introduces the proposed approach framework. This chapter presents the proposed approach along with its respective strengths (motivation for the deployment of the personalised search application).

Chapter 5 - outlines and discusses the preliminary study conducted based on the personalised search application aiming to show some early experiments.

Chapter 6 - outlines and discusses the results of the related main study. The shortcomings related to the problem of the hybridisation of the VSM and PO model are addressed in detail in this part. The solution adopted - by incorporating the dwell-based feature to improve the performance of the retrieval models - is detailed.

Chapter 7 - Concludes and outlines some interesting future directions to be addressed.

Chapter 2 Background and Related Work

The main goal of this chapter is to equip the reader with a background to the major textual retrieval models prior to outlining the key strengths and shortcomings investigated in the existing retrieval models. Various models of IR are outlined in Section 2.1 to clarify the basis of the multi-strategy approach to IR employed and developed within the project. The theories underlying different classes of term weighting functions are surveyed in Section 2.2. Section 2.3 specifies the background material of the VSM - the underlying model of this work. Section 2.4 discusses the main IR evaluation metrics, while the important points detailed in the chapter, are summarised in Section 2.5.

2.1 Information Retrieval Models

Information Retrieval is a field that deals with natural language queries and documents and attempts to address the problem of information superabundance by automatically returning to the users, only those documents that are deemed relevant to their needs (queries). Traditional information retrieval systems (Baeza-Yates and Ribeiro-Neto 1999) are based on the Boolean retrieval model (Belkin and Croft 1992). This is referred to as the exact matching retrieval approach. Modern IR systems, on the other hand, are based on the factual and statistical retrieval model, referred to as the partial matching retrieval approach (Belkin and Croft 1992). This retrieval model is also known as the Best Match (BM) model and it includes the statistical retrieval models (i.e. the VSM and the probabilistic retrieval model) and as will be described in Section 3.3, the linguistic and knowledge-based models (i.e. commonly based on expert system techniques or Artificial Intelligence (AI)). Some of these models and particularly the VSM will be discussed in the following sections with a particular focus on how weight of terms can automatically be assigned by modern IR.

2.1.1 Boolean Model

Information Retrieval systems based on the Boolean model in which queries are posed in the form of a Boolean expression of terms (Manning, Raghavan and Schütze 2008) are classical models which detect documents that match the Boolean type query by using Boolean logic. Simple logical operators (i.e. *AND, OR* and *NOT*) are employed to formulate queries used to detect a set of documents. This means that a document is either *relevant*, in which case it should be contained in the returned set; or it is ***not*** *relevant*, in which case it should not be in the returned set. This IR model (Rijsbergen 1979) is mainly based on the assumption that a document is considered relevant if it comprises a term or a set of terms matching the query keywords.

However, despite its simplicity and being easy to implement, this model suffers from two major drawbacks. A partial matching of a document is not provided in this model which makes it impractical to provide a ranked list of results on a gradual level-based matching scheme. Yet, exact and partial matching should be combined and treated as complementary to make effective use of the respective strengths of both approaches (Belkin and Croft 1992). The other drawback is that it puts pressure on the user to formulate proper queries, as it requires a certain level of knowledge to construct Boolean queries.

2.1.2 Probabilistic Model

Considering the users' information needs mentioned earlier in Section 1.1.1 - translated into query representation - IR systems similarly convert documents into document representations. The two representations differ at least by how text is tokenised, but perhaps query representation contains fundamentally less information, as when a non-positional index (i.e. postings - deferred till Section 3.1.1.1) is used (Manning, Raghavan and Schütze 2008). Thus IR systems try to determine how well documents achieve the users' information needs based on these two representations.

In the Boolean models of IR, although formally defined, the matching of semantic calculation of index terms is not precise. In IR systems, the understanding of the users' information needs is not certain since it is only based on the representations of their queries. Using the two representations (i.e. query and document), the IR systems similarly try to predict the relevance of the

content of a document satisfying the users' information needs. The theory of probability - first proposed by Robertson and Jones (1976) - provides a principled foundation for such reasoning under uncertainty (i.e. to estimate how likely a document could be relevant to an information need (Manning, Raghavan and Schütze 2008)). It is grounded on the Probability Ranking Principle (PRP (Robertson 1977)) which forms the optimal performance based on the ranking of documents performed by the probability of their relevance, which is in turn, inferred according to the terms distribution of both relevant and non-relevant documents.

The probabilistic model of retrieval assumes the independence of terms (Robertson and Jones 1976) - a recurring assumption of most of the IR models - which is mainly due to both its simplicity and the overhead cost of assembling details of the co-occurrence and interdependence of words. Another assumption in the probabilistic model of retrieval is that the relevance of a document is based on the binary query terms (i.e. not present/present) in that document. Since the details of terms distribution in both the relevant and non-relevant document sets are not known in advance, it becomes harder to determine the optimal weighting strategy (deferred till Section 2.2). Thus, this information (i.e. the probable relevance distribution) must be predicted (through term weighting scheme such as Okapi BM25 - see Section 2.2.4.4) by using the measures of term frequencies at the document level and collection level. Therefore, the concepts of term weighting to score and judge the relevance of a document - as will also be seen throughout this work - form an integral part of the IR models. A derivation of the probabilistic model along with its associated term frequencies measures are introduced in the next section.

2.1.3 Language Model

The language model (LM) is the branch of the probabilistic modelling and one of the most popular IR models. It was first incorporated into IR (Ponte and Croft 1998) for natural language processing in order to model the probability of a sequence of words. It was considered as a generative process (Ponte and Croft 1998) to be used for document ranking based on the likelihood of its providing the query terms. Equation 2.1 (Lu 2013) can be used to implement the maximum likelihood estimation to infer the likelihood of the presence of a term in an LM document.

$$\hat{P}(t/D) = \frac{tf_t^D}{|D|} \quad (2.1)$$

where

1. tf_t^D is the frequency of term in the document, and
2. $|D|$ is the document length.

Assuming document models (i.e. documents topically represented by probabilistic models) are smoothed against the collection model and commonly modelled as multinomial distributions, the Jelinek-Mercer smoothing function shown in equation 2.2 (Zhai and Lafferty 2004) can be used to blend the probability estimated from both the document and the collection (Lu 2013). Thus the data sparseness problem can be addressed and the standard LM weighting can then be decomposed into equation 2.3 (Zhai and Lafferty 2001):

$$\hat{P}_{smooth}(t/D) = \lambda\hat{P}(t/D) + (1-\lambda)\hat{P}(t/C) \quad (2.2)$$

where

$\hat{P}(t/C)$ is the collection frequency of the term divided by the total term count.

$$\log p(Q/D) = \sum_{i=c(q_i D)>0} \log \frac{p(q_i/D)}{(1-\lambda)p(q_i/C)} + m\log(1-\lambda) + \sum_{i=1}^{m} \log(q_i/C) \quad (2.3)$$

where

1. $p(q_i/D)$ is the probability of a query term in a document,
2. $p(q_i/C)$ is the probability of a query term in the whole collection,
3. λ is a parameter that controls the amount of smoothing, and
4. $\sum_{i=c(q_i D)>0} \log \frac{p(q_i/D)}{(1-\lambda)p(q_i/C)}$ is the only component influencing the rankings,

which shows that the LM weighting scheme is literally proportional and inversely proportional to the frequency of terms in the document and in the collection respectively, which is quite similar to the technique employed in the common $tf \bullet idf$ weighting scheme (see next section) although term frequency is used in the whole collection in LM rather than in the document.

It can already be noted from the IR models discussed above that the term weighting is an important issue in IR systems. Thus, the term weighting problem for IR requires a system to automatically assign weights to query-term and document-term pairs. The assigned weight should accurately reflect the term's

importance with respect to the query or document whereby more important terms are assigned a higher weight. The weighting scheme or strategy for determining the weight of different terms should thus be effective enough to provide a highly effective underlying retrieval model achieving both maximum precision and recall. Thus, the next sections give a review of scoring and weighting terms in IR models before the VSM is discussed.

2.2 Scoring and Term Weighting

In modern IR, the functions used to determine term weighting are referred to as ranking functions or term weighting schemes. Term weighting is thus a sub-field of information retrieval that studies the question of the importance of a word or a set of words in a given text (Yu, Lam and Salton 1982, Salton, Yang and Yu 1975, Salton and Buckley 1988).

2.2.1 Term Importance

The statistical approach to searching information was first proposed as term frequency occurrence to measure the term's usefulness (Luhn 1957, Luhn 1958). Luhn's invention provided a counting method for determining word significance. His technique ranked the words according to their frequencies. He employed *Zipf's law (1949)* - which states that the product of the frequency of words and the rank order is almost constant - as a null hypothesis to specify two arbitrary chosen cut-off points. These enabled him to (1) reject top-ranked terms in distinguishing documents, i.e. those terms that are too frequent; (2) reject very low ranked terms, namely those terms that are not frequent and do not significantly contribute to the document content (see Figure 2-1).

Luhn's technique actually measures the so-called *resolving power* to promote terms which are neither too frequent nor too rare. Luhn's approach further suggested using a degree of similarity (in IR systems, the relatedness between two texts is referred to as similarity) between documents' representations to search a document collection. His ideas reveal that the more two representations match in given terms and their distribution, the higher the probability of their representing similar information is.

The ideas of many of the on-going concepts of term weighting schemes which employ the occurrence of a term frequency are based on the concept of *resolving power* initiated by Luhn. Since then, subsequent term weighting scheme ideas (Yu, Lam and Salton 1982, Greiff 1998) have justified and confirmed Luhn's approach; and the statistical issue of determining a term's importance in a given document has remained vital for IR systems (Salton, Yang and Yu 1975, Buckley 1993).

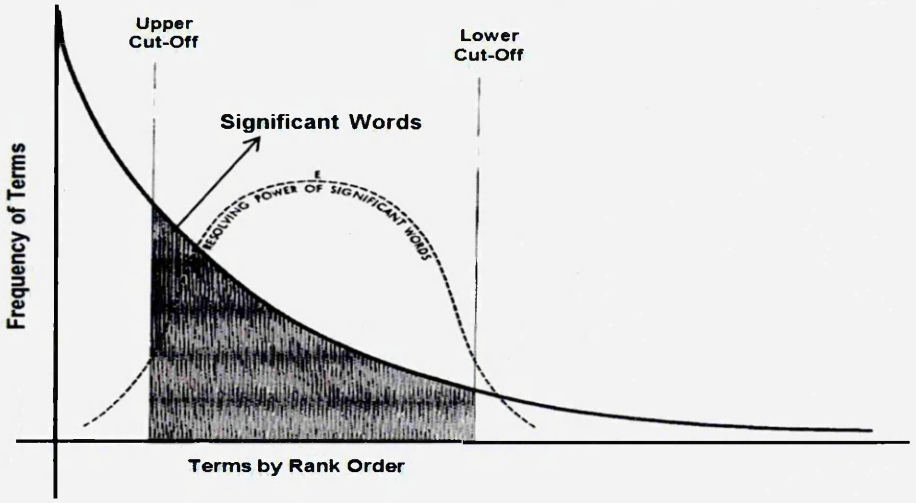

Figure 2-1 Relating frequency of term occurrence with rank order
Adapted from (Luhn 1958)

> ***Definition 2.1*** *Term weighting is the measure or ranking function utilised to detect and/or determine the importance of a term contained in the user's query while matching it with terms contained in the document representing the user's information needs.*

It was stated in Section 1.1.2 that each document presented by the system to the user consists of important terms, less important terms, etc. In order to estimate the usefulness of a document, it is crucial to differentiate between them. For example, the so-called stop-words[3] (i.e. "a", "the", "can", "and", etc.) commonly utilised in most documents, do not represent their contents, but are bound to be frequent in them as well as in the inputted queries. The expectation of matching such documents (i.e. utilising many stop-words) with the keyword queries is very high.

[3] http://www.lextek.com/manuals/onix/stopwords1.html last accessed for this project on 30/September/2015.

Furthermore, if an uncommon keyword (e.g. *peace*) is present in both the keyword query and the document, the chance that this document will be useful to the user is in reality very high. This observation indicates that different words should be given a different level of importance in the matching process. Proper approximation of term weights can significantly improve the ranking effectiveness of IR systems (Buckley 1993). Proper matches play a major role in the predicted relevance to estimate a better prediction of document relevance.

However, assigning a high weight to the more important terms and a low weight to the less important terms is a key issue in IR which is worthy of a literature review in its own right (Salton and Buckley 1988, Buckley 1993). The final estimated relevance of the document can be made proportional to the importance of the document and its term importance in the query using proper term matching between a query-term and a document. Nevertheless, it has been empirically observed that the importance of a term in textual retrieval might actually be affected by the following three main factors which are defined as a triple: a term frequency element; a term discrimination element and a form of normalisation. Various interpretations have been validated over the years (Salton and Buckley 1988, Maron and Kuhns 1960), to incorporate this triple into term importance in one way or another.

2.2.2 Term Frequency

Documents that repeatedly use a query term are potentially more relevant to the searcher than documents using that query term rarely (Maron and Kuhns 1960, Luhn 1957). For example, considering the query term '*peace*', a document containing this term '*peace*' a dozen times, for example, is theoretically more relevant to the searcher than another document which contains the word '*peace*' less than a dozen times (e.g. only two or three times).

The number of occurrences of a word in a text indicates the importance of that word (Salton and Buckley 1988, Luhn 1957) in that text. Therefore, to estimate the term weight in a text, it is imperative to use some monotonically cumulative function of the number of occurrences (i.e. frequencies) of a term. This number of term occurrences, called the **T**erm **F**requency of a term in term weights, is known as the *tf* factor (Salton and McGill 1983). The *tf* factor is thus a function that is used to calculate a term's importance in a text (i.e. document).

In IR systems (Manning, Raghavan and Schütze 2008), the term frequency for example denoted as tf_t^D - defined below - provides one measure of how well that term describes the document content. However, since each document's length is different from the other (see Section 2.2.3.2), the term frequency is often divided by the total number of terms in the document to normalise the term frequency. If for example nt and nT are respectively the number of times the term t occurs in a document d and the total number of terms in that document, then $tf_t^D = nt/nT$. Some other *tf* factors commonly used include the logarithmic *tf* factor, Okapi's *tf* factor, the augmented *tf* factor and the binary *tf* factor (Manning, Raghavan and Schütze 2008).

> ***Definition 2.2*** *Term Frequency is the number of occurrences of a keyword or term in a document - while taking into account the document length - to provide one measure of how well that keyword or term describes the document content.*

2.2.3 Inverse Document Frequency

In documents where common terms (i.e. stop-words) are used repeatedly, the matching of such a function term should contribute less towards these documents predicted relevance as opposed to the matching of uncommon terms (i.e. terms that are used in few documents (Salton, Yang and Yu 1975, Salton and Buckley 1988)). A term that occurs in all the documents in the collection is assigned a zero weight (Jones and Willet Peter 1997). To measure the importance of a term, an inverse function of the document frequency, called **I**nverse **D**ocument **F**requency is used in term weights.

Thus, while the **D**ocument **F**requency provides the number of documents containing a term, the **I**nverse **D**ocument **F**requency (*IDF* - defined below) provides the measure of the frequency rarity of the term across all documents. To calculate the *idf* factor, the total number of documents can be divided by the number of documents comprising that term, and then the logarithm can be used as shown in equation 2.4 (Maron and Kuhns 1960).

$$idf = \log \frac{N}{df} \tag{2.4}$$

where N and df are the total number of documents in a collection and the document frequency of the term respectively.

The idf factor is thus the term discrimination constituent which often aims to determine the effectiveness of a term search by using the term's characteristics in the collection as a whole. Typically, a higher weight is given to a term that occurs in fewer documents as it seems to be the better descriptor of a document. Term discrimination of an element is commonly used either directly or indirectly in most term weighting approaches to promote a term that is more likely better at identifying certain documents.

> ***Definition 2.3*** *IDF is the estimation of the rarity of a keyword or term across all documents - a term thus appearing in all the documents of the collection gets the lowest weight* (Manning, Raghavan and Schütze 2008) *- to measure the relative importance of that keyword in the whole document collection. Thus frequent terms (i.e. "of", "is"...) are weighted down so that rare terms can be scaled up.*

Some term importance functions derived from the probabilistic term relevance for weight estimation have been proposed (Croft and Harper 1979). They estimated the weight for term relevance based on the terms distribution in both relevant and non-relevant documents. They demonstrated that some reasonable assumptions can allow an idf analogous function in which the term's relevance weight is reduced to $\log(n-df/df)$, to be achieved. Some other alternatives to idf factor which were based on different importance estimators related to the theory of term discrimination were also proposed by Salton and Yang (1973) and Salton, Yang and Yu (1975).

Actually, the term discrimination theory hypothesises that if terms in the vector space produce document vectors which are not close to each other (i.e. they are apart), they are considered good discriminators of those terms as opposed to document vectors which are close to each other. Such terms thus assign a weight which, according to Manning, Raghavan and Schütze (2008) is:

1. highest when t occurs many times within a small number of documents to lend to them, a highly discriminating power;

2. lower when t occurs many times within many documents or fewer times within a document since it has less pronounced relevance; and
3. lowest when t occurs in almost all documents .

Considering Luhn's concepts of *resolving power* seen earlier based on terms with a low rank (infrequent terms), it can be interestingly noted that it is not consistent with this term discrimination theory since terms which are not frequent receive the highest weight.

2.2.3.1 The $tf \bullet idf$ Factor

The composite of the above two factors - the tf factor and the idf factor - denoted as $tf \bullet idf$ is defined below, and is called the $tf \bullet idf$ function. It is the weighting scheme which is commonly used to assign the final weight of a term (Salton and McGill 1983, Manning, Raghavan and Schütze 2008). More formally, a term's weighting score can be obtained by finding the product of both the tf and idf factors as shown in equation 2.5 (Manning, Raghavan and Schütze 2008) and it is commonly used to assign to term t a weight in document d.

$$tf \bullet idf = tf_t^D \times idf \tag{2.5}$$

Definition 2.4 *The $tf \bullet idf$ measures the importance of a keyword or term in a document with respect to a document collection or corpus. It provides a numerical statistic of a term weight which proportionally increases to the number of times the keyword occurs in the document while offsetting the frequency of the keyword in the document collection.*

2.2.3.2 Document Length

Documents that are long, apart from the fact that they use numerous different terms, also have in general a higher probability of repeating terms than short documents, thus, they might contain higher term frequency simply because they are longer (Singhal et al. 1996). As a consequence, the number of matches between the query and the document becomes high, and a retrieval preference might be based on long documents although they might be irrelevant to the user.

This dependence of the tf on the document's length is actually due to the fact that the large term frequency factors of long documents increase the overall

average weight of terms in those documents. In turn, this increases the number of individual matches to a query within a long document for query-document similarity, which results in a high overall similarity (Singhal et al. 1996, Singhal, Buckley and Mitra 1996). Therefore, document length normalisation can be factored into an IR system to address the bias of such long documents.

Given all the above definitions, it is now clear that the role of a weighting scheme is to assign a weight to each dictionary term appearing in a document to represent its relevance based on the meaning of the document it occurs in. Thus, each document can be viewed as a vector with term components (found by using equation 2.5) of the dictionary terms. The score of a document d can be grounded as the sum, over all query terms (Manning, Raghavan and Schütze 2008), of the number of times each of the query terms occurs in d. This idea can then be refined as in equation 2.6 (Manning, Raghavan and Schütze 2008) so that the $tf \bullet idf$ weight of each term in d, rather than the number of occurrences of each term t in d, is added up.

$$Score(Q,D) = \sum_{t \in q} tf_t^D \bullet idf \tag{2.6}$$

2.2.4 Document Length Normalisation

Document length normalisation known as the **N**ormalisation of **T**erm **F**requency (ntf) is a form of penalisation for the term weights in a document with respect to its length to avoid longer documents being over-weighted simply because the occurrence of these terms is higher in these documents. Normalisation techniques employed as the basic techniques for document length normalisation in most information approaches include maximum tf normalisation (Salton and Buckley 1988), cosine normalisation (Salton and Buckley 1988), pivoted document length normalisation (Singhal, Buckley and Mitra 1996) and BM25 term weighting (Robertson, et al. 1993).

> ***Definition 2.5*** *The normalisation of term frequency (ntf) is the measure used to adjust the dependence between term frequency and document length while determining the importance of a keyword within a particular document.*

It is clear at this point that the term weighting function is an important predictor used in most modern IR systems to estimate a term's importance in a text (Buckley 1993). The main reason for the current work to also take advantage of term weighting functions to determine the scores of the relevant documents based on users' input keywords and consequently their interests. An effective weighting scheme for personalising search results can be based on the combination of the observations of the above statistical scores with the following empirical observations formulated by Salton (1989):

1. multiple occurrences of a term in a document are as relevant as single occurrences (*tf* assumption);
2. rare terms are as relevant as frequent terms (*idf* assumption); and
3. long documents are not preferred to short documents (*ntf* normalisation assumption).

2.2.4.1 Maximum tf Normalisation:

Maximum *tf* normalisation is one of the well-studied techniques of normalisation used in IR. It is based on the individual *tf* weights for a document obtained from the maximum *tf* weight in the document. It is often found by using the common function shown in equation 2.7 (Salton and Yang 1973, Manning, Raghavan and Schütze 2008). An example of such normalisation is the augmented *tf* factor (Salton and Buckley 1988).

$$ntf = \alpha + (1-\alpha)\frac{tf}{\max tf \text{ in } text} \tag{2.7}$$

where α is a smoothing parameter with a value between 0 and 1 and *tf* and $\max tf$ are respectively term frequency and the maximum term weight in the text.

However, the drawback of this normalisation technique is that it is hard to fine tune since it is an unstable method which requires intensive alteration of term weightings (thus, the ranking), when the stop-word list is changed.

2.2.4.2 Cosine Normalisation:

Cosine normalisation is a standard normalisation technique used in the VSM (Salton, Wong and Yang 1975). It can be used in every document to combine the reduction of the bias of both higher *tfs* and more terms (the main reason for normalisation as mentioned in Section 2.2.4). The Cosine normalisation factor

can be defined by the expression $\sqrt{w_1^2 + w_2^2 + ...w_t^2}$ for which a Cosine correlation between two vectors - a document vector D and a query vector Q, can further be refined as shown in equation 2.8 (Salton and Buckley 1988, Salton 1989). The final normalised document-term weight is obtained by dividing the original document-term weight by this normalisation factor.

$$\cos(Q,D) = \sqrt{\sum_{i=1}^{n} w_i^2} \tag{2.8}$$

where w_i is the $tf \bullet idf$ weight for the i^{th} term, and n is the number of unique terms in the document.

The key problem in Cosine normalisation is that the higher frequencies of an individual term, increase the individual w_i values, and consequently increase the penalty on the term weights. However, a lighter penalty for the longest documents should be accounted for as the probability of relevance is assumed to be totally independent from the document length, and yet, "it is more likely that very long documents do have a slightly higher chance of being truly relevant to a query, since they have more content" (Kulp and Kontostathis 2007).

2.2.4.3 Pivoted Document Length Normalisation

Pivoted Document Length Normalisation (PDLN) is another successful term weighting scheme considered to be effective in the IR literature (Singhal, Buckley and Mitra 1996). As with BM25 (see next section), PDLN also proved to be more effective compared to Cosine normalisation in some collections, but the key disadvantage of these techniques is that they require extensive training. The PDLN scheme uses the matching function shown in equation 2.9 (Singhal, Buckley and Mitra 1996) to calculate the score of a document.

$$PIV(Q,D) = \sum_{t \in q \cap d} \left(\frac{1 + \log(1 + \log(tf_t^D))}{(1-s) + s.\frac{dl}{dl_{avg}}} . w_1 . tf_t^Q \right) \tag{2.9}$$

where

1. s is simply the normalisation parameter which is given a default value of 0.2, and it is referred to as the slope; and
2. w_1 is the idf function which is given by equation 2.10.

$$w_1 = \log\left(\frac{N+1}{df_t}\right) \tag{2.10}$$

2.2.4.4 BM25 Scheme

BM25 (Robertson, Walker and Hancock-Beaulieu 1995) is actually a broad family of scoring functions which differ slightly from each other in terms of components and parameters. It is a classical IR model which is considered as parametric (Metzler and Zaragoza 2009) with tuning parameters, as in PDLN, to allow more degrees of freedom in weighting functions since there are fewer functional restrictions. Being thus one of the most effective term weighting schemes in the field of IR, it is the most widely-used benchmark scheme against which any new term weighting schemes can be assessed. The BM25 model is normally referred to as Okapi BM25 to indicate that it was first implemented by the Okapi IR system. It derives from the probabilistic model of retrieval (see Section 2.1.2) which ranks a set of documents based on the PRP in decreasing order of likelihood of relevance, using the general form of the model as in equation 2.11 (Metzler and Zaragoza 2009).

$$S(Q,D) = P(r/q,d) \sum_{t \in Q \cap D} \log \frac{P(t/r)P(\bar{t}/\bar{r})}{P(\bar{t}/r)P(t/\bar{r})} \tag{2.11}$$

where

1. t and $\bar{t}$ represent respectively the occurrence and non-occurrence of term t.

2. $P(t/r)$ represents the probability of the event t in the relevant class of document r (where $\bar{t}$ and $\bar{r}$ denote the event 'not t' and non-relevant class of document r respectively).

Various distributional assumptions have been made for these distributions by previous researchers (Robertson and Jones 1976, Robertson, Walker and Hancock-Beaulieu 1995). Thus, the evolution through to the current BM25 model (Robertson, Walker and Hancock-Beaulieu 1995) resulted in the general form (equation 2.12) of the actual BM25 ranking function.

$$S(Q,D) = \sum_{t \in Q \cap D} \left(\frac{tf_t^D}{tf_t^D + k_1.((1-b) + b.\frac{dl}{dl_{avg}})} .w_2.tf_t^Q \right) \tag{2.12}$$

where

1. k_1 and b are tuning parameters that control the saturation of term frequency and document length normalisation; and
2. w_2 is the term discrimination element of the function found using equation 2.13.

This is the classic form scheme that represents the *idf* function (Walker, et al. 1997) and the basis for both feature extraction techniques and many term weightings. It is important to note that k_1 and b above (parameters controlling the saturation of term frequency) affect the retrieval performance. Their values are often set between 1.0-2.0 and 0.0-1.0 respectively. When $b = 0$, the document length is ignored in weighting while higher values of b allow longer documents to be more heavily penalised.

$$w_2 = \log\left(\frac{N - df_t + 0.5}{df_t + 0.5}\right) \tag{2.13}$$

2.2.5 Standard Bag-of-Words Scheme

Bag-of-words approaches started being extensively used in many studies as an effort to improve the performance of IR systems. As the significance of term weighting in IR systems was identified, these approaches became the basis of term weighting frameworks. In IR frameworks, it is common to represent any set of texts (i.e. documents or queries) as a set of weighted terms called bag-of-words (BOW) to best describe them. This can allow the computation of similarity between the set of texts by using only their BOW representations. However, specific models or assumptions based on the theories of IR remained stagnant and many attempts (i.e. in the field of machine learning, data mining etc.) rather solely adopted the development of effective term weighting schemes which were based on simple primary assumptions.

A standard framework for term weighting schemes to calculate the score ($S(.)$) of a document D with respect to a query Q) to rank documents against a query can be defined by two triples (Zobel and Moffat 1998). One triple determines the term weight (see Section 2.2.1) in the query, while another triple determines the term weight in the document. The triple defining the weight of terms in the query may be linearly condensed into a simple term frequency within-query (Sparck Jones, Walker and Robertson 2000, Singhal 2001) as shown in equation 2.14.

$$S(Q,D) = \sum_{t \in Q \cap D} (ntf(D).gw_t(C).tf_t^Q) \tag{2.14}$$

where

1. tf_t^Q is the query weight (i.e. the weight assigned to the terms that appear in the actual query) ;

2. $gw_t(C)$ is the document weight (i.e. the term discrimination constituent which aims to determine the effectiveness of a term search in the collection C as a whole), and
3. $ntf(D)$ is the normalised term frequency which aims to promote documents with a higher occurrence of the terms in its corresponding query.

Most BOW approaches, if not all, use some technique to assign terms' weights to reflect the usefulness of those terms' importance in determining the relevance of the document. Conclusively, it is the term weights that are central to the performance of existing Information Retrieval systems (Salton and Buckley 1988). Most modern IR systems, if not all, automatically assign weights to the terms in a text. Term weighting schemes are therefore vital to most current search engines, if not all, and the assignment of suitable weights to terms in both documents and queries plays a major role in the reliability of the document ranking when a computation formula to determine simple inner product similarity is applied. Although IR models might be based on different retrieval hypotheses, statistically, the results of most of these models are often very much the same. While models have an imperative impact on the improvement of their underlying theories (i.e. probabilistic, vector based, etc.), it is their performance that generally indicates the effectiveness of a given model.

It is important to note that scoring and term weighting schemes are quite intuitive and their boundless variants are used by many IR systems. The key issue is that an effective IR system always seeks to achieve a good function estimator for term importance; therefore, it is crucial to have better estimators of term importance in any proposed new models.

2.3 Vector Space Model

This section will discuss how weights of terms can automatically be assigned by modern IR systems. It further implies how important these weights applied to the document terms are to the accuracy of the retrieval system in this model, before discussing how documents are ranked in the field of VSM.

2.3.1 Preliminaries

This model was introduced (Salton, Yang and Yu 1975) to address the drawback of binary weight assignments by using common calculations of degree of similarity to provide partial matching of words. During the retrieval process, any given query, and/or document is actually transformed into a vector in a high dimensional vector space whereby terms are the dimensions employed to form an index representing the documents (Salton and McGill 1983). The construction of an index thus consists of lexical scanning to find important terms to allow computation of the occurrence of terms reduced to its common stemmed form.

In VSM, the *closeness* of any two texts can be measured using the proximity between their corresponding vectors (Manning, Raghavan and Schütze 2008). In terms of IR, two documents (i.e. textual) are semantically related when their corresponding vectors are close to each other. Given a collection of documents, their closeness to the input query can be measured by using the closeness of their vectors to the query. A desired semantic relatedness can then be achieved by ranking the documents based on their closeness to the query (Salton and McGill 1983). The VSM can give a high ranking score to a document containing a few of the query terms if the terms occur frequently in the document, but do not occur frequently in the collection. Clearly, terms that appear in a document more frequently are more indicative of that document than terms that occur scarcely (Manning, Raghavan and Schütze 2008).

To present the model with an example, imagine a scenario in which an Web searcher inputs - globally - a query keyword with three terms: *icon people peace* to search for this information. By assigning an independent dimension to each term in the text (here, a total of three), a vector in three dimensional spaces can represent any keyword from the Web information base (Salton, Yang and Yu 1975). If the length of the sub-vector in the dimension is assumed to correspond to the number of occurrences of the term, the vector's representation of the keywords for five different texts (text 1: *icon peace*, text 2: *people icon icon*, text 3: *icon people people people*, text 4: *peace people people people* and text 5: *people peace peace peace* summarised in table 2-1) in the three keyword terms will be as shown in Figure 2-2.

Table 2-1 A representation of sub-vectors' lengths and their corresponding dimensions

	Keywords/Texts				
Terms	Text 1	Text 2	Text 3	Text 4	Text 5
icon	1	2	1	0	0
peace	1	0	0	1	3
people	0	1	3	3	1

Figure 2-2 shows that the keyword corresponding to text 1 '*icon* peace' has a zero component of *people* and unit component vectors of the *icon* and the *peace* dimensions. The next text '*people icon icon*' is a vector with a zero *peace* component, but a component of *length* two in the *icon* dimension; and similarly for all texts. Considering texts 1 and 2 for example, it is clear that the vector space will have in the real world a very high dimensionality (i.e. for every query, the size of the words composing the collection of documents). However, many zero components might yield to the sparseness of keyword vectors since keywords have zero length sub-vectors for their absent corresponding terms.

For IR systems based on the VSM, the length of a sub-vector in dimension i for example, is utilised to represent the weight, or the importance, of word i in a text. Absent words in a text obtain a zero weight. By considering that two texts (i.e. documents) sharing some vocabulary in common is the key concept in measuring semantic relatedness, it is clear that, the more they use the same vocabulary, the stronger this relationship becomes. This indicates that the measure of closeness correspondingly increases with the number of keyword matches existing between two texts.

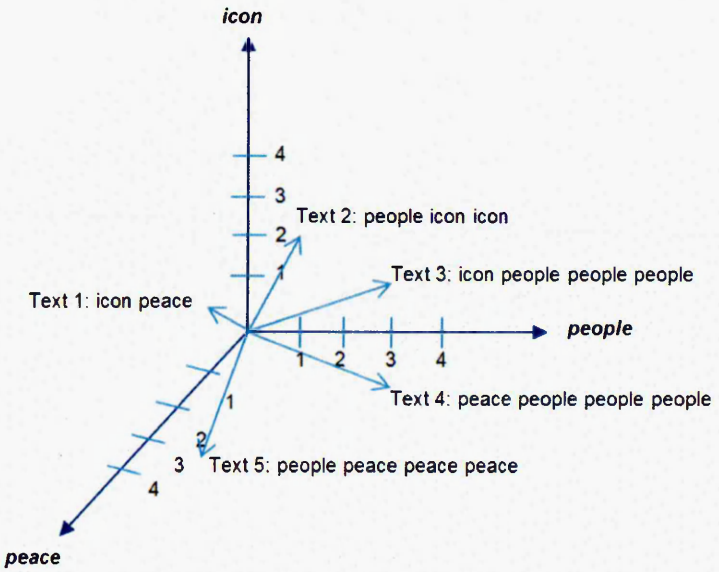

Figure 2-2 An Example of three dimensional vector space

Definition 2.6 *The number of matches: a term scores a non-zero weight in two texts, when it occurs in both texts. The more common vocabulary the texts have, the higher the number of non-zero products they share, and the greater the sum of similarity is between the two texts.*

Furthermore, if the importance of the matching terms (e.g. *peace, people*) also increases in either text, their corresponding vectors should also be regarded as closer than if the importance of the matching terms (e.g. *stop-words*) does not increase in those texts. These are the main effects for measuring closeness between texts.

Several measurements to gauge the closeness of two vectors (texts) have been proposed (Salton, Yang and Yu 1975), but the vector's inner product (Salton and McGill 1983) proved to be a measure which achieves these two effects. This means that it increases with both the importance of the matching terms and also with the number of keyword matches between texts. Given the query text vector $\vec{Q} = (q_1, q_2, ..., q_T)$ and the document text vector $\vec{D} = (d_1, d_2, ..., d_T)$ - where x_i is the sub-vector in dimension i for the vector $\vec{X}$ - in a dimensional vector space T, then the function of the vector's inner product (Axler 1997)) between $\vec{Q}$ and $\vec{D}$ can be given using equation 2.15 (Axler 1997) and further simplified into equation 2.16 (Axler 1997).

$$\vec{Q}.\vec{D} = \sum_{i=1}^{T}\sum_{j=1}^{T} q_i \times \vec{v}_i.d_j \times \vec{v}_j \tag{2.15}$$

$$\vec{Q}.\vec{D} = \sum_{i=1}^{T}\sum_{j=1}^{T} q_i \times d_j \times (\vec{v}_i.\vec{v}_j) \tag{2.16}$$

where $\vec{v}_i$ and $\vec{v}_j$ represent the unit vector in the dimensions i and j.

Since in VSM dimensions are assumed to be orthogonal (i.e. binary term independence), whenever $i \neq j$, the product is $\vec{v}_i.\vec{v}_j = 0$. Measuring closeness by using the vector's inner product function, and calling this closeness between two vectors 'Sim' (i.e. for simplicity - to represent vector similarity), the above function can be refined into equation 2.17 which satisfies both the appropriate effects of the measure related to a closeness for texts. A summary of these effects can be defined as follows (Salton, Yang and Yu 1975):

$$Sim(\vec{Q}.\vec{D}) = \sum_{i=1}^{T} q_i \times d_i \qquad (2.17)$$

Definition 2.7 *The importance of matching terms: when the matching terms are important, (i.e. they have a high q_i or high d_i or both), then the corresponding single match contributes more towards the total similarity, thus increasing the similarity between the texts.*

2.3.2 Ranking

The key challenge of document ranking is to determine those documents which are most relevant to meeting the individual user's needs in relation to the keyword entered. This problem is addressed by using similarity measures to calculate the weight (i.e. score) of the relevance of a document. In VSM, each term is given a weight score to reflect its importance in the document/document collection using the statistical distributions of the terms in the documents and in the collection (Salton and McGill 1983).

When a user enters a query in a natural language, it is converted into a weighted term vector so that a numeric similarity between the vector for every document in the collection and the query vector can be computed (Salton 1989) using the inner product function. Assuming D_i is the vector representation of document i and Q is the query vector; based on the above, equation 2.17 can be employed and refined as in equation 2.18 to compute the numerical similarity between them.

$$Sim(Q, D_i) = \sum_{common\ terms\ t_j} q_j \times d_{ij} \qquad (2.18)$$

where t_j is an existing term in both the document and the query, and d_{ij} is the weight of term t_j in document i, and q_j is its weight in the query.

The similarity can be built from the sum over all such terms t_j existing in both the document and the query to measure a document's relevancy to the query. The ranking of all the documents in the results list is then based on the order of their decreasing similarity to the query. The more similar a document vector is to a query vector, the more likely the document is relevant to that query. Thus the documents at the top of the results list have a higher degree of term weight, and are more interesting (i.e. relevant) to the users' information needs.

It is important to note that most IR systems are simply based on the typical VSM form of retrieval (Salton, Yang and Yu 1975). In this model, the query terms which are referred to as keywords in this thesis, are weighted according to the likely importance they might have in measuring the document content by matching them to the document terms. Those documents which have higher occurrence of the so-called weighted terms are then scored more highly. These term scores in each document are then aggregated to give a final score for the document.

Before a text is converted into a weighted term vector, text pre-processing is performed. This is the removal of the commonly utilised words (i.e. *stop-words*) and the conversion of morphological variants of words having similar semantic interpretations - known as stemming - to their base form. The stemming for the words "*fishing*", "*fished*", and "*fisher*" for example should convert the words to the root word, "*fish*".

In brief, the VSM technique can be grouped into three main modules: document indexing (to extract the content bearing query-terms from the document text), weighting of those indexed terms (to identify the relevant documents) and document ranking (according to a similarity measure with respect to the input query). Assuming binary term independence and considering these modules summarisation, the VSM is often seen as a technique that can determine the following:

1. the weight of each indexed term across the entire document to determine the importance of the term in both the document and in the document collection;
2. the weight of each index keyword within a given document - in the context of that document only - to find out how important the keyword is within a single document; and
3. the ranking of each document based on the extent to which the document is related to the input query.

2.4 Evaluation

In order to compare the effectiveness of two different IR models, it is necessary to use some empirical measures. In the above sections, the approach used for the empirical measure of the provided different models was not given. Most IR experimentation uses the Cranfield method which was popularised by the TREC[4] conference. This method uses a set of test queries along with a collection of documents known as a *test collection* (i.e. each query has a predefined set of its relevant documents) to determine the effectiveness of a ranking technique. To test the effectiveness of a technique, documents for the test queries are retrieved and their degree of recall and precision is measured (see next section) using that technique against the predefined set of relevant documents. The values of recall and precision are often averaged across all the queries to determine the overall average values. Judgement of the technique (i.e. good or not) is therefore based on the number of queries (i.e. most queries vs. fewest queries) achieving the highest recall and precision.

2.4.1 Objective evaluations

To determine the performance of a search engine in providing personalised search results, it is important to evaluate the effectiveness of its ranking techniques. To facilitate the study of a general applicability of any new technique, objective evaluations (stated below) across different document collections and across a set of queries proved to be one of the biggest strengths (Salton 1992, Cleverdon, Mills and Keen 1966) in the field of IR. These evaluations are based on both system coverage and accuracy.

> ***Definition 2.8*** *The main objective of any new IR model is system coverage, that is achieving 100% recall - the measure often used to determine the effectiveness with which any new IR model achieves this objective. Given a query, the target of any IR system is to find all relevant documents and display a list of results of the most useful documents.*

[4] http://trec.nist.gov/

Definition 2.9 *The main objective of any new IR model is system accuracy, that is achieving 100% precision - the measure often used to determine the effectiveness with which any new IR model achieves this objective. Given a query, the target of any IR system is to filter all non-relevant documents and display a list of results that does not contain extraneous documents.*

The definitions of precision and recall are always hard to be perceived especially with the first read through as the two appear to be the same. To introduce these measures with an example, imagine a query for which an IR system returns a list of only ten documents, of which three are relevant and the others are non-relevant. If there is a total of fifty documents in the information base, the recall is 20%(10/50) and the precision is 30%(3/10). These recall and precision calculations are often computed using equation 2.19 and 2.20 respectively (Manning, Raghavan and Schütze 2008).

$$recall = \frac{returned \cap relevant}{relevant} \tag{2.19}$$

$$precision = \frac{returned \cap relevant}{returned} \tag{2.20}$$

where *returned* and *relevant* are the sets of documents returned and relevant respectively for that particular query.

2.4.2 Mean Average Precision (MAP)

Recall and precision are determined based on the fact that there is in the whole collection, a set of retrieved documents and non-retrieved documents. All documents are presented according to how well they achieve rank prediction, but the retrieved documents and non-retrieved documents cannot be distinguished in practice, therefore, the quality of the ranking has to be evaluated in the entire collection. Thus, the most common technique used is average precision, which is calculated by using equation 2.21 (Salton and McGill 1983, Manning, Raghavan and Schütze 2008).

$$MAP = \frac{1}{R}\sum_{r=1}^{N} P(d_r) \times \mathrm{Re}l(d_r) \tag{2.21}$$

where

1. N is the number of documents in the collection,
2. $P(d_r)$ is the precision at rank r,
3. $\mathrm{Re}l(d_r)$ is the binary relevance judgment of the document at rank r and,
4. R is the number of relevant documents ($R = \sum_{r=1}^{N} \mathrm{Re}l(d_r)$) in the collection

2.4.3 F-Measure

In some IR systems, recall may be more important than precision or vice versa depending on the users' interests. For example, in a legal or medical setting, practitioners are interested in obtaining all relevant documents based on a specific information need, therefore recall is more important than precision; while a traveler may be interested in one relevant document for a particular topic, thus for this user's particular information needs, precision is more important than recall. The standard F-Measure can be calculated using equation 2.22 (Rijsbergen 1979) to combine recall and precision so that the user's interest can be altered.

$$F_\alpha = (1+\alpha) \times \frac{precision \times recall}{\alpha \times precision + recall} \tag{2.22}$$

where α is a measure of the interest for the recall. The higher the value of α, the higher the interest for recall is. The F-measure might be recovered by setting α to 1.

2.4.4 Normalised discounted cumulative gain

Discounted Cumulative Gain (DCG) is a measure of ranking quality (Järvelin and Kekäläinen 2000) which is derived from an original measure known as Cumulative Gain (CG) to measure the effectiveness of the IR applications (i.e. algorithms). In CG, the position of a result is not included in the consideration of the usefulness of a set of results. It is thus the sum of the graded relevance values - in a search results list - of all results. The CG at a particular rank position P is often calculated using equation 2.23 (Manning, Raghavan and Schütze 2008).

$$CG_p = \sum_{i=1}^{P} rel_i \tag{2.23}$$

where rel_i is the graded relevance of the result at position i.

DCG can be employed to measure the *gain* (i.e. usefulness) of a document by using equation 2.24 (Manning, Raghavan and Schütze 2008), and it includes, as opposed to CG, the document's position in the results list. The gain from the results list (i.e. top to bottom) is added with the gain of each result discounted at lower ranks, based on the assumption that, the lower the ranks, the more relevant the documents are (i.e. more useful).

$$DCG_p = rel_1 + \sum_{i=2}^{P} \frac{rel_i}{\log_2(i)} \tag{2.24}$$

However, DCG alone cannot be used to compare the performance of a search using one query to the next from the results lists since queries vary in length. So for a chosen value P a normalisation across queries of CG at each position for a chosen value is applied. This process is commonly known as Normalised Discounted Cumulative Gain (NDCG) and its function is shown in equation 2.25 (Manning, Raghavan and Schütze 2008). All the documents of a result's list are sorted by relevance, providing the Ideal DCG (IDCG) or the maximum possible DCG till the position p.

$$nDCG_p = \frac{DCG_p}{IDCG_p} \tag{2.25}$$

All the $nDCG$ values for all the queries are then averaged to give the measure of the overall average performance for the ranking algorithm; when DCG and $IDCG_p$ are equal (i.e. in which case the $nDCG_p$ gives a value of 1), a perfect ranking algorithm is achieved.

2.5 Summary

This chapter has provided the key features of the Vector Space Model of IR and shown how documents are ranked in the field of VSM. It has also shown that the VSM can be applied to determine query-document similarity to provide documents with high similarity in order to separate them from those that are distinct in content. The chapter has introduced how modern IR systems automatically assign weights to document terms (i.e. index term) to describe their semantics while assuming the terms to be independent of each other.

Furthermore, this chapter has outlined the term feature extraction techniques and many term weightings along with the three main factors affecting them. The chapter has shown that term weighting is a central facet of most retrieval models. The main emphasis in this chapter is on the importance of term weighting and its relation to proper matches while modelling the relevance of digital documents and in turn, their ranking. The explicit theoretical models of retrieval can sometimes even be skipped (e.g. BOW) in order to rely only on term weighting and overall performance.

This chapter has also introduced how the ranking effectiveness of an IR can be measured (i.e. to provide a reliable PSE - a rank list of search results by their order of relevancy based on the user's individual information needs). With this foundation, subsequent chapters will explore how the similarity between terms specified by the user in a keyword form with semantically related terms can be improved through semantic-based search (i.e. employing profile ontology approaches) to address the drawback of term independence. Table 2-2 summaries the defining characteristics of the above mentioned IR models and lists its key advantages and disadvantages.

Table 2-2 Defining Characteristics, Representation, Advantages and Disadvantages of the Stated IR Models

Method	**Defining Characteristics**	**Formal Representation of Information Needs**	**Advantages/Disadvantages**
Boolean Model	Documents Representation • Set of keywords • No Rank	Query	Limitations • No partial matching between documents and query (no rank) • Pressure on users to formulate proper queries
Probabilistic Model	Documents Representation • Binary term vectors • Rank is based on PRP (the probability of the document being relevant to the query), thus • Ensures an optimal Ranking • Example of models built on probabilistic include the LM and standard BOW	Query	Advantages • Best-match retrieval • Optimal Ranking Limitations • Term Independence • use of tuning parameters
VSM	Documents Representation • Term vectors • Rank is based on the similarity metric (i.e. cosine similarity) between documents and query • No optimal Ranking • Example of models built on VSM includes most of IR models as it is the base model	Query	Advantages • Best-match retrieval • No tuning parameters Limitations • Term Independence • No optimal Ranking

Chapter 3 Personalised Information

After introducing approaches to measuring documents' relevancy and similarity employed in keyword-based mechanisms of IR, this chapter investigates further similarity measures to estimate the relatedness between concepts associated with query keywords before describing personalisation approaches. The chapter starts with an introduction to the basic concepts required to understand the search technologies in Section 3.1. The traditional retrieval process is presented in Section 3.2. The concepts of the search solution are detailed in Section 3.3 before introducing the personalisation of the search results in Section 3.4. A survey of state-of-the-art in personalisation approaches is provided in Section 3.5 while Section 3.6 summarises the chapter.

3.1 Introduction

The basic concepts of document retrieval were surveyed in the previous section. It was argued that users employ search queries to express their information need while interacting with a Web document retrieval system. The underlying retrieval engines then retrieve results relevant to the search queries. The typical conversation between a SE and a user is the actual search. The search uses two processes (i.e. indexing and querying) to interact with the SE data structures. All the classic models, Boolean, vector space, probabilistic etc. seen in the previous chapter, start by creating a document collection by capturing on the Web as many documents as possible. The search engines achieve this by using a Web crawler and a document indexer.

The crawler - a software agent which is designed to traverse the Web - is given a starting set of URLs to retrieve their corresponding pages. It uses any of the search techniques (e.g. breadth-first or depth-first) to extract their out-going

links and terms. These links are fed back and the process is repeated whereas the terms obtained are mapped to the normalisation features (i.e. stemming).

The indexer uses a set of typical keywords called index terms to identify each document. These index terms are simply words whose semantic references serve as a mnemonic device to recall the document's main themes. Thus an index term - explained below - can be used to index and summarise document content. An inverted index term is then created with a list of Web pages in which it occurs. IR models differ in their performance depending on how queries and documents are represented as well as how similarity and relevance are defined, both of which play a major role in retrieving relevant documents.

It is important to note that the crawler-indexer architecture used by most SE is typically centralised. Although crawlers run on a local system, they traverse the Web and send updated requests or new pages to a remote Web server, where they are indexed. The index is thus used in a centralised manner to respond to queries from different places on the Web. The ranking is then based on some variations of a Boolean or vector model (Yuwono and Lee 1996).

For most SE, the ranking is, as with the searching, done on the basis of the index only without accessing the text. For most indices, a variant of data structure known as inverted file is used (Baeza-Yates and Ribeiro-Neto 1999). An inverted file is a list of sorted vocabularies (i.e. words), with each one having a set of pointers to refer to the pages where its elements occur. The compression process can be adopted to reduce the size of the index. Thus, a binary search on the sorted list of words is performed on the inverted file. When multiple words are involved in the search, the results are combined using aggregation to generate the final result.

3.1.1 Data Structure

All search applications (i.e. Web searches like Google and Bing, E-commerce companies like Amazon and Best Buy or Expert searches like LexisNexis and card catalog) possess at the core of their SE, some type of highly optimised data structures that allow the documents to be scored and retrieved. These structures mechanics - how does the querying and indexing interact with the SE data structures - play an important role in leveraging the functioning of a SE so that a smart relevant search experience can seemingly be provided.

3.1.1.1 The Inverted Index

The inverted file data structure (Zobel and Moffat 2006) is one of the most common methods employed by the search application in indexing textual documents and it is used both to retrieve them and to provide textual document evaluation in the search process. For any given term, the inverted file is used to provide access to the list of documents that contain the term. The inverted file, also known as an inverted index, is a type of index term which is similar to the physical index of a book. It simply obeys byte-by-byte term matches during a search. It is composed of two main components (non-positional indexes): (1) the term dictionary, which is a sorted listing of all terms occurring in a given field across a set of documents; and (2) the postings list, which is a corresponding list of documents containing a particular term in the term dictionary. These components can be illustrated through an example. Taking the case of a set of documents shown in Figure 3-1 - term 0 (leadership) occurs in $\{d_3\}$, term 1 (Nobel) occurs in $\{d_1\}$, term 2 (Oman) occurs in $\{d_2, d_3\}$, term 3 (peace) occurs in $\{d_0\}$, term 4 (Qaboos) occurs in $\{d_0,d_1,d_2\}$ and finally term 5 (Sultan) occurs in $\{d_1,d_3\}$ - its term dictionary and postings list can be shown as in Figure 3-2.

0. *Qaboos, Icon of peace*
1. *Sultan Qaboos prix Nobel*
2. *His Majesty Qaboos and the people of Qaboos of Oman*
3. *The Sultan of Oman and his parameters of leadership*

Figure 3-1 Example Documents

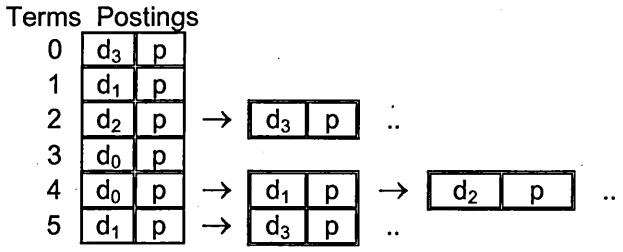

Figure 3-2 Term Dictionary and Postings List for the Stated Example

It can be noted that both components (i.e. the term dictionary and its corresponding postings list) are just mappings. The first one - term dictionary - when sorted lexicographically - is simply a map from the term to its corresponding original number. While the second one - posting list - is another map from the term numbers to a list of numbers which corresponds to the original documents. With these components in place, it is clear that any

documents matching the query terms can be quickly retrieved. Taking the above example with a searcher who is looking for a set of documents containing the term *Sultan*, which is given a term identifier of 5, the postings list associated with this term identifier indicates that it occurs in the list that refers to documents 1 and 3, while documents 0 and 1 do not contain that term.

To tune the search relevance other components which might be used and are normally associated with the inverted index include among others, document frequency, term frequency, term positions, and stored fields. It is important to note that the process of moving the data into a data store includes, according to Jones and Willet (1997), **e**xtraction (capturing data from its warehouse), **t**ransformation (i.e. converting data into token - a format amenable to the data store destination) and **l**oading (indexing or placing data into those data structures) the information (ETL). The enrichment process is optional in order to add to the documents, any additional information useful for relevance.

Although resource management and computation performance are focused upon during the indexing rather than the relevance process, some indexing decisions can influence relevance, specifically the pieces of data to be indexed and the data structure types that can be used (Zobel and Moffat 2006). However, while placing field data into core data structures, indexing is, in general, the process of storing data (i.e. saving the data) in the SE, though it has a different meaning from storing. For instance, while indexing can be considered in the inverted index as the update process for the extracted tokens (see Section 5.1.2) to support that field to be searched on so that the indexed field is searchable; storing is the process of preserving the un-tokenised document (i.e. original data) in its stored fields data structure.

3.1.2 Top-k Retrieval

Top-k retrieval is the retrieval process of the k documents which mostly match a given query. This technique typically relies on the inverted indices used. A list of all documents containing the term related to each query term is present. The documents in the list follow a descending order (see Table 3-1) of the weight of the term - determined by using any weighting scheme (i.e. $tf \bullet idf$ or *BM25)* - in each document. In VSM, the top-k documents represent the textual similarity between the query-document vectors, ordered according to the similarity score

and derived from this ordered list; a process that is clearly unfeasible considering the huge size of the information base available on the Web. It can thus be addressed, by identifying the user's information needs.

Table 3-1 Inverted Indices for keywords People, Peace and Icon

People		Peace		Icon	
d_1	0.42	d_3	0.49	d_2	0.52
d_4	0.38	d_1	0.47	d_4	0.48
d_2	0.36	d_5	0.44	d_1	0.47
d_5	0.33	d_2	0.36	d_5	0.46
d_3	0.28	d_4	0.26	d_3	0.39

3.1.3 Gathering and Representing Interest

In most SE, interactive IR models are based on relevance feedback approaches (Salton and Buckley 1997). The feedback mechanism is one of the most essential components in the IR systems. The main aim of relevance feedback technique is to identify the user's information need so that this knowledge can be exploited to adapt the search results. The main advantage of this approach is to simplify the information seeking process. Three main types of relevance feedback (Rocchio 1971) are described in the following sub-sections. They include explicit feedback, implicit feedback and pseudo (blind) feedback.

> ***Definition 3.1*** *A feature used to determine if the results returned in response to the users' queries meet their information needs is referred to as relevance feedback. Thus, relevance feedback allows determination of whether or not the search results are relevant to the user's input keyword.*

Relevance feedback has been shown (Manning, Raghavan and Schütze 2008) to be very effective at improving the relevance of results. A document containing all the words contained in the query (Manning, Raghavan and Schütze 2008) may not necessarily be relevant, it is rather said to be relevant if it addresses the user's stated information need. Relevance is specifically measured relative to the information needs rather than the query. Users should thus be involved in the retrieval process so that the final result set (Manning, Raghavan and Schütze 2008) can be improved. The idea behind relevance feedback is that both relevant and non-relevant search results hold strong evidence related to the intentions of the users - who can make judgements, which are in turn, used by the retrieval system to learn how to improve the

search accuracy. The obtained relevance judgment is then incorporated into the retrieval process with the aim of improving the search accuracy.

IR systems adopt different approaches to incorporate the relevance feedback information obtained into their retrieval functions (see Section 3.1.4). Some IR systems do it directly, while others expand the user's query (see Section 3.1.5) based on the relevant and non-relevant documents to better represent the user's intentions. Here, the discussion of relevance feedback and query expansion is based on the ranking model and the reader is referred to (Radecki 1988) for the relevance feedback and query modification found in Boolean models.

3.1.3.1 Explicit Feedback

In this feedback approach, users are involved to explicitly mark the degree of relevancy of the search results (some), thus creating their profiles. Considering systems which try to minimise the users' involvement, and its usage as a source for personalisation techniques, this approach is not viable. The practical value of relevance feedback has also raised concerns as it was shown that users are often unwilling to be involved (White, Ruthven and Jose 2002). However, some studies (Salton and Buckley 1997), attempted to use explicit feedback techniques in order to observe users' performed actions, and in turn, to investigate which actions could most likely be representative of their interests.

3.1.3.2 Implicit Feedback

Unlike explicit feedback which explicitly involves the user in rating the results, in implicit feedback approaches (Rocchio 1971), the system unobtrusively employs the user's interactions vis-à-vis the search results. Using the feedback obtained, the system then tries to automatically infer those documents that are relevant and those documents which are not relevant. This mainly captures the user's behaviour. It heuristically assumes that the documents clicked from the search results are relevant documents (Dou, Song and Wen 2007, Teevan, Dumais and Horvitz 2007); documents which are ranked higher (i.e. lower rank position) and are not clicked, should be considered as non-relevant.

By mining the users' interaction data implicitly, users' intentions can be inferred to allow more relevant information to be retrieved (White 2004). However, although implicit feedback might be more convenient for the user, it is difficult to

implement and less accurate (Salton and Buckley 1997, Joachims, et al. 2005). The advantage of implicit feedback is that it might be collected in much larger quantities, at much lower cost, and without burden on the user. Implicit feedback has further shown to improve the retrieval performance by 31% (Agichtein, Brill and Dumais 2006) compared to models that do not incorporate any feedback into their systems.

3.1.3.3 Pseudo Feedback

This feedback approach (Croft and Harper 1979) differs from the above two approaches as it simply assumes that some top-ranked documents are relevant and can thus be utilised to adapt the initial search query. Therefore, the user does not need to explicitly mark the search results to provide relevance assessments. This approach is also known as blind or ad-hoc relevance feedback. Considering the description of relevance feedback above, its usage as a source for personalisation approaches is questionable.

3.1.4 Adjusting the Retrieval Function

To incorporate relevance feedback directly into the retrieval function, a weighting function scheme (i.e. BM25) might be modified to include this relevance feedback (Teevan, Dumais and Horvitz 2005b); thus, the weight w_i of a term t_i in document d might be given by equation 3.1 shown below:

$$w_i = \log \frac{(r_i + 0.5)(N - n_i - R + r_i + 0.5)}{(n_i - r_i + 0.5)(R - r_i + 0.5)} \tag{3.1}$$

where

1. N and R are the total number of documents in a collection and the total number of relevant documents respectively,
2. n_i and r_i are the number of documents containing the term t_i and the number of relevant documents containing the term t_i respectively.

3.1.5 Query Expansion

A common technique to automatically refine the search queries is based on relevance feedback. In order to expand the query so that relevance feedback can be incorporated into the retrieval function, the system tries to acquire a presumably improved version of the original query by modifying its original representation. Additional search terms are then added to the original query - in a query expansion - based on the statistical co-occurrence of these terms

(Crouch and Yang 1992). As defined below, the process might involve adding terms or even related term concepts:

> ***Definition 3.2*** *Query expansion is the process of adding terms and/or related concepts to an original query in order to change the query so that result relevance can be improved.*

In the VSM, the feedback is often incorporated by using the Rocchio framework (Rocchio 1971). In this framework, the original query vector is moved closer to the centroid vector of the relevant document vectors (positive centroid) to construct a new query vector using equation 3.2. Similarly, it is moved farther away from the non-relevant document vector's centroid (negative centroid). The query is thus augmented with terms which best differentiate relevant documents from the non-relevant ones. In a retrieval experiment, term vectors for all relevant documents retrieved were added and term vectors for all irrelevant documents were subtracted to refine search queries. Thus, terms are aligned with an increasing and decreasing weighting during the process.

$$\vec{q}_E = \alpha\vec{q}_O + \frac{\beta}{n_1}\sum_{1=1}^{n_1}\vec{R}_i - \frac{\lambda}{n_2}\sum_{i=1}^{n_2}\vec{S}_i \qquad (3.2)$$

where

1. $\vec{q}_E$ and $\vec{q}_O$ are the refined query vector and the original query respectively,
2. $\vec{R}_i$ and $\vec{S}_i$ are the i^{th} relevant document vector and the i^{th} non-relevant document vector,
3. n_1 and n_2 are respectively the number of relevant and non-relevant documents in the collection,
4. α, β and λ are the controlling parameters for the influence of relevant and non-relevant documents on the refined query vector.

It is also possible in Rocchio to simply move the query vector closer to only the positive centroid in some versions. The new query vector might often be truncated to encompass only k terms having the highest weights for the sake of efficiency. To avoid over-fitting especially on a small sample, a relatively high weight is normally put on the original query. The feedback performance might often be significantly affected by the relative weight of the original query vis-à-vis information extracted from feedback relevance. Setting an optimal weight in pseudo feedback is even harder since there are no training data sets for tuning the weights.

3.2 Fundamentals of the Traditional Retrieval Process

Traditional IR techniques treat entity descriptions as self-contained units (Thiagarajan, Manjunath and Stumptner 2008) due to their reliance on the keyword-based search mechanism and term independence assumption (see Section 1.1.1). The explicit relationships of these entity descriptions are often ignored (Thiagarajan, Manjunath and Stumptner 2008) in these approaches. They mostly employ a search technique - which is implemented by a Boolean or the VSM (or any other BOW) to select information of a specific nature for the users' queries - which retrieves solely those documents matching the keywords specified by the user. The following are the major problems which are evident in these techniques:

Some other documents might contain the semantic information desired by the searcher (i.e. related to the query concept), without containing the searchers' specified keywords (Vallet, Fernández and Castells 2005, da Costa Pereira and Tettamanzi 2006, Castells, Fernandez and Vallet 2007, Knappe 2006). The word synonymy (i.e. dictionary words having the same meaning) and word polysemy (i.e. individual words in the dictionary having more than one meaning) are such possibilities. It is thus clear that a simple match technique (i.e. the keyword-based search) will result in either returning many non-relevant items or missing out some relevant items (Blair and Maron 1985).

3.3 The Concept Search Solution

Word Sense Disambiguation (WSD) can be adopted (Navigli 2009) to address the above mentioned limitations. WSD is a process of moving from words - which are associated with this natural language ambiguity, to concepts (word senses) which are expressed in unambiguous formal language. The reader is referred to Sanderson (2000) for an overview of the existing approaches of concept-based (i.e. a sense) IR. The aim of WSD techniques is to associate in documents corpus words with their corresponding atomic lexical concepts existing in a linguistic database so that these documents can be indexed with the associated concepts, referred to as concept search. This allows determining of documents based on query concepts rather than documents sharing the same vocabulary items with the input query (Giunchiglia, Kharkevich and Zaihrayeu 2009).

There are many linguistic databases which have been used in concept-based IR and WordNet (Miller 1995) is such an example of the concept-based IR approach (Stokoe, Oakes and Tait 2003). It might particularly be used as a WSD for information related to words co-occurring with a word in a document and frequency of senses containing the given word (Stokoe, Oakes and Tait 2003). The idea behind a concept search is to use in the match process the concept (i.e. concept search query) instead of the term itself so that the information retrieved in response is relevant to the concept contained in the query text (Giunchiglia, Kharkevich and Zaihrayeu 2009). The degree of similarity between the query concepts and the concepts of the results returned for that query is often referred to as concept search relevance. In general, the more similar the concepts of the query and the concepts of the results are, the more relevant the search results are considered to be.

However, it is sometimes possible to find a given word which does not have its corresponding concept in the lexical database due to lack of background knowledge (Giunchiglia, Shvaiko and Yatskevich 2006). Furthermore, if the direct overlap of the exact concepts representing the semantics is not taken into account (Knappe 2006, Giunchiglia, Kharkevich and Zaihrayeu 2009, Kantardzic 2011), there is still no guarantee that the right matches will be determined by simply performing a concept-based rather than a keyword-based similarity comparison. Therefore, it is crucial to integrate the term dependency (i.e. term relations) to achieve consistent improvements. It is also intuitive that "the document where the matched query terms occur closely to each other is more likely to be relevant than the document where the matched query terms are isolated" (Song et al. 2011). Thus, the development of a search mechanism that can guarantee the integration of term relation is the central problem in order to extend the inherent relationships of semantic similarity (see definition 3.4) between two texts. The syntactic search (i.e. the keyword-based search) and the semantic search (i.e. concept search query) might be combined to obtain the final results.

3.3.1 Related Concepts

IR research and technology could actually be divided into two broad categories: statistical and semantic. Those IR systems falling into the statistical category often find results based on the statistical measures of the close matching

between the query and the document. These were discussed in detail in the previous chapter. Those IR systems falling into the semantic category often implement some degree of semantic analysis based on the syntactic of natural language text (i.e. computational linguistics); however, the retrieval process still relies on statistical methods (Greengrass 2000) in these systems. The concept of the Semantic Web (Berners-Lee, Fischetti and Foreword By-Dertouzos 2000) was conceived with the aim of solving the problems related to searching, retrieving, representing and maintaining the abundance of data sourced by the World Wide Web (WWW). Using this concept, a hybridisation of intelligent agents and ontologies establishes the foundation of the Semantic Web. Ontology is made able to define metadata so that one complete glossary is built to clearly describe the data available in the large document repositories on the WWW. For a complete description of ontology structure, the reader is referred to Staab and Rudi (2013).

The Semantic Web is provided by various metadata being integrated together. These metadata further describe the contents of documents and their general concepts. The contexts related to these documents are thus perceived through the Semantic Web. Metadata are therefore processed descriptions of Web information resources which are related to the conceptualisations of ontologies or the domain of application. The use of ontology structure is thus a fundamental key of developing the conceptual relationship of these metadata.

An ontology is built up from hierarchical concepts which describe classes/sets of objects. Semantics are thus formed by the roles/relations between these objects and axioms which can further describe concepts or relations and set constraints (Baader, Horrocks and Sattler 2009). The term ontology is borrowed from philosophy (Brost 1997) to mean a systematic account of being or existing. In AI systems, what exists is that which might be represented. When domain knowledge is represented in a declarative formalism, then the represented set of objects is called the universe of discourse. A knowledge-based driver representing knowledge is reflected by the representational vocabulary of such sets of objects with the relationships describing them.

Ontology is often considered to be an important element of the semantic web (Khan 1999). Its main goal is to represent human-beings' knowledge. It is used

to provide the formalisation of variant information. The word ontology is perceived by different people in different contexts: glossaries and data dictionaries, schemas and data models, thesauri and taxonomies, formal ontologies and inference among others. The use of ontologies has been empirically proven to be one of the increasingly popular methods to effectively mediate information access as well as providing personalised search results (Haav and Lubi 2001). It is a basis for the construction of a user model (Middleton, De Roure and Shadbolt 2001) in several personalised systems (Dicheva and Aroyo 2000). Based on different contexts, ontology can be generally defined as follows:

> ***Definition 3.3*** *An ontology is a taxonomy of concepts needed to describe tasks in the topic addressed. Each concept and all its attributes are defined in natural language words. This ontology then defines the data structures that Natural Language Processing (NLP) can use in sentences.*

An ontology should be thus viewed as a tool that provides a shared and common understanding of a domain which needs to be communicated among people as well as heterogeneous application systems. Such a tool aims at capturing consensual knowledge to be used and reused among groups of people and across software applications. An ontology is a formalism for representing knowledge about a field of discourse.

In the AI community, the ontology of a program is thus defined by describing a set of representational terms. The entities' names in the universe of discourse (i.e. classes and relations) are associated with human readable texts through definitions to describe (1) what these names mean, (2) formal axioms constraining the interpretation and focusing the well-formed use of these terms. Ontologies are viewed as formal logical theories through which not only the terms and relationships between them are defined, but also the context which they are formally applied in and related facts and relationships are implied.

From the linguistic perspective (e.g., pre-defined thesauri like WordNet), ontologies express various relationships (e.g is_a, instance_of) between concepts rather than formally and explicitly describing a concept meaning. Ontologies are further viewed as schemas, taxonomies and object models in communities (i.e. databases) which do not explicitly define important constraints.

Therefore, ontologies express sets of representational terms simply known as concepts over which the interrelationships describe a target world.

In brief, an ontology is often defined as an explicit specification of concepts and relationships existing between terminology and context (Brost 1997); It might further be defined as a formal, explicit specification of a shared conceptualisation (Brost 1997, Studer, Benjamins and Fensel 1998). Conceptualisation here refers to an abstract model of a phenomenon in the world through identification of the relevant concepts of that phenomenon; while *explicit* implies that both the types of concepts used along with the constraints on their use are unequivocally defined. In this case, it refers to the machine-readable nature of the ontology; whereas *shared* indicates that the consensual knowledge is captured, that is, an ontology should be accepted by a group and not be the private possession of some individual.

Ontological techniques are being employed in retrieval systems to improve precision and recall (Guarino, Masolo and Vetere 1999). Analysis of a few examples of the uses of ontology in practical contexts includes Open Biomedical Ontologies (OBO)[5], Gene Ontology (GO)[6] and WordNet (Miller 1995) among others. Both OBO and GO use semantically related terms, and are among the surveys that proved to achieve better recall. WordNet on the other hand proved (Khan 2000) to achieve better recall by using query expansion mechanism with a generic ontology which permits a query to be matched to relevant documents without containing any of the original query terms (Thiagarajan, Manjunath and Stumptner 2008). However, Voorhees (1994) proved that this technique is promising for complete topic statements rather than queries with fewer contexts (i.e. short queries).

Ontology-based models can allow the extraction of relevant concepts which can both identify and describe documents and serve as the documents' metadata (Khan 1999). However, the key issues in ontology construction or the use of ontology-driven methods are both the indexing and extraction of semantic concepts from the keywords to pinpoint appropriate concepts that describe and identify those documents deemed relevant to users.

[5] http://obofoundry.org/
[6] http://geneontology.org/

3.3.2 Semantic Search

Semantic search was first coined by the IR community to extend the classical IR (Croft 1986) and expand the inherent relationships regarding semantic similarity between texts. Since then, many of the proposed approaches have used an informal knowledge representation structure (i.e. thesauri) with little or no formal reasoning support to codify explicit semantics. With the Semantic Web however, formal frameworks were proposed to represent and reason about knowledge.

In the semantic Web perspective, semantic search is seen as a data retrieval task. Semantic search is typically an information search which is based on both the searcher's intent and the contextual meaning of search terms, rather than searching the information based on the dictionary meaning of the individual query words/terms (i.e. keyword-based). Semantic search is a search based on the broad concept of the query to return relevant results (Giunchiglia, Kharkevich and Zaihrayeu 2009). It is based on concept matching rather than term matching. It is thus defined as a semantic search related to a sub-set of the context matching. Semantic relationships between concepts can actually be captured by ontologies such as WordNet, ODP, Wikipedia, etc. The latter can allow discovery of the inherent relationships between descriptions of entities (Thiagarajan, Manjunath and Stumptner 2008).

3.3.3 Semantic Similarity

A number of studies (Khan 1999, Vallet, Fernández and Castells 2005, Castells, Fernandez and Vallet 2007, Thiagarajan, Manjunath and Stumptner 2008) demonstrated that, although Cosine similarity is the most commonly used similarity measure for its simplicity and accuracy, it is crucial to extend it to address the challenges of keyword-based techniques. Based on the description of ontology provided that far, it is clear that ontologies must be adopted in order to include semantics and perform Cosine similarity on the ontology concepts in order to describe texts rather than describing words/terms. In this way, the bag-of-concepts representation can be used in computation rather than BOW representations (see Section 2.2.5). Every term can then be mapped to a semantic concept in the semantic mapping process. Compared to the term vector similarity measure, it was also proved that using the bag-of-concepts

format for pre-processing documents prior to computing the Cosine similarity positively affects the accuracy of results (Gabrilovich and Markovitch 2007).

> ***Definition 3.4*** *"Semantic similarity relates to computing the similarity between concepts which are not lexicographically similar" (Varelas, et al. 2005). Semantic similarity can thus be defined as a concept-based relatedness between two texts whereby the likeness of compared objects is determined based on a similarity score reflecting the semantic relation between the meanings of the content of both texts. The higher the similarity score, the stronger this relationship is.*

The learning process (see Section 4.1) relies on ontologies as a means to establish the related concepts of the query-term and to grasp users' interests to improve search quality (Cantador, et al. 2008). In order to adopt query expansion in a bag-of-words/bag-of-concepts representation so that those terms that are related to the original terms in the corresponding description of texts are included, a mechanism known as spreading might be utilised. It was proved that spreading can be adopted to overcome polysemy problems in WSD (Tsatsaronis, Vazirgiannis and Androutsopoulos 2007) as well as the inherent relationship challenges in ontology mapping (Mao 2007) and personalised multimedia access (Cantador, et al. 2008).

Spreading is the process of integrating to an entity description terms' concepts by referring to WordNet and Wikipedia ontologies. The spreading approach can thus be taken into account and be built on the notion of considering related concepts and the user's preference learning mechanism can be established to derive a personalised search. Compared to a normal search, a personalised search (see next section) is known to be of better quality (Cantador, et al. 2008).

3.4 Personalised Search Results

Personalising search results is based on the fact that knowing the individual user's interests and preferences is the best way to improve the ranking of the results to be returned by a search engine, since different users tend to have different interests and preferences. The purpose is to improve retrieval accuracy which is often achieved, as mentioned earlier in Section 1.1.1, by ranking the search results with relevant documents at the lowest ranks. Since users

express their information needs in search queries, interpreting users' information needs correctly - and consequently their intentions - is thus a requisite of any IR system during interaction with users.

Furthermore, personalising search results might help to fulfil the users' information seeking tasks while minimising their efforts. There have been in the literature many attempts (Mobasher 2007) - which often differ from what follows - to personalise search results: (1) the kind of information that can be used to derive the user's interests and/or preferences; (2) approaches employed to infer users' interests and/or preferences (i.e. by requesting the users to indicate information about themselves explicitly or by acquiring users' interests implicitly by using their interactions with the system); (3) where to store this information (i.e. the server side or the client side); and (4) how to use the acquired information to improve retrieval accuracy.

According to Jameson (2008), personalisation techniques include Recommender Systems (RS), learning personal assistant, adaptation to situational impairments, ability-based user interfaces (i.e. exploratory search) and algorithm-based approaches (i.e. personalised search). Once the users' interests have been gathered, these techniques might be employed to exploit this feedback. Due to its nature, this work focuses on personalised search (James et al. 2002) and neglects other paradigms. However, content-based filtering, although it is a part of RS, will be briefly included (see Section 3.4.1.4) as it is an essential component of the current approach. For more details related to recommender systems, the reader is referred to Pazzani and Billsus (2007).

The difference between those approaches focusing on interaction between user and system and personalised search can be made clear through search mechanisms. For instance, James et al. (2002) defined two personalised searching techniques: their first approach is based on query augmentation. Here, the system expands the users' queries by considering the context of their searches (see Section 3.1.5). Their second approach is based on result processing. Here, the search result is analysed and modified further to better reflect the users' contexts based upon the augmented queries.

All approaches to personalisation in general and content-based filtering personalisation systems in particular rely on data collection to reflect accurately the users' interests during their interactions with articles and applications (i.e. users' behaviours). Not only do personalised systems differ in the algorithms employed to rank documents in their order of relevancy, but also in the techniques employed to construct the users' profiles utilising this underlying data (Susan, Jason and Alexander 2004). The following sections will describe how to identify the users' information needs and consequently to acquire their interests.

3.4.1 Personalised Search

Teevan, Dumais and Horvitz empirically demonstrated (2010) that the difference between the performances of a personalised and non-personalised search engine is quite significant. For instance, they argue that there is great potential for personalisation. To provide a personalised search, most modern search engines rely on ranked lists which tend to rank the retrieved documents by their relevancy score. This score is derived using the term weighting functions (see Section 2.3.2) to reflect the importance of each term in the document/document collection using the statistical distributions of the terms in the documents and in the collection.

3.4.1.1 User Modelling for Web Information Retrieval

Considerable research has been done into building user models to increase the value of search results in IR (Croft, Cronen-Townsend and Lavrenko 2001). The users' models allow the result ranking to be emphasised for the same keyword queries issued by different users whose intentions are different. User modelling is the central key to any personalisation system (Croft, Cronen-Townsend and Lavrenko 2001). A user model allows the formation of a representation of individual users so that their interests and preferences can be featured. Users' models might be designed in a form of user profile (Mobasher 2007) or agent (e.g. WebMate (Chen and Sycara 1998)) to assist users while browsing, and they are generally described by keyword vectors, weighted or un-weighted to allow standard text processing procedures to be applied.

3.4.1.2 Topics of User Interest

The problem of contextual retrieval can be defined as "combining search technologies and knowledge about query and user context into a single framework in order to provide the most appropriate answer for a user's information needs" (Allen, et al. 2003). For queries which are broad or ambiguous in nature, inferring the users' interests might serve to learn their experiences. However, the benefit of such an approach might vary for recurring queries as compared to new queries; which then motivates a differentiated usage of long-term users' information (Tan, Shen and Zhai 2006).

"There are many ways of representing people's interests, including explicit user profiles, implicit profiles based on search logs (i.e. browsing histories), and richer implicit profiles based on the full content of documents" (Teevan, Dumais and Horvitz 2010). Profiles can be employed to both contextualise users' search queries within their interests and to re-rank the retrieval results (Cantador, et al. 2008, Pazzani and Billsus 2007). Documents might thus be ranked based on the correlation between the documents' content and the users' interests.

Moreover, it has been empirically proven (Morris, Ringel Morris and Venolia 2008) that users' implicit relevance feedback within one search session cannot be very representative of real life search situations. Their findings demonstrated that Web users often perform search tasks which span more than one session. In fact, 73% of their respondents in the related survey reported that they normally perform multi-session tasks which are distributed over several days. Therefore, to assist users with their interaction over multiple sessions, it might be compulsory to keep track of their long-term feedback (Mostafa 2005).

3.4.1.3 User Profile

Users' profiles contain the information collected about users during their interactions with the system (i.e. feedback). This information represents their interests and/or preferences and it can be used to model the users or to distinguish a particular user from other users so that personalisation services can be provided. For Web searching, this information (referred to as users' profiles) is often employed by the system as relevance feedback in order to decide relevant documents, and in turn which documents are ranked at the lowest ranks.

In fact most personalisation systems are based on some type of user profile (Susan, et al. 2007). User profiles might be built by employing the appropriate weights of keywords which can be determined through weighting functions (i.e. $tf \bullet idf$ or $BM25$), weighted concepts from existing ontologies, or even semantic networks and association rules. For the purpose of this work, the current project will focus on weight-based profiles with the goal of leveraging the data captured during users' interactions with the Web.

Keyword based profiles might be provided by users through explicit feedback (see Section 3.1.3.1), or they might be implicitly constructed by extracting relevant keywords from the documents of interest (see Section 3.1.3.2). Once the keywords have been extracted, they are weighted using a weighting scheme (i.e. $tf \bullet idf$) and the weight of each keyword can serve to represent the degree of interest of the document (Robertson and Jones 1976, Belkin and Croft 1992). Each profile is then represented as a feature vector using the keywords as features. The search results might be represented by a weighted vector and a similarity metric (i.e. Cosine similarity) might be applied to filter the results so that the closest vectors to the profile are passed to the user.

Concept based profiles on the other hand use concepts instead of keywords and the profile is then represented as a feature vector using the concepts instead of keywords. Here the features (i.e. concepts) are given weights to represent the user's interests in the corresponding concept. Thus, the richer the semantic in the ontology, the more accurate the user profiles would be. As with keywords, both explicit and implicit feedback might be used to assign weights to various concepts. The current approach towards the adoption of semantic-based modelling is described in the next chapter.

In order to build users' profiles, some sources of information related to the users need to be collected explicitly by asking users to provide their information (e.g. the commercial system MyYahoo[7]); or implicitly by observing users' actions. The latter approach involves using types of information available including how recently a page was visited, its frequency of visits, the dwell time related to a page and whether or not a page is bookmarked (Montebello, Gray and Hurley

[7] https://my.yahoo.com/ last accessed for this purpose on 14/June/2012.

1998). Users' browsing histories (Matthijs and Radlinski 2011) can serve to represent the documents of interest. Therefore, users' surfing behaviour are employed to create users' profiles (Lieberman 1995, Lieberman 1997, Barrett, Maglio and Kellem 1997, Sakagami and Kamba 1997), similar to the current research as will be seen in Section 4.3.1.2.

3.4.1.4 Content-Based Filtering

Information Filtering (IF) and the Information Retrieval (IR) employ many of the same techniques (Belkin and Croft 1992), but they slightly differ because IF is typically based on users' interests, whereas IR is mainly based on the user's entered query. While in IR the users express their queries in natural language, in IF users' profiles might represent the users' long-term interests.

Two different types of conceptual profile might be created based on this approach: A user's profile consisting of sets of terms based on the users' browsing histories and an item profile consisting of sets of terms (i.e. extracted from the information base). The filtering engine uses some measures of similarity between the respective profiles in order to select and rank the relevant items to be provided to the users based on their profiles (these are of interest to the users). In content-based filtering, items (i.e. documents) are selected and ranked based on the similarity of users' profiles vs. items' profiles. The other filtering approaches are rule-based filtering and collaborative filtering systems; but these are beyond the scope of the current work, the reader is referred to Pazzani and Billsus (2007) for more details.

However, it is more advantageous to extend the content-based filtering approaches by implementing query expansion through ontologies (Castells, Fernandez and Vallet 2007). The incorporation of ontologies both for representation of items and for user profile generation has previously been investigated in many research studies (Khan 1999, Dai and Mobasher 2000, Susan, Jason and Alexander 2004). Most of these surveys demonstrated that by incorporating ontologies into content-based filtering methods, compared to traditional content-based approaches, the systems were improved in terms of precision and recall. Ontology can even provide a means for selected content extraction which in turn can determine the keywords to be identified in documents (Khan 1999).

Most approaches to personalisation are often perceived as extensions of IF approaches in which navigational profiles, as in the current work, or historical rating of past users are used as input into the pattern discovery algorithms to generate users' models (Mobasher 2007). By joining in turn these users' models (i.e. user behaviours) with the profiles of active users, users' interests can be obtained based on which a PRM can be developed to provide personalised search results. The subsequent sections of all the proposals presented in this survey are based on this viewpoint.

3.5 State-of-the-Art Personalisation Approaches

There have been several research investigations (White, Ruthven and Jose 2002, Joachims, et al. 2005, Fox et al. 2005, Radlinski and Joachims 2005) into the implementation of implicit feedback techniques in retrieval systems. For instance, most of these surveys and evaluations indicated that compared to systems that do not apply relevance feedback, their implicit models were able to improve retrieval performance by adapting users' search queries. Their approaches were based on textual retrieval models by taking the most widely used VSM as representative of keyword search to adopt the query expansion and user profiling in order to improve the performance of IR systems. Many of these attempts have, however, dealt solely with term-weighting in an intuitive standard framework using a BOW approach (i.e. assuming term-independence) to improve the performance of such systems. They focused on adopting useful term-weighting schemes based on a few underlying assumptions and they have deliberately neglected to develop specific models or a theory of IR a priori.

However, this is not to argue that term-weighting schemes developed from such techniques do not have an underlying theory (as they obviously have one), nor is it true to say that the nature of IR from these schemes cannot be interpreted any more. Certainly, the theoretical understanding of retrieval can be improved based on useful schemes grounded in learning approaches. They could ultimately lead to an insight into and accuracy of document ranking that might have implications for designing a personalised search in order to improve a user's search experience. Unfortunately, the paucity of theoretical evaluation might also hinder the adoption of other learning approaches. Collecting implicit information from real users collected in a controlled laboratory experiment might

provide a solution to addressing this problem as their behavioural data would enable examination of relevance based on individual users' needs, and in consequence, derive a proper theoretical understanding of retrieval.

A recent study by Liu, Belkin and Cole (2012) contributed to the literature by developing specific models of document usefulness generated by examining users' behavioural measures as predictors of document usefulness to build prediction models. The implicit relevance feedback learned from users' behaviours employed the dwell time related to the document. Kelly and Belkin (2004) also used dwell time, named *display time*, to present the results of a naturalistic study into how behaviour could be used as implicit feedback for the relevance of a document. As per Collins-Thompson et al. (2011), they incorporated the dwell parameter into computing document relevancy to investigate the reading proficiency of users. They noted that dwells provide a valuable new relevance signal for personalised Web search. Via search logs from a commercial Web SE, Hassan and White (2013) used dwells and result clicks to estimate users' satisfaction. They demonstrated that longer dwells are highly correlated with users' satisfaction as opposed to the findings of Guo and Agichtein (2012) who reported that post-click behaviour is more significant than dwell in inferring search result relevance.

While much of this research led to an improvement in PRM, their algorithms mine browsing history to analyse users' search activities in terms of user dwells rather than re-ranking the search results. The exception is Agichtein, Brill and Dumais (2006) who measured a wide range of user behaviours to demonstrate that the ordering of top results in a real Web search setting can be significantly improved by using dwell as a user's implicit feedback. Xu, Jiang and Lau (2011) also identified that the accuracy of existing algorithms is affected by the lack of a finer granularity in the representation of dwells. A portion of their work is the closest to the current work with regard to addressing these issues by employing user dwells captured at document level to derive an individual user's interests. However, unlike their work, in order to fine-tune the document relevance, the current work combines both the basic keyword-based search and semantic-based search approaches with this feature, producing a system that is characterised by both coverage and accuracy.

3.6 Summary

This chapter has explored several areas which lay the groundwork for the development of the proposed framework. The important areas related to Web based personalisation were covered with particular emphasis on developing a reliable personalised ranking model for individual users. This chapter has outlined the adoption of relevance feedback with query expansion through the integration of concept terms to address the keyword-based drawback in matching the user's information needs. The chapter has further outlined the need to improve the theoretical understanding of retrieval by considering for the learning approaches of personalised search, different implicit relevance feedback rather than relying on that obtained by using only standard term weighting schemes.

The proposed framework, described in the next chapter, combines the elements from each of these areas in an attempt to develop a reliable and accurate personalisation system.

Chapter 4 Identifying Users' Interests

In the previous chapters, several challenges in determining users' interests were identified. To address this problem, this chapter proposes a retrieval model which is built upon different similarity measures between terms. Section 1 gives a brief introduction to the approaches adopted to exploit implicit relevance feedback. Section 2 defines the implicit indicators of relevance used to identify users' information needs. Section 3 presents the framework proposed for a personalised IR system. Section 4 presents the evaluation scheme employed to validate the adopted approaches along with its corresponding experimental results. Section 5 provides a summary of the work.

4.1 Introduction

A number of surveys which implemented implicit feedback in an attempt to improve the performance of retrieval systems were presented in the previous chapter. For instance, Kelly and Teevan (2003) and White (2004) also empirically demonstrated in their surveys that users' intentions can be learnt by implicitly mining their interaction data. Relevant documents based on the user's particular needs can thus consequently be retrieved. The current study builds upon these ideas to identify individual users' interests and preferences which are then used as a basis for learning relevant documents based on which ranking function can be crafted, leading to the generation of a relevance-focused personalised search. To address the relevance problem with regards to terms which are not directly expressed in the users' inputted keywords (i.e. queries), query expansion technique through integration of terms from ontology is adopted.

To obtain the personalisation of search results it was seen in Section 3.4 that the process of personalisation might involve (1) a data collection phase - this is

related to obtaining the users' information during their interaction with the system; (2) a construction phase of users' profiles, in which data collected in the form of raw Web log files needs to be transformed into users' profiles that can be processed and used as input for the next phase. After this transformation, these profiles are employed, and are in turn used to derive the users' information needs and establish their interests in relevant documents; (3) the users' interests are transformed into users' models required for the next phase - the learning phase; this learning phase employs the users' profiles - derived from users' models - to filter the available information based on the similarity computation between the documents and these users' profiles. The rank algorithm takes into account the active users' profiles together with the learned patterns to develop a PRM. This allows ranking of the search results in decreasing order of the users' interests for automatic personalisation. The main argument is that implicit relevance feedback can be employed by the PRM to rank the documents that users are interested in at the lowest ranks in order to assist them in retrieving relevant documents immediately. Three main issues are thus investigated:

1. Firstly, whether users' interests can be identified through implicit interactions in digital web documents. The main challenge that will be addressed is how query keywords and their related concepts can be used to identify users' individual interests (i.e. relevant documents).
2. Secondly, how acquired feedback is preserved over time in order to include representation of both the users' interests and modelling.
3. Finally, how this feedback can be exploited to identify the salient features that describe those contents matching users' interests.

4.1.1 Platform for Interaction Behaviours in Search Sessions

This model is constructed by employing the user's implicit interaction with the system. Users simply need to start working by entering their queries via the proposed system which starts learning about the users' preferences based on their navigational activities in a variety of documents they select. The system starts deriving the user interests and applies them to personalise the search results as will be illustrated here. The system is also adaptive; meaning that it observes the users' interests (i.e. information needs) in search results and

automatically adjusts the weights of different documents while incorporating the information in order to refine each user's interest profile.

Since it is postulated that search logs hold information related to different activities performed by the users, any users' interaction activities would generate succinct behaviours allowing profiling of them. Most previous user modelling approaches for Web search relied heavily on per keyword-based interests; they ignore the integration of users' navigational behaviour (i.e. usage data) with semantic contents which are expressed in ontological terms so as to enrich users' models, and hence fail to identify the individual users' interests (i.e. relevant documents) of topics not directly expressed in the keyword. The current work develops a personalised search based on how well a document matches both the query issued and the user's interest profile enriched with semantic content. The system first extracts information about the users' interests using the HTML documents that users click to build and maintain their interest profiles. These interest profiles are updated and improved by learning from users' feedback as shown in the following steps of the personalised search solution. When users enter their search queries (i.e. keywords), the system processes the keywords as follows to sort the documents in the decreasing order of their final scores based on which the ranked list of the documents is returned to the user.

1. It forwards the queries to the search engine (here Google[8]) which in turn returns a number of documents which are relevant to the queries. Each document is associated with a score (i.e. *the relevance score*) determining its weight; the higher the score, the more important and relevant the document will be to the query. This relevance score is calculated based on term weighting schemes (see Section 2.2) for each term contained in the document (i.e. its term features).
2. It retrieves the user's interest profile to extract an *interest vector* which contains for each of its elements, a word and a score (i.e. the *interest score*); indicating the degree of interest towards the document.

[8] https://www.google.co.uk/ last accessed for this purpose on 14/June/2012. Google was used because of its popularity. Other Search engines are not included so that the consistency can be maintained in the ranks analysis comparison in the experiments.

3. It determines an interest score based upon the match between the document and the user's interest vector for each of the top-k documents (here, k = 30, first three pages (Silverstein, et al. 1999, Keane, O'Brien and Smyth 2008, Wang, et al. 2013)) returned by the SE.
4. It adds new keywords into a user's interest profile and updates the weights of existing keywords to their corresponding interest vectors.
5. It then computes a final score for each document by combining the relevance score and the interest score related to each document.
6. It makes adjustments to the search results according to the user's response (i.e. relevance feedback). It uses an adjustable parameter called *personalisation degree* based on how much the interest score affects the final score depending on each individual user. The personalisation degree is increased or decreased depending on how much personalisation leads to relevant or non-relevant document retrieval for the user.

4.2 User Information Needs

Document relevance is the hardest parameter to evaluate. The previous chapters demonstrated that the documents of interest (i.e. relevant documents) to a user need to be determined. These documents are characterised as satisfying the users' information needs. That is, they return the search results that mostly satisfy the users' information needs. It was seen that implicit profiles based on search logs (i.e. browsing histories) can serve to represent such documents. These documents contain the terms/words and they can thus be represented by the weights of their corresponding terms (i.e. term features). This work formulates generic term-document frequency that utilises the users' implicit feedback to test this assertion.

> ***Definition 4.1*** *For the purpose of the current work, a feature is considered as an attribute of text content (i.e. document content, query content) which is used to make decisions related to the relevance of that text content with respect to a user's interests.*

Thus, to determine a relevant document means to extract its important features which can be properly used to measure factors which are important to a user who is searching for such a document. These features are then used by search

engines to craft the ranking predictors which are often combined together in the ranking functions to improve the retrieval process.

Users' profiles are stored after being collected over time during their interaction with the system. Given a set of users' Web search logs, any documents clicked are archived for each user and the user representations are determined based on these documents. More specifically, a user-document association matrix is constructed and content-based filtering is borrowed to learn a dimensional factor model wherein the interests and/or preferences of the user are determined by using the term frequency factor. To populate the user-document matrix values Cosine similarity scores for each user-document pair (that represent the user's interest towards the document) are employed.

The author argues that retrieval performance will improve in terms of precision and recall, with the use of PRM based on implicit relevance feedback (White, Ruthven and Jose 2002, Kelly and Teevan 2003, Joachims, et al. 2005, Radlinski and Joachims 2005, Fox et al. 2005). Moreover, the accuracy of the retrieval performance can be improved if concepts associated with the users' queries are included in the implicit feedback (Dai and Mobasher 2000, Susan, Jason and Alexander 2004), thus allowing to bridge the semantic gap in order to generate more accurate results.

Based on the above claim, it is imperative to determine which implicit feedback features can be employed in interactive Web retrieval to infer relevant documents; it is also vital to employ this implicit relevance feedback to provide automatic query expansion which can infer relevant documents which are not directly expressed in the query, but are related to its concepts.

The main aim of this chapter is to identify the principal salient features which can describe the content of the documents and search queries of users. Once they are identified, they can be employed both to compute the degree of interest a user might have in the available documents, and to re-rank the documents so that the search results are ranked accordingly based upon the users' interests and preferences.

4.2.1 Data Collection

Assuming there is a set m of users represented by $U = \{u_1, u_2, ...u_m\}$ and a set of n documents represented by $D = \{d_1, d_2, ...d_n\}$, a profile for user $u \in U$ can be represented as an ordered pair of n-dimensional vectors by the equation 4.1 (Mobasher 2007).

$$u^{(n)} = \langle (d_1, s_u(d_1)), (d_2, s_u(d_2)), ...(d_n, s_u(d_n)) \rangle \quad (4.1)$$

where each $d_j \in D$ and s_u is the function for user u which assigns (possibly null as will be seen below) interest scores to documents.

By using $m \times n$ matrix, a conceptual database UP for all users' profiles might be represented as $UP = [s_{u_k}(d_j)]_{m \times n}$, where $s_{u_k}(d_j)$ is the degree of interest related to the document d_j for the user u_k. According to Mobasher (2007), a personalisation system can formally be thought of as a mapping $PS : P(UP) \times U \times I \rightarrow R \cup \{null\}$, which assigns interest values to each user-document pair. Since the mapping PS is often not defined across the whole domain of pairs of user-document, the interest scores for each given user need to be predicted. When the system is not able to predict the interest score, a null value is produced for the PS mapping. This prediction can be represented as $PS(UP, u_k, d_j) = s_{u_k}(d_j)$ for the database of users' profiles UP, a given target user $u_k \in U$ and a target document $d_j \in D$.

In the current personalisation system, a content-based filtering approach is adopted; the other filtering approaches are rule-based filtering and collaborative filtering systems and are beyond the scope of the current work, it is sufficient to know that this work focuses on personalised search through a content-based filtering approach. Given that in content-based filtering, there is normally one single profile for each target user u_k ; the interest score $s_{u_k}(d_j)$ for the document d_j is derived by employing the Cosine similarity measure (i.e. any similarity measure could be adopted) between this document d_j and the user profile UP.

4.2.2 User Profiling

In this study, the required users' profiles are constructed implicitly, and users' behaviours are monitored and tracked intrusively over multiple searches (i.e. search logs) in order to identify the users' searching patterns. This includes keeping track of which documents the users have visited so far. Such profiles can then be used as personalised search inputs to generate the PRM.

4.2.3 Information Sources, Pre-processing and Modelling

The fine-grained navigational behaviour of Web users is the clickstream data which is automatically captured from the Web and application servers in log files (Mobasher 2007) and it is one of the most important sources. Each log entry may contain (but is not limited to) the fields related to the IP address of the client, the resource requested, the time and date of the request etc. (Mobasher 2007). These data are usually captured during user activity interaction with the system (i.e. browsing histories) and they are often used as implicit feedback.

In knowledge discovery, user-centric data representation is often generated through data pre-processing (Mobasher 2007) with the aim of creating user-centric data models (i.e. users' profiles). This includes extracting and transforming features and attributes used to represent each document. These user attributes (i.e. explicit or implicit) are employed to determine individual users' interests in various digital web documents (i.e. the functions $s_{u_k}(.)$). Assuming that each document visited is tracked, these attributes might include (but are not limited to) the document-ID (to uniquely represent the document), the document metadata (e.g., keywords or document content) and the time stamp. In the context of the current work, the objects of personalisation are represented by the abstract documents $d_j \in D$ as shown earlier in equation 4.1, and the IP addresses of each machine were used to distinguish among unique participants, and thus to identify their profiles. It is however, often recommended to use the IP address along with client-side cookies for mapping log entries onto the set of unique users (Mobasher 2007), but due to privacy concerns, client-side cookies are often disabled.

Assuming that a UP contains the respondent's navigational sessions representing his/her online activities in his/her session(s), the sequence of

activities that he/she performed while visiting a link can be reconstructed to form a heuristic session, in the clickstream data. Assuming a set of n links is represented by $L = \{l_1, l_2, ...l_n\}$, and a set of v users' navigational sessions is represented by $S = \{s_1, s_2, ...s_v\}$, where $s_j \in S$ is a sequence of ordered pairs of q-length, the so-called heuristic session can be represented as $s = \langle (l_1^s, w(l_1^s)), (l_2^s, w(l_2^s)), ..., (l_q^s, w(l_q^s)) \rangle$ where each $l_i^s = l_j$ for some $j \in \{1, ..., n\}$, and $w(l_i^s)$ is the weight associated with the Web document clicked on link l_i^s in the session s.

4.2.3.1 Keyword-Based Features

Since each document $d_j \in D$ can represent an HTML document in the context where the focus is to capture the implicit feedback related to the document clicked, equation 4.1 can be used to represent UP. A Boolean match is first performed to determine those documents matching the query from the user's navigational activities; thus a binary related to the occurrence or non-occurrence of a document in the session can be considered as a simple weight of that document; alternatively, a function of the document dwell in the user's session can also be employed.

Each document d_j can then be represented as an attribute vector of k-dimensional features where k is the total number of features extracted; and the feature weight associated with the document is represented by its corresponding dimension in a feature vector which is given by: $d_j = \langle fw_j(f_1), fw_j(f_2), ..., fw_j(f_k) \rangle$, where $fw_j(f_p)$ is the weight of the pth feature in $d_j \in D$, for $1 \leq p \leq k$. When the features extracted are the textual content of pages represented in BOW (i.e. a set of pairs, denoted as $\{t_i, w_i\}$, where t_i is a term describing the content of the page (i.e. document) such that $t_i \in d_j$, and w_i is its weight in denoting its importance with regard to that content), then the normalised $tf \bullet idf$ term values can be used to determine the feature's weights. Each document can thus be represented by sets of term-score pairs (e.g., sport (cricket; 0:54); (baseball; 0:39); (soccer; 0:45)), so that UP is represented as a feature vector using the terms of documents as features (see Section 4.3.1.2) and the user profile can thus be viewed as a vector in the space of content features.

4.2.3.2 Semantic-Based Features

The spreading approach (Crestani 1997) can be adopted in order to trace the users' interests (i.e. conceptual search) based on the input keywords which are not directly expressed. It is adopted here to perform the automatic query expansion (Thiagarajan, Manjunath and Stumptner 2008, Devi and Gandhi 2015) by appending terms that are conceptually related to the original set of terms in documents. There are potentially many overlaps between the current research and all the above-mentioned studies aimed at providing semantic similarities through ontologies, in terms of the classification technique employed to enrich the users' models (i.e. spreading process). However, this project extends the work of Thiagarajan, Manjunath and Stumptner (2008), and employs fuzzy ontology values (Lucarella and Morara 1991) during this process to enhance the semantic similarity measure between terms and the semantic similarity search.

Moreover, this project applies both term weight (detailed in chapter 5) and dwell score statistics (detailed in chapter 6) directly as a dimensional feature to enrich the users' models (Al-Sharji, Beer and Uruchurtu 2013, Al-Sharji, Beer and Uruchurtu 2015). For instance, not only was it shown in these surveys that the performance of the PRM improved, but it was also demonstrated that it could be used as a complementary feature for the system to rely on when the keyword feature proved unsuccessful in identifying the relevance of documents.

Given ontology O and a term t_i, a spreading process might employ the ontology O to spread document d_j, to determine the terms that are related to t_i, so that any terms related to the original terms of the document can be included. Denoting these terms as $\mathrm{Re}lO(T_i)$, the results of spreading the document d_j, is an expanded document $\hat{d}_j$ such that the set of terms $\hat{d}_j = \{t_1,...,t_n,t_{11},...,t_{nm}\}$ and $d_j \subseteq \hat{d}_j$ where $\forall t_{ij}/t_{ij} \in \mathrm{Re}lO(T_i)$ and a path exists from t_i to t_j. This spreading process is an iterative process; and the terms from the previous iterations that are related to the original terms are joined to the document at the end of the iteration by employing ontology terms based on fuzzy ontology values (assuming that they determine the semantic similarity between those terms and

the query keywords) for the integration as described below. The termination of the iterative spreading process thus takes place:

1. when there are no related terms to spread the document with, or simply when $\forall t_i \in d_j / \mathrm{Re}lO(T_i) = \theta$. As will be seen in Section 5.2.2.3, spreading a document with all the related terms might lead to a document with noise and unrelated terms;
2. at the end of the tenth iteration, a maximum number of iteration to integrate the terms was set to ten in this project similar to Wang, Liu and Bell (2010) to avoid a large number of iterations.

Assuming a set of terms $q = \{t_1, t_2, ..., t_n\}$, a set of concept-terms $\mathrm{Re}lO(T) = \{t_1, t_2, ..., t_X\}$ related to the query terms $t_n \in q$ and a document $d = \{t_1, t_2, ..., t_m\}$. Assuming there exists a pair of keywords $\{t_j, t_k\}$ which appear consecutively in q and initialised to 1 for consecutive keywords frequency denoted as c_{jk}. The selection of $\mathrm{Re}lO(T) = \{t_1, t_2, ..., t_X\}$ terms to be integrated into $d = \{t_1, t_2, ..., t_m\}$, will be based on equation 4.2 which is determined by taking the highest fuzzy ontology value denoted as F_{jk} and defined below. If a term $t_i \in \mathrm{Re}lO(T)$ is found to be consecutive with either term t_j or t_k in document d, then the frequency count is incremented. The fuzzy ontology value F_{jk} for the pair of keywords $\{t_j, t_k\}$ is thus defined as the ratio of c_{jk} to square of total number of the concepts keywords extracted from the ontology. For more detail related to fuzzy ontology concepts, the reader is referred to Zadeh (1965) and Bai and Wang (2006), it is sufficient here to know that the defined approach is based on the principle that the vectors arising from any two concepts can be given a non-orthogonal component which is proportional to the highest fuzzy ontology values. For each document $\hat{d}_j = \{t_1, ..., t_n, t_{11}, ..., t_{nm}\}$ integrated with ontology terms and represented by an N-dimensional vector, the semantic similarity measure can then be defined as the usual Cosine similarity measure (Mabotuwana, Lee and Cohen-Solal 2013) between the vectors of this document with any other text (i.e. user profile).

$$F_{jk} = \frac{c_{jk}}{X^2} \tag{4.2}$$

4.2.3.3 Cosine Similarity Measure

For the purpose of this work, in order to compute the vector similarities determining the user's interest in a particular document, the common Cosine similarity measure is adopted (Manning, Raghavan and Schütze 2008) as the similarity measurement technique to represent the user model since it proved to be an effective measure (Manning, Raghavan and Schütze 2008).

Given a user profile $UP = s_{u_k}(d_j)$, and a document $d_j = \{t_1,...,t_n,t_{11},...,t_{nm}\}$ for a given search (document containing a set of texts where each t_i is a k-dimensional vector in the space of content features), the binary Cosine similarity denoted (Manning, Raghavan and Schütze 2008) as $Sim(UP,d_j)$ can be determined using equation 4.3. Such similarity between the two sets of texts clearly indicates the relevance of the document in the keyword-based approach which can be applied to the respective vectors. It should be recalled at this point that sharing some vocabulary in common is the key concept in measuring semantic relatedness (see Section 2.3.1) and definition 2.7 for justification of this measurement.

$$Sim(UP,d_j) = \frac{|UP \cap d_j|}{|UP| \times |d_j|} \tag{4.3}$$

where $|UP \cap d_j|$ represents the number of keywords in both UP and d_j, and $|UP|$ and $|d_j|$ are respectively the number of keywords in the user profile and the document.

4.2.3.4 Semantic Similarity Measure

Given now a user profile with a set of texts $UP = s_{u_k}(d_j)$ and a document $\hat{d}_j = \{t_1,...,t_n,t_{11},...,t_{nm}\}$ for search (expanded document containing a set of texts where each t_{ij} is a k-dimensional vector in the space of content features), cosine similarity denoted as $Sim(UP,\hat{d}_j)$ can now be determined using equation 4.3 to represent the user's interests.

4.3 Personalised Information Retrieval

In order to create the users' profiles (i.e. their observed behaviour) based on which the system can infer user-specific information; the system needs to collect relevant data. The proposed system described in the following section allows implicit data to be saved on the client side (Cassel and Ursula 2001) with the aim of providing personalised search results based on content-based systems across Web repositories. The targeted implicit feedback features are the documents' terms to derive the *relevance score* and the *interest score* related to each document. These content features can then be integrated with user models to perform the personalisation process. Both features can be employed to determine the relevancy of documents by extracting the feedback tracked over multiple searches through the web search history of users.

For Web personalisation, Mobasher (2007) argues for a direct approach to integrate the content and usage data. The process involves transforming each user profile in UP (see Section 4.2.1), into an "**enhanced content profile**" which contains the semantic features of the underlying documents. This mapping of each document or page in the UP to the content feature(s) extracted from documents, is actually performed as part of the data pre-processing (see Section 4.3.1.1). The whole range of this mapping is then the full feature space. This transformation is done by multiplying the user-item matrix UP with the item-feature matrix which produces a new matrix $UP_{feature}$ represented as $UP_{feature} = \{\hat{t}_1,...,\hat{t}_n,\hat{t}_{11},...,\hat{t}_{nm}\}$ where each $\hat{t}_i$ is a k-dimensional vector over the feature space. This concept vector, which reflects the **user's interests** in particular concepts or topics related to that user profile, is then represented. A variety of algorithms can now be applied to the new user data to teach these features to the SE to allow it to compute the ranking function.

Some extensive strategies, plugins, projects and products to collate the data that can be transformed from primary data to a SE include FriendFeed[9] and MyLifeBits[10]. While these projects offer solutions to the problems of how to capture data from their warehouses, they do not reveal the approaches of

[9] http://friendfeed.com/ last accessed for this purpose on 14/June/2012.
[10] http://research.microsoft.com/en-us/projects/mylifebits/ last accessed for this purpose on 14/June/2012.

exploiting different information streams to provide personalised information to users. Thus, the generic framework outlined in this chapter as a guideline to implementing systems that support personalised search is itself part of the contribution of this chapter. The implementation is an algorithm-based approach which directly controls the process by rolling in-house code to extract documents from their warehouses and to capture the users' behaviours in order to build search documents directly (Mobasher 2007).

4.3.1 System Description

The proposed approach has been implemented and a pre-evaluation was performed to ensure its validity and reliability at least for the primary study. The implementation is Java based requiring the components Java Runtime Environment (JRE) and Java Development Kit (JDK) version 1.7.0 and NetBeans 7.1 as the backend with Apache Tomcat as the Servlet container. The design of the system interface is very simple and displays two links: The *Google web Search* link which interacts with the GSE home page and the *Personalised web Search* link which opens the home page of the proposed PSE. The system is developed with two main applications: the Web-server application and the client-server application.

The web-server application is equipped with an engine responsible for retrieving and saving the information matching the users' keyword entered (i.e. queries). It allows the user's input keyword to be forwarded to both the Google search engine[11] and the WordNet[12]. The implementation allows all searching activities (i.e. raw Web log files) to be captured and stored locally for profiling providing thus the first module of this work: *the Log File Creation (LFC)*.

The client-server application is responsible for all offline applications including computing the relevance score, transforming the raw Web log files into user profiles, managing the cosine similarity measures (i.e. learning algorithms) to compute the interest score and performing the ranking algorithms. The client server application supports all modules of this implementation (described in detail in the primary study): Vector Space Modeling, Profile Ontology Model and Combined Web Search Model.

---

[11] https://www.google.co.uk/ last accessed for this purpose on 14/June/2012.
[12] http://wordnetweb.princeton.edu/ "About WordNet." WordNet. Princeton University, 2010, last accessed for this purpose on 14/June/2012.

4.3.1.1 Pre-processing

Pre-processing includes the process of transforming the raw clickstream data captured into a set of user profiles which require other tasks related to the removal of stop-words. In the current study, this was manually incorporated according to the standard of text Retrieval Toolkit[13] prior to applying Porter's (1997) stemming algorithms to bring each word to its root word. These are then converted into feature vectors where the features are the terms in the documents.

As the current project deals with HTML documents only, the content of each HTML document is extracted by processing the first three Web pages related to each user's query. Figure 4-1 and Figure 4-2 show an example of the HTML code of a Web page (i.e. document) line and the extracted text respectively. Documents in pdf format could be extracted with PDFBox library[14], while the metadata and text from documents in other formats (i.e. doc) might be extracted using Apache Tika Library[15]. Regardless of the method, the end product extracted needs to be a set of documents that will be passed on to the SE in an open standard format (i.e. JSON[16]) using human-readable text to transmit data objects of the attribute-value pairs. Here, these are concerned with a collection of typed fields (i.e. relevance score) containing various values used off-line for the ranking function. All phases including data collection, pre-processing, as well as pattern discovery and evaluation, are often performed in off-line mode and the deployment of knowledge is in real-time mode (Mobasher 2007).

```
<link rel = ' 'stylesheet " href= ' 'include/style\_0.css"
type=' ' text/css"> <script language=' ' JavaScript"
src=' 'include/item.js"></script>
<script language=' 'JavaScript"
src=' 'include/fw\_menu.js"></script>
<span class=' ' bodytext">
The espousers of openness and modernisation adoption encompass regional and
international position, security, social dignity, peace keeping and maintaining heritage,
Qaboos's annual tours.
</span>
```

Figure 4-1 HTML Code of a Document

[13] http://www.lextek.com/manuals/onix/stopwords1.html last accessed for this purpose on 14/June/2012.
[14] https://pdfbox.apache.org/ last accessed for this purpose on 14/June/2012.
[15] http://tika.apache.org/ last accessed for this purpose on 14/June/2012.
[16] http://www.freeformatter.com/json-formatter.html last accessed for this purpose on 14/June/2012.

Espouse open modern adopt encompass region international position security social dignity peace keep maintain heritage Qaboos annual tour.

Figure 4-2 Document Text Extracted after Pre-processing

4.3.1.2 Profile Construction

Assuming d_j represents the above extracted document, the original HTML extracted can be represented as a simple set t_i of its n terms: $d \equiv \{t_1, t_2, ..., t_n\}$ i.e. $d \equiv \{Majesty, Qaboos, People, Oman\}$ for document 3 given in Figure 3-1. Taking the whole collection into account (Salton and McGill 1983, Baeza-Yates and Ribeiro-Neto 1999), each term t_k of the document $d_j \in D$ - is the index term or the feature whose relevance is computed (i.e. measured). Such measurement uses a numeric weight w_k which is asserted by building a function of the frequency of the term determined by the $tf \bullet idf$ to indicate its importance in its corresponding document. This leads to the representation of a vector of pairs $d \equiv \{(t_1, w_1), (t_2, w_2)..., (t_n, w_n)\}$ for each document. To discern each $d_j \in D$ with relation to the user, the higher the weight w_k, the more important the term t_k is.

The Index Term Database (ITD) is then built based on all terms extracted and it consists of all index terms (thus the features) of all documents $d_j \in D$ for the search the user is involved with. Consequently, a generic Term-Document Matrix is produced (see Table 4-1). The elements a_{ij} of this Matrix represent the weights w_{ij} of the generic index terms t_i (i.e. features) in the document d_j for a document collection of n documents and m ITD.

Table 4-1 Term-Document Interaction Matrix

ITD	d_1	d_2	...	d_n
t_1	w_{11}	w_{12}	...	w_{1n}
t_2	w_{21}	w_{22}	...	w_{2n}
:	:	:	:	:
t_m	w_{m1}	w_{m2}	...	w_{mn}

For each session, a vector related to the total set of documents is analogously created. Subsequently for the targeted documents related to each query, each document is presented as a vector which can be used to create the user's

profile. To form the final profile for a given user u_k, one session (representing short-term interests) or more sessions (representing long-term interests) associated with that user, can be aggregated, where each document d_j is a document clicked and the interest function value $s_{u_k}(d_j)$ is a function of the weight of the associated documents clicked $w(l_q^s)$; the collection of these profiles comprises the $m \times n$ matrix UP that is used to perform a personalisation process. Given a user profile UP containing p interest vectors for a user u_k, an overall interest vector is often determined by combining all interest vectors for that user (Doug and John 2016). Assuming T_i is the set terms in the i^{th} ($i \in [1, p]$) interest vector, the set of terms of the overall interest vector T can be found as $T = \bigcup_{i=1}^{p} T_i$. For every term $t \in T$, its overall interest vector, $s_u(t)$ for user u_k, can be calculated as $s_u(t) = \sum_{i=1}^{p} s_i(t) \bullet w_i$, where $s_i(t)$ is the score of t in the i^{th} interest vector ($s_i(t) = 0$, if $t \notin T_i$) and w_i is the actual weight of the i^{th} interest vector (see Section 5.1.2), thus making a vector in the space of content features to provide the user profile.

4.3.1.3 Query Processing and Ranking

Users' queries expressed in keywords to represent their information needs (see Section 1.1.1) can be considered as short documents. Therefore, for each user u_k, a BOW representation for each query issued by the user in a particular session must also be created and compared with its set of corresponding documents. This comparison is based on the similarity between both the query and the targeted documents. Thus, equation 4.4 and equation 4.5 are applied to calculate the cosine similarity measure between the query vector and the vectors of the matching documents (relevance score); and the query vector, the vectors of the matching documents and the vectors of the matching user profile (interest score) respectively.

$$Sim(q_i, d_j) = \frac{|q \cap d_j|}{|q| \times |d_j|} \tag{4.4}$$

where $|q \cap d_j|$ represent the number of keywords in both q and d_j and $|q|$ and $|d_j|$ are respectively the number of keywords in the query and the document.

$$Sim(q_i, d_j) = \frac{|q \cap UP \cap d_j|}{|q| \times |UP| \times |d_j|} \quad (4.5)$$

where $|q \cap UP \cap d_j|$ represent the number of keywords in q, UP and d_j; while $|q|$, $|UP|$ and $|d_j|$ are respectively the number of keywords in query, the user profile and the document.

In brief, the highest similarity values determined by equation 4.5 are used to represent the most similar documents between the query and the available documents when both keyword-based and semantic-based features are considered; and these are the most interesting to the users.

While many studies have concentrated either on the combination of the keyword-based and semantic-based features, or on employing the dwell-based features to address the document relevance problem, the current study proposes a combination of all three features. This allows the ranking functions to produce results with maximum relevance to improve personalisation of Web document searches.

To merge these features together, a combination technique employed is model-based and is described in the next section and in both Sections 5.1.4 and 6.3 depending on the approach being implemented.

4.3.1.4 Search Result Personalisation

The personalisation of search results lies to a large degree in merging the models that provide them. Much research in the field of IR has made considerable effort to study the methods of combining several sources of ranking (Castells, et al. 2005, Shaw and Fox 1995). While many of these studies adopted models of parameters to normalise the degree of individual personalisation in a linear combination of relevance scores to combine the rankings, it has been claimed that "automatic preference extraction techniques have an unavoidable risk of guessing wrong preferences, the negative effects of which increase with such a parameter. Even when the extraction is most successful, there is considerable risk of contradicting explicit user requests if the parameter is too high" (Castells, et al. 2005).

In the current research, both relevance and interest scores are naturally combined at document level, thus allowing linear combination at the model level

to provide the document relevancy based on all models together. Firstly, it avoids any potential overfit caused by matching the query against the same document (i.e. in both relevance score and interest score) and the subsequent increase in weighting score which a document might be given. Secondly, this has the advantage of reduction in computational time when the proposed method is applied to real-life Web searches.

In the current experiment, the so-called CombSUM formula (Fox and Shaw 1994) was adopted to combine the resulting relevance scores of both models, although several methods for combining the individual similarity values were explored and found to work better (Shaw and Fox 1995). In this strategy, the similarity values are merged at the document level and results are combined based on the *Similarity Merge* (here, the CombSUM). Thus for both approaches (i.e. keyword-based and semantic-based), the documents retrieved from the sub-collection runs are merged for their corresponding query based on *Similarity Merge*.

4.4 Profile Evaluation

Automatic personalisation implies the creation of users' profiles and automatic update by the system with possibly only minimal explicit control (Mobasher 2007) by the users. The approach adopted to create the actual users' profiles is currently being implemented as part of the proposed automatic personalisation approach without involvement of the users. Such an approach must be evaluated experimentally for both its validity and reliability (deferred till chapter 5). Evaluation is, however, one of the most crucial components of any IR system including the proposed system.

In most IR systems which are based on implicit feedback, the users' profiles store information which characterises the individual user in a format that allows efficient usage of that information (i.e. those profiles) by those systems. With the aim of analysing the reliability of the proposed approach in identifying users' interests through implicit profiles, it is essential to make an evaluation of the performance bounds of these profiles. As the current system is not yet validated, it is not possible to make a direct comparison with existing methods to evaluate these performance bounds.

Furthermore, considering the diversity of users' information needs based on the same input keyword (i.e. query), it is clear that the relevancy of documents is very subjective; therefore, users need to be involved in judging the coverage and accuracy of the system with respect to the search results it returns. Given the format of these users' profiles, such evaluation by profiles' owners who are not highly expert is questionable, although some findings (Gauch, Chaffee and Pretschner 2003, Vallet, Fernández and Castells 2005, Castells, Fernandez and Vallet 2007) recruited non-expert humans who were employed for both the data collection and evaluation of their systems.

On the other hand, scientific research requires findings which are reproducible and independently verifiable. In IR surveys including the current research, the related scientific findings are behavioural patterns. To validate these patterns, evaluation with ground truth (i.e. surveying the individuals' featuring patterns) is required to prove that users' intentions are captured. Given this diversity of users' information needs and their corresponding degree of document subjectivity, the challenge is clearly the lack of ground truth to capture behavioural patterns, although there exists a readily available variety of metric systems (see Section 2.4) to assess IR models. Some proven scientific surveys (Lamiroy and Sun 2011) addressed this evaluation problem without surveying the users, but the presentation of their evaluation techniques is not detailed enough to allow other researchers to replicate the evaluation of novel search techniques in order to test their systems.

Therefore, to evaluate the performance of the proposed approach, inspired by Robertson (1981), it was decided to develop a pre-evaluation plan whereby the data is analysed in such a way that only those questions related to the *raison d'être* of the project can be answered. Such a pre-evaluation can be used to assess how well the computational method - specifically the learning algorithm and the ranking algorithm - finds patterns and predicts the outcome; thus, based on the obtained outcome evaluation, a judgement of the performance of the computational method can be made.

4.4.1 Experimental Setup

Predefined queries (i.e. topics from TREC and Kaggle[17]) are often used for IR systems to evaluate both the effectiveness and reliability of any novel Web retrieval technologies. Such predefined queries are artificial queries (Robertson 1981) usually employed in experiments to play the role of data that can reflect the experiences of users in real-world enterprise searches. Adopting these ideas, a list of six (Kage and Sumiya 2006, Ahn, et al. 2008) predefined queries (i.e. artificial queries) was created, including *Old Oman*, *Jupiter facts, Insomnia, Global Warming Hit, USA wars*, *Prophet Mohammad SAW* (see APPENDIX A).

For the purpose of this work, these predefined queries are of general interest with two-fold goals: (1) they are used as topics for a vision application scenario (Robertson 1981) rather than topics which are intended to create a real-life application scenario; (2) they are used in real-life experiments for further evaluation (see Section 6.4). As such, they are currently employed to measure ranked lists (i.e. document relevancy) based on generated users' profiles if such queries are issued. Until a user study is conducted based on a full-scale version to validate this evaluation, the system is considered to be in its infancy, although most of its components are already in use.

For each topic, a test document collection and a suite of information needs were identified which henceforth will define the set of interests employed in the current experiment when the test runs are implemented. Some proven scientific surveys (Ahn, et al. 2008) were based on 18 searches to measure ad-hoc information retrieval effectiveness in a standard way, therefore a total of 18 different information needs were derived although as a rule of thumb (Manning, Raghavan and Schütze 2008), a sufficient minimum reasonable size is often considered as 50 information needs. However, Robertson (1981) warns of the danger of generalising these kinds of experiment results based on a very limited selection of documents rather than a sample. As will be seen in the next chapter, a primary study involving 50 users will further validate these pre-evaluation measures. This pre-evaluation can also benefit the system's scalability to be determined since the tasks will scale up with these growing data sets.

[17] https://www.kaggle.com/ last accessed for this purpose on 30/September/2014.

In order to apply evaluation metrics to the IR system proposed, the required datasets need to include a set of keyword queries and a set of documents that are deemed relevant to these queries. Thus, for each query-document pair, a set of relevance judgments (i.e. ground truth judgment of relevance) was deduced. Cross-validation of the identified information needs was performed by three independent study coordinators (Manning, Raghavan and Schütze 2008) to validate this encoded ground-truth information which might also be referred to as the *Gold Standard* (Manning, Raghavan and Schütze 2008). Having in place these three desiderata involving the total environment of IR (Keen 1981), a valid test could be performed. Table 4-2 provides a summary of the datasets used in the actual pre-evaluation.

Table 4-2 Statistics on Pre-evaluation Dataset

Statistics	Value
# of Queries (Topics)	6
# of Relevance Judgments/Topic	3
# of Relevant Documents Extracted	18

It is important to note that while the pre-evaluation provided in the current validation can be replicated for comparative purposes, the author does not claim exhaustive replicability of such data given the dynamic nature of the content of HTML documents.

4.4.1.1 Assessments

Since the current project's *raison d'être* is to push the most relevant items to the top of the ranked list, the assessment will be chiefly directed at calculating the overall averages (see Table 4-3) of precision till rank 10. The system will thus be evaluated with respect to definition 2.9 to test if the experimental system performs better at the objective level, that is how well the relevant documents are pushed to the top (i.e. first 10 ranks) of the ranked list. Assuming that $L_t[1,k]$ is the top-k of ranked list returned by the system and S_t is the set of relevant documents, *precisionl@k* and *recall@k* (see Section 2.4.1) can be respectively computed as $|L_t[1,k] \cap S_t| / |L_t[1,k]|$ and $|L_t[1,k] \cap S_t| / |S_t|$. However, the ranking position is vital here since all 3 documents (i.e. targeted documents) are relevant. Therefore, the precision of a topic for which its relevant documents are selected from position rank 1, then 2 and finally 4, their corresponding

precision[18] might be respectively calculated as 1/1=1, 2/2=1 and 3/4=0.75. The system will rather be tested to establish only some approximate ideas of its quality and feasibility, than measuring its performance in any very refined sense (Robertson 1981).

Table 4-3 Experimental Results - Document-Level Performance

Document	$P\_D_1\_Q_n$	$P\_D_2\_Q_n$	$P\_D_3\_Q_n$	Sim_Precision	Document	$P\_D_1\_Q_n$	$P\_D_2\_Q_n$	$P\_D_3\_Q_n$	Sim_Precision
1	0.60	0.50	1.00	0.75	10	0.50	0.75	0.60	0.60
2	0.60	0.50	1.00	0.75	11	0.43	0.60	0.50	0.60
3	0.50	0.60	0.60	0.75	12	0.60	0.75	0.75	0.75
4	0.60	0.50	0.75	0.75	13	0.60	0.60	0.60	0.75
5	0.50	0.60	0.75	0.60	14	0.60	0.75	0.75	0.60
6	0.60	0.75	0.60	0.75	15	0.50	0.50	1.00	0.60
7	0.60	0.75	0.75	0.75	16	0.60	0.60	0.75	0.75
8	0.75	0.60	0.60	0.75	17	0.50	0.75	0.60	0.75
9	0.60	0.43	0.50	1.00	18	0.33	0.75	0.60	0.60

The PSE was applied to issue each of the 6 queries (i.e. topics) on behalf of Web users (Wang and Jin 2010) to return the search results so that they could be collected, studied and analysed. From the interaction of the system with the above topics, internal images of the users' needs (i.e. users' profiles) were built up and a set of search results (i.e. related documents) was generated. As relevance judgements on the output of these searches are indicated, they can be used to test the tool. Several test runs were implemented on the basis of the topics and the search process, and the output relevance judgements which were available. It is important to mention that both the search process and the relevance judgements might affect the evaluation of the results; therefore, in terms of both the search process involving real users and their relevance judgements, this information is incomplete. However, although paradoxical, keeping the artificiality of all aspects of the tests is not the main purpose, but rather using genuine searches with relevance judgements while adhering to the philosophy of the proposed framework. Thus, these methods sound appropriate

[18] https://www.cs.utexas.edu/~mooney/ir-course/slides/Evaluation.ppt last accessed for this project on 30/September/2015.

in this context of the limited objects of the tests being performed (Robertson 1981).

4.4.1.2 Results

"The subject of how to analyse retrieval test data has been, with the problem of relevance, one of the two most highly debated topics in the field" (Robertson 1981). Here, rather than exploring the evaluation criteria which are directly associated with a document, this analysis aims to examine the whole process (i.e. fully-fledged version) in shaping the topic (user's needs); this assessment will be validated in the preliminary studies conducted on real-life experiments in the next chapter. Inspired by Robertson (1981), the experimental design was made quite simple and was based on straightforward rules controlling the searching part of the system, thus allowing replication of the searches or the order in which the systems are tested. Therefore, each topic will be searched against the system, provided that certain aspects of the respective experiments are made as realistic as possible to allow selecting from the prototype in a manner appropriate to other artificialities, objectives and resources.

Using the above ground truth, the evaluation metric values of precision (see Table 4-3) were calculated at the document-level to validate the scores of the PSE in identifying the user's interests. The visual representation of the system performance is shown in Figure 4-3. It is important to note that these precision values are based on the *Similarity Merge* (i.e. CombSum) the two models.

As can be seen from Figure 4-3, while poorer performance was obtained for the first run referred to as $P\_D_1\_Q_n$ in Table 4-3 (represented in blue in Figure 4-3), its highest precision value = 0.75 and lowest precision value = 0.33. These values improved for the second run referred to as $P\_D_2\_Q_n$ in Table 4-3 (represented in red in Figure 4-3) and further improved for the third run referred to as $P\_D_3\_Q_n$ in Table 4-3 (represented in green in Figure 4-3). The highest precision value = 0.75 and lowest precision value = 0.43 for the second run, thus indicating the effects of users' models. The best performance was achieved when users' models were completely constructed with highest precision value = 1.00 and lowest precision value = .50.

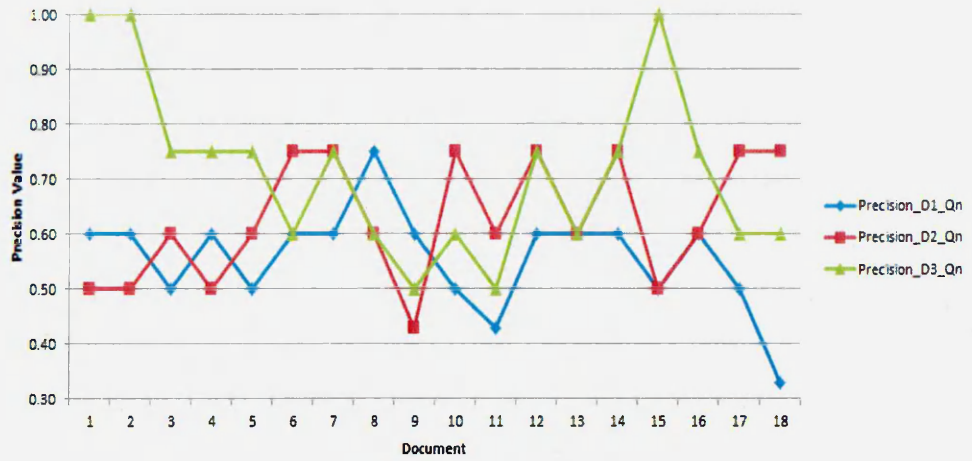

Figure 4-3 Documents Matching Artificial Queries

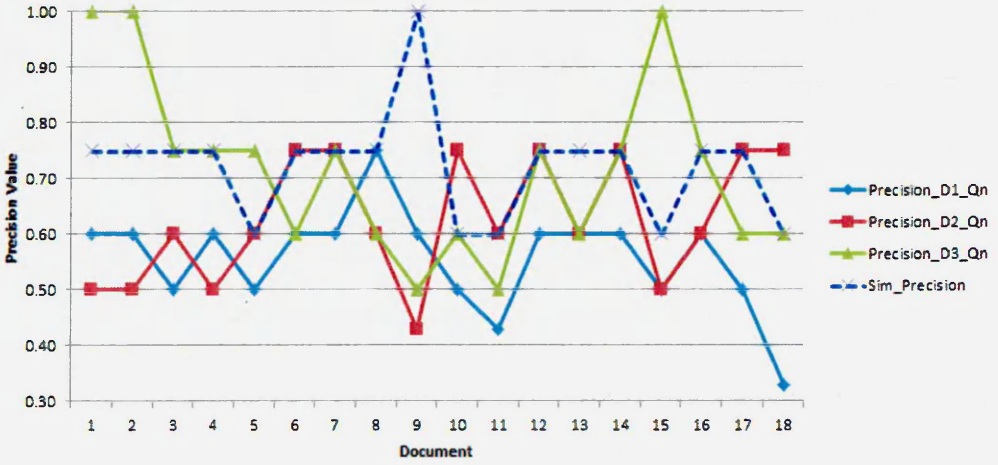

Figure 4-4 Effect of Implicit Feedback

For instance, some folders corresponding to the content of documents first clicked for each query (See APPENDIX A) including the folder created at the cold start (when users' profiles were still empty) were deleted. The same queries (topics) were issued after the deletion of the folders representing the corresponding documents first clicked to allow the same documents to be re-clicked. Clearly, this time, the users' models were already initiated and their corresponding users' profiles were represented. As can be seen from Table 4-3, some changes were observed in the order of the ranked lists as a result of this simulation process. The precision values were calculated, and they are referred to as Sim_Precision in Table 4-3 and graphically represented by the dotted line in Figure 4-4. The improvement is clear (highest precision value = 1.00 and lowest precision value = .60) and the findings are now mostly consistent with the performance obtained in the previous run represented by $P\_D_3\_Q_n$ in Table 4-3. This clearly shows that the system used the implicit relevance feedback

through users' built profiles. As can be seen, a few results did not change after the simulation process. Each term/keyword contributing to the users' profiles was checked one by one. It was then observed that the *interest score* played the major role, whereby the lower the similarity in the users' profiles, the fewer the changes were made, a clear evidence of the claim made in Section 4.2 of this chapter.

Apart from showing an improvement in performance, the results of this simulation of the users' profiles were also used to determine the accuracy of their learning process by using the relative error between the tested system and the GSE. The precision till rank 10 was calculated for GSE based on the ranked lists for all three documents related to the queries. The results of the simulated profiles were employed to calculate the relative error (Qiu and Cho 2006) by taking the difference between (1) the results obtained after simulation and results obtained after complete creation of users' profiles; and (2) the results obtained when the GSE was employed. They are both graphically plotted as shown in Figure 4-5, from which it is clear that the relative error in the proposed system is significantly smaller than in the GSE, except for the representation of only one document.

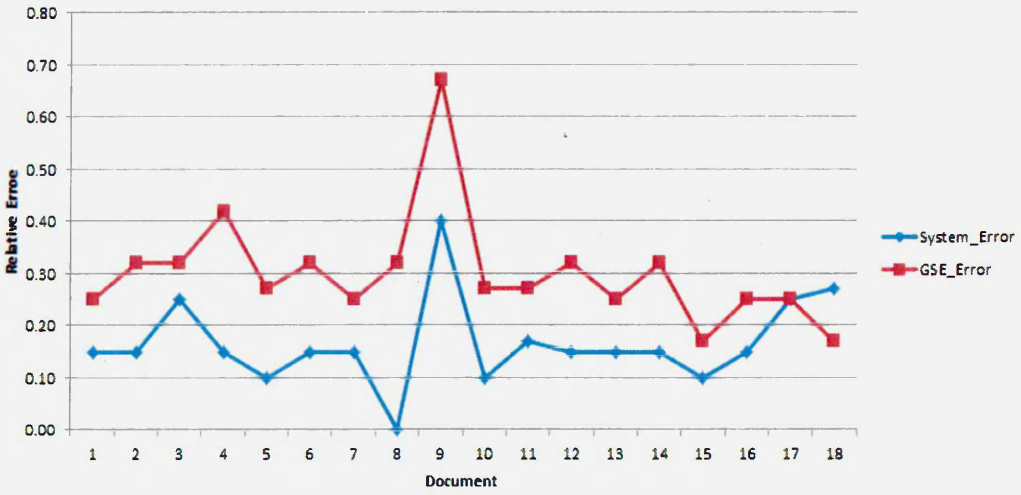

Figure 4-5 Comparative of Relative Error in Ranked Lists' Accuracies of both PSE and GSE

These pre-evaluation results indicate that the proposed system is better in terms of high precision for the ranked lists. These findings support the expected results and show that implicit relevance feedback addresses the evaluation criteria adequately when addressing the concept; therefore, they can be used to improve the retrieval results of the system. With this encouraging success, the

system can now be deployed to begin the primary study (see next chapter). Of course, considered alone, the volume of data collected here is not sufficient to support the reliability of the proposed system. It is possible that the collection of documents used for the test was selected to match the queries, thereby sacrificing the quality. While this test provided some information about the quality of the ranked list, both its quality and quantity are critical for a reliable report to be prepared. The quality of ranked lists is based on the documents displayed at the lowest ranks and document relevance is the hardest parameter to evaluate.

4.5 Summary

In this chapter, the concept and implementation of the proposed system were shown. A PRM was presented to provide personalised search results effectively based on users' interests and preferences, and employing users' feedback in their profiles which were built implicitly. The proposed approach has addressed the problem of the keyword-based approach by introducing semantic content which is expressed in ontology terms.

To evaluate the performance of the proposed system, document level precision was adopted because a ranked list of documents returned by the system as a response to queries was the main issue. Even though interesting and encouraging results were obtained, these results simply revealed the static side of the proposed approach. However, such results are considered as static pictures which were only generated for the pre-evaluation of the approaches based on the log data tracked from artificial queries and an incomplete sample size. In the aspect of this experiment, interactions from real users based on a reliable sample size were missing. Therefore, in the next chapter, rather than using a fully-fledged version as a whole, a user study based on real-life experiments will be conducted on a growing dataset with real users who will perform real search tasks to allow real world problems to be investigated for different models of the proposed approaches as closely as possible.

Chapter 5 Preliminary Study

In the previous chapters, the approaches adopted for determining the users' interests were introduced. A pre-evaluation of the performance of such approaches was performed on a small-scale retrieval experiment by using artificial queries. The findings of the pre-evaluation indicated that implicit feedback can be reliable in providing personalised search results. In real-life experiments, of which the results are presented here, users' search logs were gathered as stimuli to further study and validate these findings. Section 1 introduces the objectives of the experiments. The design of the study is given in Section 2 while Section 3 presents the discussion of the result findings. The summary of the findings are given in Section 4.

5.1 Introduction

The proposed personalised search approach was implemented and experiments on real-world data were performed to evaluate its performance. The approach, as in most other personalised searches, is algorithm-based whose function attempts to provide users with a relevant collection of documents representing their individual needs as well as the context of their activities (Micarelli, et al. 2007) represented by some form of models. The search results are tailored based on the users' preferences expressed in the form of queries while as in many previous studies (James et al. 2002, Teevan, Dumais and Horvitz 2005a), filtering is based on users' profiles and re-ranking is adopted to develop a PSE.

Considering the problem definition given in Section 1.1.1 stating that effective IR systems (i.e. those with good predictor functions) provide ranked lists with the less relevant documents below the relevant documents - the ranking contents for a user's query (i.e. search) is based on how much these contents

really satisfy the information needs - therefore it is important to describe how the relevancy of documents is interpreted in order to transform a PSE into an apparently smart system that understands the needs of users. The objectives thus put the focus on representing users' interests and preferences in a formal way, such that the validity of different models it contains can be checked with regards to the relevance of documents (i.e. search results) when handling different keywords and their related concepts. It is thus important to understand the corresponding algorithms and their functions for the sake of assessing the value of the document modelling features employed.

5.1.1 Log File Creation Module

To employ the search log data of each user as raw data (Teevan, Dumais and Horvitz 2005b, Teevan, Dumais and Horvitz 2010), the Log File Creation module algorithm was developed. Its implementation is based on Web document modelling devised by Micarelli, Sciarrone and Marinilli (2007). When the user u_k begins the initial search - by entering a search keyword related to his/her information needs and clicking on a document - the algorithm annotates the document's full-text content and saves it in a special folder to provide the user's Log File which was referred to, along with the search keyword entered (i.e. query), as raw Web log files in Section 4.1. These documents can be represented as $d_i (i = 1,...,n)$ where n is the number of documents clicked so far.

As was seen in the previous chapter, both contents (i.e. the content of the document and the content of the query) serve as raw data to model the user and represent the user's profile. Therefore, for each query entered, the implementation allows the subsequent contents of each document, in an iteration process, to be instantly saved in their special folders on the client side with the complete browsing details including the content of both the keyword entered and the link visited as well as the time of visit. Basically, both the user query and the user model - except for the first document clicked because the user profile is empty - will work together to filter and rank the documents (see Section 5.1.4). To obtain the complete log database, all the user's logs are combined, and each of the log files is identified by the user ID (i.e. annotated by the IP address) U_i where $0 < i \leq Q$. That is, if there are Q numbers of users, then the complete log database will have the details of all the Q individual log files

(see Table 5-1[19]). The content of each folder is then used as the raw Web log files to other modules described in the next two sections.

Table 5-1 Small Sample of a User's Log File Created

Keyword	URL clicked	Time Stamp	URL Position
Jupiter facts	http://space-facts.com/jupiter/	10:02:08	1
Insomnia	http://en.wikipedia.org/wiki/Insomnia	10:08:10	2
Insomnia	http://sleepfoundation.org/insomnia/home	10:12:56	8

5.1.2 Vector Space Modeling

Vector Space Modeling is the main component of this project which includes the ETL process (see Section 3.1.1.1) and the complete analysis process. This involves the conversion of the field values (i.e. text till now) into elements called tokens. Tokens are nothing more than words which can be matched to tokens extracted from a future search (Doug and John 2016); they are considered a match, solely when they match exactly. These tokens obtained from the analysis process are the dominant features used with full-text search for matching a query with documents in the index. The analysis also comprises the complete pre-processing including recasting the field values in representations (i.e. removal of stop-words and stemming process).

The implementation here allows the raw data to be extracted from their sources (i.e. where the content is saved, here, the log database). This raw data is then converted into search fields according to the documents described in Section 4.2.3 to encode the features of their content. These documents are further improved by adding in new fields with external information (i.e. semantic-based concepts). For each document, a score representing how well the document matches the query is calculated to provide the ranking function. A relevant document is obtained by identifying all the term features of its content. Document relevance can thus be measured according to how well the document content satisfies the user's information needs. Therefore, a ranking function can specify the document's relevance.

Ranking functions often take in information from (1) the keyword query and (2) each matching document of the query. Thus, equation 5.1 is used to allow a Boolean match based on the simplest version of term weights to be performed

[19] Content saved are the HTML Documents corresponding to these links, accessed at that time.

before everything else to determine the documents matching the query. With such matching documents in hand, the ranking functions can thus be used to score each matching document. Taking the four documents illustrated earlier in Section 3.1.1.1 as an example, an ITD (see Section 4.3.1.2) for five of its index terms in the Boolean Model is represented as shown in Table 5-2.

$$tf(q_t, d_j) = \begin{cases} 1 & iff \ \ q_t \in d_j \\ 0 & otherwise \end{cases} \tag{5.1}$$

where d_j is the matching document, q_t is the query term and tf is its presence.

Table 5-2 Boolean Representation of Example of Collection of Documents

ITD	Terms	d_0	d_1	d_2	d_3
t_1	*icon*	1.0	0.0	0.0	0.0
t_2	*Oman*	0.0	0.0	1.0	1.0
t_3	*prix*	0.0	1.0	0.0	0.0
t_4	*Sultan*	0.0	1.0	0.0	1.0
t_5	*Qaboos*	1.0	1.0	1.0	0.0

Every document $d_j \in D$ is then considered as a vector of real numbers (i.e. points) in an m-dimensional space with $m = |ITD|$, where every term t_i represents a dimension. By enforcing the rules of formation of weights as described below, the term features - determined by the weight w_{ij} - reflecting the relevance of the document are extracted. Such weight is calculated by considering the whole collection D of documents (see Section 4.3.1.2) from which the actual document d_j is retrieved (Baeza-Yates and Ribeiro-Neto 1999). These weights obviously represent the relevance of their corresponding documents such that for every term $t_i \in d_j$, as mentioned earlier in Section 2.3.2, the value of the weight w_{ij} is calculated based on the correspondent index term. Such weight therefore represents the term feature expressed as $w_{ij} > 0 \ \ iff \ \ t_i \in d_j$ and $t_i \notin all$ documents of the collection D; which leads for example to the four vectors represented in Table 5-3 considering the documents 0 to 3 given earlier in Section 3.1.1.1. In this implementation, this weight uses the statistics of the index (i.e. inverted index) for matched terms so that a numerical weight for each term is computed.

Table 5-3 Representation of Example of Collection of Documents by the VSM

Document	w_1	w_2	w_3	w_4	w_5
d_0	1.000	0.000	0.000	0.000	0.333
d_1	0.000	0.500	0.500	0.500	0.333
d_2	0.000	0.000	0.000	0.000	0.333
d_3	0.000	0.500	0.000	0.500	0.000

At each iteration of Vector Space Modeling, the ranking functions calculate scores for fields; according to (1) definition 2.3 (using equation 2.5), (2) how often a particular query-term occurs or the term frequency based on definition 2.4 (using equation 2.6) in those fields and (3) cosine similarity between the document and the query (based on equation 2.17). Within each field, term scores (which form the user profile) are combined by summing them together although this combination could be performed by considering the highest value of the scoring field (Doug and John 2016) providing the relevance score. Each user profile is instantly built by analysing the contents of the log database for each corresponding user ID (see Section 4.2.3.1). The user profile obtained is then employed to calculate the interest score based on equation 4.5.

Two more procedures which are performed in this algorithm[20] include (1) storing all tokens from the analysis into SE data structure to allow the documents to be indexed for fast retrieval. (2) Storing the original text (i.e. un-tokenised) and other fields' values (deferred till chapter 6) for ranking calculation purposes since they have to be returned in the search results.

5.1.3 Profile Ontology Model

The ranking function is tuned by combining different ranking predictors together in the ranking function to solve the relevance problem. Current implementation allows the concept terms associated with each term-query to be integrated into each document from WordNet to build a Profile Ontology (PO). These terms are simply mapped to generate a map from the extended set of terms to the original numbers they correspond to when sorted lexicographically (see Section 3.1.1.1). The PO module is basically the algorithm that expands each document with ontology terms and conducts the analysis process. First, each field value related to both the query and the documents is converted into tokens. The

[20] After the analysis process is completed.

analysis thus involves full-text of all the documents, including the query, documents saved and documents $\hat{d}_j \in D$ (where $0 < j \leq k$, in the current work, $k = 30$ since ranking will be performed on the first three pages of any future search).

The algorithm currently supports term frequency so that POs are processed iteratively. At each iteration of term integration, fuzzy ontology values between each term-query and its concept terms are calculated based on description provided on Section 4.2.3.2. This allows the terms with the highest semantic similarity to be selected so that the interest score is thus calculated based on equation 4.5.

5.1.4 Combined Web Search Model

This module provides an overall score for both the VSM and PO module. It is important to note that this score here is the overall score of both modules together. Therefore, assuming the relevance and interest scores provided by the VSM and the PO are represented by $g_k(x)$ and $g_o(x)$ respectively, then the overall score $gf(x)$ for a document x representing the combined Web search model can be calculated by employing equation 5.2.

$$gf(x) = g_k(x) \times (1 - p) + g_o(x) \times p \tag{5.2}$$

where p is the parameter controlling the individual personalisation degree.

In the current study, adjusting p freely between zero and one as an ad-hoc value or as a learnt *optimal* value which can be used for fusing all search results is questionable. For instance, it will contradict the desire of providing individual user's needs and gives control of the power of the retrieval model to the fusion (i.e. ignoring different personalisation degrees) rather than individual interest. Thus, in the current study, p takes on the dwell value related to each document (see Section 6.3 for details of its implementation). Rather than "taking the *vagueness* in user requests and system responses as an approximation of the uncertainty in the search" (Castells, et al. 2005) to gauge the impact of personalisation degree, the value of the dwell is employed.

The experiment results (see Section 6.4) show that by using the dwell value to control this parameter, the degree of personalisation can be improved. To the

best of the researcher's knowledge, this study is the first of its kind to use the dwell directly as the parameter controlling the individual personalisation degree in the linear combination of the final score. Searchable fields are indexed for computation of the overall score in such models, thus opening new opportunities to widen the applicability of this feature to initiate new research questions.

In order to claim that the performance of the proposed retrieval system is increased by combining the two models, as the focus is to bias the search result to the individual user, the similarity values used to provide the ranked lists need to be empirically reliable from the two retrieval runs based on both the keyword-based and semantic-based approaches derived by the above two modules: Vector Space Modeling and Profile Ontology Model. In the following section, the effects of the proposed approaches for each model will be investigated as closely as possible using real search tasks to allow the quality of each model to be validated.

5.2 The Study Design

The main question being addressed in this experiment is the same as the one dealt with in the pre-evaluation performed in chapter 4, except that a real life collection of search logs is employed thus allowing the datasets by query owners to grow. Therefore, as seen in Table 5-4, three conditions (baseline, Vector Space Modeling and the Profile Ontology Model) are defined to scrutinise the result findings based on three different models. Real user data extracted from the log files of the installed system were used. Employing such log files, a dataset was thus constructed containing the following attributes:

Table 5-4 Three Different Experiment Conditions

Condition 1	Condition 2	Condition 3
Non-Personalised (Baseline)	Personalised (Based on VSM)	Personalised (Based on PO)

- User ID
- Query ID and Query Keywords
- Documents Retrieved (top 30, first three pages)
- Documents clicked
- User Profile (Weighted Terms)

These attributes have provided for each subject, every query issued along with its corresponding retrieved documents, the user profile, and the relevance measure of the documents retrieved within each time frame. This can allow a snapshot of all the subjects to be re-constructed including their queries, their user models as well as their retrieval results. These offline experiments were organised in such a way that they were conducted based on the ranked lists of both the baseline system and the experimental system which first included a version of the system without the filtering functionality based on the PO model. Clearly this version will simply be performing the base search function (keyword-based filtering) which is triggered by the user queries without support for the query augmentation (the concept related terms). This organisation has allowed the researcher not only to ensure that when it was included, the PO model increased the overall precision in terms of a personalised search, but any differences in the value of its integration could also be detected.

5.2.1 Experiments and Results

The goal of the study is not actually formalised as a hypothesis, however, it investigates if at the operational level, the experimental system performs better. In other words, it investigated if the features based upon both the VSM and the PO models perform better than the baseline. That is, to determine whether the results returned by the experimental system (both condition 2 and condition 3) will be characterised by both coverage and accuracy. Thus, the overall precision for the ranked lists generated by both systems[21] will be calculated based on the above three conditions.

To investigate if the personalised search solution really improves search quality compared to its underlying non-personalised search engine, the objectives of the experiment were therefore based on the following: (1) Firstly, how the real data collected during interaction between users and the system can affect the performance of the proposed solution (i.e. to test if personalised search performs better). In other words, this will determine the usefulness of the acquired feedback preserved over time in the form of user profiles to include the representation of both the users' interests and modelling. (2) Secondly, whether the identified salient features describing the contents matching the

[21] Experimental here is referred to as PSE and baseline is referred to as GSE.

users' interests provided in the proposed models improve the performance of the proposed solution. The experiments conducted will thus be described and their result findings will be presented in the rest of this section.

5.2.1.1 Experimental Data

The lab-controlled experimentation carried out involved 50 users, who all work as IT researchers in a private sector in Oman. They are all professional researchers with web search experiences; hence rich log files could be obtained (Mobasher 2007). The data sample was collected over a period of two weeks (Dou, Song and Wen 2007). This data sample included 1261 keywords with their URL selected, along with the time stamp for each selected URL. The useful primary data obtained included 729 keywords. The collection of the 1261 keywords covered the period from 19th of May, 2012 to 30th of May, 2012 and the participants were encouraged to use the system any time from 07:30am to 04:30pm in the laboratory where the system was installed in each participant's machine. They were offered lunch daily during the two weeks of the collection. The participants consisted of 38 females and 12 males with an average age of 24.7 years.

5.2.1.2 Experiment Methodology

The most typical evaluation methods used in existing personalised search research is to conduct user studies (Dou, Song and Wen 2007, Wang and Jin 2010) based on large-scale evaluation to examine their search logs. While this approach does not put any constraint on either the large sample size of participants, or on the number of test queries, biasing the approach towards self-selected queries is crucial when examining relevance based on individual users' needs. Such evaluation enables the researcher not only to obtain the users' identity information in order to match their search logs with their interest profiles, but also to evaluate real collections, not to mention that relevance can be explicitly specified by the owners of the queries.

The key challenge here is to determine which documents from the ranked lists are regarded as useful and relevant to their corresponding search queries by an individual user. Following Teevan, Dumais and Horvitz (2007), the current approach employed a lab-controlled experiment using self-selected queries (i.e. 1261 keywords) to address the issue. Users' identity information (established

through the IP address) was used to match their search logs with their interest profiles built from the interaction with the system and their clicking decisions are considered as relevance judgments (see Section 3.1.3.2). That is to say that, if a given user $u_k \in U$ clicked the document d_j after issuing a query containing the word t, then the document d_j will be considered useful and relevant to t for user u_k, and documents that are not retrieved, are judged as non-relevant.

With this information in hand, in order to evaluate the search accuracy of the PSE, a strategy inspired by Wang and Jin (2010) was adopted although their method was based on a large-scale evaluation based on a collection of online social systems. The set of document $d_j \in D$ containing the word t selected by u_k needs therefore to be checked whether they are highly ranked in the ranked list generated by the personalised search solution. As such, the major drawback of this evaluation approach is the high false negative rate since a document not clicked may still be considered useful by u_k (Dou, Song and Wen 2007, Teevan, Dumais and Horvitz 2007). Despite this drawback, it remained possible to derive a fairly accurate idea about the performance of the PSE.

5.2.2 Experimental Set up

The experimental results are presented in this subsection. For simplicity, the proposed personalised search approach is referred to as Personalised Search while the normal search approach which employs the Google search engine is not personalised, and it is referred to as baseline (see Table 5-4). There are two main sets of experiments: (1) Implicit Feedback vs. No-Feedback. The experimental results of this first set of experiments represent the documents' relevancies and are presented in Table 5-5 and visualised in Figure 5-1. (2) Keyword-Based vs. Semantic-Based. The experimental results of this second set of experiments are graphically presented in Figure 5-2.

5.2.2.1 Implicit Feedback vs. No-Feedback.

The findings of the experiment performed in the pre-evaluation indicated that implicit feedback can be reliable in providing personalised search results. However, search results are highly variable over information needs and different documents; thus, the average performance needs to be calculated over fairly large test sets.

In this experiment, it will be investigated how a large sample of real data collected during interaction between users and the system can affect the performance of the personalised search. This includes investigating how useful the acquired feedback is when preserved over time in the form of user profiles. If the experimental system can generate results with higher precision in terms of result ordering, then it performs better.

For search systems, as mentioned several times in this thesis, system performance is often assessed in terms of search results, by its ability to push relevant documents in the lower ranks (see definition 2.9). Thus, to compare the performances of two systems - here, experimental and baseline systems - ranked lists of search results obtained by the user need to be considered for both systems. The one that is better able to *push* relevant documents to the top of the ranked lists of search results is the better. Table 5-5 gives the overall precision obtained at rank 5 and 10 of both systems. It is important to recall that precision is obtained by dividing the number of relevant documents - for each user - among top 5 or top 10 documents by 5 or 10 accordingly. Here, results to the first page (i.e. 10 documents) are considered (see Section 4.4.1.1).

Table 5-5 Document-Level Performance

Precision	Baseline System	Experimental Systems[22]			
		VSM	P (paired t-test)	PO	P (paired t-test)
System @ Rank 5	0.79	0.83	0.006%	1.00	0.005%
System @ Rank 10	0.56	0.75	0.50%	0.85	0.78%

From Table 5-5, the overall averages of the precision at rank 5 and at rank 10 for the experimental system when employing the PO approach, clearly indicate that out of 5 documents, the system can rank more than 4 documents based on their relevancy (1.00*5 = 4.70 and 0.85*5 = 4.25) to the query. While the performance of the system is more or less constant at rank 5 by employing the VSM approach, it is poorer at rank 10, since out of 5 documents, it can only rank 3.75 (0.75*5 = 3.75) documents. The worst performance can be observed from the baseline, as its overall averages of the precision at rank 5 and at rank 10 indicate that having 5 documents, the system is able to rank, based on their relevancy to the query, less than 4 documents (0.79*5 = 3.95) and less than 3 documents (0.56*5 = 2.80) respectively.

[22] Both the Keyword-based and Semantic-based approaches are represented by their corresponding modules (i.e. VSM and PO) in Table 5-5 due to limited space.

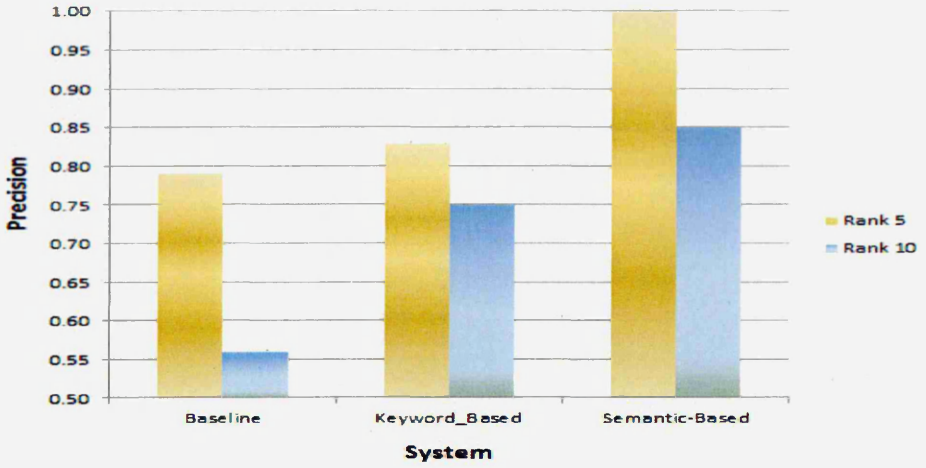

Figure 5-1 System Performance

Overall, the experiments showed that, personalised search outperforms the baseline with a statically significant (paired t-test) difference between them, which is consistent with the results obtained in the previous chapter when artificial queries (see Section 4.4.1.2) were employed.

5.2.2.2 Keyword-Based vs. Semantic-Based.

The goal of this experiment was to use the same idea with the same dataset to study whether Semantic-based approach (i.e. integrating Ontology terms) is superior to relying on the Keyword-based approach with regards to personalised search. Here, the quality of the personalised search was examined according to both condition 2 and condition 3 given above in order to establish whether the PO module in the user model indeed leads to a better understanding of the user's needs and preferences or not. Here, it should be recalled that, spreading mechanism was employed in PO module to incorporate the concept terms into the documents; therefore, PO is the VSM expanded with an integration of content semantics which are expressed in ontology terms so that an enriched user model is generated. This experiment will test the effects of combinations of keywords from the ontology terms with the keywords from the query to enrich the user model, in other words, to extend the VSM in order to generate ranked lists.

Each of the participant collections was thus indexed individually as document vector files and Figure 5-2 shows a representation of the distribution of document ITD indices (here, the values of interest vector - denoted as $s_u(t)$ - see Section 4.3.1.2) according to different combinations of the query

keywords[23] with its related concepts[24] mixtures. Here, $kxny$ means x keyword(s) and y concept(s) or ontology terms are employed in the user model. For example, $k2n2$ and $k2n3$ are respectively the keywords employed in the iterations in which two and three ontology terms are integrated into the user model for the second keyword of the query. The threshold interest vector values are the values represented by $kxn\theta$, meaning that only VSM is employed and no ontology terms have yet been added to the documents.

As can be seen from Figure 5-2, the PO layout showed the best results when a document is integrated with 3 and 4 keywords (at $kxn3$ and $kxn4$) regardless of the original number of terms (i.e. keywords) contained in the query. The presentation given here is related to only one query, but statistical evidence (ANOVA p value = 6.80%) indicated that out of 729 keyword queries, this observation is consistent across more than 650 keyword queries.

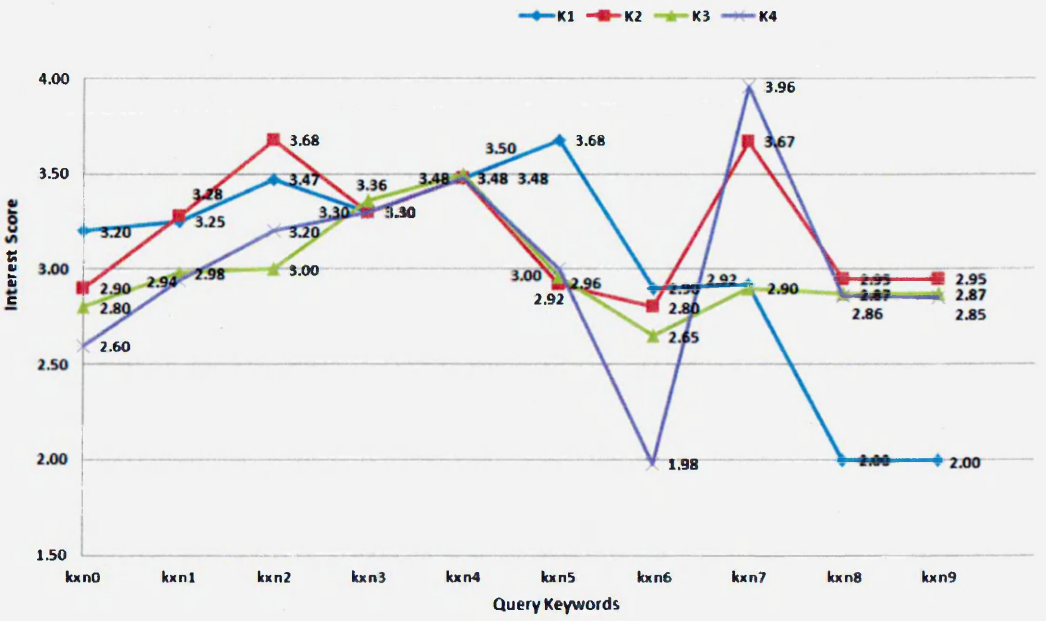

Figure 5-2 Comparisons of Mixtures of Query Keywords with Ontology Terms

However, expanding the document with one keyword or 2 keywords, 5 keywords and 7 keywords showed some slight improvements for most documents. On the other hand, integrating the document with 6 and 8 keywords

[23] According to Jansen, Spink and Saracevic (2000), on average, a query contains 2.21 terms.
[24] Thiagarajan, Manjunath and Stumptner (2008) demonstrated that the computation process of terms' weights turns monotonic after the third iteration, while in the current project, it turns monotonic after the document is expanded with the eighth term concept.

showed worse performance (represented by *kxn6* and *kxn8* in Figure 5-2), which might be due to the inclusion of keywords which are not related to the original term meaning. As was suggested in Section 4.2.3.2, a mechanism to control this noise is required.

Overall, VSM alone showed poorer performance than profile ontology. However, expanding the query by spreading the document with 3 or 4 keywords can improve the performance of the VSM.

5.3 Discussions

A series of two different web search experiments was performed using different keywords from real users. For each search session, a list of personalised webpage re-ranking over the search results returned by Google was generated. The evaluation metric parameter of precision was adopted to measure the ranking quality of the personalised search engine in order to determine the relevance of the results according to the order of relevance.

It is clear from the findings that personalised search gives better results in terms of higher precision. The findings have also shown, as was expected and claimed, that implicit feedback is experimentally reliable in producing personalised search results with high precision. It was also shown that semantic-based approaches improved the performance of IR systems in terms of precision (i.e. document relevance score); provided that the ontology terms added to documents are filtered from noise.

5.4 Summary

In this chapter, the fundamental concepts and algorithms underlying the proposed Personalised Search Engine (PSE) were presented. PSE is an offline semantic web personalisation solution which is based on the integration of users' navigational behaviour (i.e. usage data in an educational context) with content semantics which are expressed in ontology terms so that an enriched user model is generated.

Future search documents are expanded with ontology terms to enable semantic similarity between the terms of these documents to be performed. Such semantically enhanced representation allows overcoming of the problems emerging when pure keyword-based personalisation is performed, thus allowing

a rank list to be presented according to the decreasing order of documents' relevance. The claim that semantic enrichment of the personalisation process improves the quality of the ranked list in terms of complying with the users' needs is validated by experimental results with real users. A general observation made is that, out of the two models employed by the system, the ranked list generated after semantic expansion is more accurate, yet comparing the sets of results generated by the two models with the *hybrid* one, it can be concluded that this setup is the most advantageous compared to the VSM alone.

Chapter 6 TUNING THE SEARCH APPLICATION

The previous chapters have detailed the features that can be extracted from text to be shaped into good predictors that create search fields representing them. Those chapters pointed out that IR systems use textual implicit feedback to identify relevant documents for Web users. This chapter provides a document relevance tuning approach based on a non-text implicit feedback feature. Section 1 provides the preliminaries of the underlying search application, while Section 2 outlines the details related to the proposed ranking model. A relevance-focused personalised search is presented in Section 3. The details of the experimental results and discussions are given in Section 4. The conclusions are summed up in Section 5.

6.1 Introduction

It was emphasised earlier that before designing a search application for individual users, it is necessary to gather the information and requirements related to each individual user. Such information represents both the user's needs and interests and in turn relevant documents. Once the search application has been designed, it is normally deployed and monitored in order to decide if any improvement is required. This is often the systematic approach of any IR system whose focus is to build a relevance-focused search application.

To assess the value of the user modelling features for personalised search, two experimental studies were performed in the previous chapter to test different models of the proposed personalised system on a deeper level. It was observed that the combination of the keyword-based features with semantic-based features based on the integration of only 3 and 4 concept terms from ontology terms led to a better performance of the system. To address this relevance

feedback problem, the current study incorporates the dwell time feature in order to further tune the ranking function.

Few research investigations (Kelly and Belkin 2004, Agichtein, Brill and Dumais 2006, Collins-Thompson, et al. 2011, Hassan and White 2013) proved that the dwell time is a valuable implicit feedback feature (see Section 3.5). It is incorporated here by following the strategy of Xu et al. (2008) so that a relevance-focused search application is built. The aim is to balance the predictors of all concepts integrated into the combination of the above approaches. The problem statement is formulated similarly to the Kaggle and Yandex 2014 competition[25] which was framed thus:

"Participants need to personalise search using the long-term (user history based) and short-term (session-based) user context. The evaluation relies on a variant of a dwell-time based model of personal relevance and is data-driven, as it is presently accepted in the state-of the-art research on personalised search".

6.2 Dwell-Based Ranking Model

To weigh the degree of document relevance, implicit feedback information such as the frequency of query terms and behavioural signals derived from search log data were employed so that the PRM could determine a **document's** relevance matching users' information needs expressed in a query. The main objective is to re-rank the URLs[26] of the first three pages returned by the search engine according to users' personal preferences. This section describes the ranking technique through exploring users' dwell times in their previous clicked documents over individual documents, from which the user's dwell times ranking the personalised search results are inferred. They are then employed as a feature in the ranking function to provide the individual user's interest.

6.2.1 Designing Search Application

The fields representing the users' **individual interests** and preferences identified through $tf \bullet idf$ standard measures applied to the users' search logs and the dwell times of the corresponding clicked documents can be employed to infer relevant documents of the individual user (Xu, et al. 2008). A **user's dwell times**

[25] http://www.kaggle.com/c/yandex-personalised-web-search-challenge
[26] To re-rank the URLs is to produce the ranked list based on the similarity merge of the proposed models.

on new documents can be obtained by inferring the dwell times through the estimated relevant documents. In other words, the dwell time of a document can be employed and a dwell-based score can be iteratively calculated to represent the relevance of each document in order to improve the ranking function. For the purpose of the current project, the dwell time is defined as described below.

> ***Definition 6.1*** *Time stamp is defined as the time when the click occurred. Dwell time is the document visiting time. This has previously been defined as both the interval between the page being loaded and the searcher leaving the page (Hassan, Jones and Klinkner 2010), and as the basic indicator of document relevance (Guo and Agichtein 2012). It is calculated here by taking the time difference between two successive clicks as there is no direct means of measuring either of the above.*

The dwell time (referred to as dwell for simplicity and denoted as $dwell$) is thus a feature related to text context which can be used to make decisions related to that content as well. For each document clicked and saved $d_j \in D$, the dwell[27] is first manually calculated and converted into seconds to avoid confusion as a tokeniser. It was seen that tokens are considered a match, solely when they match exactly, and tokens obtained from the analysis process are the dominant features used with full-text search for matching a query with documents in the index.

To include the dwell information in the search application, since it is a non-textual data type, the standard tokeniser needs to be converted into a keyword tokeniser so that the two data types (i.e. the actual non-textual data type and the actual textual data type) can be seemingly combined in the ranking function. This means that the analysis strategy to be used must be able to handle both data types together so that the relevance can be controlled. Therefore, an analysis strategy, described below, was formulated around this specific principle of relevance which focuses on particular analysis features that are applied to a variety of relevance problems based on textual data types.

[27] Time range used in the current study is 30 seconds to 15 minutes - a commonly-used threshold is a dwell of at least 30 seconds (Hassan, Jones and Klinkner 2010, Guo and Agichtein 2012) and manual check of the dataset indicated the longest dwell to be 7.5 minutes.

Given a non-clicked document d_x, the dwell prediction can be determined based on the textual content similarity (Xu, et al. 2008) as shown below. Assuming that two documents (i.e. their contents) are sufficiently similar, then the interest scores of these documents for user u_k will be more or less the same. Therefore, using previously clicked documents, the textual similarity between these documents and k documents (fixed as $k=30$ in the current work to consider the first three pages) can be employed to determine the dwells of the k documents. The dwell of a document d_x can be determined based on the dwell of the clicked document $d_j \in D$ having the highest similarity between d_x and set of clicked documents $D=\{d_1,d_2,..d_j...d_n\}$ (see Section 4.2.1). Without loss of generality, the d_x dwell can thus be calculated using equation 6.1 (Xu, et al. 2008). by selecting the dwell of the clicked document with the $d_j \in D$

$$dwell(d_x)=\frac{\sum_{i=1}^{k}(dwell(d_j)Sim(d_j,d_x)\delta(d_j,d_x))}{\sum_{i=1}^{k}(Sim(d_j,d_x)\delta(d_j,d_x))+\varepsilon} \quad (6.1)$$

where ε is a small positive number to avoid the divide-by-zero error and $\delta(,)$ is the function that filters out the effects of those documents whose similarity is below or above the threshold, which is defined as per equation 6.2:

$$\delta(d_i,d_x)=\begin{cases}1 & if\ Sim(d_i,d_x)>0.01\\ 0 & otherwise\end{cases} \quad (6.2)$$

To determine text similarity estimation in equation 6.1, the current study employs the binary cosine similarity measure to maintain consistency as it proved to work better in all the experiments. Therefore, the parameter used in Xu, et al.'s equation (2008) employing Tanimoto approach (Strehl and Ghosh 2000) as an extended Jaccard method to control how the similarity of the text contributes to the estimation of dwell is omitted.

In order to combine textual and non-textual data types, a filter pattern, referred to as dwell-tf-idf that captures the dwell as input while preserving the original token, was defined. Such a filter allows the analysis to be shaped to capture the most important features of the field's data, namely the meaning, so as to model the user's intent. This analysis was organised in such a way that it allows the meaning of the dwell to be delimited from text by a numeric value which is face

value in real numbers so that its meaning is in the number that composes it. It is important to note here that the dwell-value fields are dealt with rather than a free text that happens to contain a dwell-value. In this analysis, a custom string tokeniser allows the string input results line to be automatically separated from the number input results line by parsing the line as defined in the internal field separator rather than working on space as a normal tokeniser. Thus, the keyword tokeniser actually creates a single token containing the entire text of the field without changing it. Now textual and non-textual data types can be combined since the tokeniser is a keyword rather than standard.

> ***Definition 6.2*** *dwell-tf-Idf is a dwell-based filter that allows a standard tokeniser to be converted into a keyword tokeniser while creating a token containing the entire unaltered text of the field related to the dwell value.*

However, since each document's length is different from the other (see Section 2.2.3.2), taking this length into account is crucial in order to avoid the dependence of the dwell on the document's length to cause a longer dwell to be induced. Therefore, similar to document length normalisation, the dwell length normalisation needs to be factored into an IR system to address the bias of such long dwells.

6.3 Relevance-Focused Personalised Search

In some IR as in the current one, the relevance score from the query is finally allowed to interact with the numerical values produced by the relevance functions by multiplying together each function and the resulting value is then multiplied with the score of the main query as shown in equation 6.3 (Doug and John 2016). This behaviour could moreover be modified so that - an additional multiplicative (i.e. additive boost) based on either the numerical value of a field or the numerical value of a function based on other values from several more fields is also applied - the function values are summed together before finally being multiplied with the query score as shown in equation 6.4 (Doug and John 2016) in order to harness the search components together.

$$\text{TotalScore} = (\text{Relevance Score} \times \text{Interest Score}) \times \text{Dwell Score} \quad (6.3)$$

$$\text{TotalScore} = (\text{Relevance Score} + \text{Interest Score}) \times \text{Dwell Score} \quad (6.4)$$

In the current study, each dwell score obtained from the dwell-tf-idf field is thus employed to rank the documents in decreasing order of the dwell weight for the top-k list of the search results. To test its assertion several different experiments will be conducted, of which the dwell score will be employed

(1) directly as a parameter controlling the degree of personalisation to calculate the overall score $gf(x)$ for a document x by using equation 5.2 whereby p represents the dwell time score and $g_k(x)$ and $g_o(x)$ are respectively the combined relevance and interest scores from the VSM and the PO models (see Section 5.1.4). To normalise this dwell, when this implementation is employed, it is converted into its thousandths form and it takes a value between 0 and 1. Thus for example, 20 seconds is equal to 0.020; similarly, 3 minutes is equal to (3*60)/1000. It is also important to note that when p takes a value of 0, the rank based on the PO module (i.e. semantic-based) is not given a weight and the rank based on the VSM module (i.e. keyword-based) remains the main rank; whereas the rank based on the PO module might be the main rank if it happens that p is 1; and

(2) as an implicit feedback feature employed in the ranking function. To provide a relevance-focused personalised search, equation 6.3 is employed whereby the dwell-based score is incorporated into the PO model (based on both its relevance and interest scores so that the sematic features are included) in order to tune the ranking function to address the relevance feedback problem (see definition 6.3).

> ***Definition 6.3*** *A ranked list is defined as a list of retrieval documents provided by a modern search engine. These documents are ranked based on their relevance scores computed by the system. The relevance scores are often merged with interest scores which might both result from different models of the system (Doug and John 2016) to obtain one single score for the final rank. The decreasing order of this single score represents the ranked list.*

6.4 Experiments

Now that all the components could be merged together, the results of the pilot study were replicated in the main study. That is, the same dataset was used

again and the strategy of Järvelin and Kekäläinen (2000) was adopted in order to observe the ranking of documents of the proposed system with a fully-fledged version of the search application. Three sets of experiments were conducted by adopting the CombSUM (see Section 4.3.1.4) to combine the ranking scores provided by both the VSM and the PO models. Here, the dwell score is employed in the first two experiments as a parameter controlling the degree of personalisation (equation 5.2) to calculate the final scores of the corresponding models.

(1) The first set of experiments presents the NDCG obtained for both the experimental system (personalised - PSE) and non-personalised system (baseline - GSE). The results of this set of experiments are presented in Table 6-1 and graphically represented in Figure 6-1.

(2) The second set of experiments reports the evaluation metric values of precision, recall and F-measure calculated for 3 different queries based on both the experimental system (personalised) and non-personalised system. The results of this set of experiments are presented in Table 6-2 and graphically represented in Figure 6-2.

(3) In this set of experiments - referred to as PRM Pre-dwell in Table 6-3, the calculated overall average of the Mean Average Precision (MAP), precision, recall and F-measure based solely on the combination of only the VSM and PO models, excluding the dwell-based model, is presented. The results of this set of experiments are shown in Table 6-3 for comparison purposes.

6.4.1 Experiment Procedure

The experiment procedure is the same used in the pilot study with the same dataset, except that it included the system evaluation based on 5 evaluators among the 50 users who volunteered to provide a score of the links they clicked using the personalised system. These evaluators were requested to score four statements using a scale of 0 to 3 where 3 is relevant, 2 is partially relevant, 1 is somewhat relevant and 0 is irrelevant (Al-Sharji, Beer and Uruchurtu 2013). The score provided by each evaluator for 15 keywords was used as the ground truth; and the first experiment was conducted whereby the NDCG scores of the search results were calculated accordingly for all 15 keywords.

The comparison results clearly show a significant difference between the NDCG scores of the two systems. As can be seen in Table 6-1, for the personalised search engine, a total of 13 keywords have an NDCG score between 0.90 and 0.60 and only two keywords have an NDCG score below 0.60. For the non-personalised search engine on the other hand, only 6 keywords have attained an NDCG score above 0.60 with a maximum score of 0.78 and more keywords (i.e. 9 keywords) are observed to have an NDCG score which is below 0.60.

Table 6-1 NDCG Scores of 15 Example Keyword Search Results

	NDCG			NDCG			NDCG	
Keyword No	PSE	GSE	Keyword No	PSE	GSE	Keyword No	PSE	GSE
1	0.73	0.60	6	0.90	0.78	11	0.64	0.47
2	0.61	0.41	7	0.73	0.56	12	0.78	0.58
3	0.83	0.62	8	0.59	0.48	13	0.76	0.66
4	0.63	0.45	9	0.51	0.37	14	0.84	0.58
5	0.61	0.43	10	0.76	0.63	15	0.81	0.68

Figure 6-1 Ranking Order of Relevance based on NDCG: Personalised Vs. Google

In this set of experiments, the precision, recall and F-Measure of all keywords were calculated. Table 6-2 and Figure 6-2 are one of the best examples selected here for presentation, as the same queries were issued by all 5 evaluators (Al-Sharji, Beer and Uruchurtu 2013). However, the overall average of precision, recall and F-Measure is consistent with these results. As can also be seen in Table 6-2, the average of two of the measures, the precision and the F-measure is higher for the personalised technique in all 3 queries, which shows that the proposed technique achieved better results than the Google search by producing a higher number of relevant Web search results for the 3 queries.

Table 6-2 Values of the Average Precision, Average Recall and Average F-Measure of Search Results of 3 Queries

		PSE	GSE
Query 1	Avg Precision	0.80	0.58
	Avg Recall	1.00	1.00
	Avg F-Measure	0.89	0.73
Query 2	Avg Precision	0.82	0.68
	Avg Recall	1.00	1.00
	Avg F-Measure	0.90	0.81
Query 3	Avg Precision	0.80	0.60
	Avg Recall	1.00	1.00
	Avg F-Measure	0.89	0.75

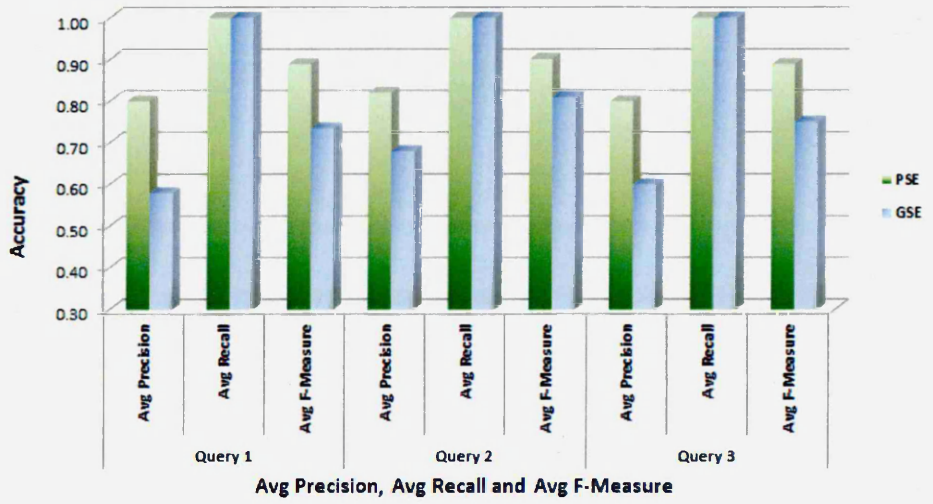

Figure 6-2 Chart Showing Average Precision, Average Recall and Average F-measure for Query 1, Query 2 and Query 3

From Figure 6-2, it can also be seen that compared to Google search, the PSE obtained better average precision and average F-measure values. While the average precision is 0.80, 0.82 and 0.80 for queries 1 to 3 respectively in the PSE, it is about 19% higher than that in the non-personalised search engine which is 0.58, 0.68 and 0.60 for queries 1 to 3 respectively. The F-measure's average is 0.89, 0.90 and 0.89 for queries1 to 3 respectively in the PSE, which is about 14% higher than the average F-measure in the non-personalised search engine which produced 0.73, 0.81 and 0.75 for queries 1 to 3 respectively. The better precision and F-measure show the effectiveness of the PSE compared to GSE.

6.4.1.1 Lesson Learned

The results obtained were consistent with the previous results conducted on individual models before merging the three models together. Although the lab-controlled experiments made the data collection easier, it was observed that users had somehow suffered from mundane realism. It was speculated that the contrived environments would result in producing similar queries to the ones presented in the findings. However, data collection generated by participants conducting typical day to day navigation will be used next before generalising the results of these findings.

6.4.2 Results Validation

A fourth set of experiments was conducted using a data set which originates from day to day rich search logs from non-commercial consenting users, who all work at the Educational Technology Centre (ETC) at the Nizwa College of Technology (NCT) in Oman. This data sample covered the period from 23rd of December, 2012 to 10th of January, 2013. The complete data sample was collected over a period of three weeks (Elsweiler and Ruthven 2007, Hu, et al. 2012, Hassan and White 2013) and it included 1004 clicked URLs with their respective query keywords and time stamps. Participants were instructed to use the system installed in their machines for the information search any time from 07:30am to 04:30pm in order to allow their browsing histories to be saved. However, the collection method for some participants was changed at their request (as they were undergoing examination, they preferred not to be distracted by the researcher), and a more convenient method was agreed with them. Therefore, their browsing histories captured were recorded (copied) by the participants themselves to a pre-defined folder in the network drive. These browsing histories of individual participant could then be collated for analysis by the researcher using a password for each individual user.

Before starting the analysis, the dwell time of each clicked URL was manually calculated. To calculate the dwell time corresponding to the first and the last clicked URL, the average length of the dwells of a session is used. It is worth noting that the final URL is the most important, as it often indicates the searcher's satisfaction. Useful data was obtained from 979 clicked URLs, and with this primary data, a set of two experiments was conducted with dwell

functionality using the dwell as an implicit feedback feature (see Section 6.3). The results are presented in Table 6-3.

Although the collection was not confined to a laboratory and the respondents were not separated into two groups, the experiment conducted in this project can still be scrutinised as a controlled experiment since the reliability of the results was increased with the PRM as an independent variable to test the rank of its returned results. The PRM was installed on each participant's machine and the collection of the search logs related to the GSE started two days later. Some days all the participants were instructed to use the PRM and some days they were required to use the GSE. The idea of collecting evaluation feedback in this way was inspired by Ageev et al. (2011), although they designed their work in the form of a game. The participants were instructed to select the PRM on the first 2 days, then to select the GSE for the next 5 days, and again to use the PRM for the last two weeks.

Since the PRM needs to use implicit feedback to learn users' interests in order to re-rank the search results according to individual user's needs, the PRM is not expected to produce a better performance than the GSE if used first - which forms the null hypothesis of this experiment. There was a total of three weeks of data collection, out of which 322 clicked URLs were obtained from GSE and 657 clicked URLs from PRM. These were the results of 157 different keywords issued by 48 participants including at least 2 predefined keywords (see APPENDIX A) per participant. It is important to note that the findings reported by Kelly and Belkin (2004) involved a sample of 7 subjects, and thus could not be reliably generalised to larger populations. However, the range and quality of the data collected in their survey involving those 7 subjects provided abundant insight into personalisation research. Moreover, a Web-based personal information management survey presented by Elsweiler and Ruthven (2007), involved 36 participants for 3 weeks. With 48 respondents involved in this study, this sample size is even more valid, especially in terms of search log data as opposed to a study related to a large-scale evaluation involving human labellers who are not query owners.

It is worth recalling that participants were asked to score four statements (Al-Sharji, Beer and Uruchurtu 2015); and to do the same for at least two keywords from the predefined queries. The data related to the predefined keywords was collected to allow for consistency of the judgment of the participants to be measured (see Section 4.4.1). The average agreement between the participants was 66.78%, which indicates that the participants' evaluations were reliable. The extent to which the returned results matched the participants' search interests and preferences based on the overall average of the Mean Average Precision (MAP), precision, recall, and F-Measure of the corresponding keywords for both GSE and the proposed PRM was determined based on this evaluation feedback. The obtained results are provided in Table 6-3.

Table 6-3 Dwell-based Experiments: MAP, Precision, Recall and F-Measure values

	MAP	**Precision**					**Recall**	**F-Measure**
		@1	@3	@5	@10	@20		
*PRM Pre-dwell	0.85	0.85	0.89	0.88	0.84	0.77	0.64	0.72
Post-dwell (day 1-2)	0.74	0.63	0.71	0.73	0.82	0.80	0.54	0.63
GSE (Day 3-9)	0.74	0.63	0.71	0.73	0.81	0.80	0.55	0.63
PRM Post-dwell (Day 10-21)	0.88	0.98	0.96	0.95	0.87	0.80	0.69	0.78

A number of interesting observations could be drawn from these results. Firstly, a MAP falling between 0.74 (74%) and 0.88 (88%) in the retrospective experiments indicates a difference in search engine performance - the proposed PRM achieved better retrieval effectiveness (14% improvement), which is consistent with the previous findings. Secondly, it can be seen from the same table that all F-Measures of PRM are higher than those of the GSE; this indicates an improvement in the accuracy of the PRM. In the lower ranks, the precision and recall of the PRM are higher than those of the GSE, which indicates that the PRM returns not only more relevant results than GSE but also most of the relevant results. It can be seen that the PRM significantly improves precision @1, but it improves less significantly for precision @3 to precision @20 indicating that PRM can be more effective in lower ranks than higher ranks.

Finally, these differences are also noticeable when the PRM pre-dwell is compared against the PRM post-dwell which is consistent with the findings in Hassan, Jones and Klinkner (2010) and Xu, Jiang and Lau (2011). Considering the PRM post-dwell results on days 10-21 and using the null hypothesis that the

PRM does not produce a better performance than the GSE, the null hypothesis was rejected. This is also clear from the post-dwell results on day 1-2 of the experiment (see Table 6-3), which indicate that the null hypothesis holds, since the same results from the GSE and PRM (apart from a difference of 0.01 - 1% - in both precision @10 and average recall which might be due to feedback error when users abandon the clicks) were obtained, an interesting property that shows the reliability of the ranking technique employed in the PRM. The values obtained for precision @1-@5 are constants to an extent for the PRM (i.e. days 10-21), which indicates that it provides static results (i.e. based on users' model feedback) for individual users unlike the Google search engine which somehow fails to provide constant values even at the top-k ranks.

6.4.2.1 Lesson Learned

The findings of these experiments assess the relationship between the results obtained when both different data sets were used. It was observed that implicit feedback increases the performance of IR in all experiments. This is inconsistent with the researcher's expectations that contrived environments would result in similar queries to those presented in the findings. In fact, the results indicated that users still tend to issue the queries based on their long-term individual interests even though they might sometimes relocate their short-term interests. After relocating, the users still begin to adhere to their long-term interests, evidence that implicit feedback is valuable in identifying a user's information needs while crafting the ranking function.

The findings of the last set of experiments allowed the researcher to validate the effectiveness of incorporating into the ranking process the proposed *Dwell-tf-idf* scheme since they showed a significant improvement in the search results within the top-k rank.

6.5 Summary

This chapter has presented a technique for developing a relevance-focused personalised search application using dwell time features to (1) identify individual users' interests and preferences and extract search information from web pages based on individual users' needs; 2) employ the dwell-keyword search as the representation of individual users' search information using a customised pre-processor analyser called *Dwell-tf-idf*; 3) use the dwell-keyword

as a feature to develop a reliable predictor function to determine the relevance ranking for ranked lists characterised by coverage and accuracy. The findings were demonstrated to be effective enough to provide a personalised ranking model with a low level of non-relevant information (high precision) while displaying most of the relevant information (high recall) thus allowing users to retrieve *immediately* and *exactly* what they need based on their queries.

Two different sets of data collected from real life search logs returned by the GSE from consenting participants were employed. These enabled the researcher to identify users' individual preferences using the implicit feedback acquired through their profiles (built by employing documents previously clicked) in order to re-rank the top-k list of search results. The effectiveness of the proposed PRM was validated through a controlled experiment which revealed a 14% improvement in the overall average MAP in the top-k rank, interestingly consistent with the overall average F-Measure (14% improvement) based on a second dataset which further empirically validated the proposed search application.

Chapter 7 Conclusions

This dissertation has presented a novel Personalised Ranking Model to design and implement a reliable search application using the VSM technique. Two main features including textual (i.e. keyword-based approach) and non-textual (i.e. dwell-based approach) were laid out to identify individual users' interests and preferences and extract search information from web pages based on individual users' needs. The keyword-based approach was extended in an attempt to debug the relevance problem and expand the query so that semantic-based features (i.e. through profile ontology) are included to address the problems of words' independence and provide a solution to polysemy and synonymy.

Before conducting a user study, a pre-evaluation experiment was performed to assess the reliability of the features employed in attempting to provide predictor functions to determine the relevance ranking. This led to the conclusion that the proposed PSE could improve the performance of the ranked lists. A pilot study experiment was then conducted to validate the findings of the pre-evaluation experiment. The experiment assessed both approaches (i.e. models) which attempted to construct the PSE with semantic-based approach and without (i.e. VSM). The experiment results led to the conclusion that not all the keywords integrated into the documents could generate user models which provided better results, while some concepts were not related to the user's query and did not lead to improvements in the system.

A dwell-based search was included in an attempt to further fine-tune a relevance-focused search application for each individual user in order to improve the degree of personalisation. As compared to any traditional search system, it was empirically demonstrated that by using human subjects

conducting real-life searches on the Web, the strength of the PSE lies in pushing significantly relevant documents to the top of the lists (i.e. lower ranks) of the search results.

A number of research questions were posed in their corresponding chapters in order to raise awareness, challenges and suggest future research directions in the area of IR in general and the personalisation of information in particular.

7.1 Main Findings

The results of the main study were presented in two interrelated stages although they were not theoretically distinguished: (1) system performance and (2) user satisfaction. As stated in the analysis, compared to the GSE (used as a baseline personalised search system), the proposed PSE was able to return document sets with higher precision. The results of the pilot study were replicated in the main study. That is, the same dataset was used again and the ranking of relevant documents was observed again.

Compared to GSE, based on the system with higher performance, more relevant documents were pushed to the top of the ranked lists thereby allowing individual users to obtain *immediately* (i.e. achieving coverage) and *exactly* (i.e. achieving accuracy) what they need when they enter their queries and interact with the system. The findings also demonstrated that employing the users' judgement of the system led to higher precision and recall which subsequently demonstrates higher levels of user satisfaction.

The effect of the PO-based user modelling showed up differently; the ranked lists were more precise than the ranked lists generated by the VSM-based model alone with almost the same level of system coverage and accuracy. This difference between the two models can be interpreted as contributing to different areas of advanced IR such as coverage of semantic-based information and accuracy of detecting information.

Despite all these positive results for the proposed PSE, it is not easy to recommend that users abandon GSE, and negative reactions towards change always exist. Of course, Google is very popular and even big search engines still cannot compete with it, let alone a complementary tool to address some shortcomings of the ranked lists-based systems.

7.2 Limitations and Delimitations

The analytical and empirical superiority of the retrieval effectiveness of the proposed personalisation search which employs users' profiles through keyword-based search techniques expanded with semantic-based search techniques has been demonstrated. As with many other personalisation approaches, the current project suffers from a number of limitations including, but not limited to:

Boosted Ranking Function: the learning mechanism used in the current project does not consider parameter weights, for instance giving more weight to other fields such as headings in the HTML, incorporating temporal information which might also yield further substantial improvements.

A Concept Filtering Mechanism: the key problem in using the spreading technique to enrich a document with concepts related to the keywords query is that it does not identify exactly the appropriate concepts that may match with it. For instance, concepts are integrated straightforwardly into the documents after some similarity measures are employed to select ontology terms with highest similarities in the iteration process. This approach might affect the precision or recall, or both. Therefore, it might be very advantageous to consider a mechanism for the appropriate selection of ontology terms features to be mapped to the Web documents that match the users' exact information needs.

Stemming Mechanism: stemming is an important process of analysis in IR whose main role is to ensure that the appropriate level of detail of the meaning of each word is correctly captured. Unfortunately, many research studies including the current one simply adopt the stemming algorithms of Porter (1997) or Krovetz (1993) and ignore the effect of these algorithms in personalisation search. For example users might want the information '*Omani*' relating to a citizen of Oman rather than to the country Oman. The Porter (1997) stemming algorithm employed here simply normalises all the words to their root, which of course does not return the result '*Omani citizen*'. Therefore, Named-entities (Bier, Ishak and Chi 2006) can be employed to discriminate, say between a country and a person.

7.3 Future Work

The crux of the current innovation is the deployment of an expanded VSM with both an ontology-based model and dwell-based model that assists users in obtaining information which is characterised by both high precision and high recall. Although these techniques were mainly used to compute similarity between documents and users' profiles, they can be generalised to all other content matching scenarios. The author is confident that the implementation of the developed project is completely feasible and its fundamental conceptual framework is viable from a technical standpoint.

Any future work, therefore, will be in the direction of expanding the algorithms to focus on encoding semantics into the documents to describe their content rather than solely integrating the ontology terms into the documents to create enriched users' models and users' profiles. This is a hot topic in the area of query expansion which was recommended as an open research question (see Section 4.2.3.2). Also, it is worth considering incorporating more novel approaches into the system. This includes designing a more intuitive interface for enhanced searching and implementing more flexible integration modes (i.e. using ElastiSearch and Solr (Doug and John 2016)) while combining queries and the document model.

In order to yield personalised search results, the proposed framework, as in many other personalised search approaches or artificial intelligence systems, tends to rely solely on algorithms without involving the users' efforts; however, it might be more beneficial to incorporate into the IR problem, smart user interfaces using exploratory search approaches. This can allow users to view the system's internal calculation of the relevance so that they can explicitly feedback to the system their specific viewpoint of relevance. It is important to clarify that this internal computation needs to be displayed with a short document summary and its corresponding keyword highlighted to help the user make the right decision. Thus, during the interaction between the two, the system can also be helped by the user, to bring together the *best-of-the-two-worlds* in order to obtain more accurate search results.

This project could further be expanded based on a number of interesting open research questions posed as stated below, which might be equally beneficial to other researchers for further exploration:

1. With reference to inverted index (see Section 3.1.1.1) data structure, identifying an optimisation technique to store as little information as possible for limiting storage like a unique identifier;
2. With reference to query expansion approach (see Section 3.1.5), an optimisation technique for the weight of the original query terms with the new terms to provide effective pseudo feedback;
3. With reference to integrating ontology terms into a document (see Section 4.2.3.2), devising a mechanism to control and correct the query expansion matching the users' information needs thereby guaranteeing that recall is improved during the phase without degrading precision as a result of this process;
4. With reference to inferring a dwell document (see Section 4.2.3.2), devising an optimisation technique to control the dependence of the dwell on the document's length causing an increase of dwell score of those documents.

7.4 Concluding Remarks

Despite the fact that algorithm-based personalised search is a hot topic for research in IR and continues to be so, most of the existing projects have concentrated on keyword-based searches (i.e. tf-idf), thus making personalisation search an under-studied area.

Certainly, any theoretical understanding of retrieval can be improved based on useful schemes grounded in learning approaches. They could ultimately lead to an insight into and improved accuracy of document ranking that might have implications for designing a personalised search in order to improve the user's search experience. Unfortunately, the adoption of other learning approaches has been hindered by the paucity of theoretical evaluation. Collecting implicit information from real users in a controlled laboratory experiment could be one solution to addressing this problem as their behavioural data would enable a fuller examination of relevance based on individual users' needs, and in consequence, derive a proper theoretical understanding of retrieval.

REFERENCES

Ageev Mikhail, Guo Qi, Lagun Dmitry, Agichtein Eugene (2011). Find it if you can: a Game for Modeling Different Types of Web Search Success Using Interaction Data. In: *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp. 345-354.

Agichtein Eugene, Brill Eric and Dumais Susan (2006). Improving Web Search Ranking by Incorporating User Behavior Information. In: *Proceedings of the 29thAnnual International ACM SIGIR Conference on Research and Development in Information Retrieval,* pp.19-26.

Ahn Jae-Wook, Brusilovsky Peter, He Daqing, Grady Jonathan, and Li Qi (2008). Personalized web exploration with task models. In *Proceedings of the 17th international conference on World Wide Web*, ACM, pp. 1-10.

Allen J., Aslam Jay, Belkin Nicholas, Buckley Chris, Callan Jamie, Croft W. B., Dumais Sue & Zhai, B. C (2003): Challenges in Information Retrieval and Language Modeling. In: *SIGIR Forum*, Vol. 37(1), ACM Press, pp. 31-47.

Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth (2013). Enhancing the Degree of Personalisation Through Vector Space Model and Profile Ontology. *In: Computing and Communication Technologies, Research, Innovation, and Vision for the Future* (RIVF), IEEE RIVF 2013 *International Conference*, IEEE, pp. 248-252.

Al-Sharji Safiya, Beer Martin and Uruchurtu Elizabeth (2015). A Dwell Time-Based Technique for Personalised Ranking Model, *In: Database and Expert Systems Applications,* Springer International Publishing, pp. 205-214.

Axler Sheldon Jay (1997). *Linear Algebra Done Right*. Vol. 2. Heidelberg: Springer International Publishing.

Baader Franz, Horrocks Ian and Sattler Ulrike (2009). Description Logics. In: *Handbook on Ontologies. Staab Steffen and Rudi Studer (Ed.), International Handbooks on Information Systems.* Springer Berlin Heidelberg, pp. 21-43.

Baeza-Yates Ricardo and Ribeiro-Neto Berthier (1999). *Modern information retrieval.* Vol. 463, ACM Press New York.

Barrett Rob, Maglio Paul P and Kellem Daniel C. (1997). How to Personalize the Web. In: *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 75-82.

Belkin Nicholas J. and Croft W. Bruce (1992). Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM, Vol.* 35(12), pp. 29-38.

Berners-Lee Tim, Fischetti Mark and Foreword By-Dertouzos Michael L (2000). *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its Inventor.* Harper Information.

Bier Eric A., Ishak Edward W. and Chi Ed (2006). Entity Workspace: an Evidence File that Aids Memory, Inference and Reading. *Intelligence and Security Informatics*, Springer, pp. 466-472.

Blair David C. and Maron Melvin E (1985). An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System. *Communications of the ACM*, Vol. 28(3), pp. 289-299.

Brost WN (1997). PhD Dissertation: Construction of Engineering Ontologies, Center for Telematica and Information Technology, University of Twenty, Enscheda.

Buckley Chris (1993). The Importance of Proper Weighting Methods. In: *Proceedings of the Workshop on Human Language Technology,* Association for Computational Linguistics, pp. 349-352.

Cantador Iván, Miriam Fernández, David Vallet, Pablo Castells, Jérôme Picault, and Myriam Ribiere (2008). A Multi-Purpose Ontology-Based Approach for Personalised Content Filtering and Retrieval. In: *Advances in Semantic Media Adaptation and Personalization*, Springer Berlin Heidelberg, pp. 25-51.

Cassel Lillian N and Ursula Wolz (2001). Client Side Personalization. In: *Proceedings of the 2nd DELOS Network of Excellence Workshop on Personalisation and Recommender Systems in Digital Libraries,* CiteSeer, *http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.24.4013&rep=rep1&type=pdf*

Castells Pablo, Fernandez Miriam and Vallet David (2007). An Adaptation of the Vector-Space Model for Ontology-Based Information Retrieval. In: *IEEE Transactions on Knowledge and Data Engineering,* Vol. 19 (2), pp. 261-272.

Castells Pablo, Fernández Miriam, Vallet David, Mylonas Phivos, and Avrithis Yannis (2005) Self-tuning Personalized Information Retrieval in an Ontology-Based Framework. In: *On the Move to Meaningful Internet Systems: OTM 2005 Workshops*, Springer Berlin Heidelberg, pp. 977-986.

Chen Liren and Sycara Katia (1998). WebMate: A Personal Agent for Browsing and Searching. *In: Proceedings of the 2nd International Conference on Autonomous Agents, ACM,* pp.132-139.

Cleverdon Cyril W, Mills Jack and Keen EM (1966). Factors Determining the Performance of Indexing Systems, Vol. (1): Design, Cranfield College of Aeronautics.

Collins-Thompson Kevyn, Bennett Paul N., White Ryen W., De La Chica Sebastian and Sontag David (2011). Personalizing Web Search Results by Reading Level. In: *Proceedings of the 20th International Conference on Information and Knowledge Management*, ACM, pp. 403-412.

Crestani Fabio (1997). Application of Spreading Activation Techniques in Information Retrieval. In: *Artificial Intelligence Review*, Springer International Publishing, Vol. 11(6): pp. 453-482.

Croft W. Bruce (1986). User-Specified Domain Knowledge for Document Retrieval. *In: Proceedings of the 9th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 201-206.

Croft W. Bruce and Harper David J (1979). Using Probabilistic Models of Document Retrieval without Relevance Information. *Journal of Documentation,* Vol. 35 (4), pp. 285-295.

Croft W. Bruce, Cronen-Townsend Stephen and Lavrenko Victor (2001). Relevance Feedback and Personalization: A Language Modeling Perspective. In: *DELOS Workshop: Personalisation and Recommender Systems in Digital Libraries,* Vol. 3, pp. 13-19.

Crouch Carolyn J and Yang Bokyung (1992). Experiments in Automatic Statistical Thesaurus Construction. In: *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp.77-88.

Da Costa Pereira Célia and Tettamanzi Andrea GB (2006). An Ontology-Based Method for User Model Acquisition. In: *Soft Computing in Ontologies and Semantic Web*. Springer, pp. 211-229.

Dai Honghua Kathy, and Mobasher Bamshad (2000). Using Ontologies to Discover Domain-Level Web Usage Profiles. In*: Proceedings of the Technologies: The EC-Web 2000 Conference, Lecture Notes in Computer Science* (LNCS) 1875, Springer, pp. 165–176.

Devi M. Uma and Gandhi G. Meera (2015). An Enhanced Ontology Based Measure of Similarity between Words and Semantic Similarity Search. In: Emerging ICT for Bridging the Future, Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Vol. 1, Springer International Publishing, pp. 443-454.

Dicheva Dirina and Aroyo Lora (2000). An Approach to Intelligent Information Handling in Web_Based Learning Environments. In: *Proceedings of International Conference on Artificial Intelligence*, CSREA Press, pp. 1327–1333.

Dou Zhicheng, Song Ruihua and Wen Ji-Rong (2007). A Large-Scale Evaluation and Analysis of Personalized Search Strategies. In: *Proceedings of the 16th International Conference on World Wide Web,* ACM, pp. 581-590.

Doug Tumbull and John Berryman (2016). Relevant Search with Applications for Solr and Elasticsearch. Manning 2016.

Elsweiler David and Ruthven Ian (2007). Towards Task-Based Personal Information Management Evaluations. In: *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp. 23-30.

Fox Edward A. and Shaw Joseph A. (1994). Combination of Multiple Searches. In NIST special publication SP. pp. 243-243.

Fox Steve, Karnawat Kuldeep, Mydland Mark, Dumais Susan, and White Thomas (2005). Evaluating implicit measures to improve web search. *ACM Transactions on Information Systems (TOIS) 23 (2)*, ACM, pp. 147-168.

Fox, Edward A and Shaw Joseph A (1994). Combination of Multiple Searches NIST Special Publication SP, pp. 243-243.

Gabrilovich Evgeniy and Markovitch Shaul (2007). Computing Semantic Relatedness Using Wikipedia-Based Explicit Semantic Analysis. In: *International Joint Committee on Artificial Intelligence (IJCAI)*, pp. 1606-1611.

Gauch Susan, Chaffee Jason and Pretschner Alexander (2003). Ontology-based personalized search and browsing. In *International Journal of Web Intelligence and Agent Systems, Vol.* 1(3), pp. 219-234.

Giunchiglia Fausto, Kharkevich Uladzimir and Zaihrayeu Ilya (2009). Concept Search. *The Semantic Web: Research and Applications*. Springer Berlin Heidelberg, pp. 429-444.

Giunchiglia Fausto, Shvaiko Pavel and Yatskevich Mikalai (2006). Discovering Missing Background Knowledge in Ontology Matching. In: *ECAI*, Vol. 141, pp. 382-386.

Greengrass Ed (2000). Information Retrieval: A Survey. University of Maryland, Baltimore County, Technical Report Tr-R52-008-001. *Html.-Fulltext at Http://Www. CSEE. Umbc. Edu/Cadip/Readings/Ir.*

Greiff Warren R (1998). A Theory of Term Weighting Based on Exploratory Data Analysis. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp. 11-19.

Guarino Nicola, Masolo Claudio and Vetere Guido (1999). OntoSeek: Content-Based Access to the Web. *Intelligent Systems and their Applications, IEEE*, Vol. 14 (3), pp. 70-80.

Guo Q. and Agichtein E (2012). Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-Click Searcher Behaviour. In: *Proceedings of the 21st International Conference on World Wide Web*, ACM, pp. 569-578.

Haav Hele-Mai and Lubi Tanel-Lauri (2001). A Survey of Concept-Based Information Retrieval Tools on the Web. In: *Proceedings of the 5th East-European Conference ADBIS,* pp. 29-41.

Hassan Ahmed and White Ryen W (2013). Personalized Models of Search Satisfaction. In: *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management,* ACM, pp. 2009-2018.

Hassan Awadallah, Jones Rosie and Klinkner Kristina Lisa (2010). Beyond DCG: user behaviour as a predictor of a successful search, *Proceedings of the third ACM international conference on Web search and data mining*, ACM. pp. 221-230.

Henry Lieberman (1995): Letizia: An agent that assists web browsing. In: *International Joint Committee on Artificial Intelligence (IJCAI)'s Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann*, IJCAI, Vol. 8, pp. 924–929.

Hu Yunhua, Qian Yanan, Li Hang, Jiang Daxin, Pei Jian and Zheng Qinghua (2012). Mining Query Subtopics from Search Log Data. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp. 305-314.

James Pitkow; Hinrich Schütze; Todd Cass; Rob Cooley; Don Turnbull; Edmonds Andy; Adar Eytan and Thomas Breuel (2002). Personalised Search. *In: Communications of the ACM*, Vol. 45(9), pp.50-55.

Jameson Anthony (2008). Adaptive Interfaces and Agents. *In: Human-Computer Interaction: Fundamentals, Evolving Technologies* and *Emerging Applications*, Sears & J. A. Jacko (Eds.), (2nd Ed.), Boca Raton, FL: CRC Press, pp. 433–458.

Järvelin Kalervo, and Kekäläinen Jaana (2000). IR Evaluation Methods for Retrieving Highly Relevant Documents. In: *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp. 41-48.

Joachims Thorsten, Granka Laura, Pan Bing, Hembrooke Helene, and Gay Geri (2005). Accurately Interpreting Clickthrough Data as Implicit Feedback. In: *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.154-161.

Jones Karen Sparck and Willet Peter (1997). Overall Introduction. In: Jones Karen Sparck and Willet Peter (Ed.), Readings in Information Retrieval, Vol. 24 (5), pp. 1-8.

Jones Karen Sparck, Walker Steve and Robertson Stephen E (2000). A Probabilistic Model of Information Retrieval: Development and Comparative Experiments: Part 2. *In: International Journal of Information Processing & Management,* Vol. 36 (6), pp. 809-840.

Kage Tomoyo and Sumiya Kazutoshi (2006). A Web Search Method Based on the Temporal Relation of Query Keywords. *Web Information Systems–WISE 2006.* Springer Berlin Heidelberg, pp. 4-15.

Kantardzic, Mehmed (2011). Data Mining: Concepts, Models, Methods, and Algorithms. Hoboken, N.J, John Wiley.

Keane Mark T., O'Brien Maeve and Smyth Barry (2008). Are People Biased in their Use of Search Engines? In: Communications of the ACM, Vol. 51(2), pp. 49-52.

Keen Michael E (1981). Laboratory Tests of Manual Systems. In: *Karen Sparck Jones (Ed.). Information Retrieval Experiment*. Butterworths Publisher, CiteSeer, Vol.1, pp. 136-155.

Kelly Diane and Belkin Nicholas J (2004). Display Time as Implicit Feedback: Understanding Task Effects. In: *Proceedings of the 27th Annual International SIGIR conference on Research and Development in Information Retrieval*, ACM, pp. 377-384.

Kelly Diane and Teevan Jaime (2003). Implicit Feedback for Inferring User Preference: a Bibliography. In: *ACM SIGIR Forum*, Vol. 37(2), pp. 18-28.

Khan Latifur (1999). Structuring and Querying Personalized Audio Using Ontologies. In: *International Multimedia Conference, Proceedings of the 7th ACM international Conference on Multimedia (part 2)*, pp. 209-210.

Khan Latifur and Dennis McLeod (2000). Disambiguation of Annotated Text of Audio Using Onologies. In: Proceeding of ACM SIGKDD Workshop on Text Mining.

Knappe Rasmus (2006). PhD Dissertation: *Measures of Semantic Similarity and Relatedness for Use in Ontology-Based Information Retrieval.* Roskilde University, Denmark.

Krovetz Robert (1993). Viewing Morphology as an Inference Process. In: *Proceedings of the 16th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 191-202.

Kulp Scott and Kontostathis April (2007). On Retrieving Legal Files: Shortening Documents and Weeding Out Garbage. In: *TREC*.

Lamiroy Bart and Sun Tao (2011). Precision and Recall Without Ground Truth. In: *9th IAPR International Workshop on Graphics Recognition (GREC)*.

Leacock Claudia and Martin Chodorow (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In: *Christiane Fellbaum (Ed.), WordNet: An Electronic Lexical Database,* MIT Press, Vol. 49(2), pp. 265-283.

Lieberman Henry (1997). Autonomous interface agents. In *Proceedings Of The ACM SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 67-74.

Liu Chang, Belkin Nicholas J., and Cole Michael J (2012). Personalization of Search Results Using Interaction Behaviors in Search Sessions. In: *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, pp. 205-214.

Lu Kun (2013). An Insight into Vector Space Modeling and Language Modeling. In: *Proceedings of Iconference*, pp. 717-721.

Lucarella Dario and Morara R. (1991). First: Fuzzy Information Retrieval System. *In: Journal of Information Science*, Vol. 17(2), pp.81-91.

Luhn Hans Peter (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development, Vol.* 1 (4), pp. 309-317.

Luhn, Hans Peter (1958). The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development,* Vol. 2 (2), pp. 159-165.

Mabotuwana Thusitha, Lee Michael C. and Cohen-Solal Eric V. (2013). An Ontology-Based Similarity Measure for Biomedical Data–Application to Radiology Reports. *Journal of Biomedical Informatics, Vol.* 46 (5), pp. 857-868.

Manning Christopher D., Raghavan Prabhakar and Schütze Hinrich (2008). Introduction to Information Retrieval. Cambridge University Press, Cambridge.

Mao Ming (2007). Ontology Mapping: An Information Retrieval and Interactive Activation Network Based Approach. In: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC, Springer Berlin Heidelberg, pp. 931-935.

Maron Melvin E. and Kuhns John L (1960). On Relevance, Probabilistic Indexing and Information Retrieval. In: *Journal of the ACM (JACM),* Vol. 7 (3), pp. 216-244.

Matthijs Nicholas and Radlinski Filip (2011). Personalizing Web Search Using Long Term Browsing History. *In: Proceedings of the 4th ACM international conference on Web search and Data Mining*, pp. 25-34.

Metzler Donald and Zaragoza Hugo (2009). *Semi-Parametric and Non-Parametric Term Weighting for Information Retrieval.* Springer-Verlag Ed., LNCS 5766, pp. 42-53.

Micarelli Alessandro, Gasparetti Fabio, Sciarrone Filippo and Gauch Susan (2007). Personalized Search on the World Wide Web. In: *Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.)*. The Adaptive Web, Methods and Strategies of Web Personalization. Springer-Verlag Ed., LNCS 4321, pp. 195-230.

Micarelli Alessandro, Sciarrone Filippo and Marinilli Mauro (2007). Web Document Modeling. In: Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.). *The Adaptive Web, Methods And Strategies of Web Personalization*. Springer-Verlag Ed., LNCS 4321, pp. 155–192.

Middleton Stuart E., De Roure David C. and Shadbolt Nigel R (2001). Capturing Knowledge of User Preferences: Ontologies in Recommender Systems. In: *Proceedings of the 1st International Conference on Knowledge Capture*, ACM, pp. 100-107.

Miller George A (1995). WordNet: a Lexical Database for English. In: *Communications of the* ACM, Vol. 38(11), pp. 39-41.

Mobasher B (2007). Data Mining for Web Personalization. In: Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.). *The Adaptive Web: Methods and strategies of Web Personalization*. Springer-Verlag Ed., LNCS 4321, pp. 90–135.

Montebello Matthew, Gray W. A. and Hurley Stephen (1998). Evolvable Intelligent User Interface for WWW Knowledge-Based Systems. In*: Proceedings IDEAS'98 of Database Engineering and Applications Symposium*, IEEE, pp. 224-233.

Morris Dan, Ringel Morris Meredith and Venolia Gina (2008). Search Bar: A Search-Centric Web History for Task Resumption and Information Re-Finding. In: *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, pp. 1207-1216.

Mostafa Javed (2005). Seeking Better Web Searches. In: *Journal of the American society for Information Science*, Vol. 292(2), pp. 66-73.

Navigli Roberto (2009). Word Sense Disambiguation: A Survey. In: *ACM Computing Surveys (CSUR), Vol.*41 (2), Article 10.

Pazzani Michael J., and Billsus Daniel (2007) Content-Based Recommendation Systems. *In: P. Brusilovsky, A. Kobsa, and W. Nejdl (Eds.): The Adaptive Web, Methods and Strategies of Web Personalization.* Springer-Verlag Berlin Heidelberg, LNCS 4321, pp. 325-341.

Ponte Jay M. and Croft W. Bruce (1998). A Language modeling approach to Information Retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information Retrieval,* ACM, pp. 275-281.

Porter Martin F (1997). An algorithm for Suffix Stripping. In: Jones Karen Sparck and Willet Peter (Ed.), *Readings in Information Retrieval, Vol.* 24 (5), Morgan Kaufmann, pp. 313-316.

Qiu Feng, and Cho Junghoo (2006). Automatic Identification of User Interest For Personalized Search. In: *Proceedings of the 15th International Conference on World Wide Web*. ACM, pp.727-736.

Radecki Tadeusz (1988). Trends in research on Information Retrieval - The Potential for Improvements in Conventional Boolean Retrieval Systems. *In: International Journal of Information Processing and Management,* Vol. 24(3), pp. 219-227.

Radlinski Filip and Joachims Thorsten (2005). Query Chains: Learning to Rank from Implicit Feedback. In: *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, ACM, pp. 239-248.

Rijsbergen Cornelis J. V (1979). Information Retrieval. In: *Journal of the American Society for Information Science, Vol.* 30 (6), pp. 374-375.

Robertson Stephen E (1977). The Probability Ranking Principle. In: J*ournal of Documentation, Vol.* 33 (4), pp. 294-304.

Robertson Stephen E (1981). The Methodology of Information Retrieval Experiment. In: Karen Sparck Jones (Ed.), *Information Retrieval Experiment.* Butterworths Publisher, CiteSeer, Vol.1, pp. 9-31.

Robertson Stephen E. and Jones K. Sparck (1976). Relevance Weighting of Search Terms. *Journal of the American Society for Information science, Vol.* 27 (3), pp. 129-146.

Robertson Stephen E., Walker Steve and Hancock-Beaulieu Micheline M (1995). Large Test Collection Experiments on an Operational, Interactive System: Okapi at TREC. *In: International Journal of Information Processing & Management,* Vol. 31 (3), 345-360.

Robertson Stephen E., Walker Steve, Hancock-Beaulieu Micheline, Gull Aarron and Lau Marianna (1993). Okapi at TREC. In: *Text Retrieval Conference,* pp. 21-30.

Rocchio Joseph John (1971). Relevance Feedback in Information Retrieval. In: Gerald Salton (Ed.), the Smart Retrieval System: Experiments in Automatic Document Processing, Prentice Hall, pp. 313-323.

Sakagami Hidekazu and Kamba Tomonari (1997). Learning Personal Preferences on Online Newspaper Articles from User Behaviors. In: *Computer Networks and ISDN Systems*, Vol. 29(8-13): pp. 1447-1455.

Salton Gerard (1989). *Automatic Text Processing: The Transformation, Analysis and Retrieval of information by Computer*. Addison-Wesley Press.

Salton Gerard (1992). The State of Retrieval System Evaluation. *In: International Journal of Information Processing & Management,* Vol. 28 (4), pp. 441-449.

Salton Gerard and Buckley Christopher (1988). Term-Weighting Approaches in Automatic Text Retrieval. *In: International Journal of Information Processing & Management,* Vol. 24 (5), pp. 513-523.

Salton Gerard and Buckley Christopher (1997). Improving Retrieval Performance by Relevance Feedback. In: Jones Karen Sparck and Willet Peter (Ed.), *Readings in information retrieval, Vol.* 24 (5), Morgan Kaufmann, pp. 355-363.

Salton Gerard and McGill Michael J (1983). Introduction to Modern Information Retrieval.

Salton Gerard and Yang Chung-Shu (1973). On the Specification of Term Values in Automatic Indexing. *Journal of documentation,* Vol. 29 (4), pp. 351-372.

Salton Gerard, Wong Anita and Yang Chung-Shu (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM,* Vol. 18 (11), pp. 613-620.

Salton Gerard, Yang Chung-Shu and Yu Clement T (1975). A Theory of Term Importance in Automatic Text Analysis. *Journal of the American Society for Information Science,* Vol. 26 (1), pp. 33-44.

Sanderson Mark (2000). Retrieving with Good Sense. Information Retrieval, (White Rose Research for this Paper: Http://Eprints.Whiterose.AC.UK/4573/). Vol. 2(1), published in Elsnews, pp. 49-69.

Shaw Joseph A and Fox Edward A (1995). Combination of Multiple Searches *NIST Special Publication SP,* pp. 105-108.

Silverstein Craig, Marais Hannes, Henzinger Monika and Moricz Michael (1999). Analysis of a very Large Web Search Engine Query Log. *In: ACM SIGIR Forum,* Vol. 33(1), ACM, pp. 6-12.

Singhal Amit (2001). Modern Information Retrieval: A Brief Overview. *IEEE Data Engineering Bulletin,* Vol. 24 (4), pp. 35-43.

Singhal Amit, Buckley Chris and Mitra Mandar (1996). Pivoted Document Length Normalization. In: *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* ACM, pp. 21-29.

Singhal Amit, Salton Gerard, Mitra Mandar and Buckley Chris (1996). Document Length Normalization. *In: International Journal of Information Processing & Management,* Vol. 32 (5), pp. 619-633.

Song Ruihua, Liqian Yu, Ji-Rong Wen, and Hsiao-Wuen Hon (2011). A Proximity Probabilistic Model for Information Retrieval. Technical Report, Microsoft Research.

Staab Steffen and Rudi Studer (2013). In: *Handbook on Ontologies. Staab Steffen and Rudi Studer (Ed.), International Handbooks on Information Systems.* Springer Science & Business Media.

Stokoe Christopher, Oakes Michael P. and Tait John (2003). Word Sense Disambiguation in Information Retrieval Revisited. In: *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp.159-166.

Strehl Alexander and Ghosh Joydeep (2000). Value-Based Customer Grouping from Large Retail Data Sets. *AeroSense* International Society for Optics and Photonics, pp. 33-42.

Studer Rudi, Benjamins V. Richard and Fensel Dieter (1998). Knowledge Engineering: Principles and Methods. In: International Journal of *Data & Knowledge Engineering,* Vol. 25 (1), pp. 161-197.

Susan Gauch, Jason Chaffee and Alexander Pretschner (2004). Ontology Based User Profiles for Search and Browsing. In: *Web Intelligence and Agent Systems*, Vol. 1 (3), pp. 219-234.

Susan Gauch; Mirco Speretta; Aravind Chandramouli and Alessandro Micarelli (2007). User Profiles for Personalized Information Access. In: Peter Brusilovski, Alfred Kobsa, Wolfgang Nejdl (Ed.). *The Adaptive Web*: *Methods and Strategies of Web Personalization*. Springer-Verlag Ed., LNCS 4321, pp. 54-90.

Tan Bin, Shen Xuehua and Zhai Chengxiang (2006). Mining Long-Term Search History to Improve Search Accuracy. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* ACM, pp. 718-723.

Teevan J., Dumais, S.T. and Horvitz E (2010). Potential for Personalization. *ACM Transactions on Computer-Human Interaction*, Vol. 17(1), Article 4.

Teevan Jaime, Dumais Susan T, and Horvitz Eric (2007). Characterizing the value of personalizing search. *Proceedings of the 30th annual International ACM SIGIR conference on research and development in information retrieval*. ACM, pp. 757-758.

Teevan Jaime, Dumais Susan T. and Horvitz Eric (2005a). Beyond the Commons: Investigating the Value of Personalizing Web Search. In: Proceedings of the Workshop on New Technologies for Personalized Information Access (PIA), pp. 84-92.

Teevan Jaime, Dumais Susan T. and Horvitz Eric (2005b). Personalizing Search Via Automated Analysis of Interests and Activities. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 449-456.

Thiagarajan Rajesh, Manjunath Geetha, and Stumptner Markus (2008). Computing Semantic Similarity Using Ontologies. Hewlett-Packard (HP) Development Company, L.P, Labs Technical Report HPL-2008-87.

Tsatsaronis George, Vazirgiannis Michalis and Androutsopoulos Ion (2007). Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri. *In: International Joint Committee on Artificial Intelligence (IJCAI), Vol. 7*, pp. 1725-1730.

Vallet David, Fernández Miriam and Castells Pablo (2005). An Ontology-Based Information Retrieval Model. In: *The Semantic Web: Research and Applications.* Springer, pp. 455-470.

Varelas Giannis, Voutsakis Epimenidis, Raftopoulou Paraskevi, Petrakis Euripides G.M. and Milios Evangelos E. (2005). Semantic Similarity Methods in WordNet and their Application to Information Retrieval on the Web. In Proceedings of the 7th annual ACM International Workshop on Web Information and Data Management, ACM, pp. 10-16.

Voorhees Ellen M (1994). Query Expansion Using Lexical-Semantic Relations. *In SIGIR'94*, Springer London, pp. 61-69.

Voorhees Ellen M and Harman Donna (2000). Overview of the Sixth Text Retrieval Conference (TREC-6). *In: International Journal of Information Processing & Management,* Vol. 36(1), pp. 3-35.

Walker Steve, Robertson Stephen E., Boughanem Mohand, Jones Gareth JF, and Jones Karen Sparck (1997). Okapi at TREC-6 Automatic ad hoc, VLC, Routing, Filtering and QSDR. In: *TREC,* CiteSeer, pp. 125-136.

Wang Hongning, He Xiaodong, Chang Ming-Wei,Song Yang,White Ryen W., and Chu Wei Chu (2013). Personalized Ranking Model Adaptation for Web Search. In: P*roceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, pp. 323-332.

Wang Qihua and Jin Hongxia Jin (2010). Exploring Online Social Activities for Adaptive Search Personalization. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, pp. 999-1008.

Wang Ying, Liu Weiru and Bell David A. (2010). A Structure-Based Similarity Spreading Approach for Ontology Matching. In: Scalable Uncertainty Management. Lecture Notes in Artificial Intelligence (LNAI), Amol Deshpande and Anthony Hunter (Eds), Springer, pp. 361-374.

White Ryen (2004). PhD Dissertation: Implicit Feedback for Interactive Information Retrieval. 2nd Ed., University of Glasgow.

White Ryen W., Ruthven Ian, and Jose Joemon M (2002). The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. In: *Advances in Information Retrieval*, Springer Berlin Heidelberg, pp. 93-109.

Xu Songhua, Jiang Hao and Lau Francis (2011). Mining User Dwell Time for Personalised Web Search Re-Ranking. In: *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, Vol. 3, AAAI Press, pp. 2367-2372.

Xu Songhua, Zhu Yi, Jiang Hao, and Lau Francis CM (2008) A User-Oriented Webpage Ranking Algorithm Based on User Attention Time. In *AAAI*, Vol. 8, pp.1255-1260.

Yu Clement T, Lam K and Salton Gerard (1982). Term Weighting in Information Retrieval Using the Term Precision Model. In: *Journal of the ACM (JACM),* Vol. 29 (1), pp. 152-170.

Yuwono Budi and Lee Dik L (1996). : Search and Ranking Algorithms for Locating Resources on the World Wide Web. In: *Proceedings of the 12th International Conference on Data Engineering,* IEEE, pp. 164-171.

Zadeh, Lotfi Aliasker (1965). Fuzzy Sets. In: I*nternational Journal of Information and Control*, Vol. 8(3), Elsevier, pp. 338-353.

Zhai Chengxiang and Lafferty John (2001). A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval. In: *Proceedings of the 24th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, pp. 334-342.

Zhai Chengxiang and Lafferty John (2004). A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS), Vol.* 22 (2), pp. 179-214.

Zipf George Kingsley (1949). Human Behavior and the Principle of Least Effort. Addison-Wesley Press.

Zobel Justin and Moffat Alistair (1998). Exploring the Similarity Space. In: *ACM SIGIR Forum,* ACM, pp. 18-34.

Zobel Justin and Moffat Alistair (2006). Inverted Files for Text Search Engines. In: *ACM Computing Surveys (CSUR), Vol.* 38(2), Article 6.

APPENDIX A

A list of the 6 predefined Topics (artificial queries) and their 18 corresponding relevant Documents meeting 3 users' needs (last accessed for this purpose on 10th, May, 2012). The order of enter for each topic is 1, 2 and 3. For simulation of users' profiles, the folders related to the documents given by the links below - crossed - are the ones that were deleted.

Query	Link
Old Oman	1) ~~https://en.wikipedia.org/wiki/History_of_Oman~~ 2) http://www.omanisilver.com/contents/en-us/d20_Oman_old_photos.html 3)http://www.alrahalah.com/2011/05/ahmad-ibn-majid-15th-century-ce-%E2%80%93-9th-century-ah-the-lion-of-the-seas/
Jupiter Facts	1) ~~http://theplanets.org/jupiter/~~ 2) http://solarsystem.nasa.gov/planets/jupiter/basic 3) http://www.space.com/7- -largest-planet-solar-system.html
Insomnia	1) ~~http://www.webmd.com/sleep-disorders/guide/insomnia-symptoms-and-causes~~ 2) http://www.nhs.uk/conditions/insomnia/pages/introduction.aspx 3) http://www.counselling.cam.ac.uk/selfhelp/leaflets/insomnia
Global Warming	1) ~~https://www.dosomething.org/us/facts/11-facts-about-global-warming~~ 2) http://www.conserve-energy-future.com/various-global-warming-facts.php 3)http://www.wwf.org.uk/what_we_do/tackling_climate_change/climate_change_explained/?
USA Wars	1) ~~http://www.loonwatch.com/2011/12/we-re-at-war-and-we-have-been-since-1776/~~ 2) https://en.wikipedia.org/wiki/Timeline_of_United_States_military_operations 3) http://www.infoplease.com/ipa/A0931831.html
Prophet Mohammad	1) ~~http://www.religionfacts.com/muhammad~~ 2) http://www.islam-guide.com/ch3-8.htm 3) http://www.islamreligion.com/articles/2626/who-is-prophet-muhammad/

APPENDIX B

This appendix presents both the information sheet and the consent form related to experiment documents in Section 5.2 and Section 6.4.1. These include:

B.1: Ethical Approval Form

B.2: Information Sheet (for the First Data Collection)

B.3: Consent Form

Sheffield Hallam University shuspace

ETHICAL APPROVAL FORM

ETHICAL APPROVAL (please tick):

√ *(Standard approval)* This project does not require specific ethical approval.

☐ *(Category approval)* In my opinion this work falls within the category of ..... projects which has been previously approved by the FREC and it does not therefore need individual approval.

☐ *(Approval awaited)* This project must be referred to the FREC for individual consideration – the work must not proceed unless and until the FREC gives approval.

I can confirm that I have read the Sheffield Hallam University Research Ethics Policy and Procedures document and agree to abide by its principles (please tick) ☐√

Signed ..... Name Safiya Al Sharji ..... Date 27/09/12
Student / Researcher/ Principal Investigator (as applicable)

Signed ..... Name ..... Date 27/9/12
Supervisor or other person giving ethical sign-off

Note: University Research Ethics policy available from the following web link:
http://www.shu.ac.uk/research/researchhallam.html

Sheffield Hallam University shuspace

Information Sheet for Research Project Title

Providing Personalised Information Based on Individual Interests and Preferences.

Researcher: Al Sharji (Telephone Number: 99889849)

1. What is the project's purpose?

The objective of the research is to study the experience of online information-seeking to provide an enhanced personalised Web search to users. The research is intended to identify the users' background knowledge of the topics of interest to them, and to capture this to build users' profiles which in turn can be used to match searchers' needs with their interests and provide information in a personalised manner. The goal is to tailor Web searching to the needs of individual users and provide them with personalised and adapted information based on their interests and demands in order to improve the efficiency of Web searches. To do so, it is necessary to analyse the browsing pattern of individual users over multiple searches.

2. Why have I been chosen?

In this research, data collection is through browsing observation. You have been selected on the basis of your positive response to the invitation e-mail which was sent earlier and the fact that you are currently undertaking searches for information on the Web.

3. Do I have to take part?

Taking part in the research is entirely voluntary and any refusal to agree to participate will involve no penalty or loss of benefits. You may also discontinue participation at any time of the observation. If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

4. What will happen to me if I take part?

The information collected on each participant when he/she is browsing the Web, will be kept in his/her user's profile. The information captured will only be used for academic purposes and it is guaranteed that it will remain anonymous. It will never be possible to personally identify any participant. Individual information (such as name, gender, age and so on) will not be revealed under any circumstances. This is completely up to you. Your records will only be used in ways that you agree to. For instance, the following points for which your consent is needed have to be indicated to you:

- In any use of your profile, your personal information will not be identified.
- Any anonymised profile can be studied, transcribed and analysed by the researcher only for the research aims.
- The anonymised profiles can be used for scientific publications and/or meetings.
- The anonymised profiles can be shown in presentations to scientific or non-scientific groups.

Please be assured that confidentiality is highly protected by the current survey. Any transcribed observations will be kept in the University data archive with no identifying information. The personal information collected about you (via email once you sign the consent form) is only for the purpose of discerning patterns in the data collected, but it will never be used to identify you personally. The data collected will be anonymised before being kept in the University data archive and accessed only by the researcher and the supervisors of this research and will never be made available to other parties or be made public.

5. What do I have to do?

If you consent to the information on this sheet, you are kindly requested to sign the consent form. Please be ensured that you can withdraw at any time even after signing the consent form.

6. What are the possible disadvantages and risks of taking part?

There will be no possible disadvantages or risks whatsoever from participating in this study. Even though the study will be on all your browsing histories, the main purpose is to improve the efficiency of surfing the Web for information seeking. The survey is not biased towards any particular kind of information surfed by the user. The main focus of this study is on the online searching experience and not about the specific queries you have searched for. As participation is voluntary, you may choose to discontinue it at any time.

7. Will my taking part in this project be kept confidential?

This research complies with the Data Protection Act 1998. All the information that is collected about you during the course of the research will be kept strictly confidential. You will never be identified in any reports or publications.

8. What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?

There will be some general details about you (i.e. name, age, gender and education), that are required to be asked for the purpose of discerning patterns in the data collected.

9. What will happen to the results of the research project?

The results of this study will be published mainly in a PhD thesis and may also be published in academic papers. All participants in this research will not be personally identified in any of these publications.

10. Who has ethically reviewed the project?

This project has been ethically approved via Sheffield Hallam University's ethics review procedure.

11. What are the possible benefits of taking part?

Whilst there are no immediate benefits, i.e. monetary benefit, for those people participating in the project, it is hoped that this work will help to provide first hand evidence of the current situation on online information seeking in the educational context and will help to produce a methodology that may be advantageous for this group of online users.

12. What happens if the research study stops earlier than expected?

If the research had to stop for some unexpected or accidental reason, all the information you contributed to this research would be destroyed or managed by the researcher's supervisors, and would not affect you in any way.

13. Who is organising and funding the research?

This is a post-graduate research funded by the Ministry of Manpower in Oman with the aim of fulfilling the requirements of the PhD in Information Retrieval at Sheffield Hallam University, UK.

14. What if something goes wrong?

If you have any enquiries or complaints about any aspects of this research please use the following information to e-mail the researcher or the research supervisors:

Name:	Safiya Al Sharji (researcher)
E-mail:	Safiya.M.Sharji@student.shu.ac.uk
Address:	Communication & Computing Research Institute, Science Park, Unit 12 Sheffield Hallam University,
Tel:	+44 (114) 225 6283 (current mobile number 99889849)

Name:	Dr. Martin Beer (Supervisor)
E-mail:	M.Beer@shu.ac.uk
Address:	Room 9404, Communication & Computing Research Institute, Sheffield Hallam University
Tel:	+44 (114) 225 6917

Name:	Dr. Uruchurtu Elizabeth (Supervisor)
E-mail:	E.Uruchurtu@shu.ac.uk
Address:	Room 9226, Communication & Computing Research Institute, Sheffield Hallam University
Tel:	+44 (114) 225 6939

Should you feel that your complaint is not being dealt with satisfactorily, you may also contact the Communication & Computing Research Institute, Secretary's Office, Sheffield Hallam University, Arundel Street, S1 2NU, Sheffield, UK, Tel: +44 (114) 225 6741.

Please keep this information sheet, it is your copy.

Lunch time: 12:30-14:30. Location: New Building, Al Canteen (Room N 345). You will be required to show your signed consent form.

Thank you for your participation.

Sheffield Hallam University shuspace

Consent Form for Research Project Title

Providing Personalised Information Based on Individual Interests and Preferences.

Researcher: Al Sharji (Telephone Number: 99889849)

		Please tick Box
1	I confirm that I have volunteered to take part in the above research study.	☐
2	I confirm that I have read and understood the information sheet related to the above research study and I have been given the opportunity to ask questions.	☐
3	I understand that this permission is voluntary and that I am free to withdraw at any time, without giving any reason. In such a case, it will not affect my legal rights.	☐
4	I understand that my anonymised profile can be transcribed, studied, analysed, used for scientific publications or meetings and can be shown to scientific or non-scientific groups.	☐
5	I wish to receive a summary sheet of the experimental findings.	☐

Name of Participant	Signature
Researcher	Signature

APPENDIX C

This appendix presents both the information sheet and the consent form related to experiment documents in Section 6.4.2. These include:

C.1: Information Sheet (for the Second Data Collection)

C.2: Consent Form

Sheffield Hallam University shuspace

Information Sheet for Research Project Title

Providing Personalised Information Based on Individual Interests and Preferences.

Researcher: Al Sharji (Telephone Number: 99889849)

1. What is the project's purpose?

The objective of the research is to study the experience of online information-seeking to provide an enhanced personalised Web search to users. The research is intended to identify the users' background knowledge of the topics of interest to them, and to capture this to build users' profiles which in turn can be used to match searchers' needs with their interests and provide information in a personalised manner. The goal is to tailor Web searching to the needs of individual users and provide them with personalised and adapted information based on their interests and demands in order to improve the efficiency of Web searches. To do so, it is necessary to analyse the browsing pattern of individual users over multiple searches.

2. Why have I been chosen?

In this research, data collection is through browsing observation. You have been selected on the basis of your positive response to the invitation e-mail which was sent earlier and the fact that you are currently undertaking searches for information on the Web.

3. Do I have to take part?

Taking part in the research is entirely voluntary and any refusal to agree to participate will involve no penalty or loss of benefits. You may also discontinue participation at any time of the observation. If you decide to take part you will be given this information sheet to keep and be asked to sign a consent form.

4. What will happen to me if I take part?

The information collected on each participant when he/she is browsing the Web, will be kept in his/her user's profile. The information captured will only be used for academic purposes and it is guaranteed that it will remain anonymous. It will never be possible to personally identify any participant. Individual information (such as name, gender, age and so on) will not be revealed under any circumstances. This is completely up to you. Your records will only be used in ways that you agree to. For instance, the following points for which your consent is needed have to be indicated to you:

- In any use of your profile, your personal information will not be identified.
- Any anonymised profile can be studied, transcribed and analysed by the researcher only for the research aims.
- The anonymised profiles can be used for scientific publications and/or meetings.
- The anonymised profiles can be shown in presentations to scientific or non-scientific groups.

Please be assured that confidentiality is highly protected by the current survey. Any transcribed observations will be kept in the University data archive with no identifying information. The personal information collected about you (in the attached form) is only for the purpose of discerning patterns in the data collected, but it will never be used to identify you personally. The data that will be collected will be anonymised before being kept in the University data archive and accessed only by the researcher and the supervisors of this research and will never be made available for other parties or be made public.

5. What do I have to do?

If you consent to the information on this sheet, you are kindly requested to sign a consent form. Please be ensured that you can withdraw at any time even after signing the consent form. If you choose to transfer the profile captured by the installed system by yourself to the pre-arranged folder so that it can be collated for the analysis, please indicate this on the consent form so that your request is fully taken into account.

6. What are the possible disadvantages and risks of taking part?

There will be no possible disadvantages or risks whatsoever from participating in this study. Even though the study will be on all your browsing histories, the main purpose is to improve the efficiency of surfing the Web for information seeking. The survey is not biased towards any particular kind of information surfed by the user. The main focus of this study is on the online searching experience and not about the specific queries you have searched for. As participation is voluntary, you may choose to discontinue it at any time.

7. Will my taking part in this project be kept confidential?

This research complies with the Data Protection Act 1998. All the information that is collected about you during the course of the research will be kept strictly confidential. You will never be identified in any reports or publications.

8. What type of information will be sought from me and why is the collection of this information relevant for achieving the research project's objectives?

There will be some general details about you (i.e. name, age, gender and education), that are required to be asked for the purpose of discerning patterns in the data collected.

9. What will happen to the results of the research project?

The results of this study will be published mainly in a PhD thesis and may also be published in academic papers. All participants in this research will not be personally identified in any of these publications.

10. Who has ethically reviewed the project?

This project has been ethically approved via Sheffield Hallam University's ethics review procedure.

11. What are the possible benefits of taking part?

Whilst there are no immediate benefits, i.e. monetary benefit, for those people participating in the project, it is hoped that this work will help to provide first

hand evidence of the current situation on online information seeking in the educational context and will help to produce a methodology that may be advantageous for this group of online users.

12. What happens if the research study stops earlier than expected?

If the research had to stop for some unexpected or accidental reason, all the information you contributed to this research would be destroyed or managed by the researcher's supervisors, and would not affect you in any way.

13. Who is organising and funding the research?

This is a post-graduate research funded by the Ministry of Manpower in Oman with the aim of fulfilling the requirements of the PhD in Information Retrieval at Sheffield Hallam University, UK.

14. What if something goes wrong?

If you have any enquiries or complaints about any aspects of this research please use the following information to e-mail the researcher or the research supervisors:

Name:	Safiya Al Sharji (researcher)
E-mail:	Safiya.M.Sharji@student.shu.ac.uk
Address:	Communication & Computing Research Institute, Science Park, Unit 12 Sheffield Hallam University,
Tel:	+44 (114) 225 6283 (current mobile number 99889849)

Name:	Dr. Martin Beer (Supervisor)
E-mail:	M.Beer@shu.ac.uk
Address:	Room 9404, Communication & Computing Research Institute, Sheffield Hallam University
Tel:	+44 (114) 225 6917

Name:	Dr. Uruchurtu Elizabeth (Supervisor)
E-mail:	E.Uruchurtu@shu.ac.uk
Address:	Room 9226, Communication & Computing Research Institute, Sheffield Hallam University
Tel:	+44 (114) 225 6939

Should you feel that your complaint is not being dealt with satisfactorily, you may also contact, the Communication & Computing Research Institute, Secretary's Office, Sheffield Hallam University, Arundel Street, S1 2NU, Sheffield, UK, Tel: +44 (114) 225 6741

Please keep this information sheet, it is your copy.

Thank you for your participation.

Sheffield Hallam University shuspace

Consent Form for Research Project Title

Providing Personalised Information Based on Individual Interests and Preferences.

Researcher: Al Sharji (Telephone Number: 99889849)

		Please tick Box
1	I confirm that I have volunteered to take part in the above research study.	☐
2	I confirm that I have read and understood the information sheet related to the above research study and I have been given the opportunity to ask questions.	☐
3	I understand that this permission is voluntary and that I am free to withdraw at any time, without giving any reason. In such a case, it will not affect my legal rights.	☐
4	I understand that my anonymised profile can be transcribed, studied, analysed, used for scientific publications or meetings and can be shown to scientific or non-scientific groups.	☐
5	I wish to receive a summary sheet of the experimental findings.	☐
6	I understood what data will be collected for my profile, and I opt to copy the profile into the appropriate folder by myself.	☐

Name of Participant	Signature
Researcher	Signature