Implicit Feedback System For The Recommendation Of Relevant Web Documents

Akuma, Stephen Shiaondo Cyril

Award date:
2016

Awarding institution:
Coventry University

IMPLICIT FEEDBACK SYSTEM FOR THE RECOMMENDATION OF RELEVANT WEB DOCUMENTS

STEPHEN S. C. AKUMA

A thesis submitted in partial fulfilment of the University's requirements for the Degree of Doctor of Philosophy

March 2016

Faculty of Engineering and Computing

COVENTRY UNIVERSITY

Abstract

The web is constantly updated with information, leading to the problem of information overload. The generic information retrieval systems retrieve a relatively large amount of irrelevant web resources, forcing information seekers to spend a significant amount of time searching for their information needs. Learners of a particular domain are usually faced with common problems and they visit similar sources of information when searching for their information needs.

The key idea of this research is to capture and record previous web documents visited by a homogeneous group of learners with their associated user behaviour which is inferred from the reading time, mouse and key activity, and to utilise this information to optimise the recommendation of relevant documents to users.

Several user studies were conducted to investigate relevance feedback parameters. An investigation was carried out to examine the relationship between user-generated implicit indicators and user perception of relevance. Thirteen users were given fifteen web documents to read and rate according to their perception of relevance based on given tasks. The results show a positive co-relationship between explicit relevance feedback such as user ratings and implicit relevance feedback such as reading time.

The second study was focused on user searching behaviour and it builds on the results of the preliminary study. A plugin in Firefox browser was used to capture and log several implicit relevance feedback indicators, explicit ratings of document familiarity, difficulty and relevance from 77 users. A number of implicit relevance feedback indicators were correlated with user explicit relevance feedback such as ratings. A predictive function model was developed based on the captured implicit and explicit relevance feedback. The effect of task type, document familiarity and

document difficulty on user behaviour was also examined. The predictive function model was validated through an eye gaze study and standard evaluation metrics. The results of the eye gaze study indicate that the predictive model derived from implicit indicators can be used in place of an eye gaze.

Furthermore, a prototype system for domain-specific implicit feedback was developed and evaluated. The results show that supplementing user queries with implicit feedback considerably improves the relevancy of returned results from a domain-specific search engine.

To the Blessed Trinity (God the father, God the Son and God the Holy Spirit)

Acknowledgements

I would first like to thank the Nigerian State who through her parastatal, Tertiary Education Trust Fund (TETFund), provided the financial aid for this academic journey. Let me also thank Benue State University, Makurdi for facilitating the process for the financial support.

My Director of study, Dr Rahat Iqbal, my first supervisor Prof. Chrisina Jayne and my second supervisor, Dr Faiyaz Doctor, provided me with the clear direction, guidance, constructive feedback and encouragement throughout my study period. I am sincerely grateful for their unquantified impact on my overall graduate experience. Dr Rahat Iqbal introduced me to this area of research when he supervised my Masters dissertation, and he assisted me with all the necessary support and advice needed for this programme.

I would like to thank my siblings and all my friends for their encouragement throughout this academic journey. Let me thank in a special way my friends on Facebook who supported me with their virtual presence during the periods I felt lonely and worn out. I would also like to thank all the students who participated in the user studies without which the investigation would not be completed.

Let me finally thank my parents who shaped my life into what I am now. They encouraged and supported me emotionally throughout this programme. One of the reasons for my doggedness despite some challenges during this programme is to put smiles on your faces and make you a proud dad and mum.

Table of contents

Table of Figures

Table of Tables

Published work

This thesis includes research which has been published in AIAI, EANN conference proceedings and Computers in Human Behaviour Journal. I am the primary author of the publications. The papers are:

- Akuma, S., Iqbal, R., Jayne, C. and Doctor, F. "Comparative analysis of relevance feedback methods based on two user studies", Computers in Human Behavior 60 (2016) 138 -146

- Akuma, S., Jayne, C., Iqbal, R., and Doctor, F. (2015) "Inferring Users' Interest on Web Documents Through Their Implicit Behaviour", 16th International Conference, EANN 2015, Rhodes, Greece, September 25-28 2015.Proceedings

- Akuma, S., Jayne, C., Iqbal, R., and Doctor, F. (eds.) (2014) 10th IFIP WG 12.5 International Conference, AIAI 2014. 'Implicit Predictive Indicators: Mouse Activity and Dwell Time'. held September 19-21 at Rhodes, Greece

I

Introduction

CHAPTER 1

1 Introduction

1.1 Introduction

There has been a rapid growth in computer technology in the last decades. As this growth in technology advances, the complexity of organising and retrieving information increases. The present challenge in the field of Information Retrieval (IR) is to assist users to find relevant documents for their information needs. This challenge is because an enormous amount of information is updated on the web and a user might be interested in only a single document in the vast collection of web resources (Alhindi et al. 2015). Users are interested in documents that are relevant to their current task. One of the ways users acquire information in the vast information space is through the use of a search engine. Users' problem statements are normally represented in a search engine through queries, but users have little or no training on how to formulate effective queries. Importantly, due to the generic nature of the search engines, the user queries do not adequately represent their problem statement, making it difficult for them to retrieve relevant information. Few studies report that up to 62% of searchers do not find satisfaction with the generic search tools (Balakrishnan and Zhang 2014, Delphi-Group 2004, Iqbal et al. 2015). Due to the gap in information feedback process in the traditional search engines like Google (Núñez-Valdez et al. 2015), efforts have been made to augment users' queries with implicit and explicit feedback parameters obtained from their

interaction with an information retrieval system (Iqbal et al. 2015). This involves capturing users' interest in a particular domain through their interaction with their browsers.

To assist users to retrieve relevant web resources in a specific context for their current needs, personalisation is found useful (Grzywaczewski et al. 2013, Iqbal et al. 2015, Zemirli 2012). Personalisation enables dynamic injecting of web resources, arranging the web resources to users' satisfaction and suggesting contents in a way that will be relevant to them (Romero et al. 2009). It can be achieved through users' implicit behavioural characteristics or explicit suggestions, or both (Xu, Jiang and Lau 2010). Personalisation considers either users' subjective perspective or common behaviour exhibited by a group of users through the use of contextual tools. When such tools are used to link users to their web experience, it is referred to as web personalisation (Mobasher, Cooley and Srivastava 2000). In E-commerce, personalisation enables sellers to suggest products to buyers based on their demographics as well as buying habits. Similarly, in an education setting, personalisation of relevant web resources according to learners' need is closely related to Adaptive Hypermedia where hyperlinks are recommended to learners based on their previous interaction with the web. The core foundation for personalisation of information services is user modelling (Huai 2011). Through user modelling, a concise description of the user and their interests is obtained (Huai 2011). Both implicit and explicit relevance feedback approaches are used to develop user profiles. Implicit personalisation involves profiling users' interest through observing their behaviour as they interact with a system while explicit personalisation involves profiling users by requesting some specific information from them, which is normally intrusive and does not pay off shorter term (Balakrishnan and Zhang 2014, Claypool et al. 2001, Iqbal et al. 2015).

In order to use a non-intrusive approach to personalised information retrieval, an in-depth analysis of user web behaviour and their preferences needs to be carried out. It involves interpreting user behaviour and estimating the relevance of web documents viewed by users (Kelly and Teevan 2003, Kumar and Ashraf 2015). This technique uses implicit indicators (mouse movement, mouse click, time spent, scroll movement, keystroke and so on) to estimate users' interest on web documents (Jawaheer, Weller and Kostkova 2014, Leiva and Huang 2015, Pasi 2014). Although attempts have been made by researchers to use eye gaze as an implicit indicator to measure users' interest (Buscher et al. 2012a, Guo and Agichtein 2010, Gwizdka 2014, Lopez et al. 2015), it is yet to be applied in the 'real world' due to the expensive cost of eye trackers.

The remaining part of this chapter is arranged as follows: Section 1.2 presents the background. Section 1.3 is the motivation for this research. The research question is presented in Section 1.4. The aim and objectives of this research are stated in Section 1.5. Section 1.6 gives an introduction to the research methodology. The contribution of this research is briefly discussed in Section 1.7. This chapter concludes with Section 1.8 which gives the layout of the remaining part of the thesis.

1.2 **Background**

The volume of information on the internet is constantly growing, and the process of accessing relevant information is difficult and time-consuming (Brusilovsky and Tasso 2004). The traditional IR system retrieves web resources based on user query input. The process of retrieving information in a traditional IR system begins when a user enters a query consisting of text (terms) in a search engine; the search engine measures the similarity between the query terms and the terms contained in the

documents and then returns a list of documents relevant to the query. Research suggests that most users of the web usually enter queries vaguely which may not translate their problem statement clearly; additional measures are needed to capture users' interest in order to supplement their queries.

The most common and consistent approach to capturing users' interest about a piece of information is by asking them to explicitly state it (Claypool et al. 2001). Such explicit statements of what users' think about a piece information can be done by their preference information (Takano and Li 2009). Although the explicit rating is the most used and consistent approach to personalization, it alters reading and browsing patterns (Claypool et al. 2001).

To build a robust non-intrusive feedback system, users' perceived interest can be captured implicitly through their sequence of actions as they browse; this removes the cost of rating by the users (Zhang et al. 2010). Users dwell and focus more on documents that are interesting, useful or relevant to their current situation (Buscher et al. 2012a). Information relating to user interest can be obtained from such dwelling activities. Although the explicit rating is commonly used and trusted by many, it is not always reliable as presumed (Claypool et al. 2001); thus the solution is to unobtrusively obtain users' suggestions especially in the context of learning (Shi et al. 2013). The advantages of an implicit feedback approach over an explicit feedback approach also include:

- A large amount of data can be collected without interrupting the users.
- User feedback can be captured at any time.
- With implicit feedback, users need not examine and rate items explicitly.
- Bias in rating is eliminated through an implicit method.

Several implicit measures for capturing users' interest have been studied in previous research. The indicators mostly studied include dwell time (also called reading time), the amount of scroll movement, mouse movement, mouse clicks, mouse distance, copy and paste, printing, highlighting, emailing, and bookmarking. When these implicit indicators are used in isolation, they tend not to capture users' interest (perception of relevance) as opposed to when they are combined (Balakrishnan and Zhang 2014). The assumption used in this thesis is that users will view documents that they find interesting and their degree of interest can be estimated by collecting and analysing their behaviour on the visited documents. This research builds on previous studies in the area of implicit feedback. In this thesis, the phrase "Interest" and "perception of relevance" are used interchangeably.

1.3 **Motivation**

There is a lack of intelligent adaptive systems that can assist a homogeneous group of users to retrieve relevant documents efficiently and accurately. There is also a gap in research in solving the information overload problem and the development of effective recommender systems. The present retrieval systems are generic in nature and do not take into consideration a user's context. The average user of general search engines spends a significant amount of time searching for their information needs based on their current task. For instance, students studying the same module are often faced with the same search problems as previous cohorts. The process of visiting similar web documents in a search engine like Google is often repeated by all students taking that module. Relevant documents previously visited by learners of a common domain are not captured and shared, even though sharing the documents will significantly optimise learners' performance and minimise

the time they spend on their search activity. Some attempts have been made by previous research to address the problem of information overload by developing models of information seeking behaviour. This research focuses on the predictive behaviour of web users that can be used to personalise information retrieval. The goal of this work is to develop a prototype implicit feedback system to assist learners in a particular domain in their search activity, by accurately recommending relevant documents to them based on their previous interaction with the system.

1.4 Research Question

How can we predict the relevance of web documents for a task-specific domain based on users' activity?

a. Can specific task situations be used to derive a predictive function model from implicit indicators that will signify that a web document is relevant?
b. To what extent do document familiarity and document difficulty affect users' behaviour?
c. Does the context of the task affect users' behaviour?
d. Can we use previous learners' perceived relevant web documents to optimise recommendation of the relevant document in a given domain?

1.5 Research Aim & Objectives

The purpose of this research is to investigate the correlation between explicit and implicit relevance feedback parameters in order to develop a model for the prediction of relevant web documents. The following objectives have been set to achieve this aim:

1. To examine and review existing research related to relevance feedback.
2. To conduct a user study to capture user explicit ratings and various implicit indicators.
3. To investigate whether document familiarity and difficulty affect user implicit and explicit feedback parameters.
4. To unfold the effect of task type on user behaviour.
5. To examine the implicit interest indicators in relation to the explicit ratings in order to develop a predictive function model.
6. To evaluate and validate the predictive model using standard evaluation metrics and eye gaze.
7. To develop and evaluate a prototype system for the recommendation of relevant web documents to learners.

1.6 Research Approach

In order to achieve the research aim and objectives and to answer the research question, the following research approach was adopted.

- Stage 1: This research begins with a problem formulation and a review of existing literature in the area of information retrieval. It also investigates the previous approaches used for the development of recommender systems.

- Stage 2: This step reviews previous work on specific implicit indicators and relevance feedback systems, and how they can be effectively used to predict document relevance.

- Stage 3: An investigation of how task type, document familiarity and document difficulty affect user behaviour is carried out.

- Stage 4: In this step, a new methodology is proposed which addresses how best to use implicit indicators to enhance prediction of relevant documents based on a user profile. A user study is conducted which leads to the derivation of an implicit predictive function model for retrieving relevant web documents.

- Stage 5: This step evaluates the predictive strength of the new model through standard evaluation metrics and user study. Eye tracking experiment is employed in this evaluation phase. Also, the predictive strength of several classifiers is compared.

- Stage 6: A conceptual framework is created for the development of a context based implicit feedback system that can assist learners in their search activities. A prototype system is developed and evaluated.

1.7 **Main Contributions**

The main contributions of this thesis include:

1. A model to estimate document relevance is developed. Here, the relationship between implicit and explicit relevance feedback parameters obtained from a large number of users in different task situations is interpreted and a predictive model to estimate document relevance is derived based on assigning weight to various relevance feedback parameters.
2. Validation of the predictive model with standard evaluation metrics and eye gaze tracker. It is argued that the predictive model derived by combining implicit indicators is more effective in predicting document relevance than predictive models of a single indicator. Also, it is shown that there is no significant difference between the eye gaze measures and the predictive model.

3. A number of classifiers are used in this research and a comprehensive comparison is shown within the context of this research. More precisely, the thesis demonstrates that K-nearest neighbour and J4 Decision tree classifiers have a higher accuracy to classify documents based on user behaviour.
4. The thesis has also examined and unfolded the extent at which task type, document familiarity and document difficulty affect user behaviour.
5. A framework for a domain-specific feedback system is developed and a prototype system is implemented. The evaluation of the system shows that supplementing user queries with implicit feedback parameters can improve the quality of search results.

1.8 Thesis Layout

The organisation of this thesis is in four parts:

Part I: Introduction

This part comprises the introduction, the background and the state of the art approaches related to this issue being addressed in this thesis. It focuses on the main concept of the thesis which motivates subsequent chapters. It discusses the different feedback approaches used for information retrieval and recommender system. A particular focus is on relevance feedback approaches.

Part II: Implicit Evidence

This comprises Chapters 3 and 4. In Chapter 3, three user studies are conducted, thereby following comprehensive methodologies. Part II concludes with Chapter 4 in which the results of the studies as described in Chapter 3 are presented and discussed. A model to implicitly estimate document relevance is developed and evaluated.

The predictive strength of several classifiers is also compared in this chapter.

Part III: Implicit Feedback System

This part consists of Chapter 5 and it presents the proposed context-based relevance feedback system for the recommendation of relevant web documents. It also presents the implementation and evaluation of the prototype system, showing how the system can improve the recommendation of relevant web documents.

Part IV: Conclusion

Part 4 concludes the thesis. It summarises the work described in this thesis and presents the contribution. The limitation of the thesis work is also stated and the planned work for the future is highlighted.

CHAPTER 2

2 State of the Art

2.1 Introduction

This chapter discusses the background knowledge which is needed to address the problem specified in Chapter 1 and it reviews related work in the area of relevance feedback, discussing what has been achieved so far in the area and the challenges at hand. The goal of this chapter is to review the literature to investigate the state of the art approaches so as to gain an insight on how the problem of information overload and personalisation have been addressed by the previous research. The chapter also investigates perceived limitation of the existing approaches and identify the gap in the knowledge.

Some important concepts about information seeking and retrieval are explained in Sections 2.2, 2.3, 2.4, and 2.5. Section 2.6 briefly explains explicit feedback measures of user interest. Implicit feedback measures, which are the main focus of this work, are explained in detail in Section 2.7. A detailed review of previous research in the field of implicit indicators begins from Section 2.7.1 to 2.7.2. Section 2.8 discusses relevance feedback systems. The content of this chapter explores and presents the motivation behind this work, and it relates to other chapters of the thesis.

2.2 **Information Seeking Behaviour**

Information seeking can be said to be a form of problem solving (Marchionini 1992) while information retrieval is said to be all the processes involved in providing users with documents that satisfy their information needs or query (Baeza-Yate and Riberro-Neto 1999). When dialogue is involved in the process of retrieving these needed documents in a dynamic way, we say the process of information retrieval is interactive. Most of the previous studies in IR have centred on retrieval algorithms more than user approaches towards retrieval. Effective information retrieval also depends on the user activities towards retrieval. Since it is practically impossible to know exactly what exists in a user's head, it is important to investigate the possible behaviour we can use to predict what a user is interested in. The motive behind information seeking is to acquire information that a user need. The need is subjected only to the individuals who interact with a system to find it. An external observer can only discover this need by monitoring the user behaviour or by obtaining a report from the user.

Considerable research has been carried out in the area of information seeking and retrieval (Kuhlthau 1993, Wilson 1999) to assist users to find relevant web resources for their current needs. Users' way of searching for information has changed from frequently accessing library books to general surfing of the internet (Ajiboye and Tella 2007, Chapman and Ivankovic 2002, Siddiqui 2011). The dynamic and hypertext nature of the web means users' online behaviour is constantly changing. Information seeking behaviour involves the manner and process through which people obtain information to augment their knowledge for self-development and other reasons (Ajiboye and Tella 2007). Users can easily get 'lost in hyperspace' due to the increasing volume of documents on the internet - this affects their ability to get their needed information in

minimal time. Users information needs can be easily met if they can have quick access to documents that are relevant to their current needs.

The way users seek information online is said to vary (O'Day and Jeffries 1993). Hearst (2009) dynamic model explains why users vary in their information seeking behaviour. He emphasised that due to the dynamic nature of humans when they seek information, they usually begin their search process with a particular goal then change from the initial goal as the search intensity increases. The study by O'Day and Jeffries (1993) affirms Hearst's dynamic model. They found that an average searcher is influenced by other goals while searching for a particular initial task. Barry (1998) groups the factors that affect user online behaviour as:

- Personal: This includes all factors that relate the reader to the information. It entails how a reader understands a document and whether the document is new to the reader.
- Quality: This relates to the source of the document and how the document is presented. Users tend to be attracted to documents with clear presentation and from reputable sources.
- Content: This relates to the accuracy of a document. It also entails its availability and references to other useful and relevant documents.

One way of keeping a user to a particular search goal is by suggesting appropriate recommendations that are of interest to the user. Most of the present IR feedback systems are designed in such a way that recommendation/feedback is based on how a user query matches the terms in a document. Little or no consideration is given to users' past interactions with documents (Busby 2003, Iqbal et al. 2015). Since searchers find it difficult to construct effective queries through the use of keywords, their ability to retrieve relevant documents is limited. There is, therefore, a need to capture and model users' past behaviour and perceived interest on previous documents visited in a given context.

2.3 **Personalised Models**

The branch of information retrieval that focuses on user behaviour and experiences as well as cognitive, physical and affective behaviour is called Interactive Information Retrieval (IIR) (Kelly 2009). IIR covers the area of Library science, psychology, traditional information retrieval and human-computer interaction (HCI). Its essence is to study the interactions between a user and a system, and the information obtained from the system is used to personalise information retrieval. Such interactive behaviour is obtained from users cognitive activities like reading time, mouse and key activity.

Personalisation is a way to solve the problem of being 'lost in hyperspace', caused by document overload on the web. The web is complex with multiple domains and users may differ in their domain of interest. To connect users to their specific domain and needed information, a domain based approach of personalisation should be employed (Kelly 2004). Personalised models are constructed by observing users' behaviour and topical interest. It connects users to their preferred documents. The goal is usually to build an effective personalised interaction model and incorporate it into an information retrieval system (Limbu et al. 2009). Contextual factors like task type, task difficulty and topic familiarity affect users' behaviour (Liu, Belkin and Cole 2012). Capturing and representing users' task and behaviour without some level of privacy intrusion is yet to be realistic (Koene et al. 2016). Hence in this thesis, some implicit predictive indicators that are consistent with a community of web users are examined. These indicators can be captured unobtrusively from users with their consent and modelled.

2.3.1 User Modelling

To offer assistance to learners in their searching activities, there is a need to understand their web behaviour and how best to capture and represent their interest in a system. The system should be able to use previous interactions to improve its recommending strength. The method for modelling users can be complex or simple depending on the aim of the recommendation. Nguyen et al (2009) describe user modelling in two phases: knowledge acquisition and adaptation. The phase of knowledge acquisition gathers and organises the documents while adaptation retrieves and presents relevant documents to users according to their profile and preferences.

2.3.1.1 Profile Construction

User modelling can be done by asking users to explicitly state examples associated with their interest or by automatically capturing user interest through the use of specific implicit indicators (Huai 2011). Early studies in user modelling required users to explicitly state their interest (Luhn 1958). Modern studies are proactive and are based on observing users' browsing behaviour while they do some activity (Zemirli 2012). A current and popular study is the analysis of server-side data obtained from commercial websites (Zemirli 2012). The data logged are usually descriptive but can be modelled for some situations. TREC is an evaluation infrastructure that facilitates the lab-to-product transfer of technology. User modelling for the development (construction) of user profile has the following features depending on the goal of the recommender system:

- *Personal details**:*** When a system requires profiling, there is a need to collect user personal information. These personal details could be gender, age or country of the user.

- *User History:* For a prediction to be carried out, a system must have a record of past user actions. In the context of implicit feedback, user history can be obtained through records of their previous behavioural characteristics.
- *Preference:* Users may be asked to explicitly state their interest. The challenge with this feature is that a user's interest is dynamic and it may change over time.

These three features are considered in the design and implementation of the implicit feedback system in this work.

2.4 **Vector Space Model**

Vector Space Model (VSM) is a popular model used in information retrieval (Busby 2003). It represents documents as vectors, and similarities are determined by the dot products of two vectors. The vector space model algorithm ranks similarity between documents by comparing the user query with document keywords (Baeza-Yate and Riberro-Neto 1999). Keywords have a value, which indicates the level of relevancy to a document. The procedure for VSM is divided into three phases. The first phase is document indexing where the document is split into units called tokens. In the second phase, each term of the document is given a weight to enhance retrieval of relevant documents. The third phase ranks the document in relation to the query based on the similarity measure.

The vector space model uses two factors (Term Frequency (TF) and Inverse Document Frequency (IDF)) to give weight to a term in a document (Salton, Wong and Yang 1975). The term frequency is the rate at which a term occurs in a document and weights are assigned to terms based on their frequency. The efficiency of TF is however affected by common words like "is", "the", "a". This limitation can be overcome by the Inverse Document Frequency. The IDF calculates the number of

documents that contain each term and reduces the weight of terms that occur in many documents. The TF-IDF scheme gives high weight to terms that occur often within a document but do not commonly occur in the collection of documents (Salton and Buckley 1988). It is given as:

$$wd_t = tfd_t \times idf_t \quad 2.1$$

Where wd_t is term weight t found in document d
tfd_t is the frequency factor for the term t in document d
idf_t is the IDF for term t

$$idf_t = \log(D/docFreq_t) \quad 2.2$$

Where D is total number of documents
idf_t is term t inverse document frequency
$docFreq_t$ is all the documents that has term t

2.4.1.1 How VSM determines relevant document

After documents are indexed by a search system, the documents are then ranked based on their similarity with a query. The vector space model calculates the similarity by comparing the angle of deviation between each document vector and a query vector (the query vector is the same type of vector with the documents) so as to rank documents according to the angle they make with the query. In practice, the cosine similarity between two vectors is calculated. The cosine coefficient calculates the angle between the query vector and the document vector. It is calculated by multiplying the weight of each term from the vectors and dividing each by the length of the vectors (Salton, Wong and Yang 1975) as shown in Equation 2.4. This normalises the vectors.

Figure 2.1: Illustration of document and query relationship on the VSM (Source: Salton, Wong and Yang (1975))

Some materials have been removed from this thesis due to Third Party Copyright. The unabridged version of the thesis can be viewed at the Lanchester Library, Coventry University.

Where q is the query term vector and d is the document term vector, $d.q$ is the intersection (Salton and Buckley 1990).

2.5 **Contextual Measures**

The main aim of measuring users' characteristics in a search process is to evaluate the extent of user individual differences and how the knowledge of users' intentions, interest and context can improve recommendation (Bennett et al. 2015). In IR, contextual measures describe the state in which information seeking activities occurs. Although the present search engine has provided some support to users, like the ranking of documents according to a user query, spelling and query suggestion, interpretation of the suggestions remains the duty of the user. Contextual retrieval approach combines different search technologies and user context to meet users' information needs (Limbu et al. 2009). Multi-purpose personalisation approaches have been developed

over the years but this has been limited to laboratory use. Research (Busby 2003, Iqbal et al. 2015) has shown that such generic personalisation approach does not adequately capture users' interest in a real life situation. To create an efficient and robust system that will personalise user interaction with the web, contextual approaches need to be employed to capture users' information seeking behaviour and incorporate it into the system. Although contextual retrieval is important, it has not been fully implemented because of the difficulty in developing the right instrument to capture user's knowledge, intelligence, cognitive style, personality, memory, document familiarity, task or domain of interest. Attempts have been made to use pre-task and post-task questionnaires to predict user satisfaction (Crescenzi, Capra and Arguello 2013), but it is yet to effectively capture user context; for instance, users might be instructed to use a 6-point rating scale to state their familiarity with a given web document but it might not be possible to obtain information about how much understanding the user has to a current topic. Also, domain knowledge can be measured based on experience or previous study but such classification cannot lead to a concise interpretation of the results of a study. However, careful task-based experimentation can be used in place of questionnaires to capture these contextual factors while users browse the web.

2.5.1 Task

Research in information retrieval in the last decade has focused on user interest and how context can be used to improve users' search experience. When an information system makes use of user context, it improves the quality of task outcomes (Bennett et al. 2015). Tasks are activities that people make an attempt to accomplish to meet a particular goal. These have been viewed in a different perspective by researchers (Järvelin and Ingwersen 2004). They influence humans' social and

psychological behaviour. Information seeking task refers to the general problems of a user which can be met by searching any related resources like documents, people or information system.

Researchers in the area of information seeking have investigated the effect of task type on users' information seeking behaviour (Kuhlthau 1993, Wilson 1999) and they inferred that the needs of information seekers keep changing as they progress from one task to another task. This is viewed under different streams; identifying an accurate task stream is difficult because the demarcation of boundaries is not clear enough. However, common streams have classified different tasks along features like information gathering against fact-finding (Kellar, Watters and Shepherd 2007). Studies have also shown that such differences in task streams affect users' implicit behaviour (Liu, Liu and Belkin 2013). Kellar et al (2007) say that information gathering tasks are more complex among the task types and require longer completion time and page views. Another study on the effect of the task as a contextual factor on reading time conducted by White and Kelly (2006) found that under a task specific context, dwell time can be used to measure document usefulness and improve the performance of an implicit feedback system. Liu and Wu (2008) also reported similar findings; they examined whether task type is a good contextual factor that can be used for document prediction. Their result shows that task type assists in inferring document usefulness at the initial task stage and also influences the total time spent while performing the task. Kelly and Belkin (2004) studied 7 users' behaviour in a naturalistic setting for 14 weeks and they reported that reading time is best used for measuring document usefulness in a task specific context.

Researchers (Cole et al. 2011, Järvelin and Ingwersen 2004, Li and Belkin 2008) have also investigated the effect of task difficulty on user

behaviour. Cole et al (2011) used eye movement patterns to examine the relationship between user interaction behaviour and task difficulty. They inferred that their technique can accurately use task difficulty patterns to differentiate between tasks types. A similar contextual research was carried out by Järvelin and Ingwersen (2004). The study by Li and Belkin (2008) investigated the effect of task difficulty on user behaviour (reading time and 'hits' per query). They asked users to perform 6 tasks and rate the tasks according to difficulty. They infer that reading time is a good measure of user behaviour but it cannot predict task difficulty. The research by Cole et al (2011) and Ingwersen & Jarvelin (2005) centred on contextualisation involving a single user while Alhindi et al (2015) adopted the approach of contextualization of a group of users. They used profile-based summarization of similar documents in a particular domain. This thesis suggests a novel approach which examines the behaviour of users of a particular domain like Alhindi et al (2015), but multiple indicators are considered for examination.

2.5.2 Domain Knowledge

The idea of recommending relevant information to users based on their behaviour has been studied for decades in IR community. One of the limiting factors of actualizing the personalisation approach is the inconsistency in user behaviour (Iqbal et al. 2012) and the diverse subject areas (Li and Belkin 2008). Previous studies have used features like query input, search techniques and dwell time to examine how domain knowledge affects searchers' behaviour. In some cases, only a few indicators (reading time and search efficacy) were examined in relation to topic difficulty, task type and topic familiarity (Kelly and Cool 2002). The research by Kelly and Cool (2002) investigated the effect of topic familiarity on user behaviour. They inferred that users search efficiency increases and their dwell time on documents decrease when they are

familiar with the search topic. Bhavnani (2001, 2002) conducted a study to examine novice and experts search behaviour. They recruited 5 healthcare experts and 5 shopping experts for the study. Their findings suggest that while novice users begin their search with the general search engines, domain experts go straight to websites that will provide them with their needed information. Although a study by Hsieh-Yee (1993) on novice and expert searchers suggest that domain knowledge has an effect on only experienced searchers. A similar study was conducted by Hembrooke et al (2005) to investigate the effect of domain knowledge on how searchers enter and reformulate queries. Their result shows that domain experts entered complex and longer queries than domain novices. Zhang et al (2005) used engineering domain to study the relationship between user search success and domain knowledge. They gave graduate and undergraduate engineering students 200 engineering terms to state whether they are familiar with the terms. Their findings show that the domain experts rated that they are familiar with more terms than the novices. Liu, Liu and Belkin (2013) study confirm that users with different level of topic familiarity differ in their search behaviour. Unlike previous small-scale studies, a large scale log analysis of searchers' behaviour on four domains was carried out by White et al (2009). They developed a model that can predict domain experts based on how they search for information.

As briefly described above, dwell time, queries and search techniques have been used by previous research to investigate whether domain knowledge affects user behaviour. Most of the previous research in domain knowledge were based on library resources and not web resources. In this work, an experimental approach is used to unfold the effect of domain knowledge with a focus on document familiarity on user behaviour.

2.5.3 Relevance

One of the fundamental concepts in Information Retrieval theory is Relevance. There is no single definition of the concept of relevance despite its key role in IR and not much is known about the factors that affect relevance judgement (Xu and Chen 2006). However, two definitions are broadly accepted. They include:

a) Topicality: Topicality or topic-appropriateness relates to whether a piece of information has a topic bearing to a user query. It relates to when a query matches a topic that has useful information.

b) User-Utility: This relates to how useful a piece of information is to a user who submits a query. That is when a query result returns useful information to the user. This definition was affirmed by Su (1994, 1992). She conducted a study to evaluate an interactive IR system and she found that the best evaluating measure of performance for an interactive IR system is the user satisfaction.

Relevance feedback enables the system to recognise a user and personalise results according to a user's previous interest (Teevan, Dumais and Horvitz 2005). Since it is not known how the human mind filters what is relevant from what is not, relevance is mostly measured by user explicit actions in the field of IR (Gwizdka 2014), but an average user considers this method intrusive. A non-intrusive and objective approach of capturing relevance is to implicitly infer the users' interest through their movement of input devices, reading time, eye gaze and brain signals from biometrics and psychological sensors (Rocchio 1971). Relevance feedback information is employed for constructing user profile in a contextual retrieval environment. Although much research has been done in this area, recommending what is relevant to a particular user is still a challenge.

This thesis attempts to address this challenge by investigating user search behaviour in a task-specific context and developing a prototype system for the effective recommendation of relevant web documents.

2.6 **Explicit Feedback Measures**

Explicit measures are the simple and direct approach to collecting user-interest data. The type of explicit approaches commonly used include: “comment” (Núñez-Valdéz et al. 2012), “product review” (Aciar et al. 2007), “tagging” (Wei et al. 2016), “think aloud” (Fattahi et al. 2016), and “explicit rating” (Balakrishnan, Ahmadi and Ravana 2016, Wei et al. 2016). Comments and product review are mostly used in E-commerce recommender system while tagging is used in a content-based recommender system. In think-aloud approach, users are asked to verbalise their thoughts as they use the system, while in the explicit rating, users are given a scale of preference to rate their thoughts about a system. Explicit ratings are mostly used by E-commerce sites to collect data from buyers (Zemirli 2012). Users are usually asked to give feedback about an item they purchase. Most E-commerce sites use this measure to recommend related items to users. Explicit measures are mostly used for relevance feedback (Claypool et al. 2001); its limitation lies in the fact that users are always ‘forced’ to update their needs, which makes data collection difficult. Another limitation is the cognitive overload this measure put users into (Claypool et al. 2001). To eliminate the cost of rating and to reduce the cognitive overload, there is a need to unobtrusively capture user data. This approach is called implicit feedback.

2.7 **Implicit Feedback Measures**

Implicit feedback measures are normally used in place of an explicit feedback measure to unobtrusively capture user's interest. Considerable research has been done to improve the quality of information retrieval by using implicit feedback (Buscher, Dengel and Van Elst 2008, Iqbal et al. 2012). Implicit feedback approach uses implicit indicators to replace explicit rating for the development of recommender systems (Ding, Liu and Tao 2010, Jawaheer, Weller and Kostkova 2014). It is used to unobtrusively estimate users' interest on web documents. Although implicit feedback is widely available, it is considered a secondary option to explicit feedback (Jawaheer, Weller and Kostkova 2014) and it is noisy and less accurate as compared to the explicit method (Claypool et al. 2001). Current research investigates the best way of replacing explicit feedback measures with implicit feedback approaches. For instance, in a controlled setting, mouse and scroll movement has been found to have some correlation with the explicit rating, but it is somewhat difficult to interpret in the real world (Buscher et al. 2010). An advantage of this measure is that a large amount of data can be collected ubiquitously without restricting a user to a particular place.

The predictive strength of a number of implicit indicators has been investigated in the field of implicit feedback. The implicit indicators previously investigated include: dwell time which is also called reading time, mouse clicks, mouse movement, amount of scroll movement, mouse distance, copy and paste, printing, highlighting, emailing and bookmarking (Balakrishnan and Zhang 2014, Kim and Chan 2005, Zemirli 2012). Unlike explicit rating which is intrusive, expensive and alters user browsing pattern, implicit measures remove the cognitive cost of rating and these are not intrusive (Zemirli 2012). The next two sub-sections focus on commonly used implicit feedback measures and eye-gaze-based implicit feedback measures.

2.7.1 Commonly Used Implicit Feedback Measures

Dwell time was introduced by Morita and Shinoda (1994) as a behavioural characteristic to substitute for explicit rating. They conducted an experiment with 8 users who were given a 6-week task to read articles in a newsgroup they belong to and explicitly rate them. The investigation was based on how the document length, its readability and the number of unread articles affect the reading time. Their findings suggest that users dwell more on articles perceived to be relevant to their current task and the reading time is not affected by document length. Using modified distributed software, Konstan et al (1997) repeated the study by Morita and Shinoda (1994) in a natural setting. Explicit rating and reading time was logged from participants in a recommender system trial. Their findings show that in terms of accuracy, there is no significant difference between a recommender system that is based on reading time and a recommender system that is based on explicit rating. A study by Yi et al (2014) explored the use of item-level dwell time to estimate the relevancy of web content. They found that item-level dwell time is a good indicator for a personalised recommender system. The research by Konstan et al (1997) and Morita & Shinoda (1994) restricted users search to Usenet newsgroup and only a single implicit feedback parameter (reading time) was examined.

Guo & Agichtein (2012) suggest that an aggregation of dwell time with other promising indicators, like cursor movements and scroll, can serve as better evidence of relevance. Nichols (1997) evaluated the following implicit feedback parameters - mark, reply, glimpse, query, associate, refer, repeated use, delete, save and print. Oard and Kim (1998) extended Nichols (1997) work by capturing users' useful information and making appropriate recommendation to them. They categorised user behaviour into *Minimum Scope axis* and *Behaviour Category axis*. Minimum Scope axis encapsulates the smallest scope of the object in use. It consists of

Segment, Object and Class while Behaviour Category axis represents the purpose of the observed behaviour and it comprises examine, retain, reference and annotate behaviour.

Table 2.1: Classification of implicit indicators by Oard and Kim (1998)

Kelly and Teevan (2003) supplemented the classification with a behavioural category axis called "Create", for the creation of new information. They also added find, browse, query and scroll to the "examine" group and email to the "Retain" group as highlighted in Table 2.1. Only a few of these indicators are frequently used by online users. This work focuses on the frequently used online behaviour that can be employed to assist users retrieve relevant web documents in an interactive information retrieval environment. Claypool et al (2001) developed the Web browser (Curious Browser) to study the predictive strength of implicit indicators. The web browser captured user implicit and explicit data. The implicit indicators measured were mouse clicks, scrolling, mouse movement and elapse time. They found that the amount of scroll, the reading time, and the combination of the amount of scroll

and the reading time are stronger predictors of user interest. Kim and Chan (2005) examined similar indicators to Claypool et al (2001). They conducted a study in which subjects were requested to bookmark more than 10 pages, use memo on more than 5 pages, print more than 5 pages and save more than 5 pages. They found that the dwell time and the distance of mouse movement are good indicators for measuring user interest. The study by Claypool et al (2001) and Kim and Chan (2005) used special browsers which were limited to a controlled environment. Also, information seeking task was not used for the study. The method employed by Kim & Chan (2005) compelled the users to a certain behaviour (bookmark, print, save) which might have affected their 'true' web experience. This research used instrumented browser to collect data in a more controlled and naturalistic way.

Zhu et al (2012) used the parameters of Clicks, Bookmaking, Voting and reply to adaptively model users' interest. Fox et al (2005) correlated implicit and explicit judgement and developed a predictive model using Bayesian method. They found that the aggregation of click-through, dwell time, and the way users exited a search result page or ended a session gave the best prediction for their explicit judgement of satisfaction. The study by Fox et al (2005) focused on the search engine result page and not users' post click behaviour on documents, which is the focus of this work. In the context of electronic book recommendation, Núñez-Valdez et al (2015) proposed an architecture that analysed and transformed implicit feedback parameters to approximate explicit ratings for a community of readers. Their results show that users' interest can be determined by analysing and converting their behaviour. Leiva and Huang (2015) used a client-side approach of tracking user activity to record users' cursor movements for computing relevance of search results. They infer that the 'cursor movement' capturing tool is a viable tool for understanding user behaviour.

Zemirli (2012) worked on post-retrieval documents. He developed a web browser (WebCap) that uses 'examine' and 'retention' indicators to infer users' interest in real time, and experimentally evaluated the system with 6 users. He found that WebCap was able to capture 80% of the relevant documents when compared with explicit user judgements. Similar success was reported by Shapira, Taieb-Maimon and Moskowitz (2006). Velayathan and Yamada (2007) investigated factors that affect user behaviour by integrating a number of web browsing factors of interest. They used a decision tree algorithm to classify documents perceived to be interesting and non-interesting, and they found scrolling as a common indicator of interest among users. In a relative study on aggregating implicit indicators, Balakrishnan and Zhang (2014) examined the effect of some implicit indicators on post-retrieval document relevancy. They found that a combination of text selection, dwell time, click-through and page review post-click behaviour can improve the precision of relevance feedback.

Users vary in their web behaviour. The variability of user behaviour as it relates to modelling web navigation was investigated by Juvina and van Oostendorp (2006). They found that web applications can be designed to consider indicators that have been proven to predict with significant accuracy task outcome for a group of users. This work builds on this finding as it models the most consistent implicit feedback parameters that can be used to estimate document relevance among a group of users. Most of the previous studies on implicit feedback parameters in the field of information retrieval were focused on the result page of search engines and just a handful of studies were based on developing a predictive model through the aggregation of implicit indicators generated from web documents visited by users. A task-based approach to examining user behaviour is employed in this work with a focus on user post-click behaviour. Implicit feedback parameters studied in previous

research is revisited and information tasks are used to tailor users' search activities towards a particular domain.

Section 2.7.1.1 to Section 2.7.1.9 provides insights into the commonly used implicit feedback parameters which can be employed in the later chapters of this research. The implicit indicators are sourced from user dwell time (reading time), mouse activity and key activity.

2.7.1.1 Dwell time

Dwell time, also called the active reading time, is the period at which the window or web document is in focus. Dwell time was one of the first indicators studied in relation to user's interest. Morita and Shinoda (1994) examined user dwell time as a source for measuring user interest. Their results show that users dwell more on articles perceived to be interesting. This assertion was supported by another study conducted by Huai (2011). Other researchers (Claypool et al. 2001, Kim, Oard and Romanik 2000, Lee, Park and Park 2008, Núñez-Valdéz et al. 2012) also laid emphases on dwell time as the key implicit indicator for measuring user interest. Kim and Chan (2005) added that the time users spend on a web page is related to their interest regardless of their attention to the page. Buscher et al (2009) examined the correlation of eye gaze and display time and they found that segment-level display time can be used in place of eye gaze for implicit information retrieval. Chapter 4 of this thesis evaluates the extent by which a model of 'low-cost' implicit features can substitute for user eye gaze.

Although most research say that dwell time is an important indicator of interest, a caution was suggested by Liu et al (2011). They say that dwell time alone cannot predict users' interest in all contexts but it should be considered along with user task. This assertion was affirmed by other researchers (Kelly and Belkin 2004, Velayathan and Yamada 2007). The work of Liu et al (2011) is revisited in this thesis by examining the effect

of task type on user behaviour. Research attests to the fact that longer dwell time signifies that a user is interested in the page (Balakrishnan and Zhang 2014, Guo and Agichtein 2012). However, it is acknowledged that predicting the relevance of a page will be more accurate if another post-click behaviour is aggregated with the dwell time (Guo and Agichtein 2012).

2.7.1.2 Mouse Movement

Most people move the mouse when browsing or reading a web document. Some use the mouse cursor to point to their focus or interest area as they view a web document. Experimental evidence has shown that there is a relationship between mouse movement and user interest. Zhang et al (2010) say mouse movement is a good parameter for implicit feedback. Claypool et al (2001) captured the time spent in moving the mouse and they infer that this parameter correlates with users' interest. Kim and Chan (2005) say that mouse movement is easy to detect and in most cases, it is as accurate as the time spent on a page. Mouse movement is also as good as eye gaze in predicting users' intention (Guo and Agichtein 2010).

2.7.1.3 Mouse Distance and Speed

The mouse distance and speed are derived from the mouse movement. The mouse distance was evaluated by Jung (2001) to be an important indicator of relevance. Kim and Chan (2005) say that the more the distance of a user's mouse on a page, the greater the interest. The mouse distance is calculated by the movement of the mouse cursor along the x and y coordinates on the monitor screen. The mouse speed is calculated by the movement of the mouse cursor along the x and y coordinates on the monitor screen with time.

2.7.1.4 Mouse Click

When people navigate from one web page to another, they click the mouse. Some people click frequently on documents they find useful. The count for the mouse clicks increments whenever the left mouse button is pressed. Studies have shown that mouse clicks correlate with user interest. Iqbal et al (2012) see mouse click as a good parameter to investigate users' intention and relevant data. Kim and Chan (2005) say mouse click is a good implicit indicator next to mouse movement. Huang, White and Dumais (2011) say that there is a correlation between explicit ratings and mouse clicks when hover activities are added to clicks. Claypool et al (2001) have a contrary submission about mouse clicks. They say that mouse click is not a good implicit indicator. This submission was affirmed by Takano and Li (2009). They say searchers click the mouse while browsing with no intention. Chapter 3 of this work explores this parameter to support or refute previous assumptions.

2.7.1.5 Scroll Movement

Web documents are normally longer in height than the monitor screen. When users find an interesting document, they scroll and read further down. Users also examine a search result by scrolling down the results (Huang et al. 2012). Claypool et al (2001) and Kim and Chan (2005) say that the greater the scroll bar movement of a page, the more interested the page is to a user.

2.7.1.6 Keystroke

Some people prefer using the arrow keys on the keyboard to scroll web documents on a monitor screen. Keystrokes are usually used in place of the scrollbar to scroll a web document. This indicator is measured by the number of times the user strikes the keys. Claypool et al (2001) and Kim and Chan (2005) didn't find this parameter to be a predictor of interest.

2.7.1.7 Amount of Copy

Copy and Paste is an important parameter for measuring relevance, especially for Software developers (Iqbal et al. 2012). The amount of copy is calculated as the number of times segments of text are copied from a web document (Liu, Belkin and Cole 2012).

2.7.1.8 URL Hit

Common intuition suggests that when a user is interested in a particular web document, the user will frequently visit the document. Users of a particular domain also visit a common web document that is of interest to them. The assumption in this research is that a frequent or commonly visited URL is perceived to be of interest to users.

2.7.1.9 Document Height

Although document height is not user behaviour of interest, previous research has studied it along with user behaviour (Morita and Shinoda 1994). Every web document has a particular height. The height of a web page is the vertical length of the web page calculated in pixels. This is explored along with other implicit indicators in Chapter 3.

2.7.2 Eye Gaze-Based Implicit Indicators

The modern eye trackers have a better degree of accuracy and precision in measuring gaze features compared to previous ones. This has led to increased study of gaze parameters as they relate to information retrieval (Gwizdka 2014). The previous study by Buscher et al (2012a) suggests that eye gaze is an important indicator of interest and it has a direct link to a user's visual attention. In a related study conducted by Salojärvi, Puolamäki and Kaski (2005) to explore whether eye movement is a good source of relevance feedback, they found that accurate prediction of document relevance can be deduced from eye movement. Cole et al (2011) investigated how user eye tracking information can be used to estimate users' interest/intention during information retrieval. They found that

eye gaze is a good source of implicit feedback for a users' task type. Granka, Joachims and Gay (2004) also reported that eye movement is a good indicator of interest.

Although eye gaze features are said to be the most predictive indicator of interest (Buscher et al. 2012a), they are, however, not used in the 'real world' due to the expensive cost of an Eye tracker and the unportable nature of the device. It is, therefore, necessary to substitute the eye gaze features with other implicit indicators obtained from 'cheap and available' sources. Attempts have been made by researchers to substitute the eye gaze indicator with other single implicit indicators. Huang, White and Dumais (2011) found a slight coordination between cursor movement and eye gaze. Guo and Agichtein (2010) say that regions, where mouse pointer and eye are within 100 pixels of each other, can be predicted accurately by nearly 77%. Most of the previous studies focused on finding a relationship between mouse cursor and eye gaze on Search Engine Result Page (SERP). In this work, I examined how we can validate the strength of the predictive function derived from aggregating commonly used implicit indicators with eye gaze. I show that to a reasonable degree, we can use an aggregation of non-gaze implicit indicators as a replacement for gaze-based indicators. The gaze measures studied in relation to the implicit indicators include Fixation (fixation duration and fixation count) and Heat map.

2.7.2.1 Fixation

Fixation is a condition where the eye is relatively focused on a subject of interest for a given period of time. In most gaze based systems, user interest is determined by fixation threshold (Maglio et al. 2000). Granka, Joachims and Gay (2004) say that the period of fixation depends on the user task. The fixation duration for an information task falls between 225ms and 300ms (Granka, Joachims and Gay 2004). Fixation Duration is the sum of all the individual fixation duration within a specific area of

interest of a document. Fixation Count is a number of times that a user fixates within a specific area of interests of a document.

2.7.2.2 Heat Map

This is a visualisation technique that separates different levels of fixation intensity. It shows areas that are more fixated to be denser than areas that are less fixated.

2.8 Relevance Feedback Systems

Recommender systems have been used over the years in E-commerce to suggest products for customers based on their needs. The common approach for this prediction is by collaborative filtering (Xu, Zhang and Huang 2010) - where the recommendation is done by comparing a user's profile with other similar users. There is a rapid drive in recent years for developing non-commercial recommender systems for text retrieval tasks (Zemirli 2012). Among these non-commercial recommender systems are the education hypermedia system and adaptive recommender systems. Unlike the E-commerce systems where the recommendation is primarily done by observing products purchased by a user, the education adaptive systems focus on user access of web documents and it is tailored towards learners' activities and knowledge acquisition (Brusilovsky 1998). This makes the implementation of educational recommendation systems challenging. Neji et al (2011) suggested ways by which a personalised feedback can be created from learners' activities. They suggested capturing learner's personal data, preference, browsing history, knowledge, emotion and navigation. Chen (2010) used similar learners' characteristics to those of Neji et al (2011) to propose a system that can capture learners' transition.

More than 3 billion people now have access to the internet (Internet Society 2015), and a significant amount of documents are uploaded every

day on the internet. The general search engines are designed to serve all users without putting into consideration the context or domain of the users. Search engines normally crawl web documents via their Meta tags and store them in a database, such that URLs are returned for documents matching the query entered by the user in the search engine text box (Busby 2003). Queries are however not sufficient enough to capture users' interest because the 'all purpose' search engines like Google and AltaVista use only keywords (query words) to rank documents (Busby 2003). Measures like visitor polarity have been employed by some user-controlled search engines to improve the relevancy ranking (Busby 2003) but some software has been designed by website operators to automatically increase the number of hits on their sites. There is, therefore, a need to augment users' queries with implicit feedback parameters previously obtained from their interaction with the system (White and Kelly 2006). Such supplementary information obtained from users' post-click behaviour, like the amount of time they spend on the document, the amount of copy and so on, can be used as an evidence of interest to optimise recommendation of relevant documents to users.

Previous adaptive hypermedia systems focused on producing a 'browsing agent' that will recommend relevant web resources to users through a content feedback approach. Letizia (Lieberman 1997) was developed to track users' browsing behaviour and recommend relevant web documents to users based on the links previously visited. WebACE (Han et al. 1998) extends the operation of Letizia by capturing and building a user profile with previous documents visited and the time the user spent viewing the documents. Other adaptive systems like WebMate (Chen and Sycara 1998) are based on explicit feedback which is intrusive. WebMate contains a proxy that observes users' interaction with the system. It allows users to explicitly state some examples of links they are interested in, and the system learns from them. It was used for newspaper

recommendation. WebWatcher (Joachims, Freitag and Mitchell 1997) is similar to WebMate. Users of the system are asked to enter certain keywords to represent their interest and the system learns from these keywords. It also has a function for users to evaluate whether a link was useful or not, which is then used as feedback for future recommendation. LIRA (Balabanovic, Shoham and Yun 1995) is another adaptive hypermedia system that explicitly seeks users' current interest and recommends to them documents relevant to their interest the next day.

These systems (Letizia, WebACE, WebMate, WebWatcher and LIRA) are generic in nature, attempting to fit all domains of interest, thereby limiting efficient recommendation of relevant documents to users of a particular domain. Contextualising information retrieval has a potential of helping users to find relevant and accurate information within a minimal timeframe (Alhindi et al. 2015). Context sensitive systems have been developed to improve web search. INQUIRIS2 (Glover et al. 2001) was developed as a metasearch system that asks users to explicitly state their context of interest in a given category of context. It uses the desired context along with the user query to find relevant documents in general search engines. The system proposed in this work uses queries along with implicit evidence of interest to improve the retrieval of relevant documents for a community of users. Whereas unobtrusive systems like POIROT (Ramírez, Donadeu and Neves 2000) uses keywords obtained from users' browsing history to supplement their queries and re-rank search engine results, the proposed system uses an aggregation of implicit indicators to supplement user query.

Kumar and Ashraf (2015) proposed a framework to personalise web search based on dynamic user profile, query expansion, user search history and collaborative filtering. They found that personalisation of web search is more efficient than a generic search engine. Researchers have

worked on aggregating implicit feedback parameters from users post click behaviour to improve the results of search engines. Guo and Agichtein (2012) studied how users interact with Search Engine Result Page (SERP). They estimated document relevance through user scrolling and cursor activities and found that a combination of scrolling and cursor movements predicts document relevance more effectively than using only dwell time. In a natural setting, Buscher et al (2012b) used large scale behaviour log data to investigate the effect of user-task on user behaviour on SERP. They were able to cluster users based on their scrolling, clicks, cursor movement, and text highlighting behaviour. Núñez-Valdéz et al (2012) reported that most of these implicit indicators can be used to improve the recommendation of electronic books.

Vector Space Model (Salton and Buckley 1988) as explained in Section 2.4 is the retrieval algorithm used by most search engines to evaluate the relevance of web documents. Information is retrieved based on query-document similarity. Efforts have been made to improve information retrieval by augmenting query input with previous user interaction with the system. Zhu et al (2010) applied user implicit data as a surrogate of user interest to develop a personalised information retrieval system. They used a combination of selected implicit parameters (saving, printing, favourite, viewing, click-through) to estimate user interest on documents and integrated it with the traditional search engines. Their findings suggest an improvement in information retrieval. Some of the indicators of interest employed by Zhu et al (2010) are not frequently used by online users. A similar method was employed by Balakrishnan and Zhang (2014) to improve document search results relevancy. Balakrishnan and Zhang (2014) used previous users' post click behaviour (dwell time, click-through, text selection, page review) as an additional information source to re-rank SERP. The integrated model proposed by Balakrishnan and Zhan (2014) was based on heuristics. An 'intrusive' explicit feedback

study to improve retrieval relevancy was conducted by Balakrishnan et al (2016). They derived a model by integrating three explicit feedback parameters (comment, rating and referral) and their findings indicate that search retrieval relevancy can be improved when users' explicit feedback is aggregated.

The goal of the system proposed in this work is to use implicit user behaviour to improve the recommendation of relevant web documents to users of a particular domain. The predictive model derived via experimentation in Chapter 4 is integrated with the traditional vector space model. Whereas previous research (Balakrishnan and Zhang 2014, Balakrishnan, Ahmadi and Ravana 2016) used heuristic to assign weight to implicit and explicit indicators, this work uses a multiple regression approach to deduce a predictive function to represent users' interest on documents, which is then used to improve query result re-ranking. Also, the implicit indicators of interest used in this work are statistically selected from domain specific experimentation via the use of well-defined tasks.

Chapter Summary

A comprehensive description of the background and motivation for this work is presented in this chapter. Information seeking behaviour which is the foundation for understanding user information seeking and retrieval is discussed. A brief discussion of personalisation with a focus on user modelling is carried out. Vector Space Model, one of the popular models used for information retrieval is discussed. Contextual factors that affect user behaviour are also discussed. They include tasks, domain knowledge and relevance. The extent at which these moderating factors affect user behaviour is discussed in the later chapters of this work.

This chapter also reviewed related work to this research. It began with a brief discussion of the explicit feedback parameters and it discusses the implicit feedback parameters in detail. Relevance feedback systems are also discussed. Explicit relevance rating by users is commonly employed by most feedback systems but it is intrusive and expensive (Claypool et al. 2001). This makes implicit feedback approach a useful alternative. Implicit feedback involves the use of user behavioural characteristics to infer the relevance of a document. User implicit features are commonly sourced from their search activity (time they spend on the documents, mouse and key activity). Eye gaze as a source of implicit feedback was also discussed and some eye gaze measures were listed.

This chapter is the foundation that subsequent chapters are built on. A method is proposed in Part II of this thesis and experimental analysis is used to derive a predictive function for recommending relevant web documents. This function is based on user implicit behaviour on the web as explained in this chapter.

II

Implicit Evidence

CHAPTER 3

3 Methodology and User Studies

3.1 Introduction

This chapter presents the methods employed for the user studies. Three user studies are conducted. The first study is a preliminary study to investigate user reading behaviour of the web documents. The study examines the predictive strength of dwell time and mouse activity on web documents. It focuses on understanding users' reading behaviour and whether user generated implicit indicators of dwell time and mouse activity correlate with user explicit ratings. The preliminary study is specific to reading as users were not asked to search for documents.

The second user study is a study of user searching behaviour. A task specific approach is employed to capture user implicit and explicit parameters. A pilot study is described and the outcome of the study is stated. A description of the methodology used for finding the relationship between explicit ratings and implicit indicators is presented. The two approaches for studying user behaviour (naturalistic setting and laboratory setting) are intertwined in this study. It correlates implicit and explicit feedback parameters, and it uses multilinear regression to derive a predictive model that can estimate document relevance.

The third study is a validation study. It investigates the reliability and validity of the predictive model by comparing it with eye gaze during a reading task.

3.2 User Study 1: A Preliminary Study on Implicit Predictive Indicators

To investigate the predictive strength of some implicit indicators discussed in Section 2.7.1, a preliminary study was conducted on some given web documents. An automated study was carried out and 13 participants were given 15 short documents to read and rate according to their perception of relevance to a given topic area. The study aimed to investigate if there is a correlation between users' generated implicit indicators and the explicit ratings.

3.2.1 Implicit Feedback Indicators

In this study, a number of implicit indicators were used to capture participants' interest on the given web documents. The implicit behaviour captured include *Active Time Spent on the Document (TS), Distance of Mouse Movement (DMM), Total Mouse Movement (TMM) and Mean Mouse Velocity (MMV).* The Explicit Ratings (ER) was also captured. These features are explained in Section 3.3.7

3.2.2 Study Design

The goal of the study was to capture the participants' interest in web documents via some implicit indicators and to correlate the users' interest against their explicit ratings of the given documents. The participants were 4 Ph.D. students, 1 Research assistant, 6 MSc students and 2 undergraduate students of Coventry University. Data for this research was collected by automated software developed with JavaScript. The software was injected in 15 web documents to record users' mouse activity, dwell time and explicit rating. Participants were given a task brief (see Appendix A1) to read and a consent form (see Appendix B) to complete, after which they were allowed to perform the experiment at a time of their convenience. They were to login into a website containing

links to the 15 web documents and read the documents. The implicit data was captured unobtrusively as the participants read through the documents and the data was sent to MySQL database when they rated the document by clicking on any of the buttons on the rating scale as shown in Figure 3.2. The task ended after the participants read and rated the 15 documents. Figure 3.1 shows a step-to-step schema of the task process.

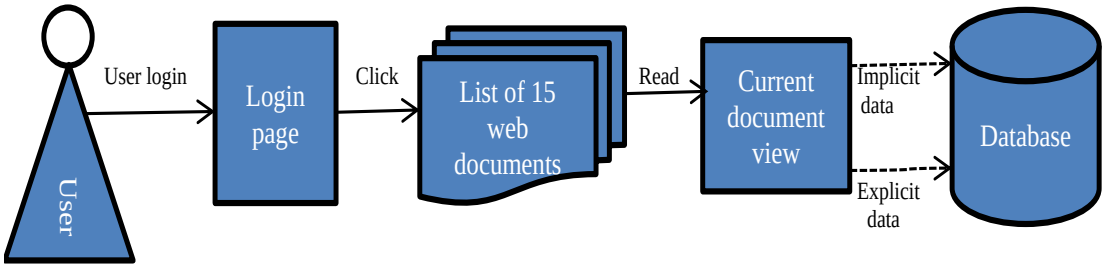


Figure 3.1: Step-to-step schema of the task process

3.2.2.1 User Task

All the participants were given the same task to perform for 60 minutes. The participants were asked to prepare a presentation on the topic, "Ethical issues in Big Data". Their task was to read through each of the 15 documents prepared for them and rate them according to how relevant the documents are to the topic. The rating was on a scale of 0 – 5. Six buttons were attached on top of the documents and labelled 0 to 5 for explicit rating of the documents as explained in Section 3.3.7. The 15 documents were of equal length containing 350 words with a font size of 20px and a font type of Arial, making the documents one screen view. The documents were created from web articles on ethical issues in Big Data. Two of the documents were however not related to the topic.

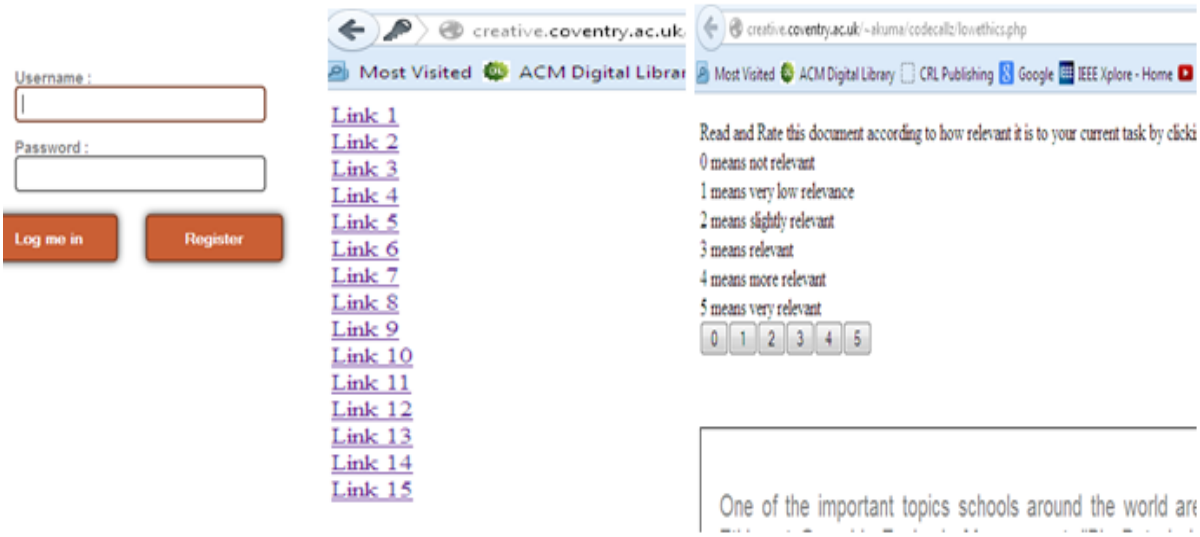


Figure 3.2: Login page, Index page with links to the 15 documents and document page with explicit rating buttons

The experiment was given a realistic feel by creating a second phase of the task which was called the presentation writing phase. The participants were told that documents would be presented to them according to how they rated them for later use in the presentation writing phase. To avoid Hawthorne effect (the alteration of behaviour by the subjects of a study due to their awareness of being observed), participants were told to do the experiment when and where they were most comfortable. The participants did not actually perform the second phase of the experiment (the presentation writing phase).

3.3 **User Study 2: Study of User Search Behaviour**

The aim of this experiment was to a use contextual approach and learners' behavioural characteristics to predict learners' level of interest in web documents in a current search session and the relevance of documents to the current task. The following goals were explored:

1. To correlate user implicit behaviour with user explicit ratings.

2. To investigate how task type affects user browsing behaviour.
3. To examine whether document difficulty and familiarity affect user behaviour.
4. To derive a predictive model from the relationship between the implicit and explicit feedback parameters.

Hypothesis

Users' interest in web documents can be inferred from their web behaviour. The degree of the behaviour on the documents determines the level of interest in the documents. This behaviour is represented by a set of implicit indicators. When these implicit indicators are aggregated, they predict users' interest more than an individual implicit indicator.

3.3.1 Pilot Study

The Pilot study captures users' behaviour as they interact with the web. It involved multiple browsing session over 6 weeks. Data was captured from 03/07/2014 to 10/08/2014. A total number of 8 students participated in the experiment and they included 4 Ph.D. students and 3 Masters Students from Coventry University and 1 Ph.D. student from London school of Commerce. The participants' subject area was in Engineering, Computing and Management. They were all given a consent form to complete before partaking in the experiment. Two of the participants were females and 6 were males. They were aged between 26 and 33 and they all had over four years searching experience. A plugin to capture their interaction with the web was installed on the Firefox browser of their computer and it ran 'invisibly' whenever the subject's browsers were in use. They were asked to read documents as they naturally do and rate the documents on a six-point rating scale according to how relevant they were to their research area. They were also informed to rate whether they were familiar with the documents they visited. I represented their search domain with their course of study since they

were asked to enable the plugin only during their study period. The implicit indicators captured include number of clicks, document height, amount of copy, amount of scroll, mouse movement on the X-axis, mouse movement on the Y-axis, dwell time, mouse distance, mouse duration count, mean mouse speed and keystroke together with the respective URLs (see Section 3.3.7).

3.3.1.1 Pilot Study Result

79 documents were captured but due to some technical issues with the plugin, the mouse movements of 17 documents were not captured. The 17 documents were removed from the dataset and 62 documents were analysed. Pearson correlation (see Section 4.2) was used to test for a relationship between the implicit and explicit feedback parameters. The result did not produce a significant correlation between the implicit features and the ratings for relevance. This might be because users were from a different domain and they had different peculiar needs. There was, however, a positive relationship (0.274) between the keystrokes and user ratings for document familiarity with a significant coefficient of 0.031 (p<=0.05). Since the experimental procedure did not produce a statistically significant correlation between the implicit measures and the explicit relevance ratings, the research was redirected to focus on a single domain area through the use of search tasks.

3.3.2 Logging

The oldest and most common approach of capturing a large amount of quantitative data in an IR system is logging. The challenges that accompany this method of data capturing is how to capture data without noise and how to prepare and interpret it. There are four types of web-based logging that are commonly used in IR. They include Server-side logging, Proxy logging, client side logging and Instrumented Web browsers.

Server-side logging: This approach is used by most commercial websites because relevant web usage data can simply be obtained without the installation of analytic software at the client side (Zemirli 2012). This method of logging captures data on a large scale but users do not have access to the server. It is mostly used by search engines to capture the IP address of the requesting machine, URL visited and timestamp. Its major drawback is that it records only data requested from a single server and navigation of web pages is limited.

Proxy logging: This is like an interface. It is placed between a server and users to store interaction between the users and the servers. Proxy service can be used to modify resources sent from the server to the user. The advantage it has over the server side logging is that it captures more data.

Client-side logging: This is normally installed on the user's computer and it captures both web data and other system data uniquely. It is the most comprehensive type of logging but involves a lot of technicalities to develop.

Instrumented Web browsers: This is designed by researchers for a particular research goal. It involves injecting a plugin in an already existing browser like Mozilla Firefox and Windows Explorer or by developing a customised browser from scratch. The advantage of this browser is that it can run unobtrusively in a naturalistic setting.

3.3.3 Procedure

This study employed both controlled and naturalistic approaches to collect data. This was done to cover for the disadvantages that exist in using only one of the approaches. One of the disadvantages of a controlled study is the issue of Hawthorne effect, where subjects alter their behaviour due to their awareness of being observed. On the other

hand, the naturalistic study normally produces some noisy data. The subjects were to choose whether to perform the study in a controlled environment (selected IT laboratories of the Department) or in a natural environment. The duration set for each group of participants to perform the tasks was 45 minutes. The participants were given a task brief (see Appendix A2) and a consent form (see Appendix B) was also given to them to complete. A brief tutorial, explaining the procedure for the experiment was carried out by the researcher. A Mozilla Firefox Portable which has an injected JavaScript plugin to capture user implicit and explicit features was given to the subjects to use in searching for answers to the given tasks. To prevent nervousness and anxiety, the participants were advised not to look at their clock during the study. They were to directly key in the URL address of their required web page or enter a query of their choice in a search engine to search for their required web page. For every web page they visit, they were to do the following:

i. Enter their User Id (See Figure 3.3) then read through the web page for information relevant to the task under consideration. They were to close the web document (current tab) after reading it. On closing the web document, they were prompted to rate the relevancy of the document as it relates to the given task (see Section 3.3.7 and Figure 3.4) then state whether the web document was difficult to understand and also state their familiarity with the document.

ii. The participants were instructed to visit and read not less than seven web pages and write a one-page report of the solution to the task under consideration.

Figure 3.3: Users prompted to enter their Id on opening a web page

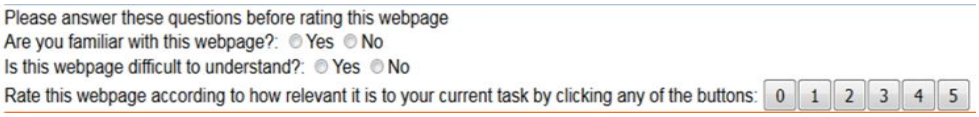


Figure 3.4: Users prompted to rate documents when closing a web page

3.3.4 Document Domain

Since the study involved searching the web for relevant documents, the document domain for the study was the World Wide Web. The users were allowed to visit any web document of their choice and find answers to the given task. The documents visited with their corresponding ratings were logged.

3.3.5 Participants

Since there is no agreeable standard for sample size for an Interactive Information Retrieval (IIR) experiment and no linear relationship between sample and population (Kelly 2009), a convenient sample size was selected that will produce a statistically significant result. A total number of 77 undergraduate students from the Engineering and Computing faculty at Coventry University were recruited for the study. Coventry University is an International University with students from different countries spread across the 6 continents of the world and the course content for Computer science in Coventry University is similar to that of other Universities. The number of participants selected for the study was large enough to make credible conclusion following the rule of thumb

(Johnson 2010). Most of the participants were recruited through their lecturers and they performed the study in some selected laboratories of the University. Other participants were recruited by email and word-of-mouth, and they were given the task brief and instruction sheet to perform the tasks in the laboratory or at their homes at a time of convenience.

Only participants above 18 years were allowed to participate. The participants had a high proficiency with the use of computers. Remuneration was not given to the participants but they were informed of the overall benefit of the research to student learning.

3.3.6 Experimental System

An instrumented browser for easy and remote use was employed for collecting and logging data. The system comprises a client side and storage side as shown in Figure 3.5. The client side has a Mozilla Firefox Portable which embeds a JavaScript plugin to capture user web data (implicit measures, explicit ratings for relevance, explicit ratings for document familiarity and difficulty) for each web document visited. In the storage side, data captured by the JavaScript plugin is sent to a central server and then transferred to MySQL database. The plug-in life cycle is dependent on the web page that calls it. When the Firefox web browser engine (Gecko) starts, it first searches for the plugin. As the user opens a web page that invokes the plugin, the browser loads the JavaScript plugin code then it initializes and creates a new instance of the plugin. The plugin then captures user web data and sends it to the central server after the user rates the web page. Once a user leaves the web page or closes the window, the plugin instance is deleted.

To ensure that only documents associated with the tasks under consideration are stored, URL addresses from Yahoo, Facebook and Google result page were excluded from logging. The data was then

extracted from the database through dedicated SQL queries and exported to CSV format for analysis. Most of the systems used by previous researchers (Claypool et al. 2001, Kim and Chan 2005, Zemirli 2012) were customised web browsers for controlled experiments in the laboratory. This system can be used in the laboratory for controlled study and remotely for naturalistic study. The difference between the experimental system in study 1 and study 2 is explained in Table 3.1.

Table 3.1 The difference between the experimental system in study 1 & study 2

S/N	Experimental System in Study 1	Experimental System in Study 2
1	Study 1 was specific to reading, not searching. JavaScript was injected in the 15 web documents provided for the participants in order to capture their implicit and explicit feedback parameters.	Study 2 involved searching. Users were to visit any documents of their choice on the web in their attempt to find answers to the given tasks. A JavaScript plugin was injected in Firefox plugin to capture their implicit and explicit feedback parameters on web documents.
2	The 15 web documents were self-created by the researcher so it was possible to inject the JavaScript directly into the documents.	The web documents were from the different website so a JavaScript plugin was injected in a Firefox browser to alter the web documents in order to capture user explicit and implicit parameters on the documents.

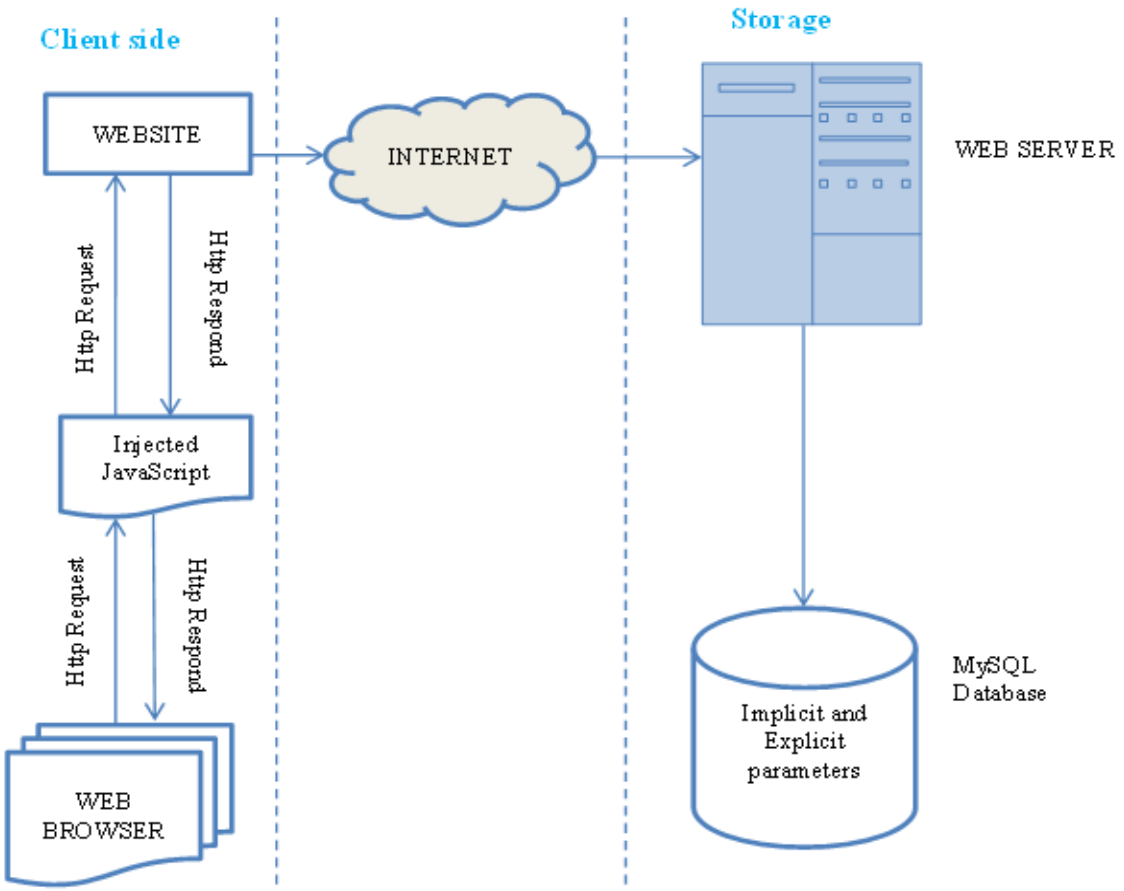


Figure 3.5: Experimental system

3.3.7 Implicit and Explicit Parameters Captured

This section discusses the implicit and explicit feedback parameters captured by the system. Both site structural data and interactive event data were captured.

3.3.7.1 Interactive Events Captured

1. *Dwell Time (DT):* The dwell time (active time on document) is the actual period at which a web document is in focus. It is the total time (in seconds) that a user spends while reading a web document in one session.
2. *Mouse Distance (MD):* As explained in Section 2.7.1.3, the mouse distance is calculated by the cursor movements along the x and y-

axes of the monitor screen for every 100ms. It is the Euclidean distance of the mouse as shown in Equation 3.1:

$$MD = \sum_{i=1}^{n} \sqrt{(x - x_i)^2 + (y - y_i)^2} \quad 3.1$$

Where x, x_i, y_i and y are the locations of the mouse cursor on the monitor screen, n is the total number of location points moved by the mouse.

3. *Mouse Movement (MM):* As explained in Section 2.7.1.2, the mouse movement is calculated as the mouse hovers along the x and y-axes of the monitor screen. The count for the movement along the x and y-axes are incremented by the change of its current value at each movement.

4. *Mouse Duration Count (MDC):* This is the total number of 100ms intervals that occurred while the mouse moved on the monitor screen.

5. *Mean Mouse Velocity (MMV):* The mean mouse velocity is the average speed covered by the movement of the mouse on the screen. The formula for computing the mean mouse speed is given in equation 3.2:

$$MMV = \left(\sum_{i=1}^{n} \sqrt{(x - x_i)^2 + (y - y_i)^2} / (t)\right) / MDC \quad 3.2$$

Where x, xi, yi and y are the locations of the mouse cursor on the monitor screen, n is the total number of location points moved by the mouse; t is the time covered (100ms). MDC is as explained above.

6. *Number of Mouse Clicks (NMC):* This the total amount of mouse clicks on the current web document. The count is incremented for each page every time the mouse is clicked by a user.

7. *Amount of Scroll (AS):* The vertical length of most web pages is longer than the monitor height. Users normally scroll the web page by either

clicking or dragging the scroll bar. The count increments anytime the scrollbar is dragged or clicked.

8. *Number of Keystrokes (NK):* This is the total number of keystrokes on a web document. The count for each page is incremented when a user strikes a key.

9. *Amount of Copy (AC):* This is the number of times text is copied to the clipboard from a web document. Anytime a text from a particular document is copied, the count for the document is incremented by one.

10. *Explicit Relevance Ratings (ER)* This is users' statement of the relevance of the current web document. A six-point rating button is attached by the Firefox plugin on each of the web documents. A user is to rate the relevance of the web document in relation to the task under consideration by pressing any of the six buttons. "0" means not relevant, "1" means very low relevance, "2" means slightly relevant, "3" means moderate relevant, "4" means more relevant and "5" means very relevant.

11. *Document Familiarity:* This is the users' statement of familiarity with the current document. The rating is done on a two-point scale.

12. *Document Difficulty:* This is the users' statement of whether the current document is difficult to understand. The rating is done on a two-point scale.

3.3.7.2 Site Structural Data

1. *Time Stamp:* This is the exact time and date in GMT when a document is loaded (open timestamp) and when a document is closed (close timestamp).

2. *Page Height:* This is the vertical length of the document measured in pixels.

3. *IP Address (IP):* This is the unique string of numbers that represents the location of each computer. It is the internet protocol address of the user machine.
4. *URL:* This is an acronym for Uniform Resource Locator. Each web document has a unique web address. It is the HTTP address of each web document visited by a user.

3.3.8 Tasks

When tasks are assigned to users without background information or context, it demotivates the users and they may consider the task as artificial (Kelly 2009). Simulated work task situations (Borlund 2003) were employed for this study. The simulated work task situation is a 'cover story' with two parts (*work task situation* and *indicative request)* that has three characteristics:

- The subjects should be able to relate and identify with the task.
- The subjects should find the topic situation interesting.
- The task situation should have an imaginative context so that subjects can apply the situation.

Apart from these three characteristics, the tasks are designed in consideration of the subject's background and the indicative request gives the subjects a direction of how they can initiate the search process.

The tasks used in this study were designed to encourage the participants to naturally search for web documents. They were designed with the intention to be interesting for the participants and enable them to easily relate to. The domain of the tasks was in the field of Computer Science. Section 3.3.8.1 discusses the classification scheme by Li and Belkin (2008) and Liu, Liu and Belkin (2013) used in categorising the task's components into a single scheme.

3.3.8.1 Task Components

The tasks were grouped into components (facets) according to their values and complexity to simulate different information seeking task situations. We selected two of the tasks with large participants for detailed analysis. Table 3.2 shows a classification of the task components.

Table 3.2: Task classification by (Liu, Liu and Belkin 2013)

Simulated Work Task Situation 1 (Mixed Task)

GIG Software Development company employed you as a consultant to provide a solution to the Company's pressing problem of developing a customised software within a minimal time frame. Some professional software developers achieved this by using the Rational Unified Process while others used the waterfall model.

Indicative request 1

Which of the approaches would you consider for a small project of few lines of code (LOC) and what stage of the software lifecycle do you consider to be the most important? State the reason for your answer in your report.

Simulated Work Task Situation 2 (Factual Task)

Google is looking for young and ambitious students of Computer science for an internship to work under the Company's Service Management Department. Consider that you are shortlisted for an interview among 2000 applicants and you are asked to search the internet and find answers to questions related to Information Technology Infrastructure Library (ITIL):

Indicative request 2

i. What are the five stages of the ITIL lifecycle?
ii. What are the differences between ITIL v1, v2 and v3 (2007)?
iii. What are ITIL processes?
iv. What are ITIL functions?
v. Who should use ITIL?
vi. When should ITIL be used?
vii. What are the differences between ITIL and ISO/IEC

3.3.8.2 Classification of the two Tasks

The grouping of the search task given to the participants as shown in Table 3.3 follows Li and Belkin (2008) and Liu, Liu and Belkin (2013) task classification.

Table 3.3: Component grouping of the two tasks

Task	Product	Goal (Quality)	Objective complexity
Task 1	Mixed	Specific goal	High
Task 2	Factual	Specific goal	High

Task 1 is considered a mixed product (Decision and Intellectual task) because it involves making a decision to solve a problem with the most efficient method (RUP or Waterfall Model). It also focuses on 'how' a problem can be solved. It asked for the most important stage of the lifecycle, making it also an intellectual task. The goal of the task is specific because participants have to find which approach is better for a few lines of code and it is of high complexity because a minimum of 7 documents was to be sourced.

Task 2 is considered a Factual product because facts are to be located and it focuses on gathering information about a thing or subject; it is specific because participants were to find specific information which was explicitly measurable. The complexity of this task is high because a minimum of 7 web documents was to be consulted.

3.4 User Study 3: Eye Gaze Measures in Relationship to Classical Implicit Indicators

The aim of this study is to use eye gaze measures (Fixation count, Fixation duration) to validate the predictive strength of the function model derived. The goal is:

1. To find the correlation of the users' gaze generated indicators on some perceived relevant web documents with their explicit relevance ratings.
2. To find the correlation of the mean explicit ratings of some documents used for this study with the explicit ratings of the same documents in the main study.

The correlation is aimed at finding the relationship between the predictive model and eye gaze indicators. Also, it examines the consistency of user rating on the documents.

Hypothesis

It is hypothesised that the higher the eye gaze measures (Fixation count, Fixation duration) the higher the explicit relevance ratings.

3.4.1 Apparatus

Gaze data was captured with Tobii TX300 desk mounted eye-tracker. It was paired with a 23-inch LCD monitor with a resolution of 1920 X 1080 pixels. The tracking frequency for the eye tracker was 300Hz and it gave room for subjects to move their heads. The accuracy was 0.4 degree of visual angle.

3.4.2 Procedure

The aggregated function model derived from the relationship between implicit indicators and explicit ratings in Section 4.5.3 was run on the dataset of the 'Mixed task' (task 1) described in Section 3.3.8. Some documents perceived to be the most relevant and some documents perceived to be the least relevant were identified from the pool of documents. These documents were then given to 9 participants to read through with an eye tracker installed on the machine and rate them on a six-point scale according to how relevant they were to the task under consideration. The participants were Coventry University students

majoring in the area of Computer Science and their eyes were calibrated on a five-point calibration scale shortly after they completed and signed the consent form (see Appendix B). A short tutorial and a task brief describing the task was also given to the participants (see Appendix A3) and each of the participants had to sequentially read through the 6 documents within 30 minutes and rate them. The participants' Fixation count, Fixation duration and heat map as described in Section 2.7.2 were captured by Tobii SDK Software.

Chapter Summary

The methodology employed in the three user studies is discussed in this chapter. It presents a step to step approach employed in the user studies. A description of the experimental system, behavioural features and the population sample is stated. The chapter also describes the task's components and the simulated tasks employed for the study. A pilot study was conducted for user study 2 to evaluate the effectiveness of the system but no significant relationship between the implicit indicators and the explicit relevance ratings was obtained, which led to a slight adjustment of the generic searching procedure to a domain specific one.

Three studies were conducted to comprehensively examine the relationship between implicit and explicit feedback parameters. Study 1 is a preliminary study on user reading behaviour. It gives insight into understanding the relationship between implicit and explicit feedback parameters. Although study 2 can be conducted independent of study 1, it builds on the results obtained in study 1. Study 3 is dependent on study 2. The predictive model derived in study 2 is validated in study 3. Figure 3.6 is a diagram that shows the connection of the three user studies.

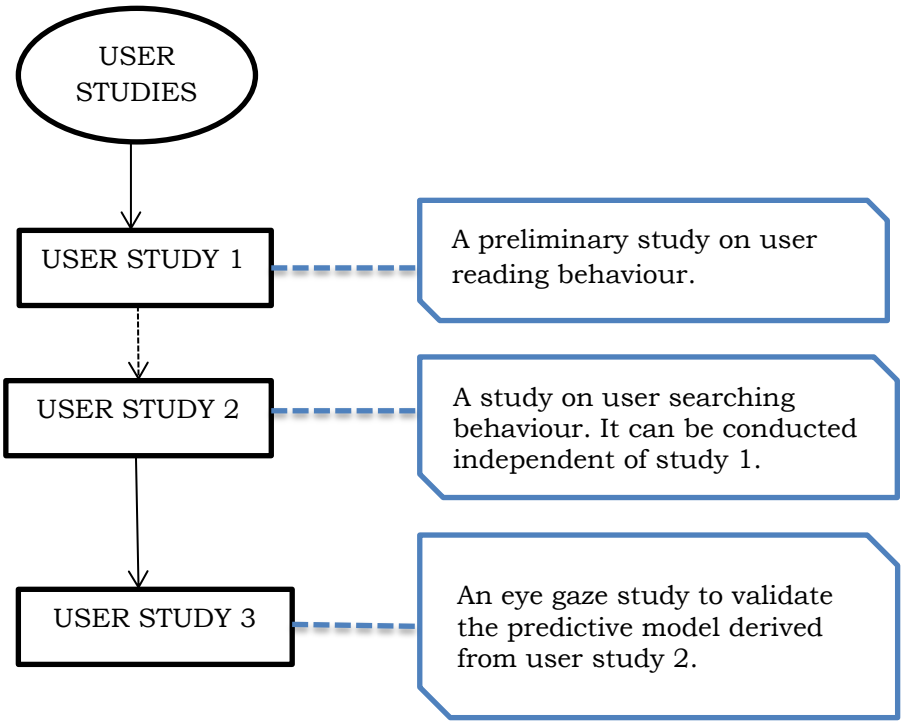


Figure 3.6: Diagram connecting the three user studies

CHAPTER 4

4 Empirical Results

4.1 Introduction

The previous chapter discussed the methodologies employed for the user studies. It also conducted an investigation into user reading and searching behaviour. The results of the user studies conducted in Chapter 3 are presented and discussed in this chapter. The result of user study 1 is presented in Section 4.3; it correlates users' generated implicit indicators with their explicit relevance ratings of selected documents. The result of user study 2 is presented in Section 4.4; it predicts learners' level of interest in web documents in a current search session and the relevance of the documents to the current task. A predictive function model was developed from the relationship between the implicit and explicit feedback parameters. The evaluation of the predictive function model and the classification analysis is presented in Section 4.5. This chapter concludes with the result of user study 3 which validates the predictive model.

4.2 Statistical Concepts

This section describes the statistical concepts used for analysing the results of the experiments. The concepts explained include Pearson correlation, Independent T-test, Chi-Square test and significance test.

4.2.1 Pearson Correlation

The Pearson correlation (denoted as *r)* generates a coefficient which measures the direction and the strength of linear relationship between two variables. The value ranges from -1 to +1 depending on the strength of the relationship. A positive coefficient means that there is a positive linear correlation and a direct relationship between two parameters. A negative coefficient means that there is a negative linear correlation and an inverse relationship between two parameters. When the coefficient is closer to 1 or -1, the linear relationship is stronger. When the value of the coefficient is zero (0), it means that there is no correlation. The hypothesis is given as:

H0: $\rho = 0$; the sample coefficient is equal to zero

Else

H_A: $\rho \neq 0$; the sample correlation is not equal to zero

To carry out Pearson correlation, the test for linearity is determined by a scattered plot and in this work, the test produced a linear distribution for all the variables considered. The test for normality is not considered since it is assumed that the mean of a sample distribution is normal when the sample size is above 25 or 30 (Hogg and Tanis 2005).

4.2.2 Independent T-Test

The independent T-Test statistically determines whether there is a significant difference between the means of two independent groups. The independent T-test is carried out on large data set regardless of the test for normality since Independent T-Test and ANOVA are not very sensitive to a large distribution of data that slightly deviate from normality (Glass,

Peckham and Sanders 1972, Harwell et al. 1992, Lix, Keselman and Keselman 1996).

4.2.3 Chi-Square Test

This test is employed to determine whether there is an association between two categorical variables. It tests for association between two nominal/dichotomous variables. The basic concept of the Chi-Square test is to determine the strength of the association between two categorical features. In this work, the Chi-Square is used for testing the relationship between:

i. User explicit relevance ratings and document familiarity ratings.
ii. User explicit relevance ratings and document difficulty ratings.

4.2.4 Statistical Significance Test

In order to determine if results from measurements are real and not random, significant testing is employed. Smucker, Allan and Carterette (2007) say that the following parts are needed for conducting a significance test:

- ***A statistical significance test:*** This is any standard measure like the Wilcoxon test, t-test or any similar test measure.
- ***P-value:*** this is the test result used for establishing confidence for the null hypothesis.
- ***Null hypothesis:*** This refers to a statement that indicates that there is no statistically significant relationship or association between two measured parameters. A rejection of the null hypothesis means that there is evidence that there is a relationship between two parameters. In this work, the null hypothesis is considered if the significance level (p - value) for Pearson correlation and that of the Independent T-Test and chi-square test is greater than 0.05 as illustrated in Table 4.1.

Table 4.1: Statistic testing and condition for a result to be significant

Statistic significant test (*p*)	Coefficient level	Condition	Action
Pearson correlation / Independent T-Test / Chi Square Test	> 0.05	There is no significant relationship	Accept null hypothesis
	< = 0.05	There is a significant relationship	Reject null hypothesis

4.3 **User Study 1 Results (Preliminary Study)**

This section discusses the results of user study 1 in which initial data was captured from 13 users based on a reading task. The data was analysed separately for each participant and as a group. Pearson Correlation was used to correlate the parameters of dwell time, total mouse distance, average mouse velocity, and total mouse movement along the X and Y axes with user explicit rating. The result of this study shows a significant correlation between the dwell time and the explicit relevance rating. The correlation between dwell time and user explicit rating was 0.21 (p<=0.05). Regression analysis was also used to rank the implicit indicators by their predictive strength. The indicators that showed much prominence in relation to the explicit ratings were the dwell time and mouse movement along the X-axis.

The dwell time also has a positive correlation with the mouse movement/mouse distance. Figure 4.1 shows a box plot of varying median of the user explicit rating and the dwell time of the participants, and it shows that the values of the ratings for 3 and 4 are the most

consistent. The inconsistencies in the other values might be due to noise in the data. The Kruskal-Wallis test on the median for each of the explicit ratings shows that the median values are not the same by rejecting the null hypothesis.

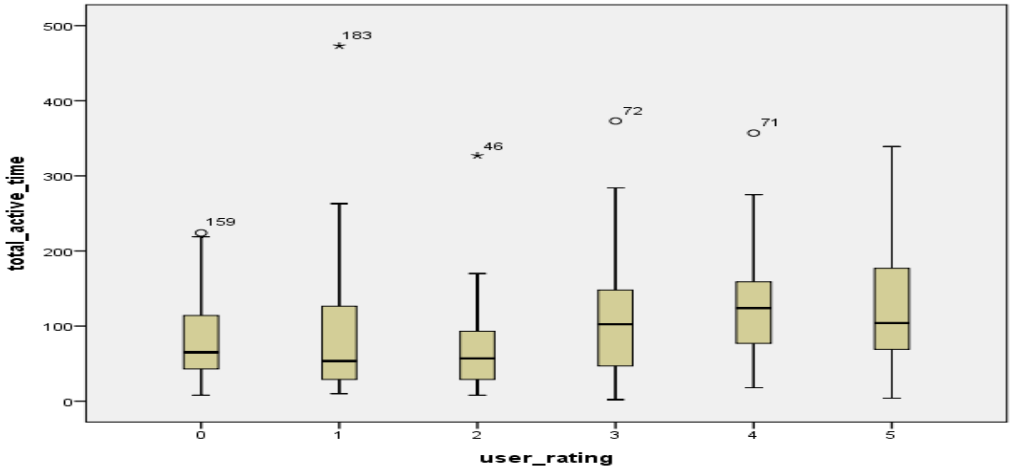


Figure 4.1: Graph showing the Boxplot for the combination of all participants' time/explicit rating relationship

The result also shows that although users vary in their reading behaviour, some of them have a similar behavioural pattern. We analysed the first two participants' data separately to find out the extent of individual differences. We discovered that they have a relatively similar pattern of dwell time and mouse activity on the documents visited as shown in Figure 4.2.

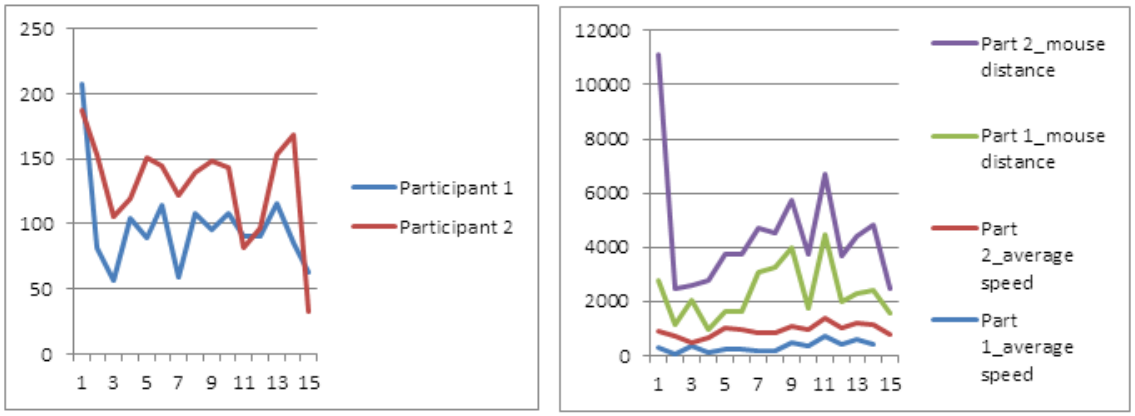


Figure 4.2: Graph showing participant 1 and 2 dwell time, mouse distance and average speed on the documents.

In order to regroup the user ratings, the six levels of ratings were then reduced to Relevant (1) and Non-relevant relevant (0). The user ratings from 0 to 2 were combined together and represented as 0 while the ratings from 3 to 5 were represented as 1 (Relevant). A Multilayer Perceptron was used to conduct further analysis on the primary data set (mouse movement along the x and y-axes, dwell time and mouse velocity time count) to predict relevant and non-relevant user rating. A 65% successful mapping between user relevance ratings and the implicit indicators after testing the trained data set was obtained.

The most promising indicator in the measure of perceived relevance is the dwell time. Participants spent more time on documents perceived to be of topical relevance. The correlation of mouse activity with the explicit rating is relative low, probably because of the length of the documents which were of 350 words and could be mostly viewed on screen at once. There is, however, evidence of a positive relationship between dwell time and explicit user rating. This conclusion is in line with previous research conducted in relation to implicit feedback (Kellar et al. 2004, Lee, Park and Park 2008, Morita and Shinoda 1994, Zhu et al. 2010). The mouse distance/movement is closely related to the dwell time. We can substitute

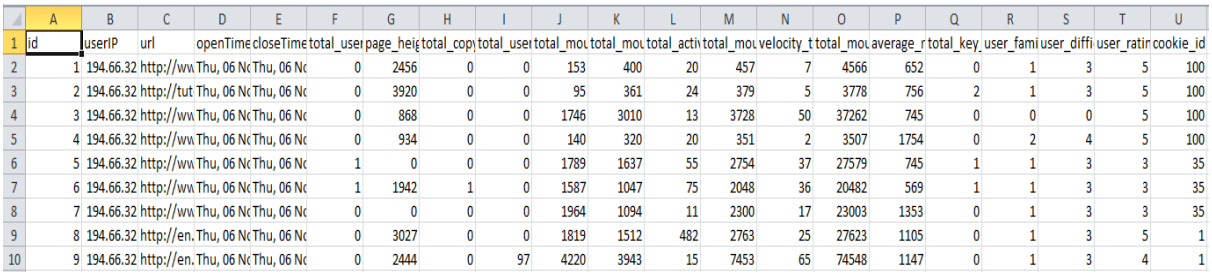
in some way the dwell time by mouse movement or mouse distance in an implicit system.

We can also infer that learners dwell more on documents that are of topical importance and interest to their current activity. The effect of the concept of prior knowledge and cognition on the reader's behaviour was not examined. It was assumed that the selected participants had limited knowledge of the task domain. Some individual behavioural differences among the participants were observed. To examine the variety in reading behaviour, two of the participants' (Participants 1&2) data was examined closely and it was discovered that some level of similarity exists in their behaviour in terms of dwell time and mouse activity. Multilayer Perceptron was used to further analyse the primary data set, with user rating as the dependent variable. A fair result of 65% mapping after training the data set was obtained. This suggests a relationship between the user behaviour and their explicit ratings.

4.4 **User Study 2 Results**

This study uses simulated tasks to limit users to a particular domain. Users freely surfed the web to find answers to a given task, and their web behaviour was captured through a JavaScript plugin in Firefox browser as discussed in the previous chapter. Web documents were collected together with their respective implicit measures, difficulty ratings, familiarity ratings and user explicit relevance ratings through a software plugin. Web documents used during the short tutorial and training were manually removed from the dataset. Documents whose active time was above 600 seconds (10 minutes) were reduced and fixed at 600 seconds. This was done on the assumption that if users are not distracted by something else, they will spend a maximum time of 10 minutes on a web page for a 45 minutes' task. A total of 343 web documents were extracted

from MYSQL database to SPSS statistics software (IBM 2013) for analysis. Figure 4.3 shows a section of the data captured on a spreadsheet.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U
1	id	userIP	url	openTime	closeTime	total_user	page_heig	total_copy	total_user	total_mou	total_mou	total_activ	total_mou	velocity_t	total_mou	average_r	total_key	user_fami	user_diffi	user_ratir	cookie_id
2	1	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	0	2456	0	0	153	400	20	457	7	4566	652	0	1	3	5	100
3	2	194.66.32	http://tut	Thu, 06 N	Thu, 06 N	0	3920	0	0	95	361	24	379	5	3778	756	2	1	3	5	100
4	3	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	0	868	0	0	1746	3010	13	3728	50	37262	745	0	0	0	5	100
5	4	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	0	934	0	0	140	320	20	351	2	3507	1754	0	2	4	5	100
6	5	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	1	0	0	0	1789	1637	55	2754	37	27579	745	1	1	3	3	35
7	6	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	1	1942	1	0	1587	1047	75	2048	36	20482	569	1	1	3	3	35
8	7	194.66.32	http://ww	Thu, 06 N	Thu, 06 N	0	0	0	0	1964	1094	11	2300	17	23003	1353	0	1	3	3	35
9	8	194.66.32	http://en.	Thu, 06 N	Thu, 06 N	0	3027	0	0	1819	1512	482	2763	25	27623	1105	0	1	3	5	1
10	9	194.66.32	http://en.	Thu, 06 N	Thu, 06 N	0	2444	0	97	4220	3943	15	7453	65	74548	1147	0	1	3	4	1

Figure 4.3: Captured data on a Spreadsheet

Every participant visited not less than one document during the experiment. The highest number of documents visited in a single task situation by a participant was 11. Pearson correlation and Chi-square test were employed in correlating implicit and explicit feedback parameters. Pearson correlation was used for finding a correlation between scale and nominal variables while the Chi-square test was used for finding a correlation between two nominal/dichotomous variables. The mean differences of the implicit features based on user ratings for relevance, familiarity and difficulty were determined by the Independent T-Test. A confidence interval of 95% was used for analysing the data.

4.4.1 Relationship between Implicit Indicators and Explicit Relevance Ratings

4.4.1.1 Pearson Correlation for Implicit Indicators and Relevancy Ratings

A Pearson correlation was run to examine the relationship between the explicit relevance rating and implicit indicators. Initial analysis of linearity shows that the implicit variables are linearly related with the explicit relevance ratings. There was positive correlation between the explicit relevance rating and the number of mouse clicks (r = 0.211),

amount of copy (r = 0.28), the amount of scroll (r = 0.123), the mouse movement along X-axis (r = 0.225) and Y-axis (r = 0.261), the dwell time (r = 0.285), the mouse distance (r = 0.254) and the mouse duration count (r = 0.238) with significance coefficients of 0.000, 0.000, 0.023, 0.000, 0.000, 0.000, 0.000 and 0.000 respectively as shown in Table 4.2. Although the relationship between the explicit relevance ratings and the mean mouse speed and number of keystrokes produced a negative correlation (an inverse relationship), the p-values of the two parameters are greater than 0.05 and the null hypothesis was accepted.

The correlation between implicit indicators and explicit relevance ratings obtained in this work is higher than those obtained in previous research. Guo and Agichtein (2012) estimated document relevance from dwell time, cursor movements and other post-click behaviour and they obtained a correlation of 0.167 for dwell time, 0.101 for mouse movement along X axis, 0.172 for mouse movement along Y axis, -0.143 for mouse speed along the X axis, -0.124 for mouse speed along the Y axis and -0.008 for amount of vertical scroll. They described the correlation for dwell time as moderate.

Table 4.2: Pearson correlation between the implicit indicators and explicit relevant ratings

Implicit Indicators	Pearson Correlation (*r*) with User Explicit Rating	Significant coefficient level (*p*)
Number of Mouse Clicks	0.211	0.000
Page Height	0.032	0.557
Amount of Copy	0.286	0.000
Amount of Scroll	0.123	0.023
Mouse Movement X	0.225	0.000
Mouse Movement Y	0.261	0.000
Dwell Time	0.285	0.000
Mouse Distance	0.254	0.000
Mouse Duration Count	0.238	0.000
Mean Mouse Speed	-0.73	0.180
Number of Keystrokes	-0.18	0.742

The result is consistent with previous assumptions. Figure 4.4 to 4.13 show boxplots of the relationship between the implicit indicators and explicit ratings of relevance. The X-axis represents the user explicit relevancy ratings are and the Y-axis represents the implicit indicators.

i. ***Active Time Spent on the Document (DT):*** In figure 4.4, the boxplot shows the minimum value (lower whisker) for rating 0 to 2 to be slightly above 1 second and the maximum value (upper whisker) for the three ratings (without the outliers) is within 100 to 150 seconds. The lower quartile is within 10 to 15 seconds and the upper quartile is between 50 to 70 seconds.

   The ratings for 3 to 5 show higher lower whisker values than that of 0 to 2 and the upper whisker values are between 350 to 550 seconds. The lower quartile is within 20 to 30 seconds and the upper quartile is within 180 to 230 seconds. Also, the median for the box plots increases with the ratings, indicating that the more time a user

spends on a document, the more relevant it is to a user. This is consistent with previous research (Kellar et al. 2004, Lee, Park and Park 2008, Morita and Shinoda 1994, Zhu et al. 2010).

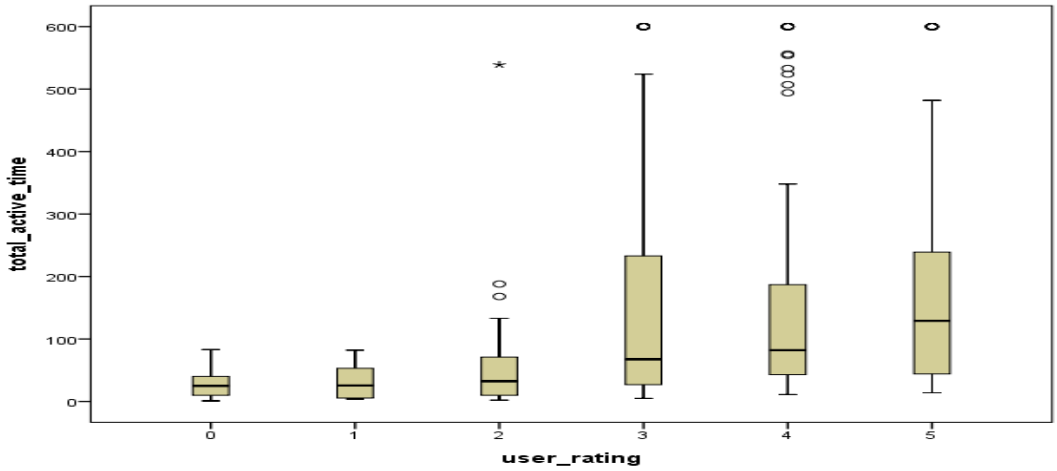


Figure 4.4: Dwell time VS Explicit Ratings

ii. ***Number of Mouse Clicks (NMC):*** As shown in Figure 4.5, the total user clicks increase with the explicit relevance ratings. The lower quartile, median, upper quartile and upper whiskers of the boxes increase progressively as the explicit rating increases. This satisfies our hypothesis that the more mouse clicks, the more relevant the document is. This is consistent with previous research by (Claypool et al. 2001, Huang, White and Dumais 2011, Kim and Chan 2005).

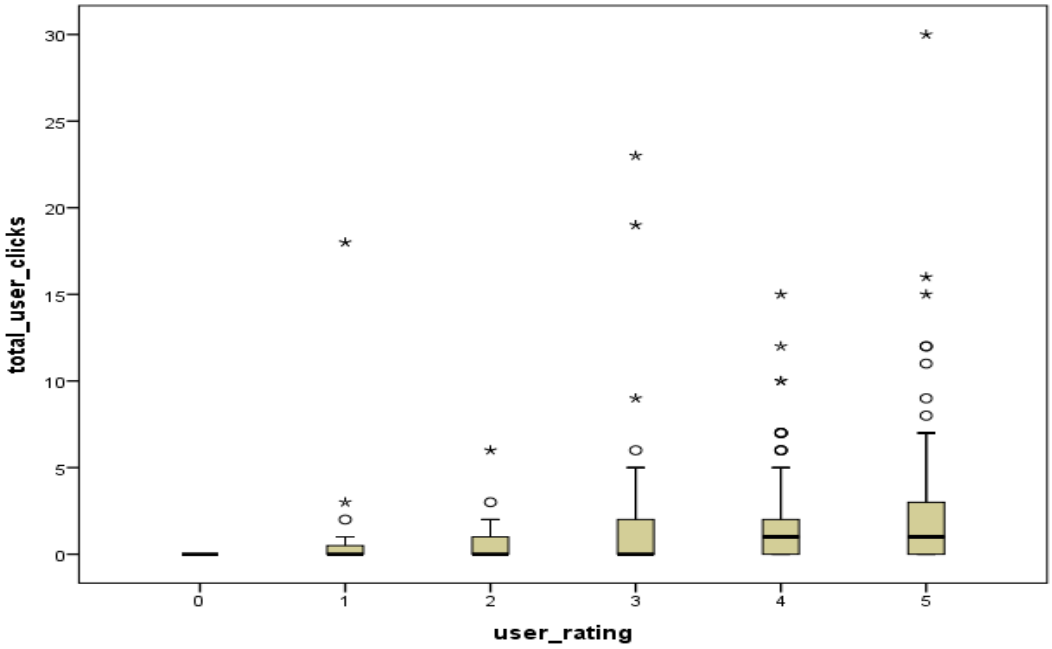


Figure 4.5: User Clicks VS Explicit Ratings

iii. ***Mouse Movement (MM):*** The box plots in Figures 4.6 and 4.7 show how the mouse movement along the Y-axis and X-axis relate to the explicit relevance ratings. The lower quartile, median, upper quartile and upper whiskers of the boxes increase progressively as the explicit rating increases. The test for significance also rejected the null hypothesis, meaning that the more the mouse movement, the more the relevance. This supports the research by Guo and Agichtein (2012).

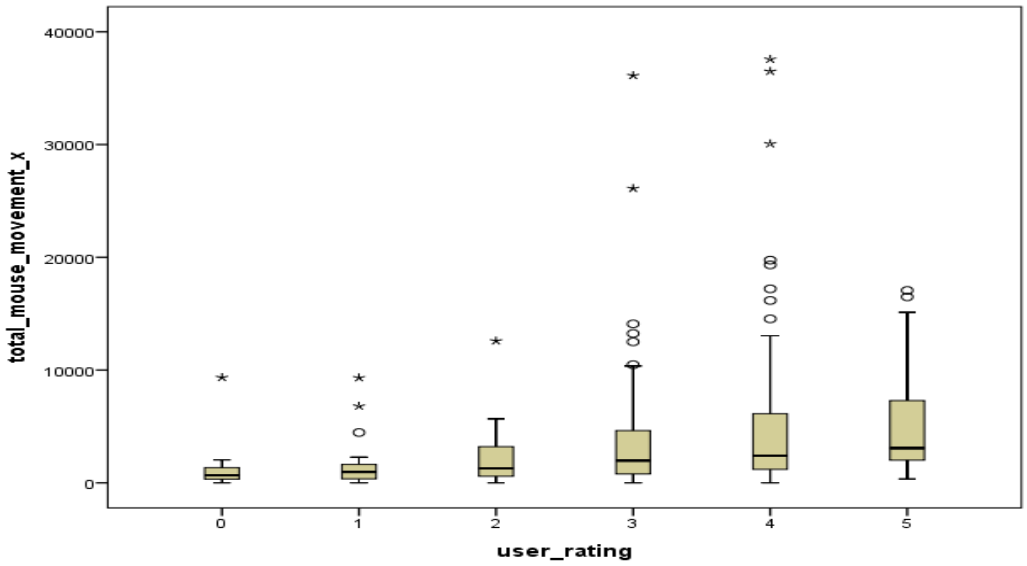


Figure 4.6: Mouse_Move_X VS Explicit Ratings

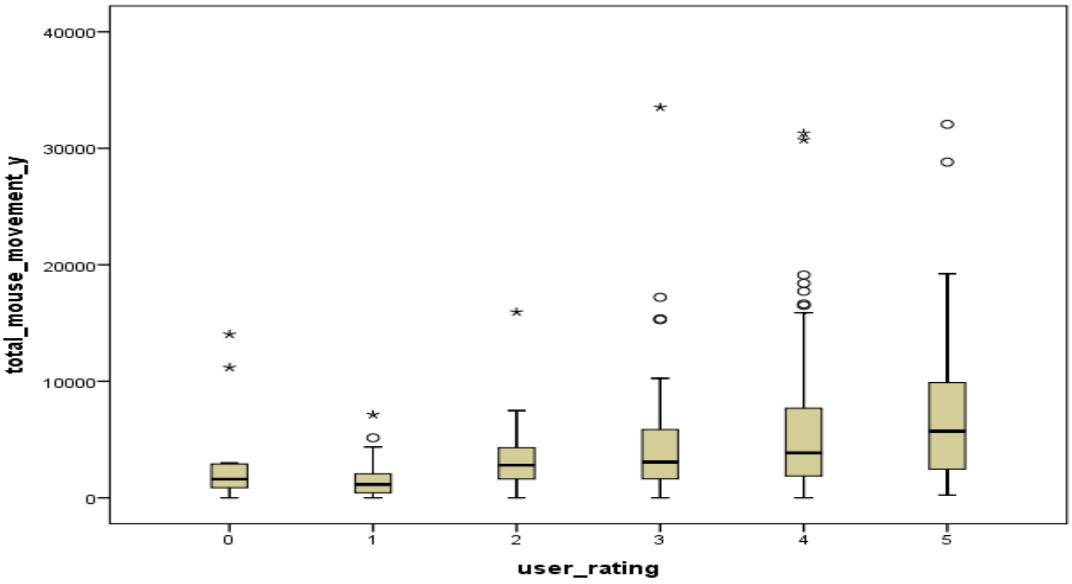


Figure 4.7: Mouse_Move_Y VS Explicit Ratings

iv. ***Mouse Distance (MD):*** As shown in Figure 4.8, the lower quartile, median, upper quartile and upper whiskers of the boxes increase progressively as the explicit rating increases. This indicates that the more distance the mouse moves, the more interesting the page is to

the user. This is in line with previous research by Zemirli (2012) and Kim and Chan (2005).

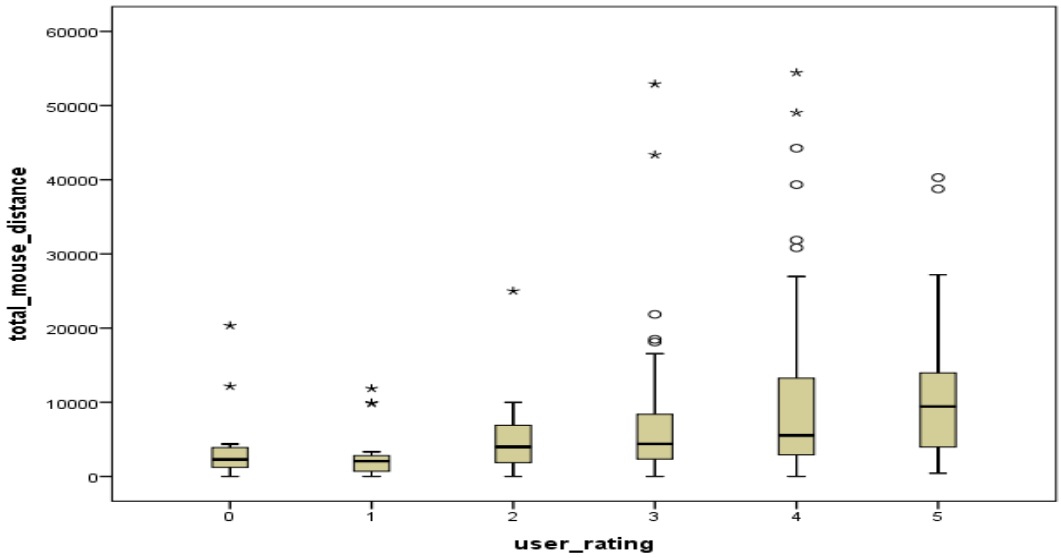


Figure 4.8: Mouse Distance VS Explicit Ratings

v. ***Amount of Scroll (AS):*** Previous research has shown that the greater the amount of scroll, the greater the interest in the document (Claypool et al. 2001, Goecks and Shavlik 2000, Kim and Chan 2005). The diagram in Figure 4.9 shows the relationship between the explicit relevance rating and the amount of scroll. It shows the boxplots increasing steadily in terms of the upper whiskers as the explicit rating increases.

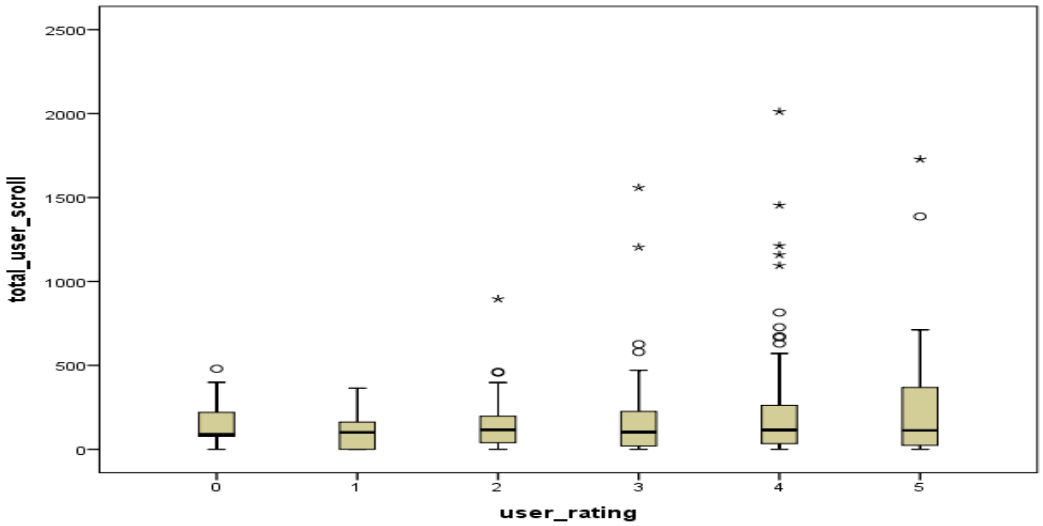


Figure 4.9: Total Scroll VS Explicit Ratings

vi. ***Amount of Copy (AC):*** Copy parameter is a measure of users' interests on documents. In Figure 4.10, the boxplot shows that the more users copy from the document, the more relevant the document is. It shows that documents that were perceived not to be relevant were not copied while documents related relevant were copied. Previous studies like that of Iqbal et al (2012) found that copy is a good measure of interest for Software Developers.

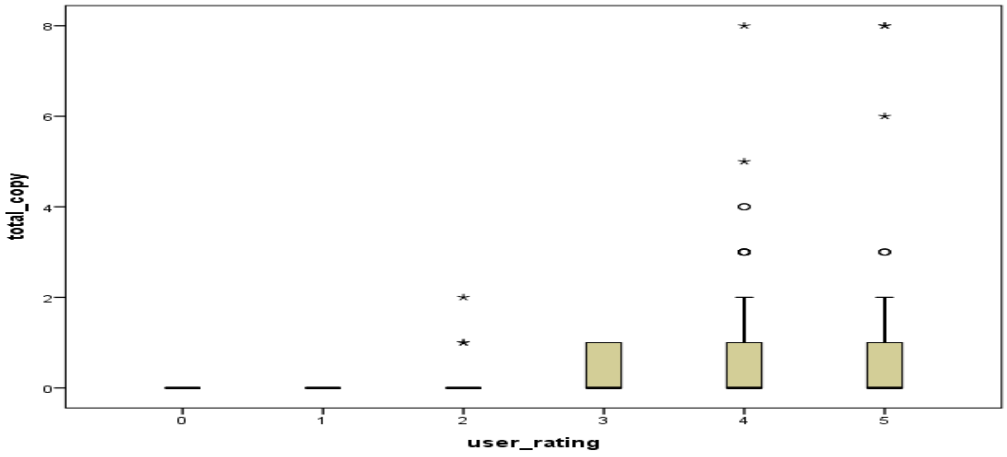


Figure 4.10: Total copy VS Explicit Ratings

vii. ***Mouse Duration Count (MDC):*** The relationship between mouse duration count and the explicit relevance rating is shown in Figure 4.11. It increases positively with the User explicit ratings. The lower quartile, median, upper quartile and upper whiskers of the boxes increase progressively as the explicit rating increases. This is consistent with the result obtained by Claypool et al (2001).

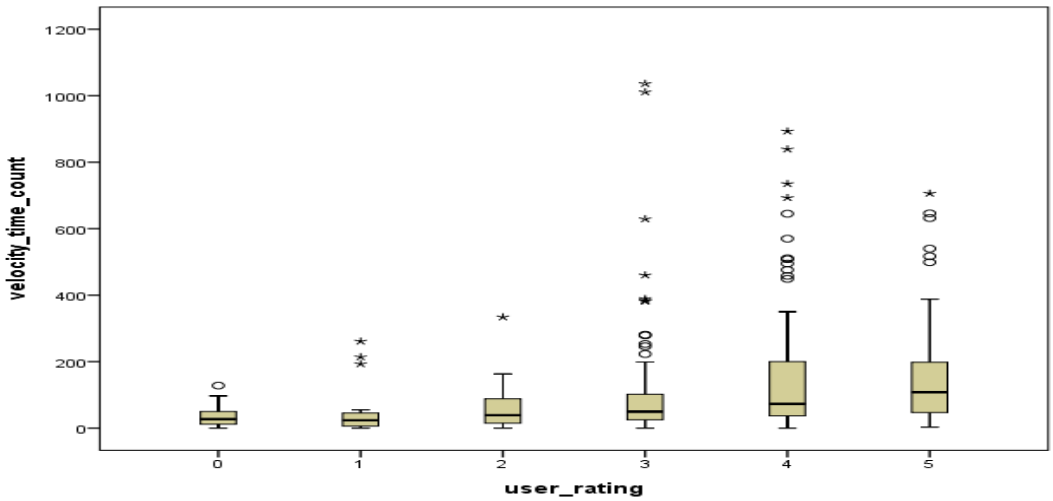


Figure 4.11: Mouse Duration count VS Explicit Ratings

viii. ***Mean Mouse Velocity (MMV):*** Low speed may indicate that a document is of interest and relevant to a user (Guo and Agichtein 2012). Although the relationship between the mean mouse velocity and the explicit ratings has a negative correlation which is consistent with the results of Guo and Agichtein (2012), it is however not statistically significant as the *p*-value was greater than 0.05. Figure 4.12 shows the non-correlated relationship between the mean mouse velocity and user explicit ratings. The box plot varies but the lower quartile, median, upper quartile and upper whiskers of the boxes do not increase/decrease progressively as the explicit rating increases.

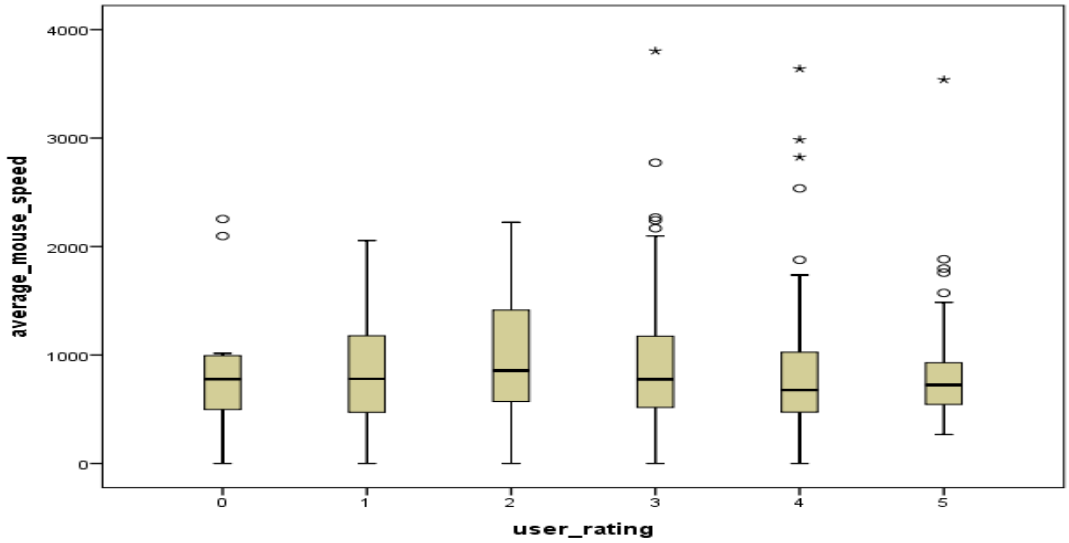


Figure 4.12: Mean Mouse Speed VS Explicit Ratings

ix. ***Number of Keystrokes (NK):*** The Keystroke was found not to be significantly related to user explicit ratings as can be seen in Figure 4.13. The boxes are not visible, showing almost the same lower quartile. Claypool et al (2001) and Kim and Chan (2005) didn't find this parameter to be a predictor of interest.

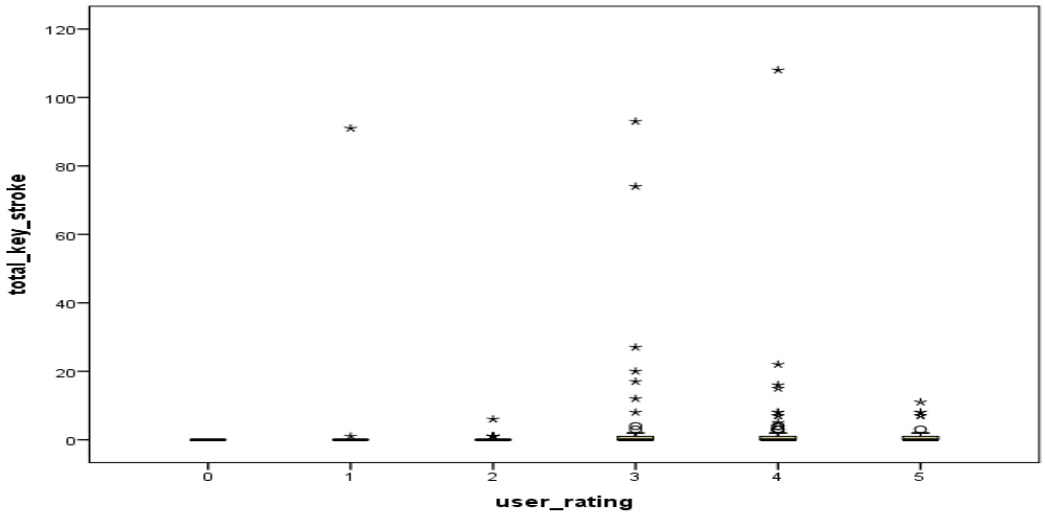


Figure 4.13: Total Keypress VS Explicit Ratings

4.4.1.2 Independent T-Test for Relevancy groupings

In order to find whether there is a statistical difference between documents rated as relevant from documents rated as non-relevant, the 6-point scale explicit ratings were merged into two: relevant and non-relevant. Ratings 0, 1 and 2 were merged as non-relevant while ratings 3, 4 and 5 were merged as relevant. Independent T-Test was used to compare the mean of the two groups of relevancy based on the implicit parameters. There was a significant difference ($p < 0.05$) in the mean for the following variables: number of mouse clicks (*p=0.000)*, the amount of copy (*p=0.013)*, the amount of scroll (*p=0.000)*, the mouse movement along X-axis (*p=0.000)* and Y-axis (*p=0.000)*, the dwell time (*p=0.000)*, the mouse distance (*p=0.000)* and the mouse duration count (*p=0.000)*. The mean of the indicators for the relevant group was higher than the non-relevant group as can be seen in Table 4.3. This signifies that the users focused on documents perceived to be relevant than those perceived to be non-relevant. This satisfies our hypothesis that the more relevant the documents, the greater the user generated implicit indicators.

Table 4.3: Comparison of implicit indicators based on relevancy groups (mean and median of relevant and non-relevant groups and *p*- values of the Independent T-Test)

Implicit Indicators	Mean Not Relevant (N=79)	Mean Relevant (N=264)	Median Not Relevant (N=79)	Median Relevant (N=264)	T-TEST (*p*)
Mouse Clicks	0.59	2.0	0.00	1.00	0.000
Page Height	4494.96	4370.76	2636.00	2856.00	0.665
Copy	0.05	0.66	0.00	0.00	0.000
Scroll	140.12	204.27	101.0	111.0	0.013
Mouse movement X	1827.35	4632.80	1015.00	2443.00	0.000
Mouse movement Y	2785.24	5540.16	2033.00	3848.50	0.000
Dwell time	46.96	164.24	27.00	86.00	0.000
Mouse distance	3980.39	8684.45	2684.00	5496.50	0.000
Mouse duration count	53.19	144.82	29.00	72.50	0.000
Mean mouse speed	927.82	853.74	806.00	742.50	0.201
Keystroke	1.38	2.29	0.00	0.00	0.473

4.4.2 Relationship between Document Familiarity and Document Difficulty on User Behaviour

4.4.2.1 Pearson Correlation for the Implicit Indicators and Document Familiarity Ratings

Only documents rated for familiarity were analysed. Among the 324 rated by the participants, 203 were rated as not familiar while 121 were rated as familiar. Pearson correlation was run on the dataset to examine if there is a correlation between the implicit indicators and the user ratings for familiarity. The result, as shown in Table 4.4, indicates that there is no statistically significant correlation between the implicit indicators and

the user ratings for familiarity. The non-correlation between the implicit indicators and familiarity might be due to the time given for the experiment; users read the documents regardless of whether they were familiar or not while performing the tasks.

Table 4.4: Pearson correlation between the implicit indicators and document familiarity ratings

Implicit Indicators	Pearson Correlation(r) with Document Familiarity rating	Significant coefficient level (*p*)
Mouse Clicks	0.008	0.882
Page Height	0.085	0.128
Amount of Copy	-0.009	0.878
Amount of Scroll	0.079	0.154
Mouse movement X	0.027	0.634
Mouse movement Y	0.031	0.583
Dwell time	0.98	0.08
Mouse distance	0.035	0.534
Mouse duration count	0.072	0.196
Mean mouse speed	0.001	0.980
Keystroke	0.057	0.305

The Independent T-Test conducted also did not reject the null hypothesis. Table 4.5 shows that although the mean ratings for familiar and non-familiar documents appeared different, it was not statistically significant.

Table 4.5: Comparison of implicit indicators based on document familiarity groups (mean and median of relevant and non-relevant groups and *p*- values of the Independent T-Test)

Implicit Indicators	Mean Not Familiar (N=203)	Mean Familiar (N=121)	Median Not Familiar (N=203)	Median Familiar (N=121)	T-TEST (*p*)
Mouse Clicks	1.70	1.64	0.00	0.00	0.882
Page Height	4194.36	4742.67	3191.0	4065.0	0.128
Copy	0.53	0.51	0.00	0.00	0.878
Scroll	173.29	217.02	107.0	110.0	0.196
Mouse movement X	3882.77	4175.37	2074.0	2300.0	0.634
Mouse movement Y	4787.45	5118.16	3034.0	3595.0	0.583
Dwell time	124.665	158.901	57.0	80.0	0.08
Mouse distance	7381.16	7994.09	4269.0	4981.0	0.534
Mouse duration count	114.51	139.63	55.0	63.0	0.196
Mean mouse speed	870.06	871.67	770.0	724.0	0.980
Keystroke	1.62	2.86	0.00	0.00	0.305

4.4.2.2 Pearson Correlation for the Implicit Indicators and Document Difficulty Rating

Unlike the test for familiarity where there was no correlation between the implicit indicators and user familiarity, in the case of document difficulty, there is a correlation between the mouse activities and the rating for document difficulty on the 317 documents rated by the users. The Mouse movement along the X and Y axes and the mouse distance correlated negatively with the ratings for document difficulty with the correlation of coefficient as -0.111, -0.115 and -0.119 respectively. This indicates that the greater the difficulty of the document, the less movement of mouse

activity. Table 4.6 shows the Pearson correlation between the implicit indicators and difficulty ratings.

Table 4.6: Pearson correlation between the implicit indicators and document difficulty ratings

Implicit Indicators	Pearson Correlation(r) with Document difficulty rating	Significant coefficient level (*p*)
Mouse Clicks	-0.057	0.310
Page Height	0.111	0.49
Amount of Copy	-0.096	0.088
Amount of Scroll	-0.052	0.359
Mouse movement X	-0.111	0.049
Mouse movement Y	-0.115	0.04
Dwell time	-0.079	0.158
Mouse distance	-0.119	0.034
Mouse duration count	-0.072	0.202
Mean mouse speed	-0.086	0.125
Keystroke	-0.007	0.897

Independent T-Test was used to compare the means of the implicit indicators for documents rated as difficult to understand and those rated as non-difficult to understand. Although the results showed a difference in mean for all the indicators measured, the mean difference for the copy indicator, mouse movement along X and Y axes of the screen, the mouse distance, the mouse duration count and the mean mouse speed were, however, significant with their *p*-values as 0.008, 0.004, 0.003, 0.002 and 0.043 respectively. Users copied and performed more mouse activities on documents they considered not difficult than on documents they considered difficult to understand, as highlighted in Table 4.7. This indicates that when users find a document difficult to understand, they

perform fewer activities on the document and move swiftly to documents they can easily comprehend.

Table 4.7: Comparison of implicit indicators based on document difficulty groups (mean and median of relevant and non-relevant groups and *p*- values of the Independent T-Test)

Implicit Indicators	Mean Not Difficult (N=267)	Mean Difficult (N=50)	Median Not Difficult (N=267)	Median Difficult (N=50)	T-TEST (*p*)
Mouse Clicks	1.79	1.24	1.00	0.00	0.31
Page Height	4251.83	5203.32	3464.00	3987.50	0.116
Amount of Copy	0.57	0.28	0.00	0.00	0.008
Amount of Scroll	198.03	159.86	113.00	108.50	0.359
Mouse movement X	4235.16	2621.54	2372.00	1542.50	0.004
Mouse movement Y	5191.97	3527.16	3318.00	2642.00	0.003
Dwell time	144.610	107.320	73.00	46.00	0.158
Mouse distance	8063.30	5258.18	5038.0	3312.50	0.002
Mouse duration count	129.10	95.76	61.00	47.00	0.202
Mean mouse speed	892.30	761.90	755.0	737.50	0.043
Keystroke	2.15	1.94	0.00	0.00	0.897

4.4.3 The Effect of Document Familiarity and Document Difficulty on Explicit Relevance Ratings

4.4.3.1 Chi-Square Test (Categorical Data Comparison) for Document Familiarity and Explicit Relevance Ratings

As shown in Table 4.8, User explicit relevance rating has a relation with the rating for familiarity. The Chi-square test shows a statistically significant relationship of 13.57 (p = 0.019). The Independent T-Test

depicts the difference in mean values between non-familiar and familiar groups based on explicit ratings for relevance. The results show that users rated documents they were familiar with higher than those they were not familiar with.

Table 4.8: Relationship between explicit relevance ratings and familiarity ratings

	Mean Not familiar (N=203)	Mean familiar (N=121)	Median Not Familiar (N=203)	Median Familiar (N=121)	Chi-Square Test with Familiarity Rating	T-TEST (*p*)
Explicit relevance ratings	3.182	3.694	3.0	4.0	13.57, p = 0.019	0.000

4.4.3.2 Chi-Square Test (Categorical Data Comparison) for Document Difficulty and Explicit Relevance Ratings

The Chi-square test did not produce a significant relationship between the rating of a document's difficulty and the rating for relevance. The Pearson Chi-Square test was 7.63 (p = 0.178). The null hypothesis of no relationship was accepted because the p-value is greater than 0.05. The result of the Independent T-Test showed a significant difference between the mean for documents rated as difficult and those rated as not difficult based on the ratings for relevancy as shown in Table 4.9.

Table 4.9: Relationship between explicit relevance ratings and difficulty ratings

	Mean Not difficult (N=267)	Mean difficult (N=50)	Median Not Difficult (N=267)	Median Difficult (N=50)	Chi-Square Test with Difficulty Rating	T-TEST (*p*)
Explicit relevance ratings	3.442	3.04	4.0	3.0	7.63, p =0.178)	0.047

4.4.4 The Effect of Task Type on User Behaviour

This section investigates the effect of task type on user searching behaviour. Two task types with a high number of participants were used as case studies in evaluating task effect on user behaviour. The tasks (Task 1 and Task 2) performed by the participants is presented in Section 3.3.8.1 and explained in Section 3.3.8.2.

4.4.4.1 Mixed task result (Task 1)

Pearson correlation and Independent T-Test were used for analysing the data. The analysis was similar to the one carried out in Sections 4.4.1 to 4.4.3. The results obtained in the Mixed task for explicit relevance ratings and document familiarity ratings were consistent with those obtained overall. In the explicit relevance rating category for the mixed task, among the 206 documents visited, 159 were rated relevant while 47 were rated non-relevant. Apart from the keystroke and mean mouse speed, all the other implicit indicators correlated significantly with the explicit ratings and the Independent T-test showed that the more a user engages in a web document, the more relevant the document is to the user as shown in Table 4.10.

In terms of document familiarity as shown in Table 4.11, among the 197 documents rated for familiarity, 106 were rated not familiar and 91 were rated familiar. There was no significant correlation between the rating for document familiarity and user generated implicit indicators. The Independent T-Test also did not produce any significant difference in mean between documents rated familiar and those rated non-familiar with respect to the implicit indicators.

In terms of document difficulty, among the 192 documents rated, 34 were rated as difficult to understand and 158 were rated as not difficult to understand. There was no significant correlation between the Implicit indicators and ratings for document difficulty, but the Independent T-test

shows significant difference between the two groups of difficulty for the following features: Mouse movement along Y-axis (p = 0.010), total distance (p =0.009) and Mouse movement along X-axis (p =0.021), as can be seen in Table 4.12. The difference in mean, as highlighted in the two groups in Table 4.12, indicates that users of Task 1 (Mixed Task) moved the mouse more on documents they considered not difficult to understand than documents they considered to be difficult. It may be that users were mindful of the time given for the experiment and they paid more attention to documents that were not difficult to understand.

The Chi-square test between the ratings for relevance and document familiarity produced a significant relationship. The T-test also showed a statistically significant difference between the mean familiar (M = 3.703) and mean unfamiliar (M = 3.038) as shown in Table 4.13. There was, however, no significant relationship between the user ratings for relevance and the ratings for document difficulty. Although the mean for the two groups of difficulty differs, it was however not statistically significant, as can be seen in Table 4.14.

4.4.4.2 Factual task result (Task 2)

Person correlation and Independent T-Test were used for analysing the 100 documents rated in the relevancy category as shown in Table 4.10. Among the 100 documents, 80 were rated as relevant while 20 were rated as non-relevant. Apart from the amount of scroll which was not significant in the Factual task, the results obtained from the Factual task are similar to those obtained for the Mixed task. This indicates that in both tasks, the mouse clicks, amount of copy, the mouse movement along X and Y axes, the dwell time, the mouse distance and the mouse duration count were significantly correlated with the explicit relevance ratings and their means for the relevant ratings were greater than the means for non-relevant ratings, as shown by the T-Test in table 4.10.

In terms of document familiarity, among the 92 documents, 63 were rated as not familiar while 29 were rated as familiar. Like the mixed task, no significant correlation between the implicit indicators and the ratings for familiarity was found. However, the T-test as shown in Table 4.11 produced a significant difference in mean for the copy parameter between documents rated familiar from those rated unfamiliar. It showed that texts were copied from documents considered to be unfamiliar than those considered to be familiar.

In terms of ratings for document difficulty, 92 documents were rated by the users. Among these, 12 were rated as difficult to understand and 77 were rated not difficult to understand. There was no correlation between the implicit indicators and the ratings for difficulty. Only the copy parameter was significantly different for the difficult and non-difficult groups as shown in Table 4.12. This suggests that in a Factual task, users copied text on documents that were rated not difficult more than documents that were rated difficult.

In the case of the relationship between the user explicit ratings for relevance and that of familiarity, the Chi-Square test showed no significant correlation between the familiarity ratings and the relevance ratings as can be seen in Table 4.13. The T-Test also did not produce a significant result. The Chi- Square Test between the ratings for relevance and that of difficulty was also not significant. The T-Test was, however, significant (p = 0.035). It showed that the greater the difficulty the lower the ratings for relevance as depicted in Table 4.14.

Table 4.10: Task Specific grouping of the relationship between implicit indicators and explicit relevance ratings

Implicit Indicators	**Mixed task**				**Factual Task**			
	Mean Not Relevant (N=47)	Mean Relevant (N=159)	Pearson Correlation (r)	T-TEST (*p*)	Mean Not Relevant (N=20)	Mean Relevant (N=80)	Pearson Correlation(r)	T-TEST (*p*)
Clicks	0.72	1.82	***0.149, p =0.033***	***0.025***	0.45	2.33	**0.270, *p* = 0.006**	**0.000**
Height	4429.94	4962.52	0.114, *p* =0.103	0.279	5012.75	3491.13	-0.156, *p*= 0.117	0.219
Copy	0.06	0.57	***0.329, p =0.000***	***0.000***	0.0	0.79	***0.268, p =0.007***	***0.000***
Scroll	136.40	223.55	***0.156, p =0.025***	***0.005***	142.0	159.96	*0.008, p =0.939*	*0.717*
Mouse movement X	1859.06	4689.75	***0.201, p =0.004***	***0.000***	2351.45	4903.51	***0.251, p =0.011***	***0.002***
Mouse movement Y	2761.60	5873.92	***0.295, p =0.000***	***0.000***	2715.05	5258.68	***0.206, p =0.038***	***0.027***
Dwell time	35.532	159.472	***0.295, p =0.000***	***0.000***	69.150	187.732	***0.242, p =0.014***	***0.001***
Mouse distance	4007.02	9083.76	***0.263, p =0.000***	***0.000***	4252.25	8551.71	***0.229, p =0.021***	***0.002***
Mouse duration count	52.87	153.79	***0.220, p =0.001***	***0.000***	57.30	146.39	***0.264, p =0.007***	***0.000***
Mean mouse speed	905.55	836.09	-0.055, *p* =0.430	0.453	893.90	892.71	-0.076, *p* =0.447	0.994
Keystroke	2.04	2.44	-0.036, *p* =0.603	0.853	4.55	1.20	-0.144, *p* =0.148	0.471

Table 4.11: Task Specific grouping of the relationship between implicit indicators and document familiarity ratings

Implicit Indicators	**Mixed task**				**Factual Task**			
	Mean Not Familiar (N=106)	Mean Familiar (N=91)	Pearson Correlation (r)	T-TEST (*p*)	Mean Not Familiar (N=63)	Mean Familiar (N=29)	Pearson Correlation(r)	T-TEST (*p*)
Clicks	1.75	1.44	-0.46, p = 0.519	0.519	2.17	1.52	-0.074, p =0.483	0.483
Height	4509.84	5062.33	0.094, p =0.187	0.187	3627.48	3910.55	0.040, p =0.708	0.708
Copy	0.42	0.53	0.074, p =0.302	0.302	0.86	0.28	-0.160, p =0.128	***0.039***
Scroll	183.76	213.74	0.057, p =0.427	0.427	157.63	126.07	-0.075, p =0.477	0.477
Mouse movement X	3934.44	4343.33	0.034, p =0.633	0.633	4799.49	3644.55	-0.114 p =0.279	0.279
Mouse movement Y	5117.63	5243.71	0.011, p =0.874	0.874	4775.65	3840.79	-0.098, p =0.353	0.353
Dwell time	116.274	151.242	0.104, p =0.145	0.145	176.143	144.103	-0.081, p =0.444	0.444
Mouse distance	7761.75	8276.45	0.028, p =0.698	0.698	8043.49	6221.14	-0.113, p =0.282	0.282
Mouse duration count	123.91	143.78	0.053, p =0.458	0.458	137.40	106.76	-0.099, p =0.347	0.347
Mean mouse speed	825.88	857.53	0.029, p =0.690	0.690	863.86	990.76	0.097, p =0.358	0.358
Keystroke	1.39	3.55	0.088, p =0.218	0.236	1.19	3.90	0.128, p =0.226	0.404

Table 4.12: Task Specific grouping of the relationship between implicit indicators and document difficulty ratings

Implicit Indicators	**Mixed task**				**Factual Task**			
	Mean Not Difficult (N=158)	Mean Difficult (N=34)	Pearson Correlation(r)	T-TEST (*p*)	Mean Not Difficult (N=77)	Mean Difficult (N=12)	Pearson Correlation(r)	T-TEST (*p*)
Clicks	1.72	1.21	-0.059, p = 0.415	0.415	2.0	1.50	-0.047, p = 0.660	0.660
Height	4697.97	5289.65	0.077, p =0.288	0.368	3632.45	4464.83	0.085, p =0.429	0.429
Copy	0.50	0.32	-0.089, p =0.219	0.219	0.77	0.08	-0.137, p =0.202	**0.003**
Scroll	213.99	137.91	-0.109, p =0.131	0.131	153.04	126.08	-0.047, p =0.665	0.665
Mouse movement X	4401.06	2678.26	-0.111, p =0.127	**0.021**	4605.95	2479.75	-0.159, p =0.136	0.136
Mouse movement Y	5523.84	3593.12	-0.132, p =0.068	**0.010**	4727.90	2335.08	-0.184, p =0.084	0.084
Dwell time	142.051	94.147	-0.108, p =0.135	0.135	178.052	99.833	-0.143, p =0.181	0.181
Mouse distance	8546.61	5401.82	-0.130, p =0.073	**0.009**	7837.13	3991.25	-0.178, p =0.095	0.095
Mouse duration count	139.13	101.74	-0.076, p =0.292	0.292	130.19	83.75	-0.113, p =0.294	0.294
Mean mouse speed	864.03	729.09	-0.093, p =0.199	0.072	954.87	655.83	-0.166, p =0.120	0.120
Keystroke	2.89	0.35	-0.078, p =0.281	0.281	2.42	0.08	-0.080, p =0.459	0.459

Table 4.13: Task Specific grouping of the relationship between explicit relevance rating and document familiarity ratings

	Mixed task				**Factual Task**			
	Mean Not Familiar (N=106)	Mean Familiar (N=91)	Chi- Square Test with Familiarity Rating	T-TEST (*p*)	Mean Not Familiar (N=63)	Mean Familiar (N=29)	Chi- Square Test with Familiarity Rating)	T-TEST (*p*)
Explicit User Ratings	3.038	3.703	**19.51, p =0.002**	**0.000**	3.508	3.483	0.629, p =0.987	0.933

Table 4.14: Task Specific grouping of the relationship between explicit relevance rating and document difficulty ratings

	Mixed task				**Factual Task**			
	Mean Not Difficult (N=158)	Mean Difficult (N=34)	Chi- Square Test with Difficulty Rating	T-TEST (*p*)	Mean Not Difficult (N=77)	Mean Difficult (N=12)	Chi- Square Test with Difficulty Rating	T-TEST (*p*)
Explicit User Ratings	3.399	3.059	4.62, *p* =0.464	0.155	3.6323	2.750	8.365, *p* =0.137	**0.035**

4.4.5 Consistency in Explicit Relevance Rating

Since the goal of this research is to obtain common consistent implicit indicators that can be used to represent users' interest, the consistency of the participants' explicit relevance ratings was also examined. Common documents visited by the users were extracted from the pool of documents visited. Among the common documents extracted, the most viewed document had 21 hits while the least had 2 hits. An investigation was carried out to examine if documents commonly visited are relevant. The explicit relevance ratings of the common documents visited have a mean of 3.21 and a median of 3.32. The explicit rating scale, as explained in Section 3.3.7, states that when a document is rated "3", the document is relevant. This indicates that common documents captured across all user populations can be used to infer relevance among a community of users. The full table of the common documents visited together with their occurrences and mean ratings are listed in Appendix C. Figure 4.14 shows a graph of the relationship between the mean ratings and document number of hits.

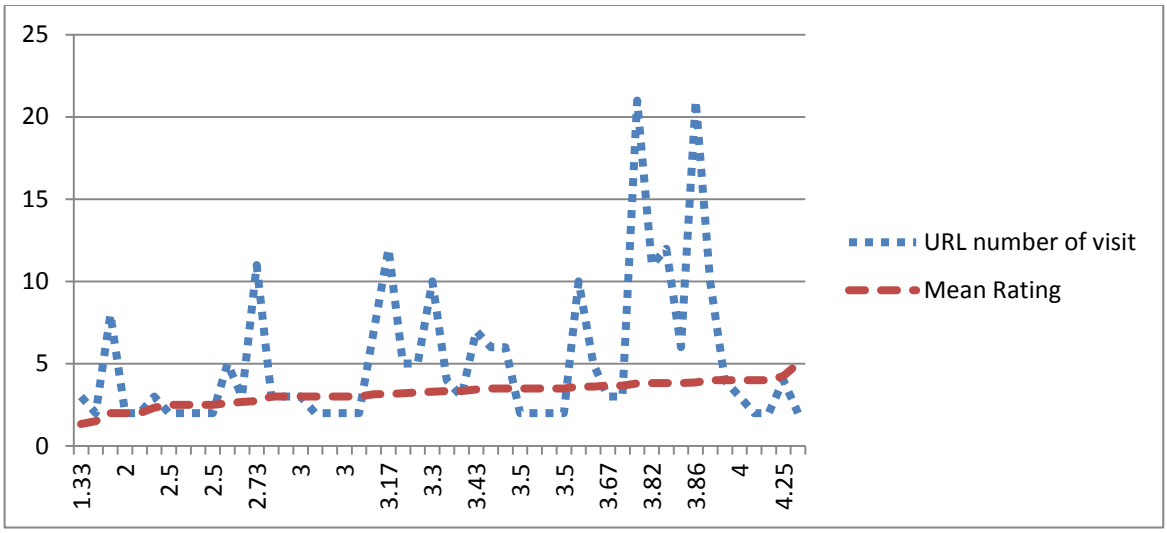


Figure 4.14: Graph of common documents visited by users with the mean ratings

4.5 **Implicit Predictive Model**

The aim of this section is to develop a predictive model from user behaviour and evaluate it using quantitative standard evaluation metrics. Linear regression machine learning technique is employed for developing the predictive model. The predictive strength of models derived from single indicators is compared with the model derived from aggregating implicit indicators. The aim is to answer the research question of whether a model derived from aggregating implicit indicators can effectively represent users' interest on web documents better than models derived from a single indicator. The standard metrics for this evaluation are the Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and Coefficient of Correlation.

The predictive strength of some classifiers is also compared in this chapter. Weka tool (Hall et al. 2009) is employed to classify documents according to relevance and to evaluate the linear regression model. The following metrics are used for comparing the classifiers: recall, precision, f-measure and accuracy.

4.5.1 Linear Regression

Linear regression is a statistical method used to determine the relationship between one or more independent variables and one dependent variable (Yan 2009). The independent variables are referred to as regressors or predictors while the dependent variable is referred to as the outcome variable. It is primarily used for prediction. Regression shows the variation of parameters and attempts to describe in detail the relationship between variables. The model for linear regression has linear regression parameters with the dependent variables represented by 'y' and the independent variables represented by x_1, x_2….x_n.

4.5.1.1 Simple Linear Regression Model

The simple linear regression determines the relationship between one independent variable (x) and one dependent variable (y). It is given as follows:

$$Y = \beta_0 + \beta_1 X \qquad 4.1$$

where Y is the dependent variable, β_0 is the intercept on the y-axis, β_1 is the slope of the line, X is the independent variable.

4.5.1.2 Multiple Linear Regression Model

This is similar to the simple linear regression but it has one dependent variable with two or more independent variables. The goal of multiple linear regression is to examine how two or more independent variables relate to a dependent variable. It is given as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_{2.....} + \beta_n X_n \qquad 4.2$$

where

$\beta_0, \beta_1 ..... \beta_n$ are regression coefficients (unknown model parameters). Y is the dependent variable while X_1, X_2 and X_n are independent variables.

There are some conditions that must be satisfied for a multiple regression test to be valid. These prerequisite assumptions for conducting multiple linear regression analysis include:

No Multicollinearity: This occurs when there is a high correlation between two independent variables. To test this, I inspected the correlation coefficient and the Tolerance/VIF values. It is assumed that if the Tolerance is less than 0.1 and VIF is greater than 10, there is a possibility of collinearity. There was collinearity between total mouse movement along the X axis, Y axis and the total distance covered by the mouse. To resolve the issue of collinearity, the mouse distance feature was removed from the independent features.

Independence of residual errors: Durbin-Watson statistic test was carried out to test if the variables were related. The Durbin-Watson statistics take values from 0 to 4. A value of 1.842 was obtained. It is approximately close to 2 and therefore indicates that there is independence of residual errors.

Linearity between dependent and independent variables: A test was carried out with scattered plot to determine if there is a linear association between the dependent and independent features. An approximately linear relationship was obtained.

Homoscedasticity of residuals: A test for homoscedasticity was also carried out to examine if the residuals are equally spread across the predicted values of the dependent variable. It was observed from the scattered plot that there was homoscedasticity.

Outlier Detection: Outliers are observations that do not follow the normal pattern of points. They are normally far away from other points. Standardised residual was used for detecting outliers. Standardised

residual cases having values that are greater than 3 or less than -3 are regarded as an outlier. All the cases were below 3 and greater than -3.

4.5.2 Evaluation Metrics for Predictive Model

Three standard methods were used for evaluating the predictive model. The metrics used include Root Mean Square Error (RMSE), Mean Absolute Error (MAE) and correlation. The best value for RMSE and MAE of a regression model is 0.0 while the strongest value for positive correlation is 1.

4.5.2.1 Root Mean Squared Error (RMSE)

RMSE is the common formula for measuring regression model error rate. It represents the sample standard deviation of the difference between observed values of the sample and the predictive value. The individual differences between the predictive values and the observed values when the same data sample is used for estimation are called residuals. RMSE aggregates the individual errors during prediction into a single predictive power. It is given as:

$$\mathbf{RMSE} = \frac{\sqrt{\sum_{i=1}^{n}(p_i - a_i)^2}}{n} \qquad 4.3$$

a = actual target, p = predicted target, n = the total number of samples

4.5.2.2 Mean Absolute Error (MAE)

This is similar to root mean squared error. It is used to measure the closeness of predictions to the eventful outcomes. MAE is the average of absolute errors obtained from the difference between the prediction and the true value. It is given as:

$$\mathbf{MAE} = \frac{\sum_{i=1}^{n}|p_i - a_i|}{n} \qquad 4.4$$

a = actual target, p = predicted target, n = total number of samples

4.5.3 Predictive Model

To aggregate the most predictive indicators of interest for a function model, multiple linear regression analysis was used to derive a predictive model of interest from user generated implicit indicators that best represent the user explicit rating. A stepwise regression method was employed for this selection. The stepwise regression automatically selects the predictive variables through a sequence of t-tests. Only the dwell time and amount of copy features were included by the stepwise regression among the nine features entered as inputs. Equation 4.5 shows the predictive model (correlation = 0.36) obtained:

$$\text{Explicit Relevance ratings} = 2.978 + 0.281(\text{Total Copy}) + 0.002(\text{Dwell Time}) \quad 4.5$$

The relationship between the implicit indicators (dwell time and amount of copy) and the explicit ratings in the predictive model has a higher correlation compared to other features examined, and only these features are sufficient to derive the predictive model. A 10 fold cross-validation with a confidence interval of 95% was used to evaluate the predictive model along with single indicator models to estimate how it will perform in "real" practice. The results show that the predictive model performed better than models with the individual indicator. The relationship between the explicit and implicit indicators in the predictive model produced a minimal error and a high correlation coefficient, which is higher than that of the individual indicators examined, as shown in Table 4.15 and Figure 4.15, This suggests that when the features are aggregated, a higher degree of accuracy in prediction of users' interest/document relevance is obtained. This result is consistent with previous findings (Balakrishnan and Zhang 2014, Claypool et al. 2001, Fox et al. 2005).

Table 4.15: Comparison of individual and aggregated predictive model

Implicit Indicators	**Predictive Model (Dependent variable = Explicit rating)**	**Correlation Coefficient (*p*)**	**Mean absolute error**	**Root mean squared error**
Clicks	0.0803 x total_user_clicks + 3.2332	0.19	1.0421	1.2846
Copy	0.3453 x total copy + 3.1912	0.27	1.0173	1.2578
Scroll	0.0006 x total scroll + 3.2522	0.09	1.0602	1.3015
Mouse movement X	0.0001 x movement_x + 3.1436	0.20	1.0416	1.2801
Mouse movement Y	0.0001 x movement_y + 3.0418	0.25	1.0155	1.2667
Dwell time	0.0022 x dwell time + 3.0656	0.26	1.032	1.261
Mouse distance	0 x mouse_distance + 3.0664	0.24	1.0257	1.269
Mouse duration count	0.0019 x velocity_time_count + 3.1369	0.22	1.0313	1.2744
Mean mouse speed	No relationship			
Keystroke	No relationship			
Aggregated Model (Dwell time and Copy)	0.2815 x total copy + 0.0018 x dwell time + 2.9784	0.36	0.9917	1.2259

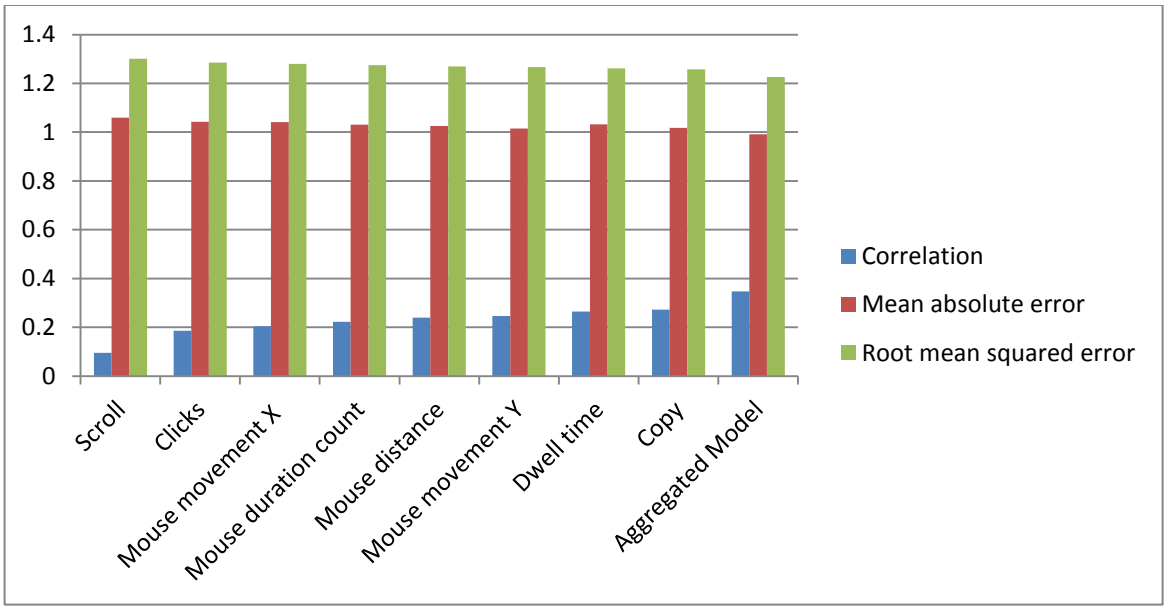


Figure 4.15: Graph showing the performance of the individual and aggregated models

As shown in Figure 4.15, the aggregated model has the lowest root mean squared error and mean absolute error. It also has the highest correlation with the explicit rating compared to the individual indicator models.

4.5.4 Held-out Testing

Held-out method was also used to evaluate the strength of the model. This method is used to "hold back" some data when training the model and to use the data that is held back to test the accuracy of the model. In order to evaluate the model, the whole dataset was divided into two parts for training and testing. 70% of the dataset was supplied for training and a predictive function model was obtained. The model was evaluated with the remaining 30% test data and the results showed an underlying relationship between the independent variables and the dependent variable, thereby limiting overfitting. Results of the training and the test set are given in Table 4.16.

Table 4.16: Held-out evaluation of the predictive model showing the training and testing result

Evaluating measures	Training	Testing
Correlation Coefficient (*p*)	0.36	0.38
Mean absolute error	1.024	1.015
Root mean squared error	1.269	1.2597

4.5.5 Classification Analysis

The focus of this section is to study some classifiers that can be used to group web documents according to relevance. To perform this comparison, the six-point explicit rating scale for relevance was merged into relevant and non-relevant. The explicit ratings for 0, 1 and 2 were merged as non-relevant and ratings 3, 4 and 5 were merged as relevant. WEKA library (Hall et al. 2009) was used to compare the predictive strength of the classifier. The following indicators were used as the input: Dwell Time, Mouse Movement, Mouse Distance, Mouse Clicks, Amount of Scroll, Mean Mouse Speed, Keystroke and amount of Copy to clipboard. The classifiers were run on the dataset to classify the data according to the two groups of relevancy. A 10 fold cross-validation was carried out to estimate the expected level of fitness of the model to the independent data set used for training the model.

4.5.5.1 Evaluation Metrics for Classifiers

The classifiers employed were evaluated with four metrics: Precision, Recall, Accuracy and F-measure. The best value for precision is 1 and the worst value is 0 (Rendle 2010). The best value for recall is 1 and the worst value is 0 (Rendle 2010). The best performance for Accuracy is 1. The best value for the F-measure is 1 and the worst value is 0. Table 4.17 describes the confusion metrics on how the accuracy of the models is derived.

Table 4.17: Table showing Confusion metrics for classification

Accuracy = $\frac{(TP + TN)}{TP + TN + FP + FN}$		Actual result/Classification	
		Yes	No
Predictive results/Classification	Yes	TP(true positive)	FP(false positive)
	No	FN(false negative)	TN(true negative)

TP: This is the number of samples of a class that are correctly classified.

FP: This is the number of samples that are not part of a class but are misclassified into the class.

FN: This is the number of samples of a class that are misclassified in another class.

TN: This is the number of samples not part of a class that are correctly classified.

Precision

The precision of a class in a given classification problem is the number of items that is correctly grouped to belong to the positive class divided by the total number of items belonging to the positive class. A precision of 1.0 for a given class means that all the elements labelled in the class are actual members of the class. It focuses less on the number of elements that are labelled wrongly but more on the exactness of the model. A Low precision suggests that the confusion matrix has a large number of false positives. The precision value is given as:

$$\text{Precision} = \frac{\textbf{number of true positives}}{(\textbf{number of true positive} + \textbf{number of false positives})} \quad 4.6$$

Recall

Recall in this context means the number of true positives divided by all the elements that are actually in the positive class. Recall can be said to

be a measure of the completeness of a classifier. When the recall is low, it suggests that there are many false negatives in the confusion matrix. It is given as:

$$\text{Recall} = \frac{\textbf{number of true positives}}{(\textbf{number of true positives} + \textbf{number of false negative})} \quad 4.7$$

F-Measure

In most classification problems, the recall and precision scores are often not discussed in isolation. They are usually combined into a single measure. A common combination measure is the F-Measure, which is the weighted harmonic mean of the precision and recall. It is given as:

$$\mathbf{F-measure} = \frac{\mathbf{2\ x\ (precision)(recall)}}{(\mathbf{precision} + \mathbf{recall})} \quad 4.8$$

Accuracy

This is the proportion of the total number of true positives and true negatives among the total number of cases investigated. It is given as

$$\mathbf{Accuracy} = \frac{(\mathbf{number\ of\ true\ positives} + \mathbf{number\ of\ true\ negatives})}{(\mathbf{number\ of\ true\ positives} + \mathbf{false\ positives} + \mathbf{false\ negatives} + \mathbf{true\ negatives})} \quad 4.9$$

4.5.6 Classifier Evaluation

4.5.6.1 Binary Logistic Model

The binary logistic model is a probabilistic model that statistically classifiers binary response. It is the type of regression commonly used when the dependent variable of a dataset is binary, that is, it takes two category variables. It measures the relationship between one or more independent variables and a categorical variable. Logistic regression is a special case of the linear regression model for predicting linear outcomes.

The logistic regression was run on the dataset and it produced an accuracy of 60.93. Other matrices measured include: MAE = 0.39, RMSE = 0.61, Precision = 0.73, Recall = 0.61, F-measure = 0.64. Logistic regression produced the lowest accuracy as compared to the other classifiers.

4.5.6.2 J48 Decision Tree

This is a supervised machine learning technique that decides the dependent variable (target value) of a new seed (sample) from attribute values of available data. The attributes are the internal nodes of a decision tree and the branches between the nodes predict the values of the attributes in the observed samples. The terminal nodes denote the last value of the dependent variable. The variables that are used for predicting the dependent variables are called the independent variables.

The building process for J48 Decision tree classifier is given as follows:

- Create decision tree based on the attribute value of training dataset. This is achieved by partitioning the training samples of the data repeatedly into many subsets.
- Try to identify each subset that contains same cases.
- Continue until cases that are 'purer' than the original cases are found.

J48 decision tree produced an accuracy of 76.97 when it was run on the dataset. Other matrices measured include: MAE = 0.35, RMSE = 0.42, Precision = 0.59, Recall = 0.77, F-measure = 0.67. The J4 classifier was higher in accuracy than the logistic regression but slightly lower than the nearest neighbour algorithm, with K as 4.

4.5.6.3 K-Nearest Neighbour

Nearest Neighbour is a supervised machine learning technique that predicts unknown data output of new instances from previously known

input instances and output values. The items are classified based on similarities of its neighbours. The similarity is normally calculated with the Euclidean distance (Ma and Kaban 2013) of the new sample to the previous samples. At K = 4, the model produced an accuracy of 78.72. Other matrices measured include: MAE = 0.32, RMSE = 0.40, Precision = 0.76, Recall = 0.79, F-measure = 0.75. When K is less or greater than 4, the accuracy of J48 classifiers was higher than the nearest neighbour. The most optimal value for nearest neighbour algorithm was obtained when K was 4.

Table 4.18 shows how the classifiers performed. In terms of the matrices measured, the K-nearest neighbour (K=4) performed better in precision, recall, F-measure, RMSE, MAR and Accuracy. It has the highest accuracy of 78.72%, indicating that most of the documents were classified correctly.

Table 4.18: Comparison of machine learning classifiers

Evaluation Metrics	Logistic Regression	J48 Decision	KNN (K=1)	KNN (K=4)
Precision	0.73	0.59	0.706	0.76
Recall	0.61	0.77	0.708	0.79
F-Measure	0.64	0.67	0.707	0.75
Accuracy	60.93%	76.97%	70.85%	78.72%
RMSE	0.61	0.42	0.477	0.40
MAR	0.39	0.35	0.324	0.32

4.5.7 User study 3 Results (Validation study)

The results of user study 2 suggest that user interest in web documents can be inferred from their behavioural activity. The aim of user study 3 is to use eye gaze measures (Fixation count, Fixation duration, Heat map) to

validate the predictive strength of the function model derived in Section 4.5.3, and to show that to a certain degree of accuracy, the model can be used in place of an eye gaze.

The data captured by the eye tracker from one of the participants was excluded from the analysis due to poor calibration which led to incomplete data (see Appendix F for the raw data). Only the remaining 8 participants' results were analysed. The correlation between fixation duration and explicit relevance ratings was not statistically significant. This is consistent with the findings of Buscher et al (2012). There was, however, a statistically significant correlation of 0.32 (p = 0.025) between the fixation count and user explicit relevance ratings.

Further analysis was carried out and the explicit relevance ratings for the 6 documents in user study 2 were correlated with the mean explicit ratings for the documents for the eye gaze study. A very strong correlation of 0.82 (p = 0.045) was obtained, showing consistency in the ratings of the documents by the participants in the user study 2 and eye gaze study. The correlation between the explicit ratings of the predictive function model used for identifying and extracting the documents from the dataset is 0.36. The correlation between the total fixation count and the explicit rating is 0.32. Considering the consistency in the ratings of the documents in the two studies, and the predictive model and fixation count producing similar a correlation coefficient to the explicit ratings (see Table 4.19), it can be inferred that there is no significant difference between the predictive model based on implicit indicators and the eye gaze in the context employed. The predictive model can be used in place of fixation count when an eye tracker is not available. Figure 4.18 is a graph showing the consistency in ratings of the selected documents in the two studies.

Table 4.19: Comparison of predictive model and gaze measure

Parameters	Predictive model	Explicit ratings (user study 3)
Fixation count		0.32
Explicit ratings (user study 2)	0.36	0.82

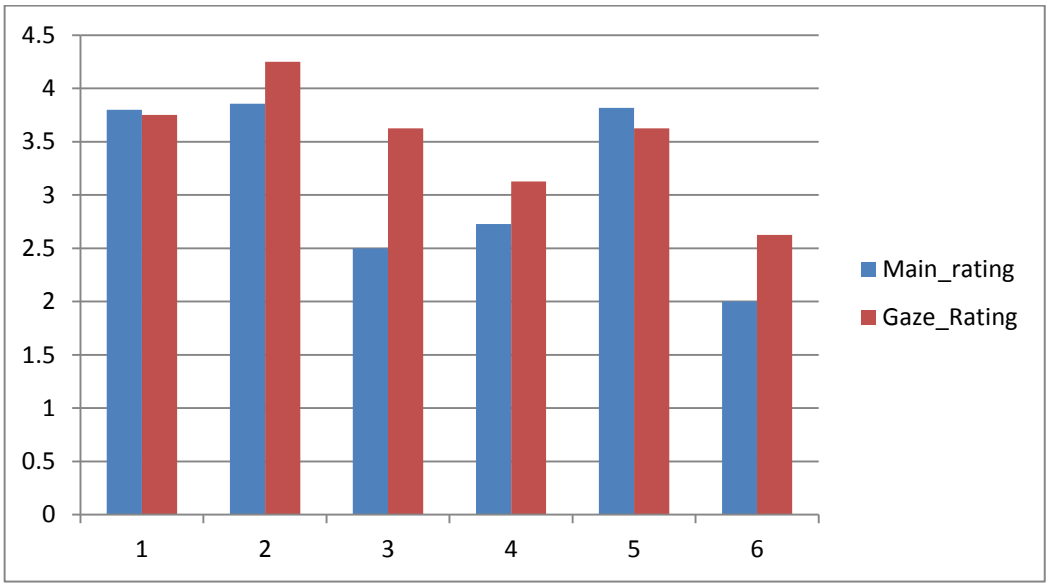


Figure 4.16: Graph showing the explicit ratings of the 6 selected documents in user study 2 and the mean explicit ratings of the documents in the eye gaze study

The qualitative results of the heat map show that documents that were explicitly rated high were denser and had high mean fixation count than documents that were rated low. Figure 4.17 to Figure 4.22 show the heat map for documents 1 – 6. It was observed that documents that were rated high were denser than documents that were rated low. This inference was not true in all cases, as can be seen in Figures 4.19 and 4.22. The ratings for these documents were low but the heat map appears denser because the lengths of the documents were relatively short compared to other documents. This suggests that the height of a document can affect the density of the heat map.

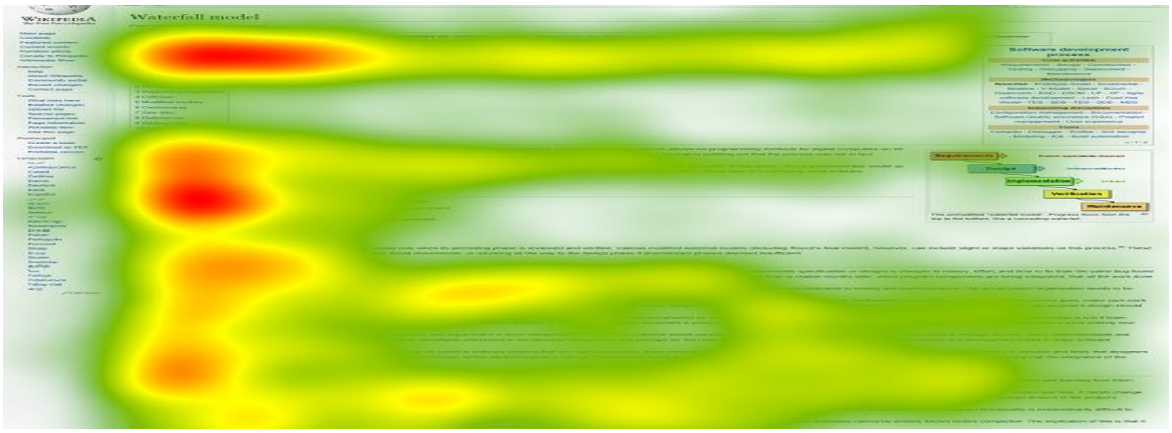


Figure 4.17: Heat Map of document 1 (explicit rating from user study 2 = 3.81)

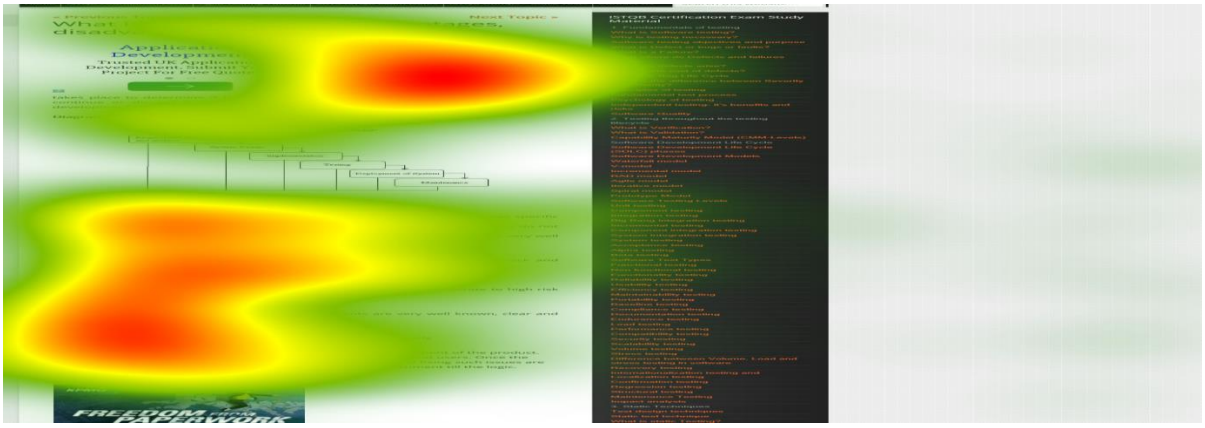


Figure 4.18: Heat Map of document 2 (explicit rating from user study 2 = 3.86)

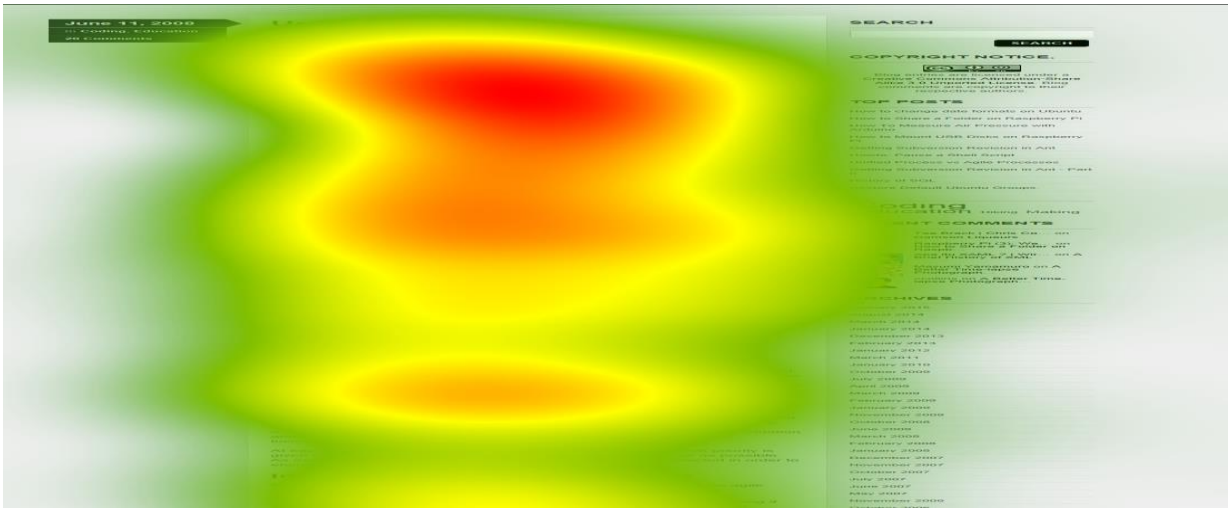


Figure 4.19: Heat Maps of document 3 (explicit rating from user study 2 = 2.5)

Figure 4.20: Heat Map of document 4 (explicit rating from user study 2 = 2.73)

Figure 4.21: Heat Map of document 5 (explicit rating from user study 2 = 3.82)

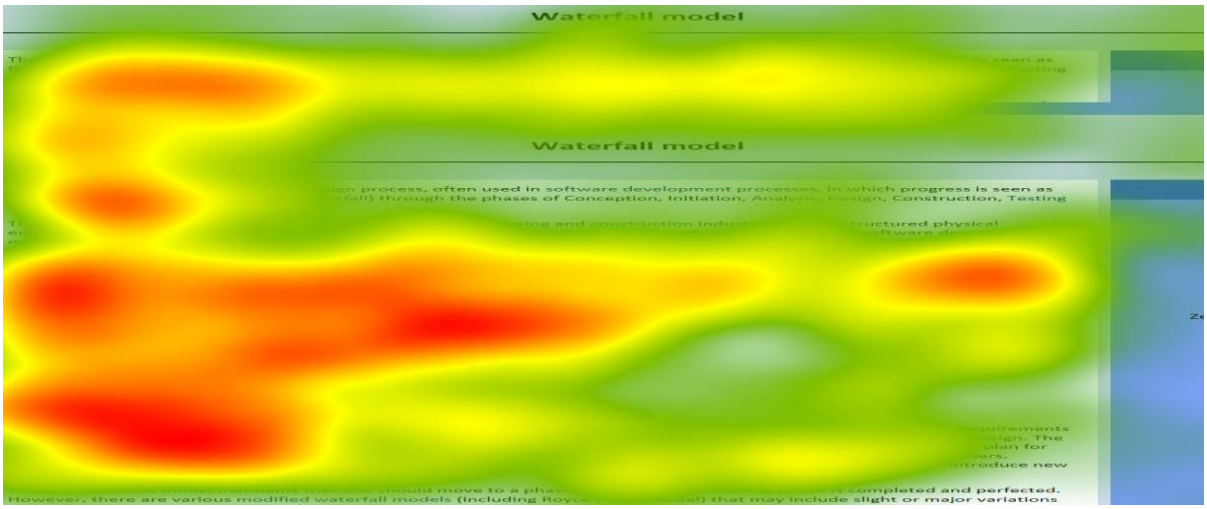


Figure 4.22: Heat Map of document 6 (explicit rating from user study 2 = 2)

Chapter Summary

This chapter discussed the results of the three user studies. The results of the preliminary study show that dwell time on a document is influenced by user-perceived relevance and topicality. The relationship between homogeneous clusters of user reading behaviour and their

explicit ratings was also found. The preliminary study was limited to some selected web documents and only a few implicit indicators were examined. Users' search behaviour was not adequately captured.

The results of the second user study showed that aside from the mean mouse speed and the keystroke, there is a significant correlation between the other implicit indicators measured with the explicit relevance ratings. There is no significant difference in user behaviour in terms of documents they consider to be familiar from those they consider as unfamiliar. The results show a significant difference between documents considered as difficult to understand and those considered as not difficult to understand in terms of mouse activities and the amount of copy. Among the behavioural features (Mouse Duration Count, Distance of Mouse Movement, Total Mouse Movement, Mean Mouse Velocity, Number of Mouse Clicks, Amount of Scroll, Number of Keystrokes, Amount of Copy and Active Time Spent on the Document) examined, the results suggest that users perform fewer mouse activities and copy less text from documents they consider difficult to understand.

The analysis of the different task types shows that in the mixed task, users moved the mouse more on documents they considered not difficult to understand. Also, users of the factual task copied text on documents they considered not difficult to understand than documents they considered difficult to understand. This indicates that users probably abandon documents they consider difficult to understand, and move swiftly to documents they can easily comprehend. The result of the relationship between implicit indicators and explicit relevance ratings for both tasks type is similar. That is, the greater the rating for relevance, the greater the Mouse Duration Count, Distance of Mouse Movement, Total Mouse Movement, Number of Mouse Clicks, Amount of Scroll, Amount of

Copy and Active Time Spent on the Document. The results also demonstrate that users rated (explicit relevance rating) documents that they are familiar with higher than those they are not familiar with, and they rated documents they could not understand lower than those they found less difficult to understand.

The chapter also established a relationship between explicit and implicit feedback parameters. A predictive model was derived from aggregating implicit indicators and standard metrics were used to evaluate the strength of the predictive model. Some classifiers were compared and K-nearest neighbour (k=4) classification algorithm produced the highest accuracy in classifying the relevancy of documents based on user behaviour.

A validating study based on eye gaze (user study 3) was conducted to confirm the predictive strength of the predictive model. It suggests that with a reasonable degree of accuracy, the predictive function model derived from 'cheap and available' sources can be substituted for an eye gaze in predicting relevant documents. These findings provide a cost effective method for understanding user web behaviour in the context of user task through the use of implicit parameters.

III

Implicit Feedback System

CHAPTER 5

5 Implementation and Evaluation

5.1 Introduction

The previous chapters found common and consistent implicit indicators that are often used by users of a particular domain while searching for their information needs. In this chapter, a prototype system for the recommendation of relevant web documents is implemented and evaluated. The requirements for the prototype system and the structure are described. The aim of the system is to return relevant web documents and minimise the number of irrelevant web documents that are retrieved. The goal is to show that the quality of search results improves when query results are re-ranked based on user implicit feedback. The system uses a vector space model (Baeza-Yate and Riberro-Neto 1999, Salton, Wong and Yang 1975) explained in Section 2.4 to find similar documents matching a query while the user interest/perceived relevance of documents is deduced from the predictive model developed in Section 4.5.3. The process uses a collaborative approach to re-rank query feedbacks in order to improve retrieval results based on query-document similarity and the interest weight of previous documents visited.

Mean Average Precision (MAP), a popular metric used by researchers in the area of information retrieval (Balakrishnan and Zhang 2014, Kelly 2009) is employed for the evaluation of the system by comparing the relevancy of documents retrieved from Google, Solr-indexed system and

the aggregated system. It measures the efficiency of the system by calculating the average precision of documents retrieved per query.

5.2 System Structure

The system needed to optimise recommendation of relevant web documents to users of a particular domain, as shown in Figure 5.1, is built on the following structure:

5.2.1 Data Collection

As explained in Section 3.3, data is collected unobtrusively from users of a particular domain through an injected plugin in the Firefox browser. The data is sent to a central server which is then stored in a database. Data can be collected independent of other subsystem and indexing can be done offline.

5.2.2 Interest Scoring

The predictive model in Section 4.5.3 is used to calculate user interest level on each document and weight is assigned to the documents based on the user interest. The implicit model is an aggregation of dwell time and amount of copy, and it estimates a user's interest level in documents.

5.2.3 Document Filtration

Apache Solr technology was used to filter documents matching inputted queries. It implements the Vector Space Model (VSM) functionality of indexing, term weighting, similarity matching and scoring (as explained in Section 2.4). Solr is an open source search platform; it is part of the Apache Lucene project and it communicates with other applications through REST-like HTTP request. The major features of Solr include real-time indexing, faceted search, full-text search, hit highlighting, dynamic clustering, rich document handling and database integration.

5.2.4 Aggregated Document Weight (ADW)

The aggregated document weight is an algorithm that computes a new score for the documents. It combines the weight of the document derived from the predictive model and the document weight computed by the vector space model. It follows this method:

$$ADW = CIW + CVW \quad (5.1)$$

where

CIW is Computed Interest Weight based on the predictive model derived from previous experimentation and analysis. It is given as:

$$CIW = 0.28(Copy) + 0.0018(Dwell\ Time) + 0.29 \quad (5.2)$$

CVW is the Computed Vector Weight of the original documents based on TF-IDF algorithm.

5.2.5 Document Re-ranking

This module sorts the documents based on the aggregated document weight in descending order for presentation to the user. It alters the original ranking which is based on only the Computed Vector Weight (VSM score) and re-ranks the documents according to the new calculated aggregated weight. It follows these steps:

1. Previous documents visited by users of a particular domain along with the associated implicit data are captured and stored in a database.
2. A user enters a query relating to a current task.
3. The TF-IDF score of the documents in the database is computed based on the query entered.
4. The implicit score of the documents in the database is computed.
5. If there are common documents in the database, the mean interest score of the common documents is returned to the document.

6. The aggregated document score (Equation 5.1) is computed.
7. The query result is re-ranked based on the aggregated weight and displayed.
8. User visits a current document and his/her implicit data is captured and stored.
9. The process begins again with a new query.

The steps are transformed into an enhanced algorithm as follows:

1: Enter user query q

2: Compute for q, the vector weight CVW of all documents D in the database

3: Compute the interest weight CIW = 0.28 (copy) + 0.0018 (Dwell Time) + 0.29 of all D

4: Considering that D = all documents

∂→D

```
for(i=1; i<= sizeD; i++) {
    for(j=1; j<= sizeD; j++) {
            if ((∂i == ∂j) && (i != j)) {
            calculate mean interest weight
            ∂i.CIW = ( ∂i.CIW + ∂j.CIW)/2;
            delete ∂j
        }
    }
```

}

5: Compute the aggregated document weight ADW = CIW + CVW for D

6: Re-rank original document list based on ADW and display result

7: Visit current document

8: Capture implicit indicators and store in database

5.2.6 Display Results

The module displays the result of the re-ranked documents. It is implemented with Html and the results are presented in descending order of relevance.

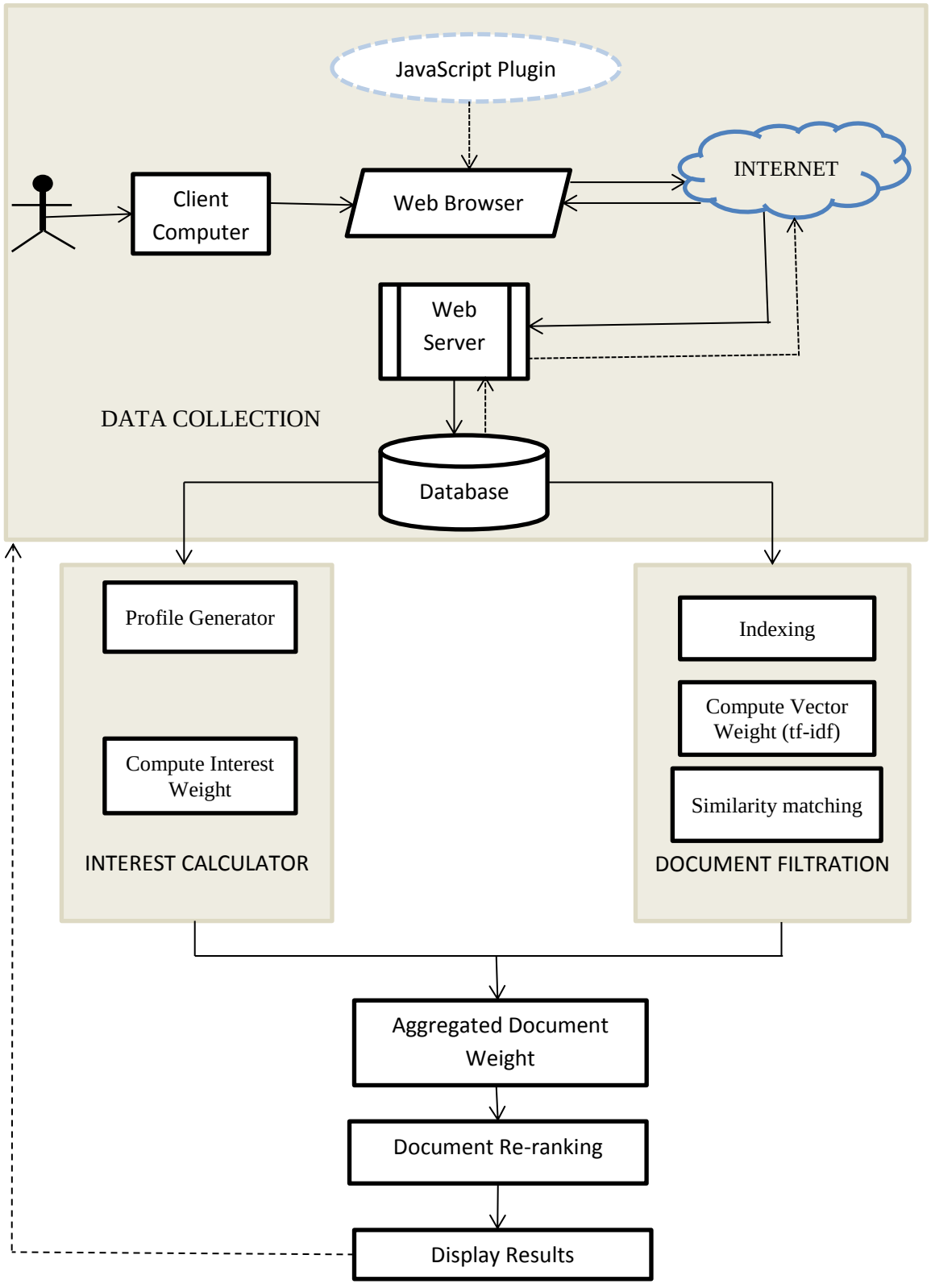


Figure 5.1: Conceptual diagram showing aggregated feedback system flow

5.3 **Evaluation Method**

The aim of the study was to conduct a comparative evaluation of the aggregated implicit feedback system with two other retrieval systems in terms of recall and precision (Mean Average Precision). The performance of the three systems was measured based on the relevant documents returned. This work compares the proposed aggregated system (which has implicit feedback) with the Solr-indexed system (domain specific without implicit feedback) and Google (generic without feedback). A similar evaluation technique was used by (Agichtein, Brill and Dumais 2006, White and Buscher 2012). A Baseline, Controlled and Experimental systems were employed for the evaluation. Users of the three systems were given a task brief containing instruction for the experiment (see Appendix A4). They were allowed to enter a single query (keywords) of their choice to search for documents that were relevant to the 'mixed' task as stated in Section 3.3.8. They were also asked to rate the first top 10 documents according to how relevant they were to the given task. The interface for the 6-point rating scale appears after a user reads the current web page and closes it as explained in Section 3.3.3. However, users were asked to manually rate the top 10 documents for Google (baseline system) after the researcher observed that some of the top 10 links returned contained video and images, and were not JavaScript enabled. The six-point ratings of the users were then merged into binary form. Ratings of 0,1and 2 were merged as 0 and labelled as non-relevant while the rating for 3,4 and 5 was merged as 1 and labelled as relevant.

Although Text REtrieval Conference (TREC) evaluation uses expert judges to assess the relevance of documents, such relevance ratings are inherently noisy due to the variability of the experts' behaviour (Smucker, Allan and Carterette 2007). Also, selecting experts to judge each of the documents used for this research in relation to the task given was not feasible. This work considers relevance judgement to be subjective to the

user accessing the web documents in relation to the current task as explained in Section 2.5.3. The user relevance rating was used for evaluating the effectiveness of the implicit feedback system in terms of precision, recall and mean average precision.

5.3.1 Experiment setup

A total number of 26 students in the Faculty of Engineering and Computing at Coventry University participated in the evaluation study for a given duration of 30 minutes. Two approaches were employed in conducting the study. Out of 26 users, 15 participated in the first approach (Approach 1) while 11 participated in the second approach (Approach 2). The participants were given a brief tutorial about the experiment and a consent form to complete (see Appendix B). Altogether, the participants entered a total number of 26 queries (See Appendix D for the queries entered by the participants). The following three systems were used for evaluation:

- ***Baseline system:*** Google is the baseline system. It is generic and non-domain-specific. Documents relating to user queries are returned based on the Google algorithm.
- ***Controlled system (Solr-Indexed system):*** The controlled system contains only the solr-indexed TD-IDF algorithm as shown in figure 5.2. The system is designed to return documents based on user query. The implicit predictive model is not integrated into the system.
- ***Experimental system (Aggregated system):*** The implicit model is integrated into the system such that documents relating to the inputted query are re-ranked according to the degree of user interest. The degree of user interest is estimated using the implicit predictive model derived. The experimental system re-ranks the documents according to the aggregated document weight, which is a

combination of the computed interest weight (CIW) and computed vector weight (CVW) as explained in Section 5.2.4.

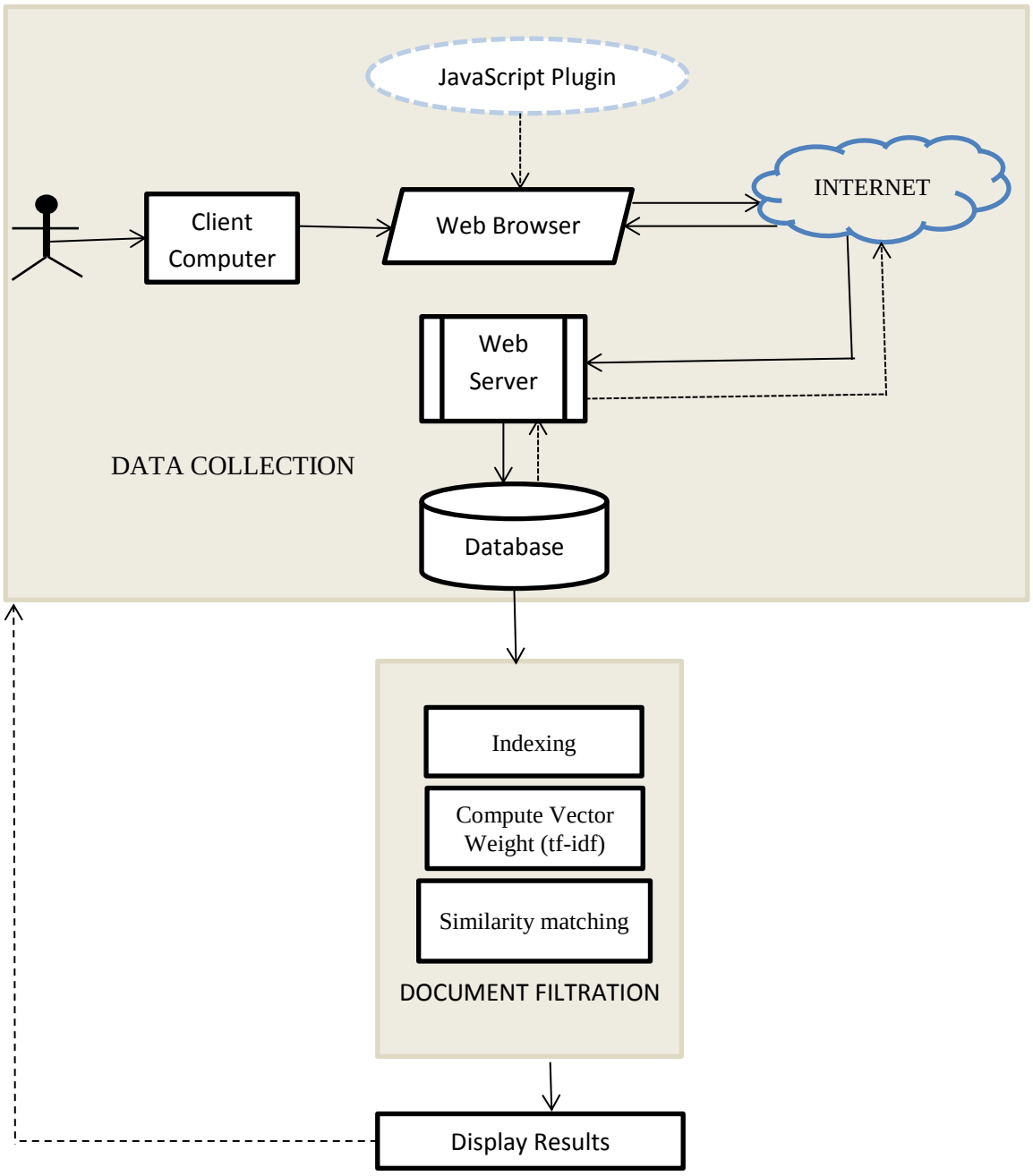


Figure 5.2: Conceptual diagram showing Solr-indexed system flow

Both the controlled and experimental system had the same pool of documents obtained from the previous study in Section 3.3. Common documents in the dataset were merged and their mean interest weight was computed and presented as a single document. This reduced the size of the dataset from 343 to 140. Documents retrieved from Google were crawled from different sources and indexed in their database. Figure 5.3, 5.4, and 5.5 show the screen for the search query "RUP vs waterfall model". Figure 5.3 shows documents returned by Google, Figure 5.4 shows the original SERP returned by the Solr-indexed system, and Figure 5.5 is the re-ranked result returned by the aggregated system. For example, the document, "Difference Between Waterfall Methodology and RUP" is ranked 3rd by Google as can be seen in Figure 5.3, It is ranked 1st in the solr-indexed system as shown in Figure 5.4 and it is ranked 2nd in the aggregated system as shown Figure 5.5. The difference in ranking of the document across the three systems underpins the variability of the systems.

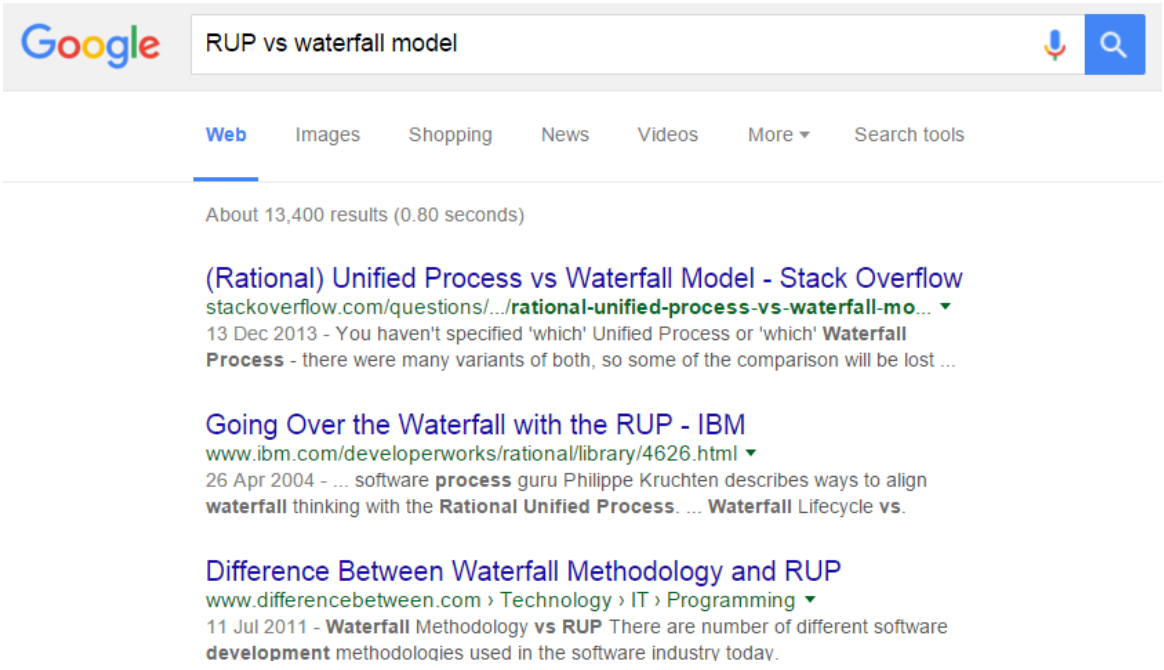


Figure 5.3: Sample interface showing search query and SERP for Google

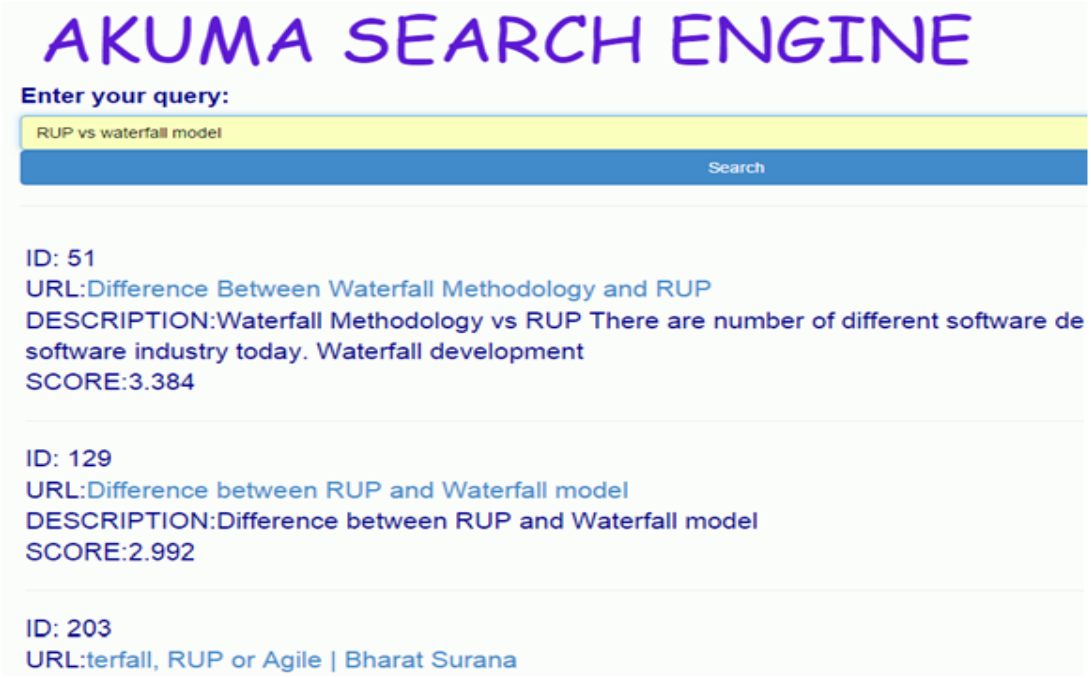


Figure 5.4: Sample interface showing search query and SERP for solr-indexed system

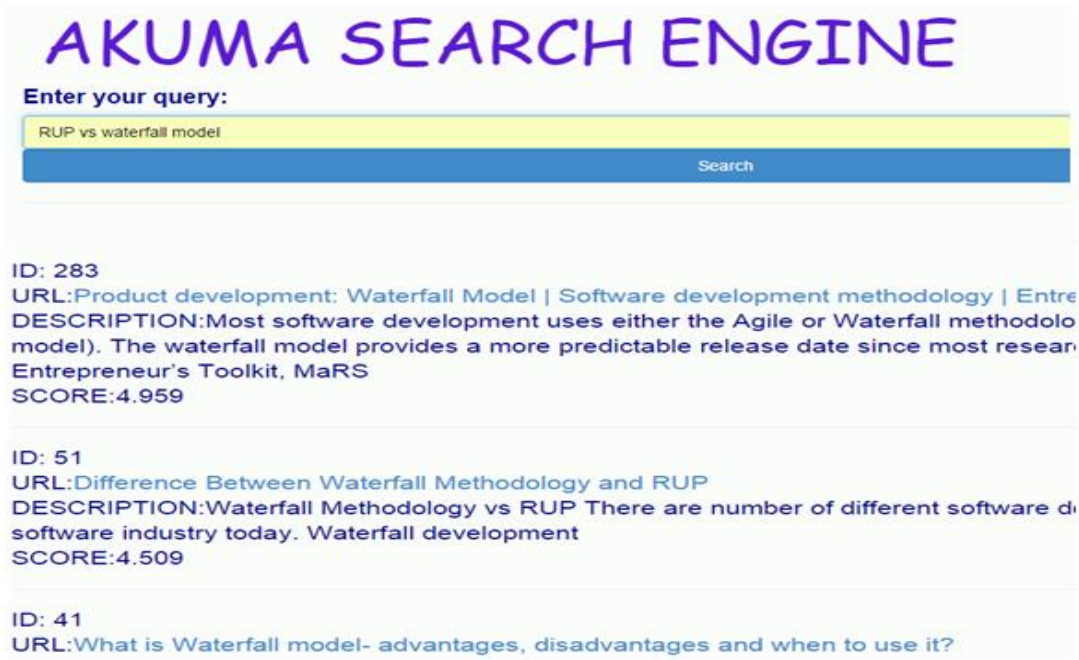


Figure 5.5: Sample interface showing search query and SERP for aggregated system

The two approaches employed for evaluation compare the performance of both the solr-indexed system and the aggregated system against the performance of the baseline Google system. Sections 5.3.1.1 and 5.3.1.2 describe the two approaches used.

5.3.1.1 Approach 1

In this approach, the A/B testing (Manning, Raghavan and Schütze 2008) was used. 15 users participated and they were randomly grouped into three sets, labelled A, B and C. Each set comprised 5 participants. The three groups of participants were given the same tasks and they visited different retrieval systems. Participants in group A performed the experiment with the baseline Google search engine; group B participants used the controlled system while the participants in group C performed the experiment with the experimental system.

5.3.1.2 Approach 2

In this approach, all 11 participants were given the baseline, controlled and experimental systems to use. They entered the same unique query in the three systems and rated the first top 10 results of each system according to how relevant they were to the given task. The participants were not told the difference between the systems in order to prevent bias in rating. Each of the 11 users rated up to 30 web pages.

5.3.2 Evaluation Metrics

The precision and recall described in this section are similar to those explained for the classification problem in Section 4.5.5. This is, however, an information retrieval problem. The relationship between precision and recall is such that as precision increases, recall decreases and vice versa. The importance of one over another depends on the context of usage. In this case, it is desired that the highest ranked documents of a retrieval system should be more relevant than documents at the bottom of a retrieval list. Therefore, high precision is needed.

The precision of an information retrieval problem measures the portion of relevant items within the total items retrieved. It involves retrieving the most relevant top-ranked documents and it is given as:

$$\textbf{Precision} = \frac{\textbf{number of relevant items retrieved}}{\textbf{total retrieved items}} \qquad 5.3$$

The recall for an information retrieval problem measures the portion of relevant items within the total relevant items retrieved. This involves the ability to find all relevant items in a given collection. It is given as:

$$\textbf{Recall} = \frac{\textbf{number of relevant items retrieved}}{\textbf{total relevant items retrieved}} \qquad 5.4$$

Mean Average Precision (MAP) represents the area under the precision and recall curve. It is a single number that is used to compare the performance of retrieval algorithm. It is the average of precision values of a retrieval list at the positions where relevant documents were retrieved (Lavrenko 2014). It is given as:

$$\textbf{MAP}(\boldsymbol{n}) = \frac{\mathbf{1}}{|\boldsymbol{n}|}\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}} \boldsymbol{AP}(\mathbf{i}) \qquad 5.5$$

where AP is the average precision and n is the number of queries used for searching (Balakrishnan and Zhang 2014). It is given as:

$$\boldsymbol{AP} = \frac{\sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{n}}(\boldsymbol{P}(\boldsymbol{l})\times\boldsymbol{R}@\boldsymbol{l})}{\mathbf{D}} \qquad 5.6$$

where P(l) is the precision at l document level, n is the number of retrieved documents, R@l states whether the document at l is relevant or

not, D is the total relevant documents for a given query (Balakrishnan and Zhang 2014, Kelly 2009).

For the purpose of result analysis, some acronyms are defined in Table 5.1

Table 5.1: Basic acronyms used for analysis

Acronym	Meaning
Top 5	This is the first 5 documents returned by a search system.
Top 10	This is the first 10 documents returned by a search system.
M_x	The mean of x, where x is either Google, Solr-indexed or Aggregated system.
SD_x	The Standard deviation of x, where x is either Google, Solr-indexed or Aggregated system.

5.3.3 Statistical Significance Testing

Paired T-test was employed in testing for significance between the average precision values of the Baseline system against the Controlled system and the Experimental system. Researchers (Cormack and Lynam 2007, Sanderson and Zobel 2005, Smucker, Allan and Carterette 2007) say that paired t-test is the most reliable test for evaluating MAP values. The requirements needed for conducting a t-test are explained in Section 4.2.

5.4 Approach 1 Results

This section compares the results of the participants in the Baseline system with the Controlled system and Experimental system in terms of the Mean Average Precision. Among the 15 participants, 5 users visited the baseline system, 5 others visited the controlled system and the remaining 5 students visited the experimental system. The precision was calculated for each of the 15 queries and the precisions at ranks where

the documents were relevant for each query were summed and averaged. The mean average precision was then computed and the results from Google showed that the result at top 10, the MAP was 0.51 and top 5, the MAP was 0.54. The MAP for the Solr-indexed system at top 10 was 0.77 and at top 5 was 0.84. The aggregated system produced an improved result. The MAP of the aggregated system for top 10 was 0.86 and for top 5 was 0.91. The paired T-test of the average precisions between the baseline system and solr-indexed system for the top 10 documents was statistically significant, it shows the mean of the solr-indexed system to be 0.26 higher than the baseline system (p = 0.015, M_{google} = 0.51, SD_{google} = 0.096, $M_{solr\text{-}indexed}$ = 0.77, $SD_{solr\text{-}indexed}$ = 0.21), there was a higher mean difference of 0.35 when the paired T-test was run between the baseline and the aggregated system (p = 0.007, M_{google} = 0.51, SD_{google} = 0.096, $M_{aggregated\_system}$ = 0.86, $SD_{aggregated\_system}$ = 0.14).

The result of the top 5 also showed a significant improvement of the solr-indexed system and the aggregated system over the baseline Google system. Google vs solr-indexed system produced p = 0.019, M_{google} = 0.54, SD_{google} = 0.11, $M_{solr\text{-}indexed}$ = 0.84, $SD_{solr\text{-}indexed}$ = 0.2 and Google vs aggregated system produced p = 0.006, M_{google} = 0.54, SD_{google} = 0.11, $M_{aggregated\_system}$ = 0.91, $SD_{aggregated\_system}$ = 0.12.

Table 5.2 and figure 5.6 shows the mean average precision of the three models in the two measured ranks.

Table 5.2: Approach 1 Mean average precision for Google, Solr-indexed and Aggregated systems

	Google	Solr-indexed	Aggregated system
MAP top10	0.51	0.77	0.86
MAP top5	0.54	0.84	0.91

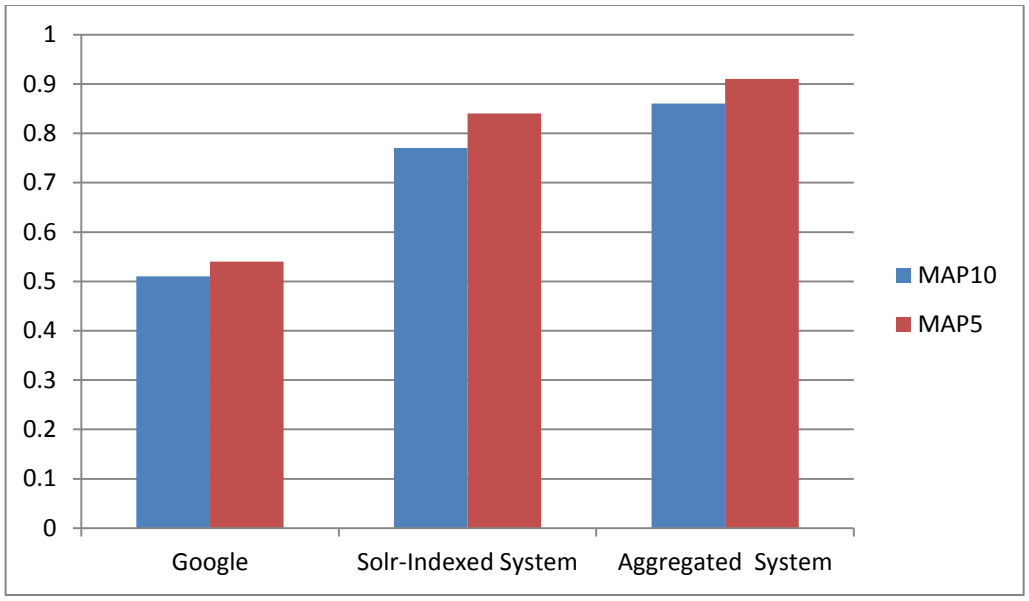


Figure 5.6: Approach 1 MAP histograms comparing Google, Solr-indexed and Aggregated systems

5.5 **Approach 2 Results**

In this approach, 11 participants accessed the three systems and each user entered the same query in the three systems and rated the 10 top results according to relevance. Their ratings for the baseline, controlled and the experimental systems were captured. The mean average precision for the three systems was computed and the results show that Google at top 10 was 0.51 and at top 5 was 0.57 while the MAP of Solr-indexed system at top 10 was 0.78 and at top 5 was 0.85. The aggregated system in this approach also produced an improved result in terms of MAP. The MAP of the aggregated system at top 10 was 0.87 while that of top 5 was 0.91, as shown in Table 5.3 and Figure 5.7. The paired T-test of the MAP of the models also showed that the aggregated model has a statistically significant improvement compared to the baseline and Solr-indexed system.

The paired T-test between Google and solr-indexed system for the top 10 documents was significant with the solr-indexed system performing

better by 0.27 MAP (p = 0.004, M_{google} = 0.51, SD_{google} = 0.083, $M_{solr\text{-}indexed}$ = 0.78, $SD_{solr\text{-}indexed}$ = 0.22) and the top 5 produced a significant improvement in MAP of the Sol-indexed system by 0.28 (p = 0.031, M_{google} = 0.57, SD_{google} = 0.14, $M_{solr\text{-}indexed}$ = 0.85, $SD_{solr\text{-}indexed}$ = 0.29). The aggregated system has a significant higher MAP over Google. For the top 10 documents, it is higher by 0.36 (p = 0.000, M_{google} = 0.51, SD_{google} = 0.083, $M_{aggregated\_system}$ = 0.87, $SD_{aggregated\_system}$ = 0.18) and for the top 5 documents, it is higher by 0.34 (p = 0.002, M_{google} = 0.57, SD_{google} = 0.14, $M_{aggregated\_system}$ = 0.91, $SD_{aggregated\_system}$ = 0.19).

Table 5.3: Approach 2 Mean average precision for Google, Solr-indexed and Aggregated systems

	Google	Solr-indexed	Aggregated system
MAP top10	0.51	0.78	0.87
MAP top5	0.57	0.85	0.91

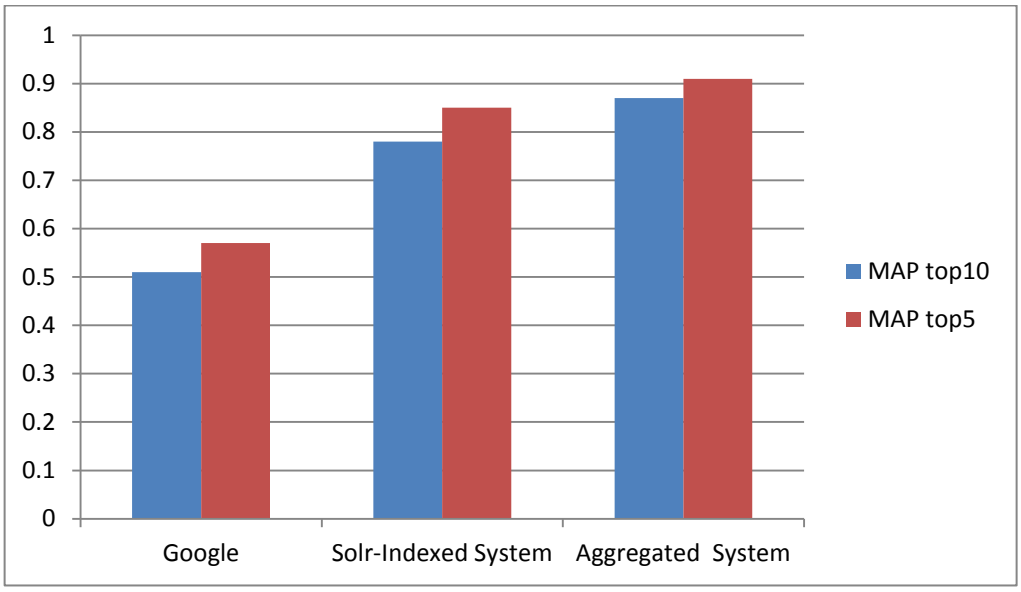


Figure 5.7: Approach 2 MAP histograms comparing Google, Solr-indexed and Aggregated systems

As shown in the two approaches, the MAP values decrease as the number of documents increases from top 5 to top 10 (see Appendix E for the

tables of the computed recall and precision values). This shows the trade-off between precision and recall. The aggregated system still performed better than both the Solr-indexed system and the baseline system at the two computed document levels. This indicates an improvement of document relevance when queries are supplemented with user post-click behaviour. This result is similar to the results reported by previous studies (Agichtein, Brill and Dumais 2006, Balakrishnan and Zhang 2014, Fox et al. 2005, Guo and Agichtein 2012, Huang, White and Dumais 2011, Joachims et al. 2007, Jung, Herlocker and Webster 2007).

Agichtein, Brill and Dumais (2006) used only a single indicator (click-through) to re-rank documents, Guo and Agichtein (2012) used scrolling and cursor movements to estimate relevance and Balakrishnan and Zhang (2014) used a heuristic to aggregate implicit indicators. In this work, experimental analysis was used to derive a predictive function model that estimates document relevance.

Users spend a significant amount of time to get relevant documents for their information need due to the growing number of documents uploaded on the web. The goal of personalisation of information retrieval is to meet users' information needs by taking into consideration users' context and behaviour (Alhindi et al. 2015, Zemirli 2012). The advantage of this over the generic system is that it helps users to find relevant documents more quickly (Alhindi et al. 2015, Balakrishnan and Zhang 2014). Although generic systems like google have personalised search feature, it uses browser cookies to store data and the data can easily be destroyed. Also, relevance is estimated by document visit (hit) and not by an implicit predictive model derived from a regression analysis. The results of this study further validate that personalisation is needed to solve the information overload problem.

Chapter Summary

This chapter discusses the proposed domain-specific implicit feedback system to improve retrieval document relevancy. A prototype system is implemented and evaluated. Apache Solr technology was used to implement the Vector Space Model functionality of indexing, term weighing, similarity matching and scoring. The system has a predictive function that aggregates copy and dwell time implicit indicators with the classical TF-IDF algorithm of the Vector Space Model. An enhanced algorithm demonstrates the working principle of the system: assigning scores to documents and re-ranking retrieval documents.

In order to evaluate the results mentioned in previous chapters, a prototype system was developed to carry out a comparative analysis between the proposed aggregated system, Solr-Indexed system and Google. The results show that when users' queries are supplemented with their post-click behaviour, it improves the original ranking of retrieval results. Two approaches were used to evaluate the system. The results in both approaches show that the aggregated system performed better than the baseline and Solr-indexed system in terms of the mean average precision.

IV

Conclusion

CHAPTER 6

6 CONCLUSION

6.1 Introduction

This research investigates user web behaviour by correlating implicit and explicit feedback approaches in different task settings. A predictive model was derived from the analysis of user implicit and explicit parameters. The predictive model derived was used to argument searchers' queries so as to optimise the recommendation of the relevant document. Part I of this work gives an introduction to the research; it also discusses the background and the state of the art in the area of information retrieval and user web behaviour. Part II describes the implicit evidence. Studies were conducted and user Implicit and explicit feedback parameters were correlated. The effect of document difficulty, document familiarity and task type on user behaviour was also examined. Multiple regression analysis was used to derive the implicit predictive model. The model was evaluated by standard evaluation metrics and validated with an eye gaze study. The predictive model derived in Part II was used in Part III to develop a prototype system for the implicit recommendation of web documents. The system combines the classical TF-IDF algorithm with the predictive model to re-rank retrieval results. An evaluation of the prototype system shows that previous relevant documents used by users of a common domain can be shared, and when users' queries are supplemented by implicit feedback parameters, more relevant documents are returned.

6.2 Contributions

This thesis describes a number of contributions to the existing knowledge of implicit feedback systems. The contributions stated below in Section 6.2.1 to 6.2.4 answers the research question and satisfy the research objectives:

6.2.1 Implicit Predictive Function

A method was developed to capture and interpret the relationship between implicit and explicit feedback parameters obtained from a large number of participants in different task situations and an implicit predictive function model was derived. The implicit model can substitute for explicit rating in estimating document relevance to a reasonable degree. It was shown that the implicit model derived from aggregating implicit indicators performed better in prediction than models with the single implicit indicator. Also, some classifiers were compared and J4 and K-nearest neighbour were found to have the highest percentage of accuracy in classifying documents based on relevancy. This finding gives clarity on the relationship between implicit and explicit feedback parameters in a task specific context.

6.2.2 Effect of Some Moderating Factors on User Behaviour

An investigation of the extent at which document difficulty, document familiarity and task type affect user behaviour was carried out. A study was conducted and the results show that document familiarity does not affect user behaviour. There is an inverse relationship between Mouse movement/distance and ratings for document difficulty. It shows that the more difficult a document the less mouse movement on the document and vice-versa. Whereas previous work focused on topic/task familiarity and difficulty in relation to user behaviour (Crescenzi, Capra and Arguello

2013, Kelly and Cool 2002, Liu et al. 2010, Liu, Liu and Belkin 2013), this work is the first to investigate the effect of document difficulty and document familiarity on user behaviour.

The results on user behaviour from the two task components examined (Mixed and Factual) show that apart from the user scroll (for document relevance), there was no difference between the Mixed and Factual tasks in terms of the explicit ratings for document relevance and document familiarity. However, in terms of document difficulty, it showed an inverse relationship with the mouse movement/distance for the Mixed task and an inverse relationship with the amount of copy for the Factual task. This study provides evidence that document difficulty and task type affect some user behaviour.

6.2.3 Eye Gaze Validation

A study was conducted to validate the predictive model with an eye gaze tracker. In this, the researcher shows that the predictive model derived by combining 'low-cost' implicit indicators is more effective in predicting document relevance than the predictive models of single indicators. Also, it was shown that the predictive model can be used in place of eye gaze measures.

6.2.4 Prototype Implicit Feedback System

A framework was created to recommend relevant documents to users based on their behaviour and a prototype system was used to evaluate the proposed approach. The system was built based on the notion that users' perception of relevance can be inferred from their searching behaviour. Whereas other studies (Balakrishnan and Zhang 2014) used heuristic to give weight to implicit indicators and incorporate it in the re-ranking algorithm, this work experimentally derived a model that assigned weight to the implicit indicators in relation to the explicit

ratings. The model was then incorporated in the re-ranking algorithm to return relevant documents to users based on their previous interaction with the system.

The evaluation of the system with real users showed that when documents are shared among common users, better retrieval results are obtained. The best result in terms of relevancy is obtained when user implicit feedback is added to the retrieval algorithm. This work validates the notion that personalisation minimises the problem of information overload.

6.3 **Limitations**

Although the aim of the research was achieved, the perceived limitations of the work are listed below:

- The size of the dataset is considerably small as compared to other web usage datasets like TREC and Cranfield. It was difficult to analyse each of the individual data because some users visited only one web document in the experiment.
- The dataset is also limited to a given domain. Although the proposed system is highly generalisable, a similar kind of experimentation may be needed to derive the implicit model for other domains.
- The capturing plugin was only able to capture documents that were JavaScript enabled. Pdf documents, images and video web resources were not captured. There is a possibility that some of the relevant documents were not captured due to the limitation of the plugin.
- One of the issues involved in capturing users' data unobtrusively is users' right to privacy. There are legal consequences when one's privacy is invaded. Even with a detailed explanation of the

genuineness of the user profiling approach, some users were still sceptical and opted out of the study. In application, the system should ask the users for explicit permission on whether to build a user profile from their searching activities.

- Since the experiment lasted only 45 minutes and users were asked to provide answers to the task immediately, non-frequently used indicators in the "Retention" category like bookmark, printing, save and email were not examined.
- The system used Vector Space Model for query-document similarity matching and it is devoid of semantic search. Some relevant documents may not be returned as a result of this.

6.4 Future work

Although the objective of this work is achieved, there are a number of areas that can be extended and improved.

- The plugin used for this work was developed with JavaScript and it is specific to Firefox Browser. The plugin can be extended to be cross browser compatible and it should be able to capture other document formats like images, pdf and video.
- Only the Mixed and Factual tasks were examined in relation to user behaviour. Future work can explore other task types in relation to user behaviour.
- Some participants visited few web documents. The individual user profile is needed for content-based analysis. Future work should include a longitudinal study to collect a large amount of data from each user in order to build user profiles for hybrid (content and collaborative) recommendation.
- Since there was only one eye tracker available, only some selected documents were used to examine the relationship

between classical implicit indicators and gaze based measures. This can be improved by conducting a study that will capture the eye gaze measures and the classical implicit indicators simultaneously.

- The 'static' approach implicit predictive model can be improved by automatically updating the model with time as users search for information.
- More implicit indicators under the "Retention" category like bookmarking, printing, saving and emailing can be examined through a long-term naturalistic study. Also, other sources of implicit evidence from electroencephalogram (EEG) recording of brain activity can be investigated.
- The proposed system returned results by similarity matching and interest level. Semantic search can be added to return more relevant results.
- Expert judges can be employed to also rate the documents viewed by the users according to topic relevance. These ratings can then be compared with the ratings of the users to evaluate the relevancy of the documents.

Chapter Summary

This research has investigated the relationship between implicit and explicit feedback parameters in different task settings. This investigation led to the development of a prototype implicit feedback system. The contribution discussed in this chapter indicates that the research objectives were achieved. The research shows that a domain-specific retrieval system returns more relevant documents to a community of users than a generic retrieval system, and when implicit feedback is added to the domain specific system, document relevancy is improved.

References

Aciar, S., Zhang, D., Simoff, S., and Debenham, J. (2007) 'Informed Recommender: Basing Recommendations on Consumer Product Reviews'. *IEEE Intelligent Systems* 22 (3), 39-47

Agichtein, E., Brill, E., and Dumais, S. (eds.) (2006) *Proceedings of the Twenty-Ninth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.* 'Improving Web Search Ranking by Incorporating User Behavior Information'

Ajiboye, J. and Tella, A. (2007) 'University Undergraduate Students' Information Seeking Behaviour: Implications for Quality in Higher Education in Africa'. *The Turkish Online Journal of Educational Technology – TOJET* 6 (1), 40-54

Alhindi, A., Kruschwitz, U., Fox, C., and Albakour, M. (2015) 'Profile-Based Summarisation for Web Site Navigation'. *ACM Transactions on Information Systems* 33 (1), 1-40

Baeza-Yate, R. and Riberro-Neto, B. (1999) *Modern Information Retrieval.* New York: Addison- Wesley

Balabanovic, M., Shoham, Y., and Yun, Y. (1995) 'An Adaptive Agent for Automated Web Browsing'. *Journal of Image Representation and Visual Communication* 6(5)

Balakrishnan, V., Ahmadi, K., and Ravana, S. D. (2016) 'Improving Retrieval Relevance using Users' Explicit Feedback'. *Aslib Journal of Information Management* 68 (1), 76-98

Balakrishnan, V. and Zhang, X. (2014) 'Implicit User Behaviours to Improve Post-Retrieval Document Relevancy'. *Computers in Human Behavior* 33, 104-112

Barry, C. L. (1998) 'Document Representations and Clues to Document Relevance. Journal of the American Society for Information Science' 49 (14), 1293-1303

Bennett, P., Collins-Thompson, K., Kelly, D., White, R., and Zhang, Y. (2015) 'Overview of the Special Issue on Contextual Search and Recommendation'. *ACM Transactions on Information Systems* 33 (1), 1-7

Bhavnani, S. K. (ed.) (2001) *Proc. TREC*. 'Important Cognitive Components of Domain-Specific Search Knowledge'

Bhavnani, S. K. (ed.) (2002) *Conference on Human Factors in Computing Systems - Proceedings*. 'Domain-Specific Search Strategies for the Effective Retrieval of Healthcare and Shopping Information'

Borlund, P. (2003) 'The IIR Evaluation Model: A Framework for Evaluation of Interactive Information Retrieval Systems'. *Information Research* 8 (3)

Brusilovsky, P. (ed.) (1998) *In: Proceedings of Workshop "www.Base & Tutoring", Fourth International Conference in Intelligent Tutoring Systems*. 'Adaptive Educational Systems on the World-Wide-Web: A Review of Available Technologies' at San Antonio, TX

Brusilovsky, P. and Tasso, C. (2004) 'Preface to Special Issue on User Modeling for Web Information Retrieval'. *User Modelling and User-Adapted Interaction* 14 (2-3), 147-157

Busby, M. (2003) *Learn Google*. Plano, Texas: Wordware Publishing Inc

Buscher, G., Dengel, A., Biedert, R., and Van Elst, L. (2012a) 'Attentive Documents: Eye Tracking as Implicit Feedback for Information Retrieval and Beyond'. *ACM Transactions on Interactive Intelligent Systems* 2 (1), 1-30

Buscher, G., White, R. W., Dumais, S. T., and Huang, J. (eds.) (2012b) *WSDM 2012 - Proceedings of the 5th ACM International Conference on Web Search and Data Mining*. 'Large-Scale Analysis of Individual and Task Differences in Search Result Page Examination Strategies'

Buscher, G., Biedert, R., Heinesch, D., and Dengel, A. (eds.) (2010) *Conference on Human Factors in Computing Systems - Proceedings*. 'Eye Tracking Analysis of Preferred Reading Regions on the Screen'

Buscher, G., Van Elst, L., and Dengel, A. (eds.) (2009) *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009*. 'Segment-Level Display Time as Implicit Feedback: A Comparison to Eye Tracking'

Buscher, G., Dengel, A., and Van Elst, L. (eds.) (2008) *ACM SIGIR 2008 - 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Proceedings*. 'Query Expansion using Gaze-Based Feedback on the Subdocument Level'

Chapman, L. and Ivankovic, H. (eds.) (2002) *VALA 2002 – E-Volvinginformation Futures. 11th Biennial Conference and Exhibition*. 'Russian Roulette Or Pandora's Box: Use of the Internet as a Research Tool' at Melbourne

Chen, L. and Sycara, K. (eds.) (1998) *Proceedings of the 2nd International Conference on Autonomous Agents*. 'WebMate - a Personal Agent for Searching and Browsing'

Chen, J., Jin, Q., Ma, J., and Huang, R. (eds.) (2010) *2010 9th International Conference on Information Technology Based Higher Education and Training, ITHET 2010*. 'An Integrated System to Assist Personalized Learning Based on Gradual Adaption Recommendation Model'

Claypool, M., Le, P., Wased, M., and Brown, D. (eds.) (2001) *International Conference on Intelligent User Interfaces, Proceedings IUI*. 'Implicit Interest Indicators'

Cole, M. J., Zhang, X., Liu, C., Belkin, N. J., and Gwizdka, J. (eds.) (2011) *SIGIR'11 - Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Knowledge Effects on Document Selection in Search Results Pages'

Cormack, V., G and Lynam, T., R. (eds.) (2007) *In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07*. 'Validity and Power of T-Test for Comparing Map and Gmap ' at New York, NY, USA,

Crescenzi, A., Capra, R., and Arguello, J. (2013) 'Time Pressure, User Satisfaction and Task Difficulty'. *Proceedings of the ASIST Annual Meeting* 50 (1)

Delphi-Group (2004) *Information Intelligence: Content Classification and the Enterprise Taxonomy Practice*

Ding, L., Liu, B., and Tao, Q. (eds.) (2010) *2nd International Workshop on Education Technology and Computer Science, ETCS 2010*. 'Hybrid Filtering Recommendation in E-Learning Environment'

Fattahi, R., Parirokh, M., Dayyani, M. H., Khosravi, A., and Zareivenovel, M. (2016) 'Effectiveness of Google Keyword Suggestion on Users' Relevance Judgment: A Mixed Method Approach to Query Expansion'. *The Electronic Library* 34 (2)

Fox, S., Karnawat, K., Mydland, M., Dumais, S., and White, T. (2005) 'Evaluating Implicit Measures to Improve Web Search'. *ACM Transactions on Information Systems* 23 (2), 147-168

Glass, G. V., Peckham, P. D., and Sanders, J. R. (1972) 'Consequences of Failure to Meet Assumptions Underlying Fixed Effects Analyses of Variance and Covariance'. *Rev. Educ. Res.* 42, 237-288

Glover, E. J., Lawrence, S., Gordon, M. D., Birmingham, W. P., and Giles, C. L. (2001) 'Web Search - Your Way: Improving Web Searching with User Preferences'. *Communications of the ACM* 44 (12), 97-102

Goecks, J. and Shavlik, J. (eds.) (2000) *International Conference on Intelligent User Interfaces, Proceedings IUI*. 'Learning Users' Interests by Unobtrusively Observing their Normal Behavior'

Granka, L. A., Joachims, T., and Gay, G. (eds.) (2004) *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Eye-Tracking Analysis of User Behavior in WWW Search'

Grzywaczewski, A., Iqbal, R., James, A., and Halloran, J. (eds.) (2013) *Proceedings of the 2013 IEEE 17th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2013*. 'Supporting Information Exchange among Software Developers through the Development of Collaborative Information Retrieval Utilities'

Guo, Q. and Agichtein, E. (eds.) (2012) *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*. 'Beyond Dwell Time: Estimating Document Relevance from Cursor Movements and Other Post-Click Searcher Behavior'

Guo, Q. and Agichtein, E. (eds.) (2010) *Conference on Human Factors in Computing Systems - Proceedings*. 'Towards Predicting Web Searcher Gaze Position from Mouse Movements'

Gwizdka, J. (ed.) (2014) *Proceedings of the 5th Information Interaction in Context Symposium, IIiX 2014*. 'Characterizing Relevance with Eye-Tracking Measures'

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I., H. (2009) 'The WEKA Data Mining Software: An Update'. *SIGKDD Explorations* 11

Han, E., Boley, D., Gini, M., Gross, R., Hastings, K., Karypis, G., Kumar, V., Mobasher, B., and Moore, J. (eds.) (1998) *Proceedings of the International Conference on Autonomous Agents*. 'WebACE: A Web Agent for Document Categorization and Exploration'

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., and Olds, C. C. (1992) 'Summarizing Monte Carlo Results in Methodological Research: The One- and Two-Factor Fixed Effects ANOVA Cases'. *J. Educ. Stat* 17, 315-339

Hearst, M. (2009) *Search User Interfaces*. Cambridge

Hembrooke, H. A., Granka, L. A., Gay, G. K., and Liddy, E. D. (2005) 'The Effects of Expertise and Feedback on Search Term Selection and Subsequent Learning'. *Journal of the American Society for Information Science and Technology* 56 (8), 861-871

Hogg, R. and Tanis, E. (2005) *Probability and Statistical Inference*. 7th edn. USA: Prentice Hall

Hsieh-Yee, I. (1993) 'Effects of Search Experience and Subject Knowledge on the Search Tactics of Novice and Experienced Searchers'. *JASIST* 44 (3), 161-174

Huai, Y. (ed.) (2011) *2011 IEEE 3rd International Conference on Communication Software and Networks, ICCSN 2011*. 'Study on Ontology-Based Personalized User Modeling Techniques in Intelligent Information Retrievals'

Huang, J., White, R. W., Buscher, G., and Wang, K. (eds.) (2012) *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Improving Searcher Models using Mouse Cursor Activity'

Huang, J., White, R. W., and Dumais, S. (eds.) (2011) *Conference on Human Factors in Computing Systems - Proceedings*. 'No Clicks, no Problem: Using Cursor Movements to Understand and Improve Search'

IBM, C. (2013) *IBM SPSS Statistics for Windows, Version 22.0. Armonk, NY: IBM Corp.* [online]

Internet Society (2015) *Internet Society Global Internet Report*

Iqbal, R., Grzywaczewski, A., Halloran, J., Doctor, F., and Iqbal, K. (2015) 'Design Implications for Task-Specific Search Utilities for Retrieval

and Reengineering of Code'. *Enterprise Information Systems*, 1751-7575

Iqbal, R., Grzywaczewski, A., James, A., Doctor, F., and Halloran, J. (eds.) (2012) *Proceedings of the 2012 IEEE 16th International Conference on Computer Supported Cooperative Work in Design, CSCWD 2012*. 'Investigating the Value of Retention Actions as a Source of Relevance Information in the Software Development Environment'

Järvelin, K. and Ingwersen, P. (2004) 'Information Seeking Research Needs Extension Toward Tasks and Technology'. *Information Research* 10 (1)

Jawaheer, G., Weller, P., and Kostkova, P. (2014) 'Modeling User Preferences in Recommender Systems: A Classification Framework for Explicit and Implicit User Feedback'. *ACM Transactions on Interactive Intelligent Systems* 4, 1-26

Joachims, T., Freitag, D., and Mitchell, T. (eds.) (1997) *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*. 'WebWatcher: A Tour Guide for the World Wide Web'

Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. (2007) 'Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search'. *ACM Transactions on Information Systems* 25 (2)

Johnson, R. (2010) *Statistics: Principles and Methods*. USA: John Wiley & Sons

Jung, K. (2001) *Modeling Web User Interest with Implicit Indicators*. [online] Master Thesis thesis or dissertation: Florida Institute of Technology

Jung, S., Herlocker, J. L., and Webster, J. (2007) 'Click Data as Implicit Relevance Feedback in Web Search'. *Information Processing and Management* 43 (3), 791-807

Juvina, I. and van Oostendorp, H. (2006) 'Individual Differences and Behavioral Metrics Involved in Modeling Web Navigation'. *Universal Access in the Information Society* 4 (3), 258-269

Kellar, M., Watters, C., and Shepherd, M. (2007) 'A Field Study Characterizing Web-Based Information-Seeking Tasks'. *Journal of the American Society for Information Science and Technology* 58 (7), 999-1018

Kellar, M., Watters, C., Duffy, J., and Shepherd, M. (2004) 'Effect of Task on Time Spent Reading as an Implicit Measure of Interest'. *Proceedings of the ASIST Annual Meeting* 41, 168-175

Kelly, D. (2004) *Understanding Implicit Feedback and Document Preference: A Naturalistic User Study*. [online] Doctor of Philosophy thesis or dissertation: Graduate School-New Brunswick Rutgers, The State University of New Jersey

Kelly, D. and Teevan, J. (2003) 'Implicit Feedback for Inferring User Preference'. *SIGIR Forum* 37 (2), 18-28

Kelly, D. (2009) 'Methods for Evaluating Interactive Information Retrieval Systems with Users'. *Foundations and Trends in Information Retrieval* 3 (1-2), 1-224

Kelly, D. and Belkin, N. J. (eds.) (2004) *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Display Time as Implicit Feedback: Understanding Task Effects'

Kelly, D. and Cool, C. (eds.) (2002) *Proceedings of the ACM International Conference on Digital Libraries*. 'The Effects of Topic Familiarity on Information Search Behavior'

Kim, J., Oard, J. D. W., and Romanik, K. (2000) *User Modeling for Information Access Based on Implicit Feedback*. University of Maryland, College Park: Technical Report: HCIL-TR-2000-11/UMIACS-TR-2000-29/CS-TR-4136

Kim, H. R. and Chan, P. K. (eds.) (2005) *WEBIST 2005 - 1st International Conference on Web Information Systems and Technologies, Proceedings*. 'Implicit Indicators for Interesting Web Pages'

Koene, A., Perez, E., Carter, C., Statache, R., Adolphs, S., O'Malley, C., Rodden, T., and McAuley, D. (2016) 'Privacy Concerns Arising from Internet Service Personalization Filters'. *ACM SIGCAS Computers and Society - Special Issue on Ethicomp* 45(3), 167-171

Konstan, J. A., Miller, B. N., Maltz, D., Herlocker, J. L., Gordon, L. R., and Riedl, J. (1997) 'Applying Collaborative Filtering to Usenet News'. *Communications of the ACM* 40 (3), 77-87

Kuhlthau, C. C. (1993) 'A Principle of Uncertainty for Information Seeking'. *Journal of Documentation* 49 (4), 339-355

Kumar, A. and Ashraf, M. (eds.) (2015) *International Conference on Computing, Communication and Automation, ICCCA 2015*. 'Efficient Technique for Personalized Web Search using Users Browsing History'

Lavrenko, V. (2014) *Evaluation 12: Mean Average Precision* [online] [10 November 2015]

Lee, T. Q., Park, Y., and Park, Y. T. (2008) 'A Time-Based Approach to Effective Recommender Systems using Implicit Feedback'. *Expert Systems with Applications* 34 (4), 3055-3062

Leiva, L. A. and Huang, J. (2015) 'Building a Better Mousetrap: Compressing Mouse Cursor Activity for Web Analytics'. *Information Processing & Management* 51 (2), 114-129

Li, Y. and Belkin, N. J. (2008) 'A Faceted Approach to Conceptualizing Tasks in Information Seeking'. *Information Processing and Management* 44 (6), 1822-1837

Lieberman, H. (ed.) (1997) *Conference on Human Factors in Computing Systems - Proceedings*. 'Autonomous Interface Agents'

Limbu, D. K., Connor, A. M., Pears, R., and MacDonell, S. G. (eds.) (2009) *ITNG 2009 - 6th International Conference on Information Technology: New Generations*. 'Improving Web Search using Contextual Retrieval'

Liu, C., Belkin, N. J., and Cole, M. J. (eds.) (2012) *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Personalization of Search Results using Interaction Behaviors in Search Sessions'

Liu, C., Liu, J., Belkin, N., Cole, M., and Gwizdka, J. (2011) 'Using Dwell Time as an Implicit Measure of Usefulness in Different Task Types'. *Proceedings of the ASIST Annual Meeting* 48

Liu, D. -. and Wu, I. -. (2008) 'Collaborative Relevance Assessment for Task-Based Knowledge Support'. *Decision Support Systems* 44 (2), 524-543

Liu, J., Liu, C., and Belkin, N. (2013) 'Examining the Effects of Task Topic Familiarity on Searchers' Behaviors in Different Task Types'. *Proceedings of the ASIST Annual Meeting* 50 (1)

Liu, J., Liu, C., Gwizdka, J., and Belkin, N. J. (eds.) (2010) *SIGIR 2010 Proceedings - 33rd Annual International ACM SIGIR Conference on Re-*

search and Development in Information Retrieval. 'Can Search Systems Detect Users' Task Difficulty? some Behavioral Signals'

Lix, L. M., Keselman, J. C., and Keselman, H. J. (1996) 'Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test'. *Rev. Educ. Res.* 66, 579-619

Lopez, S., Revel, A., Lingrand, D., and Precioso, F. (eds.) (2015) *Proceedings - International Conference on Image Processing, ICIP*. 'One Gaze is Worth Ten Thousand (Key-)Words'

Luhn, H. P. (1958) 'A Business Intelligence System'. in *Pioneer of Information Science. Selected Works*. ed. by Anon: NY: Spartan Books, 132-139

Ma, Z. and Kaban, A. (eds.) (2013) *2013 13th UK Workshop on Computational Intelligence, UKCI 2013*. 'K-Nearest-Neighbours with a Novel Similarity Measure for Intrusion Detection'

Maglio, P. P., Barrett, R., Campbell, C. S., and Selker, T. (eds.) (2000) *International Conference on Intelligent User Interfaces, Proceedings IUI*. 'SUITOR: An Attentive Information System'

Manning, D., C., Raghavan, P., and Schütze, H. (2008) *Introduction to Information Retrieval*. York, NY, USA: Cambridge University Press

Marchionini, G. (1992) 'Interfaces for End-User Information Seeking'. *Journal of the American Society for Information Science* 43 (2), 156-163

Mobasher, B., Cooley, R., and Srivastava, J. (2000) 'Automatic Personalization Based on Web Usage Mining'. *Communications of the ACM* 43 (8), 142-151

Morita, M. and Shinoda, Y. (eds.) (1994) *In Proceedings of SIGIR Conference on Research and Development*. 'Information Filtering Based on User Behaviour Analysis and Best MatchText Retrieval'

Neji, M., Ben Ammar, M., and Alimi, A. M. (eds.) (2011) *2011 IEEE Global Engineering Education Conference, EDUCON 2011*. 'Real-Time Affective Learner Profile Analysis using an EMASPEL Framework'

Nguyen, H., Santos Jr., E., Chevalier, M., Julien, C., and Soule-Depuy, C. (eds.) (2009) *Proc. Collaborative Social Inf. Retrieval Access: Techn. Im-*

proved User Model. 'Modeling Users for Adaptive Information Retrieval by Capturing User Intent'

Nichols, D. M. (ed.) (1997) *In Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*. 'Implicit Ratings and Riltering' at Budapaest, Hungary, ERCIM

Núñez-Valdez, E. R., Lovelle, J. M. C., Hernández, G. I., Fuente, A. J., and Labra-Gayo, J. E. (2015) 'Creating Recommendations on Electronic Books: A Collaborative Learning Implicit Approach'. *Computers in Human Behavior* 51, 1320-1330

Núñez-Valdéz, E. R., Cueva Lovelle, J. M., Sanjuán Martínez, O., García-Díaz, V., Ordoñez De Pablos, P., and Montenegro Marín, C. E. (2012) 'Implicit Feedback Techniques on Recommender Systems Applied to Electronic Books'. *Computers in Human Behavior* 28 (4), 1186-1193

O'Day, V. and Jeffries, R. (eds.) (1993) *In Proceeding of the INTERCHI-conference on Human Factors in Computing Systems (CHI '93)*. 'Orienteering in a Information Landscape: How Information Seekers Get from here to There' at Amsterdam

Oard, D. and Kim, J. (eds.) (1998) *In Proceedings of the AAAI Workshop on Recommender Systems*. 'Implicit Feedback for Recommendation Systems'. held July

Pasi, G. (ed.) (2014) *Procedia Computer Science*. 'Implicit Feedback through User-System Interactions for Defining User Models in Personalized Search'

Ramírez, J. M., Donadeu, J., and Neves, F. J. (2000) *Poirot: A Relevance-Based Web Search Agent*

Rendle, S. (2010) *Context-Aware Ranking with Factorization Models*. Berlin: Springer-Verlag

Rocchio, J. J. (ed.) (1971) *In the SMART Retrieval System - Experiments in Automatic Document Processing*. 'Relevance Feedback in Information Retrieval'

Romero, C., Ventura, S., Zafra, A., and Bra, P. d. (2009) 'Applying Web Usage Mining for Personalizing Hyperlinks in Web-Based Adaptive Educational Systems'. *Computers and Education* 53 (3), 828-840

Salojärvi, J., Puolamäki, K., and Kaski, S. (2005) *Implicit Relevance Feedback from Eye Movements*.

Salton, G. and Buckley, C. (1990) 'Improving Retrieval Performance by Relevance Feedback'. *Journal of the American Society for Information Science*. 44(4), 288-297

Salton, G. and Buckley, C. (1988) 'Term-Weighting Approaches in Automatic Text Retrieval'. *Information Processing and Management* 24 (5), 513-523

Salton, G., Wong, A., and Yang, C. S. (1975) 'VECTOR SPACE MODEL FOR AUTOMATIC INDEXING.'. *Communications of the ACM* 18 (11), 613-620

Sanderson, M. and Zobel, J. (eds.) (2005) *In Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '05*. 'Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability' at New York, NY, USA

Shapira, B., Taieb-Maimon, M., and Moskowitz, A. (eds.) (2006) *Proceedings of the ACM Symposium on Applied Computing*. 'Study of the Usefulness of Known and New Implicit Indicators and their Optimal Combination for Accurate Inference of Users Interests'

Shi, L., Cristea, A. I., Awan, M. S., Stewart, C., and Hendrix, M. (eds.) (2013) *19th Americas Conference on Information Systems, AMCIS 2013 - Hyperconnected World: Anything, Anywhere, Anytime*. 'Towards Understanding Learning Behavior Patterns in Social Adaptive Personalized E-Learning Systems'

Siddiqui, S. (2011) 'Information Seeking Behaviour of B.Tech. and M.B.B.S.Students in Lucknow: A Comparative Study'. *Journal of Library & Information Science* 1 (1), 55-70

Smucker, M. D., Allan, J., and Carterette, B. (eds.) (2007) *In Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management, CIKM '07*. 'A Comparison of Statistical Significance Tests for Information Retrieval Evaluation' at New York, NY, USA

Su, L. T. (1994) 'The Relevance of Recall and Precision in User Evaluation'. *Journal of the American Society for Information Science* 45 (3), 207-217

Su, L. T. (1992) 'Evaluation Measures for Interactive Information Retrieval'. *Information Processing and Management* 28 (4), 503-516

Takano, K. and Li, K. F. (eds.) (2009) *Proceedings - International Conference on Advanced Information Networking and Applications, AINA*. 'An Adaptive Personalized Recommender Based on Web-Browsing Behavior Learning'

Teevan, J., Dumais, S. T., and Horvitz, E. (eds.) (2005) *SIGIR 2005 - Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Personalizing Search Via Automated Analysis of Interests and Activities'

Velayathan, G. and Yamada, S. (eds.) (2007) *Proceedings - 2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT 2006 Workshops Proceedings)*. 'Behavior-Based Web Page Evaluation'

Wei, S., Zheng, X., Chen, D., and Chen, C. (2016) 'A Hybrid Approach for Movie Recommendation Via Tags and Ratings'. *Electronic Commerce Research and Applications* 56, 1-18

White, R. W. and Buscher, G. (eds.) (2012) *SIGIR'12 - Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*. 'Text Selections as Implicit Relevance Feedback'

White, R. W., Dumais, S. T., and Teevan, J. (eds.) (2009) *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, WSDM'09*. 'Characterizing the Influence of Domain Expertise on Web Search Behavior'

White, R. W. and Kelly, D. (eds.) (2006) *International Conference on Information and Knowledge Management, Proceedings*. 'A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance'

Wilson, T. D. (1999) 'Models in Information Behaviour Research'. *Journal of Documentation* 55 (3), 249-270

Xu, H., Zhang, S., and Huang, H. (eds.) (2010) *Proceedings - 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2010*. 'A Novel Personalized Recommendation System of Digital Resources Based on Semantics'

Xu, S., Jiang, H., and Lau, F. C. M. (eds.) (2010) *ACM International Conference Proceeding Series*. 'Observing Facial Expressions and Gaze Positions for Personalized Webpage Recommendation'

Xu, Y. and Chen, Z. (2006) 'Relevance Judgment: What do Information Users Consider Beyond Topicality?'. *Journal of the American Society for Information Science and Technology* 57 (7), 961-973

Yan, X. (2009) *Linear Regression Analysis Theory and Computing*. Singapore: Singapore : World Scientific Publishing Company

Yi, X., Hong, L., Zhong, E., Liu, N. N., and Rajan, S. (eds.) (2014) *RecSys 2014 - Proceedings of the 8th ACM Conference on Recommender Systems*. 'Beyond Clicks: Dwell Time for Personalization'

Zemirli, N. (ed.) (2012) *7th International Conference on Digital Information Management, ICDIM 2012*. 'WebCap: Inferring the User's Interests Based on a Real-Time Implicit Feedback'

Zhang, B., Guan, Y., Sun, H., Liu, Q., and Kong, J. (eds.) (2010) *2010 International Conference on Computer, Mechatronics, Control and Electronic Engineering, CMCE 2010*. 'Survey of User Behaviors as Implicit Feedback'

Zhang, X., Anghelescu, H. G. B., and Yuan, X. (2005) 'Domain Knowledge, Search Behaviour, and Search Effectiveness of Engineering and Science Students: An Exploratory Study'. *Information Research* 10 (2)

Zhu, Y., He, L., and Wang, X. (eds.) (2012) *Procedia Engineering*. 'User Interest Modeling and Self-Adaptive Update using Relevance Feedback Technology'

Zhu, Z., Wang, J. -., Chen, M. -., and Huang, R. -. (eds.) (2010) *Proceedings - 3rd International Conference on Information Management, Innovation Management and Industrial Engineering, ICIII 2010*. 'User Interest Modeling Based on Access Behavior and its Application in Personalized Information Retrieval'

Appendices

Appendix A1: *Task brief for participants in preliminary study (Chapter 3)*

Appendix A2: *Task brief for participants in main study (Chapter 3)*

Appendix A3: *Task brief for participants in eye gaze study (chapter 3)*

Appendix A4: *Task brief for participants in evaluation study (chapter 5)*

Appendix B: *Participant Consent form*

Appendix C: *Table of common documents visited by users (Chapter 4)*

Appendix D: *Queries entered by participants to evaluate the system (Chapter 5)*

Appendix E: *Tables of computed recall and precision values (chapter 5)*

Appendix F: *Tables showing raw data captured from the gaze experiment (chapter 3)*

Appendix A1

Interest-based prediction of relevant web documents

Dear participant,

Thank you for your intention to participate in this study which is conducted by Akuma Stephen (a Ph.D. student in computing). The aim of this research is to use contextual approach and learner's behavioural characteristics to predict learner's level of interest/attention on a web document and how relevant the documents are to the current task. This study is an important part of this research. Please remember that you are not the one under evaluation but the system.

The duration of the study is 60 minutes. You are to read all the given documents and prepare a presentation on ***ethical issues in big data***. The experiment is composed of two phases:

1. **Reading Phase**

In this phase, you are to read through all the given documents and judge the usefulness of each document by rating them according to the task at hand. This is just to categorise the documents according to relevance for later use.

2. **Presentation writing Phase**

In this phase, the document will be presented to you according to how you categorised them in the first phase. You will use the documents you categorise as relevant to prepare your presentation.

All data collected about you during the course of this study will be kept confidential and used only for academic purpose. Please complete the consent form before starting the experiment.

Please, kindly contact Akuma Stephen (akumas@uni.coventry.ac.uk) or Dr Chrisina Jayne (ab1527@coventry.ac.uk) if you have any question.

Note: This research is only for students who are 18 years old and above. Thank you.

PROCEDURE (STEP BY STEP)

1) Complete the consent form

2) Open the Firefox web browser
3) Type in this website:
   http://creative.coventry.ac.uk/~akuma/codecallz/index.php
4) Register with any user name & password then login
5) Read each of the listed documents and Close them after reading by clicking the “X” button on the top right-hand corner of your screen
6) Rate the documents according to how relevant they are in helping you to prepare for the presentation by typing in a number from 0 – 5.

Thank you!

Appendix A2

USER_ID:

INFERRING USERS INTEREST ON WEB DOCUMENTS THROUGH THEIR IMPLICIT BEHAVIOUR

Dear participant,

Thank you for your intention to participate in this study which is conducted by Akuma Stephen (a Ph.D. student in computing). The aim of this research is to use contextual approach and learner's behavioural characteristics to predict learner's level of interest/attention on a web document and how relevant the documents are to the current task. This study is an important part of this research. Please remember that you are not the one under evaluation but the system. The duration of the study is 45 minutes.

Login into a Computer then visit this web page: (http://goo.gl/iExZv7), you will find a compressed folder called "FirefoxPortable". Download and extract the folder then open it and double click the file "FirefoxPortable.exe". Open a new tab in the Firefox browser then freely search the web for answers to the given task below. For any web page you visit that asks for your User ID, do the following:

i. Enter the User ID as shown on top of this document and read the web page
ii. Close the current page (tab) after reading it and state your familiarity with the web document, also state whether the web document is difficult to understand.
iii. Rate the web document according to relevance to the task
iv. Open an MS-word document and write a 200 words summary of the solution to the task under consideration. Write boldly your User-ID on top of your word document then print the report.

TASK 1

GIG Software Development company employed you as a consultant to provide a solution to the Company's pressing problem of developing a customised software within a minimal time frame. Some professional software developers achieved this by using the Rational Unified Process

while others used the waterfall model. Which of the approaches would you consider for a small project of few lines of code (LOC) and what stage of the software lifecycle do you consider to be the most important?

TASK 2

Google is looking for young and ambitious students of Computer science for an internship to work under the Company's Service Management Department. Consider that you are shortlisted for an interview among 2000 applicants and you are asked to search the internet and find answers to questions related to Information Technology Infrastructure Library (ITIL):

i. What are the five stages of the ITIL lifecycle?
ii. What are the differences between ITIL v1, v2 and v3 (2007)?
iii. What are ITIL processes?
iv. What are ITIL functions?
v. Who should use ITIL?
vi. When should ITIL be used?
vii. What are the differences between ITIL and ISO/IEC

Visit and read at least 5 web pages and state reasons for your answer in your report. *(You can copy points from web pages and paste them in your Ms-word document or you can paraphrase them)*

All data collected about you during the course of this study will be kept confidential and used only for academic purpose. Please complete the consent form before starting the experiment.

Please, kindly contact Akuma Stephen (akumas@uni.coventry.ac.uk) or Dr Chrisina Jayne (ab1527@coventry.ac.uk) if you have any question.

Note: This research is only for students who are 18 years old and above.

Thank you.

Appendix A3

VALIDATING THE PREDICTIVE STRENGTH OF SOME SELECTED IMPLICIT INDICATORS WITH USER EYE GAZE

Dear participant,

Thank you for your intention to participate in this study which is conducted by Stephen Akuma (a Ph.D. student in computing). The aim of this research is to use contextual approach and learner's behavioural characteristics to predict learner's level of interest/attention on a web document and how relevant the documents are to the current task. This study is an important part of this research. Please remember that you are not the one under evaluation but the system.

The duration of the study is 30 minutes. You are to read through all the given documents and judge the usefulness of each document by rating them according to how they relate to the task below

SIMULATED TASK

[See TASK 1 in Appendix A2]

All data collected about you during the course of this study will be kept confidential and used only for academic purpose. Please complete the consent form before starting the experiment.

Please, kindly contact Akuma Stephen (akumas@uni.coventry.ac.uk) or Dr Chrisina Jayne (ab1527@coventry.ac.uk) if you have any question.

Note: This research is only for students who are 18 years old and above.

Thank you!

Appendix A4

IMPLICIT FEEDBACK SYSTEM FOR THE RECOMMENDATION OF RELEVANT WEB DOCUMENTS

Dear participant,

Thank you for your intention to participate in this study which is conducted by Stephen Akuma (a Ph.D. student in computing). The aim of this study is to validate the predictive strength of a derived implicit model and to show that the quality of search results improves when queries are supplemented by users' post-click behaviour. This study is an important part of this research. Please remember that you are not the one under evaluation but the system.

The experiment will last for approximately 30 minutes. You are to use the given retrieval system to enter one (best) query that can retrieve relevant documents relating to the given task below. Then visit and read the 10 results presented to you in a descending order. For each of the web document you visit:

i. Enter the User ID given to you and read the web document
ii. Close the current page (tab) after reading it and rate the web document according to relevance to the task. The rating is from 0 to 5

SIMULATED TASK

[See TASK 1 in Appendix A2].

All data collected about you during the course of this study will be kept confidential and used only for academic purpose. Please complete the consent form before starting the experiment.

Please, kindly contact Stephen Akuma (akumas@uni.coventry.ac.uk) or Dr Rahat Iqbal (aa0535@coventry.ac.uk) if you have any question.

Note: This research is only for students who are 18 years old and above.

Thank you!

Appendix B

RESEARCH CONSENT FORM

Full title of Project: Interest-based prediction of relevant web documents

Name, position and contact address of Researcher:

Akuma Stephen, Ph.D. student in computing, EC – Coventry

akumas@uni.coventry.ac.uk

		Please Tick
1.	I confirm that I have read and understand the information sheet for the above study and have had the opportunity to ask questions.	☐
2.	I understand that my participation is voluntary and that I am free to withdraw at any time, without giving a reason.	☐
3.	I agree to take part in the above study.	☐
4.	I understand that all information about me will be treated in strict confidence and that I will not be named in any written work arising from this study	☐

Name of Participant	Date	Signature
Name of Researcher	Date	Signature

Appendix C

URL	Number of Occurrence	% of Occurrence	Mean Rating
http://stackoverflow.com/questions/25329332/how-to-fix-this-deadlock-code-in-java	3	0.87	1.33
http://www.inc.com/articles/2000/01/16379.html	2	0.58	1.5
http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Waterfall_model.html	8	2.33	2
http://en.wikipedia.org/wiki/Information_Technology_Infrastructure_Library	2	0.58	2
http://www.resgroup.com/accounting-software-steps-efficient-budget-management	2	0.58	2
http://searchsoftwarequality.techtarget.com/definition/waterfall-model	3	0.87	2.33
http://ccollins.wordpress.com/2008/06/11/unified-process-vs-agile-processes/	2	0.58	2.5
http://smallbusiness.chron.com/difference-between-marketing-products-services-650.html	2	0.58	2.5
http://www.brighthubpm.com/agile/50473-agile-vs-waterfall-is-there-a-real-winner/	2	0.58	2.5
http://yourbusiness.azcentral.com/product-vs-service-examples-14403.html	2	0.58	2.5
http://www.webopedia.com/TERM/R/RUP.html	5	1.46	2.6
http://www.seguetech.com/blog/2013/07/05/waterfall-vs-agile-right-development-methodology	3	0.87	2.67
http://smallbusiness.chron.com/three-benefits-would-rup-bring-organization-32161.html	11	3.21	2.73
http://www.ianswer4u.com/2011/11/advantages-and-disadvantages-of.html#axzz3JVIYH0Fb	3	0.87	3
http://www.itiltraining.com/itil-benefits.asp	3	0.87	3
http://www.journaldev.com/1058/java-deadlock-example-and-how-to-analyze-deadlock-situation	3	0.87	3
http://en.wikipedia.org/wiki/Software_development_process	2	0.58	3
http://en.wikipedia.org/wiki/Unified_Modeling_Language	2	0.58	3
http://www.experts-exchange.com/Programming/Project_Management/A_6536-ITIL-v2-Vs-v3-A-Comparison.html	2	0.58	3
http://www.motorwayservices.info/	2	0.58	3
http://en.wikipedia.org/wiki/Service	7	2.04	3.14
http://stackoverflow.com/questions/20560514/unified-process-vs-waterfall-model	12	3.5	3.17
http://wiki.en.it-process-maps.com/index.php/Comparison_between_ITIL_V3_and_ITIL_	5	1.46	3.2

V2_-_The_Main_Changes			
http://www.techterms.com/definition/rup	5	1.46	3.25
http://support.it-qms.com/hc/communities/public/questions/200402081-What-is-the-difference-between-ITIL-v1-v2-v3-and-ITIL-2011	10	2.92	3.3
http://benefitof.net/benefits-of-rational-unified-process/	4	1.17	3.33
http://www.slideshare.net/rahultilloo/water-fall-model-22606242	3	0.87	3.33
http://www.tutorialspoint.com/java/java_thread_deadlock.htm	7	2.04	3.43
http://www.differencebetween.com/difference-between-waterfall-methodology-and-vs-rup/	6	1.75	3.5
http://www.waterfall-model.com/	6	1.75	3.5
http://en.wikipedia.org/wiki/Product_(business)	2	0.58	3.5
http://tutorials.jenkov.com/java-concurrency/deadlock-prevention.html	2	0.58	3.5
http://www.investorwords.com/6664/service.html	2	0.58	3.5
http://www.slideshare.net/maheshpanchal1/rup-1226744	2	0.58	3.5
http://searchsoftwarequality.techtarget.com/definition/Rational-Unified-Process	10	2.92	3.6
http://www.projectsmart.co.uk/which-life-cycle-is-best-for-your-project.php	5	1.46	3.6
http://stackoverflow.com/questions/1102359/programmatic-deadlock-detection-in-java	3	0.87	3.67
http://www.techrepublic.com/article/understanding-the-pros-and-cons-of-the-waterfall-model-of-software-development/	3	0.87	3.67
http://en.wikipedia.org/wiki/Waterfall_model	21	6.12	3.81
http://en.wikipedia.org/wiki/Rational_Unified_Process	11	3.21	3.82
http://wiki.en.it-processmaps.com/index.php/ITIL_Functions	12	3.5	3.83
http://www.businessdictionary.com/definition/product.html	6	1.75	3.83
http://istqbexamcertification.com/what-is-waterfall-model-advantages-disadvantages-and-when-to-use-it/	21	6.12	3.86
http://www.connectsphere.com/resource/articles/what-is-the-itil-service-lifecycle	10	2.92	4
http://blog.aecsoftware.com/2012/10/comparing-agile-and-waterfall-methods-of-project-management/	4	1.17	4
http://www.base36.com/2012/12/agile-waterfall-methodologies-a-side-by-side-comparison/	3	0.87	4
http://artofservice.com.au/what-is-the-difference-between-itil-v2-and-itil-v3/	2	0.58	4
http://en.it-processmaps.com/products/itil-process-map.html	2	0.58	4
http://www.tsoshop.co.uk/parliament/bookstore.asp?FO=1229332&DI=571307	4	1.17	4.25
http://www.techrepublic.com/article/10-things-you-should-know-about-itil/	2	0.58	5

Appendix D

UI	Approach 1 Queries for Solr-Indexed system
400	Rational Unified Process vs Waterfall Model for Small projects
401	software applications in unified model
402	Rational Unified Process vs Waterfall Model for Small projects
403	RUP and waterfall with LOC
404	software development techniques suitable for short software project

UI	Approach 1 Queries for Aggregated system
407	RUP vs Waterfall for small projects
408	customizing software within a short time frame
409	how rational unified ca process employed problem
410	software development methods
411	RUP and waterfall model

UI	Approach 1 Queries for Google
412	“rational unified process” vs waterfall
413	RUP and waterfall + LOC
414	Small project for RUP or Waterfall model
415	Best software development method for small project
416	RUP vs Waterfall Model for Small projects

UI	Approach 2 Queries
500	rup vs waterfall model
501	rup vs waterfall
502	the benefit of RUP over waterfall model
503	rup and waterfall
504	RUP VS. WATERFALL
505	benefits of waterfall model
506	The benefits of RUP over waterfall model
507	benefits of RUP
508	benefits of rup over waterfall model
510	what are the benefit of RUP over waterfall model
511	benefits of RUP over waterfall model

Appendix E

The results presented in this appendix is the computation of the Mean Average Precision (MAP) for each of the system evaluated. "Pos" means the position of the documents and its relevancy is denoted as 0 or 1. "0" means not relevant while "1" means relevant. The precision and Recall for each user query is computed and the average precisions (AVG) is obtained. The MAP values are then computed from the average precisions. An example of how the precision and recall is calculated is given below:

$$\text{Precision} = \frac{\text{number of relevant items retrieved}}{\text{total retrieved items}}$$

$$\text{Recall} = \frac{\text{number of relevant items retrieved}}{\text{total relevant items retrieved}}$$

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10
User 1	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**1**
Precision	0	0.5	0.33	0.25	0.2	0.33	0.43	0.38	0.33	0.4
Recall	0	0.25	0.25	0.25	0.25	0.5	0.75	0.75	0.75	1

For example the precision and recall for the document on position 6 is calculated as follows:

Number of relevant items retrieved = 2; Total retrieved items = 6; Total relevant items retrieved = 4.

Precision = 2/6 = 0.33

Recall = 2/4 = 0.5

APPROACH 1 RESULTS

1) GOOGLE

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 1	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**1**		
Precision	0	0.5	0.33	0.25	0.2	0.33	0.43	0.38	0.33	0.4	0.42	0.5
Recall	0	0.25	0.25	0.25	0.25	0.5	0.75	0.75	0.75	1		
User 2	**0**	**1**	**1**	**0**	**1**	**0**	**0**	**0**				
Precision	0	0.5	0.67	0.5	0.6	0.5	0.43	0.38			0.59	0.59
Recall	0	0.33	0.67	0.67	0.67	0.67	0.67	0.67				
User 3	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**		
Precision	0	0	0.33	0.5	0.4	0.33	0.43	0.5	0.44	0.4	0.44	0.42
Recall	0	0	0.25	0.5	0.5	0.5	0.75	1	0	0		
User 4	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**		
Precision	1	0.5	0.33	0.5	0.6	0.5	0.43	0.38	0.44	0.4	0.64	0.7
Recall	0.25	0.25	0.25	0.5	0.75	0.75	0.75	0.75	1	1		
User 5	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**		
Precision	0	0.5	0.33	0.5	0.4	0.33	0.43	0.375	0.33	0.3	0.48	0.5
Recall	0	0.33	0.33	0.67	0.67	0.67	1	1	1	1		

$MAP10_{google}$ = (0.42+0.59+0.44+0.64+0.48)/5 = 0.51

$MAP5_{google}$ = (0.5+0.59+0.42+0.7+0.5)/5 = 0.54

2) SOLR-INDEXED SYSTEM

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 1	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**		
Precision	1	0.5	0.67	0.5	0.4	0.5	0.43	0.5	0.44	0.40	0.67	0.835
Recall	0.25	0.25	0.5	0.5	0.5	0.75	0.75	1	1	1		
User 2	**1**	**1**	**1**	**1**	**0**	**1**	**1**					
Precision	1	1	1	1	0.8	0.83	0.86				0.95	1
Recall	0.17	0.33	0.5	0.67	0.67	0.83	1					
User 3	**1**	**0**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**		
Precision	1	0.5	0.67	0.75	0.8	0.67	0.57	0.50	0.44	0.4	0.724	0.85
Recall	0.2	0.2	0.4	0.6	0.8	0.8	0.8	0.8	0.8	1		
User 4	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**		
Precision	1	1	0.67	0.5	0.4	0.33	0.29	0.25	0.22	0.2	1	1
Recall	0.5	1	1	1	1	1	1	1	1	1		
User 5	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**				
Precision	0	0.5	0.5	0.5	0.4	0.33	0.29	0.25			0.5	0.5
Recall	0	0.5	0.5	1	1	1	1	1				

$MAP10_{solr\text{-}indexed} = (0.67+0.95+0.724+1+0.5)/5 = 0.77$

$MAP5_{solr\text{-}indexed} = (0.835+1+0.85+1+0.5)/5 = 0.84$

3) AGGREGATED SYSTEM

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 6	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**	**0**			
Precision	1	0.5	0.67	0.5	0.6	0.67	0.57	0.63	0.56		0.71	0.76
Recall	0.2	0.2	0.4	0.4	0.6	0.8	0.8	1	1			
User 7	**1**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**0**		
Precision	1	0.5	0.67	0.75	0.6	0.5	0.57	0.63	0.67	0.6	0.715	0.81
Recall	0.2	0.2	0.33	0.33	0.33	0.33	0.67	0.83	1	1		
User 8	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**		
Precision	1	1	1	1	0.8	0.67	0.57	0.5	0.56	0.6	0.86	1
Recall	0.17	0.33	0.5	0.67	0.67	0.67	0.67	0.67	0.83	1		
User 9	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**				
Precision	1	1	1	1	1	1	1	1			1	1
Recall	0.13	0.25	0.38	0.5	0.63	0.75	0.88	1				
User 10	**1**	**1**	**1**	**1**	**1**	**0**	**0**					
Precision	1	1	1	1	1	0.83	0.71				1	1
Recall	0.14	0.29	0.43	0.57	0.57	0.57	0.57					

$MAP10_{model}$ = (0.71+0.715+0.86+1+1)/5 = 0.86

$MAP5_{model}$ = (0.76+0.81+1+1+1)/5 = 0.91

APPROACH 2 RESULTS

GOOGLE

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 1	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**	**1**		
Precision	0	0	0.3	0.5	0.4	0.33	0.29	0.38	0.33	0.4	0.395	0.4
Recall	0	0	0.25	0.5	0.5	0.5	0.5	0.75	0.75	1		
User 2	**1**	**0**	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**0**		
Precision	1	0.5	0.33	0.5	0.4	0.5	0.43	0.375	0.44	0.4	0.61	0.75
Recall	0.25	0.25	0.25	0.5	0.5	0.75	0.75	0.75	1	1		
User 3	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**1**	**0**	**1**		
Precision	0	0.5	0.33	0.5	0.4	0.33	0.43	0.5	0.44	0.5	0.49	0.5
Recall	0	0.2	0.2	0.4	0.4	0.4	0.6	0.8	0.8	1		
User 4	**0**	**1**	**0**	**1**	**0**	**0**	**0**	**0**				
Precision	0	0.5	0.5	0.5	0.4	0.33	0.29	0.25			0.5	0.5
Recall	0	0.5	0.5	1	1	1	1	1				
User 5	**0**	**0**	**1**	**1**	**0**	**0**	**1**	**1**	**0**	**0**		
Precision	0	0	0.33	0.5	0.4	0.33	0.43	0.5	0.44	0.4	0.44	0.42
Recall	0	0	0.25	0.5	0.5	0.5	0.75	1	0	0		
User 6	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**	**0**	**0**		
Precision	0	0.5	0.33	0.25	0.2	0.33	0.43	0.38	0.33	0.3	0.42	0.5
Recall	0	0.33	0.33	0.33	0.33	0.67	1	1	1	1		
User 7	**0**	**1**	**0**	**1**	**0**	**0**	**1**	**0**	**0**	**0**		
Precision	0	0.5	0.33	0.5	0.4	0.33	0.43	0.375	0.33	0.3	0.48	0.5
Recall	0	0.33	0.33	0.67	0.67	0.67	1	1	1	1		
User 8	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**		
Precision	0	0	0.3	0.5	0.6	0.5	0.57	0.5	0.44	0.4	0.49	0.47
Recall	0	0	0.25	0.5	0.75	0.75	1	1	1	1		
User 9	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	**0**		
Precision	1	0.5	0.33	0.5	0.4	0.33	0.29	0.25	0.33	0.33	0.61	0.75
Recall	0.33	0.33	0.33	0.67	0.67	0.67	0.67	0.67	1	1		
User 10	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**1**	**1**	**0**		
Precision	1	0.5	0.33	0.5	0.4	0.33	0.29	0.38	0.44	0.4	0.58	0.75
Recall	0.25	0.25	0.25	0.5	0.5	0.5	0.5	0.75	1	0		
User 11	**1**	**0**	**0**	**1**	**1**	**0**	**0**	**0**	**1**	**0**		
Precision	1	0.5	0.33	0.5	0.6	0.5	0.43	0.38	0.44	0.4	0.635	0.7
Recall	0.25	0.25	0.25	0.5	0.75	0.75	0.75	0.75	1	1		

$MAP10_{google}$ = (0.395+0.61+0.49+0.5+0.44+0.42+0.48+0.49+0.61+0.58+0.635) = 0.51

$MAP5_{google}$ = (0.4+0.75+0.5+0.5+0.42+0.5+0.5+0.5+0.47+0.75+0.75+0.7) = 0.57

SOLR-INDEXED SYSTEM

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 1	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**0**	**1**			
Precision	1	1	1	0.75	0.8	0.67	0.57	0.5	0.56		0.87	0.95
Recall	0.2	0.4	0.6	0.6	0.8	0.8	0.8	0.8	1			
User 2	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**			
Precision	1	1	1	0.75	0.8	0.83	0.86	0.88	0.78		0.91	0.95
Recall	0.14	0.29	0.43	0.43	0.57	0.71	0.86	0.86	1			
User 3	**1**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0**			
Precision	1	1	0.67	0.5	0.6	0.5	0.57	0.63	0.56		0.76	0.87
Recall	0.2	0.4	0.4	0.4	0.6	0.6	0.8	1	1			
User 4	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**1**				
Precision	1	1	1	1	1	0.83	0.71	0.75			0.96	1
Recall	0.17	0.33	0.5	0.67	0.83	0.83	0.83	1				
User 5	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**		
Precision	1	0.5	0.33	0.25	0.2	0.17	0.14	0.25	0.22	0.3	0.52	1
Recall	0.33	0.33	0.33	0.33	0.33	0.33	0.33	0.67	0.67	1		
User 6	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**		
Precision	1	1	1	1	0.8	0.83	0.71	0.75	0.78	0.8	0.895	1
Recall	0.13	0.25	0.38	0.5	0.5	0.63	0.63	0.75	0.88	1		
User 7	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**0**	**1**		
Precision	1	0.5	0.67	0.5	0.6	0.5	0.57	0.63	0.56	0.6	0.678	0.76
Recall	0.17	0.17	0.33	0.33	0.5	0.5	0.67	0.83	0.83	1		
User 8	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**			
Precision	1	1	0.67	0.75	0.6	0.67	0.57	0.63	0.5		0.81	0.92
Recall	0.2	0.4	0.4	0.6	0.6	0.8	0.8	1	1			
User 9	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**0**		
Precision	1	1	0.67	0.75	0.8	0.83	0.86	0.88	0.89	0.89	0.88	0.89
Recall	0.13	0.4	0.4	0.38	0.5	0.63	0.75	0.88	1	1		
User 10	0	0	0	0	0	**1**	**1**	0	0	**1**		
Precision	0	0	0	0	0	0.17	0.29	0.25	0.22	0.3	0.25	0
Recall	0	0	0	0	0	0.33	0.67	0.67	0.67	1		
User 11	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**		
Precision	1	1	1	1	1	1	1	1	1	1	1	1
Recall	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1		

$MAP10_{solr\text{-}indexed}$ = (0.87+0.91+0.76+0.96+0.52+0.895+0.678+0.81+0.88+0.25+1) = 0.78

$MAP5_{solr\text{-}indexed}$ = (0.95+0.95+0.87+1+1+1+0.76+0.92+0.89+0+1) = 0.85

AGGREGATED SYSTEM

	Pos 1	Pos 2	Pos 3	Pos 4	Pos 5	Pos 6	Pos 7	Pos 8	Pos 9	Pos 10	AVG P10	AVG P5
User 1	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**			
Precision	1	1	1	1	1	0.8	0.86	0.88	0.89		0.95	1
Recall	0.13	0.25	0.38	0.5	0.63	0.63	0.75	0.88	1			
User 2	**1**	**1**	**0**	**1**	**1**	**1**	**0**	**1**	**1**			
Precision	1	1	0	0.75	0.8	0.83	0.71	0.75	0.78		0.84	0.89
Recall	0.14	0.29	0.29	0.43	0.57	0.71	0.71	0.86	1			
User 3	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0**			
Precision	1	1	1	1	0.8	0.83	0.86	0.88	0.78		0.94	1
Recall	0.14	0.29	0.43	0.57	0.57	0.71	0.86	1	1			
User 4	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**				
Precision	1	1	1	0.75	0.8	0.83	0.86	0.88			0.91	0.95
Recall	0.14	0.29	0.43	0.43	0.57	0.71	0.86	1				
User 5	**1**	**0**	**0**	**1**	**1**	**1**	**0**	**1**	**0**	**0**		
Precision	1	0.5	0.33	0.5	0.6	0.67	0.57	0.63	0.56	0.5	0.68	0.7
Recall	0.2	0.2	0.2	0.4	0.6	0.8	0.8	1	1	1		
User 6	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**1**		
Precision	1	1	1	1	1	1	1	0.88	0.89	0.9	0.98	1
Recall	0.11	0.22	0.33	0.44	0.56	0.67	0.78	0.78	0.89	1		
User 7	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**1**		
Precision	1	1	1	1	1	0.83	0.86	0.75	0.67	0.7	0.94	1
Recall	0.14	0.29	0.43	0.57	0.71	0.71	0.86	0.86	0.86	1		
User 8	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**1**			
Precision	1	1	1	1	1	1	0.83	0.75	0.78		0.96	1
Recall	0.14	0.29	0.43	0.57	0.71	0.86	0.86	0.86	1			
User 9	**1**	**1**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**		
Precision	1	1	1	1	1	1	0.86	0.88	0.78	0.8	0.96	1
Recall	0.14	0.29	0.43	0.57	0.71	0.86	0.86	0.88	0.88	1		
User 10	0	0	**1**	**1**	0	0	0	**1**	0	**1**		
Precision	0	0	0.33	0.5	0.4	0.33	0.29	0.38	0.33	0.4	0.4	0.42
Recall	0	0	0.25	0.5	0.5	0.5	0.5	0.75	0.75	1		
User 11	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**		
Precision	1	1	1	1	1	1	1	1	1	1	1	1
Recall	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1		

$MAP11_{model}$ = (0.95+0.84+0.94+0.91+0.68+0.98+0.94+0.96+0.96+0.4+1) = 0.87

$MAP5_{model}$ = (1+0.89+1+0.95+0.7+1+1+1+1+0.42+1) = 0.91

Appendix F

URL 1: http://en.wikipedia.org/wiki/Waterfall_model (mean = 3.81)

URL 2: http://istqbexamcertification.com/what-is-waterfall-model-advantages-disadvantages-and-when-to-use-it (mean = 3.86)

URL 3: https://ccollins.wordpress.com/2008/06/11/unified-process-vs-agile-processes (mean 2.5)

URL 4: http://smallbusiness.chron.com/three-benefits-would-rup-bring-organization-32161.html (mean 2.73)

URL 5: http://en.wikipedia.org/wiki/Rational_Unified_Process (mean = 3.82)

URL 6: http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Waterfall_model.html (mean = 2)

Fixation count

Participants	URL1	URL 2	URL 3	URL 4	URL 5	URL 6
Rec 01	117.0	583.0	166.0	127.0	37.0	297.0
Rec 02	361.0	359.0	137.0	68.0	21.0	298.0
Rec 03	542.0	531.0	226.0	239.0	72.0	348.0
Rec 04	12.0	163.0	172.0	26.0	8.0	31.0
Rec 05	5.0	614.0	121.0	9.0	13.0	180.0
Rec 06	68.0	12.0	0.0	1.0	0.0	0.0
Rec 07	2325.0	360.0	171.0	554.0	163.0	641.0
Rec 08	1523.0	2115.0	836.0	485.0	809.0	1407.0
Rec 09	622.0	1484.0	236.0	61.0	189.0	403.0
All Re-cordings	619.4	691.2	229.4	174.4	145.8	400.6

Fixation Duration

Participants	URL1	URL 2	URL 3	URL 4	URL 5	URL 6
Rec 01	33.9	150.8	35.0	27.6	10.2	70.1
Rec 02	99.9	92.2	44.3	19.2	5.4	135.4
Rec 03	74.0	73.8	35.3	42.3	11.6	77.4
Rec 04	0.9	19.8	21.4	2.4	0.4	4.1
Rec 05	0.7	65.6	17.1	0.3	0.9	29.4
Rec 06	11.3	1.4	0.0	0.1	0.0	0.0
Rec 07	211.0	106.4	64.0	75.1	16.4	181.5
Rec 08	451.8	856.4	217.4	180.2	127.7	579.2
Rec 09	77.0	221.0	64.7	14.9	33.3	76.0
All Record-ings	106.7	176.4	55.5	40.3	22.9	128.1

Explicit relevance ratings

Participants	URL1	URL 2	URL 3	URL 4	URL 5	URL 6
Rec 01	4	5	4	4	3	4
Rec 02	4	4	5	2	3	2
Rec 03	4	4	3	4	3	2
Rec 04	4	5	5	2	5	2
Rec 05	2	4	2	2	2	1
Rec 06	4	4	3	3	2	3
Rec 07	5	4	4	4	5	3
Rec 08	4	4	3	5	4	4
Rec 09	3	4	3	2	4	3
All Re-cordings	3.75	4.25	3.625	3.125	3.625	2.625

Note: Due to poor calibration, participant 6 data was excluded from the analysis.