博士（人間科学）学位論文

Web Information Search and Sharing:
A Human-Centric Integrated Approach

人間中心の統合的アプローチによる
ウェブ情報検索と共有

２００９年１月

早稲田大学大学院　人間科学研究科

シュティフ　ロマン

Shtykh, Roman

研究指導教員：　金　群　教授

Acknowledgments

I would like to express my sincere gratitude to my supervisor Professor Qun Jin who was kindly supporting and encouraging me through my research starting from my Master Degree course at the University of Aizu.

My gratitude extends to Professor Shoji Nishimura and Professor Hideaki Kikuchi from School of Human Sciences, Waseda University, for their kind support, encouragement and advice.

I am thankful to all the students from Networked Information Systems Laboratory who participated in discussions and user studies, and my colleagues from Media Network Center, Waseda University.

Last, but not least, I would like to thank my family in Ukraine who helped me to develop and always supported my will to study, and my wife Ayako and my son Yuriy Jin for their love and patience.

Abstract

In the situation of information overload we are experiencing today, conventional Web search systems taking one-size-fits-all approach are often not capable to effectively satisfy individual information needs. To improve the quality of Web information retrieval, information systems need to know every individual user's information needs. However, knowing and correctly applying information needs is extremely difficult, often impossible. Yet knowing multiple contexts of user information behaviour can give us some conception of conceivable information a user tries to obtain in a particular context.

Recognising the importance of human factor for overcoming problems related to information overload and the significance of knowing a user and his/her context for effective personalisation as a means to overcome these problems, we are seeking for solutions to capture contexts from user-system interaction data and employ the power of collaboration in order to achieve better search and sharing in information seeking tasks.

In our research project we set a user in the centre of information seeking task. Although fragmentarily, we endeavoured to capture users' contexts through relevance feedback obtained from user-system interactions and organise them as concepts of user interests. Concepts are stored in user profiles, which are used for personalisation of users' search experiences. We recognise that contexts are changeable and they change along with users' activities – therefore dynamism is an essential condition for organisation of contexts. In addition, in order to enforce the user-centrism of the approach we are seeking to employ 'social' elements of collaborative information search.

In the thesis we discuss a collaborative personalised search approach that makes an attempt to "understand" and better satisfy information needs for each and every searching user. The approach is realised with a Web information retrieval framework called BESS (BEtter Search and Sharing). In order to facilitate a user's information seeking activities, the framework captures his/her information contexts into dynamically changing user profiles, which are further used as a source of the user's interests and expertise and applied for search personalisation.

The thesis is organised in seven chapters.

In Chapter 1 we pose the problem of information overload and discuss its issues for information acquisition and processing. We emphasise the subjective nature of the problem and thus assume that it has to be solved with close consideration of human factor, including its social aspects.

In Chapter 2 we discuss human-centric solutions for information seeking and exploration with the main focus

on personalisation, its advances in academy and business, and its problems, and deliberate on some personalisation-related issues, privacy in particular. The survey and discussion on user profiles as the main modelling approach in personalisation in general follow. Further, we give an extensive overview of user interest inference, its methods, design and use, as an important part of personalisation process, and outline the peculiarities of the proposed approach.

In Chapter 3 we propose the BESS collaborative information search and sharing framework in attempt to incorporate the discussed user-centeredness into information seeking tasks by harnessing relevance feedback from users in order to estimate their interests, construct user profiles reflecting those interests and, finally, apply them for information acquisition in online collaborative information seeking context. We discuss the conceptual basis of the framework and show its model and architecture.

To know user contexts, which are, as we have noted, very important to achieve the goals of this research and make search and sharing user-centred, we extract concepts that reflect a user's interests and organise them in multi-layered user profiles. User profile is the central part of the proposed framework. It contains one static and one dynamic part, where the latter reflect the current, short-term and long-term user interests. The layers are dynamically updated with changes of the user's interests. Their construction and update are based on *interest-change-driven profile construction* mechanism that adopts *recency*, *frequency* and *persistency* as the three important criteria reflecting the volatility of user interests and emphasising the steadiness of persistent preferences. User modelling is discussed in Chapter 4.

Concepts for profile construction can be obtained, for instance, by grouping/classifying user relevance feedback items, such as the objects selected by a user during his/her interaction with a search system or explicitly stated user preferences. In Chapter 5 we propose *High Similarity Sequence Data-Driven clustering* algorithm for concept extraction based on similarity characteristics of uniform user relevance feedback. The algorithm is online and incremental, and does not require any prior knowledge of data to be processed. Having been evaluated against different data sets, the algorithm proves to be fast, easy to implement and produces reasonable clustering results.

In Chapter 6 we show how the dynamics of interest-change-driven user profile, its contextual information and its construction process are employed for contribution co-evaluation and search. In BESS, interacting with *objective index* data of conventional search services a user creates his/her individual repository (*subjective index*) and thus contributes to the whole community of the system. Such contributions are different from objective index data – they have an additional value inferred from users' expertise derived from their individual user profiles. Therefore, when searching subjective index repositories, a document evaluated by users with deep expertise in the topic the document belongs to get higher ranking. We propose the following two factors for users' expertise assessment – *contribution activeness* and *contribution popularity*. Furthermore, in order to narrow down the number of documents to be searched, a searcher can use *focused search*, a search on the documents contributed by users with similar contexts.

Finally, Chapter 7 concludes the thesis with the summary of the presented research and outlines future research issues.

In this study we have proposed a human-centric integrated approach for Web information search and sharing. The approach has incorporated the important user-centric elements, namely a user's individual context and 'social' factor realised with collaborative contributions and co-evaluations, into Web information search. The major contributions of this study are as follows.

1) A framework for better search and sharing has been proposed;

2) The dynamic interest-change-driven model for accurate user contextual information organisation has been devised;

3) The Similarity Sequence Data-Driven clustering algorithm for concept construction has been developed;

4) Measures to evaluate our proposed model and algorithms have been defined and introduced.

The major differences of our approach from many modern conventional personalisation solutions on the conceptual level are

1) separation of subjective value-added resources from impersonal objective data of conventional search services – creation of *'social' sub-space* for subjective information accumulation and sharing through search;

2) tight coupling of information sharing and search in the following personalisation cycle

   contribution submission through search → user modelling from contributions → contribution assessment based on the user's expertise inferred from user models → search result ranking based on the user's individual and community expertise (inferred from user models) → ……

The validity and expected quality of the personalisation cycle are ensured by several important dynamic mechanisms, such as H2S2D similarity-based online incremental algorithm and *interest-change-driven profile construction*. In spite of its relative simplicity, the proposed H2S2D method has demonstrated reasonably good clustering results in terms of accuracy and precision and has proved to be suitable for fast real-time relevance feedback processing to guarantee always-updated concepts. Further, the proposed *interest-change-driven profile construction* mechanism ensures proper organisation of volatile and persistent user interests, thus, together with H2S2D clustering, always providing accurate information about the user's context change. The experiments have demonstrated the validity and efficacy of the proposed human-centric integrated approach for Web information search and sharing. BESS is a general framework following our ideas of collaborative search and sharing, and its components implementing most methods can be separated from it to be applied to other solutions.

The achievements of our study are highly expected to contribute to the field of Web information search in general, and contextual personalisation in particular.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

CA	Contribution Activeness
CIR	Cognitive Information Retrieval
CP	Contribution Popularity
H2S2D	High Similarity Sequence Data-Driven (Clustering Algorithm)
IR	Information Retrieval
IS	Information System
LMS	Learning Management System
PIR	Personalised Information Retrieval
SFI	Slide-Film Interface
UP	User Profile
WUM	Web Usage Mining

1 Introduction

1.1 Information Overload Problem

With the rapid advances of information technologies, information overload has become a phenomenon many of us have to face, and often suffer, in our daily activities, whether it be work or leisure. We all experience the problem whenever we are in need of some information, though “people who use the Internet often are likely to perceive fewer problems and confront fewer obstacles in terms of information overload” (Beaudoin, 2008). Any of us has experienced a situation when deciding to buy a certain product, say, a washing machine, and trying to figure out its characteristics, such as availability of delayed execution, steam and aquastop functions, we browsed the Web and encountered an excessive amount of information on the product. Then we had to filter out irrelevant information, categorise and analyze the remaining part to do the best choice. Many of those who work at office acquire, filter, analyze, conflate and use the collected information – the process which requires, today more than ever, special skills and software to cope with highly excessive and not always relevant information for proper decision making.

Despite of the public recognition of the problem and the great number of publications discussing and analyzing it, information overload is often a notion slightly differing in the contexts it is applied to and findings of researchers. The words itself has many synonyms, such as information explosion or information burden, and derivatives, such as salesperson’s information overload (Hunter and Goebel, 2008), to name a few. So what is ‘information overload’?

As in the example with the washing machine purchase, information overload is generally understood as the situation when there is much more information than a person is able to process. This definition is identical to that given by Miller (1956) who considered human cognitive capacity to be limited to five to nine “chunks” of information. First of all, it is often mentioned when the growing number of Web pages and difficulties related to this are discussed[1]. Considering the growing popularity of social network systems (SNS) and user-generated content, the Web is likely to remain the primary area of concern about

[1] In this work we will tackle the information acquisition on the Web, so the notion will be considered in context of Web resources.

information overload in future. Indeed, the amount of such content grows very fast[2] and becomes even threatening for men – people are at the risk of being buried with tons of information irrelevant to a particular current information need. And since information technologies in general and the Web in particular are highly employed for most human activities today the problems raises concerns in many other technology-intensive areas of human activities. However, the problem of information overload should not be considered with regard to growing information resources on the Web only – it is much wider and multidisciplinary problem encountered in sales and marketing (e.g., Jacoby et al., 1974; Hunter, 2004; Klausegger et al., 2007; Hunter and Goebel, 2008), healthcare (e.g., Hall and Walton, 2004; Kim et al., 2007), software development (Pennington and Tuttle, 2007) and other areas.

Information overload is a complex problem. It is not just about effective management of excessive information but also, as Levy (2005) argues, requiring "the creation of time and place for thinking and reflection". Himma (2007) conducted a conceptual analysis of the notion in order to clarify it from a philosophical perspective and showed that although excess is a necessary condition for being overloaded, it is not a sufficient condition. The researcher writes:

"*To be overloaded is to be in a state that is* ***undesirable*** *from the vantage point of some set of norms; as a conceptual matter, being overloaded is bad. In contrast, to have an excessive amount of* [entity] *x is merely to have more than needed, desired, or optimal*".

Thus, being overloaded implies some result on a person, and this result is of undesirable or negative nature. Generally, conception of information overload today implies such negative effects. For instance, conducting social-scientific analysis (in contrast to Himma (2007)'s philosophical approach) Mulder et al. (2006) define information overload as

"… *the feeling of stress when the information load goes beyond the processing capacity*".

The state of information overload is *individual*, in the sense it depends on personal abilities and experiences. As Chen et al. (2007) point in their research on decision-making in Internet shopping, the relationship between information load and subjective state toward decision are moderated by personal proclivities, abilities and past relevant experiences. Also though information load itself does not directly influence an individual's decisions, its excess may negatively influence the decision quality. By conducting a series of non-parametric tests and logistic regression analysis, Kim et al. (2007) determined

[2] For instance, Technorati reports 120,000 new weblogs being created worldwide every day (Sifry, 2007)

factors which predict an individual's perception of overload among cancer information seekers. The strongest factors appeared to be education level and cognitive aspects of information seeking that proves again the individual nature of the information overload and emphasises the importance of information literacy.

Information overload is *multi-faceted* concept and have various implications as to human activities, and society in general, many of them becoming known as new researches are conducted. For instance, Klausegger et al. (2007) found that information overload is experienced regardless of the nation, with its degree somewhat differing from nation to nation, – there is a significant negative relationship between the overload and work performance for all five nations the authors investigated. It was also found that the phenomenon negatively influence the degree of interpersonal trust, which is a critical component of social capital (Beaudoin, 2008). One of its plausible and severe harmful outcomes is information fatigue syndrome which includes "paralysis of analytical capacity", "a hyper-aroused psychological condition", "anxiety and self-doubt", and leads to "foolish decisions and flawed conclusions" (Reuters, 1996). Since the problem has a subjective nature, the first countermeasure is information literacy, efficient work organisation and work habits, sufficient time and concentration (Mulder et al., 2006) – again, one's strategy will depend on one's work tasks and subjective factors. Another, and not less important, countermeasure we put the focus in our research is technological. A number of solutions as to how to reduce the negative effects caused by the phenomenon are proposed. To name a few, in order to assure the quality of information and in this way reduce the problem in folksonomy-based systems, Pereira and de Silva (2008) propose cognitive authority to estimate the information quality by qualifying its sources (content authors). To reduce excess of information in wiki-based e-learning, Stickel et al. (2008) assume every link in the proposed hypertext system having a predefined life-time and use "consolidation mechanisms as found in the human memory – by letting unused things fade away" in order to remove unused links.

For more substantial information on the overload problem, interested readers are recommended to refer to Jacoby et al. (1974), Grisé and Gallupe (1999) or Himma (2007). But to summarise, though simplistically, we reflected the principal and essential components of the phenomenon in Figure 1.1:

- excessive amount of information;
- subjective and objective information processing capabilities conditioned by experience,

proclivities, etc. and environment, situation, etc. respectively;

- individual's psychological and cognitive state.

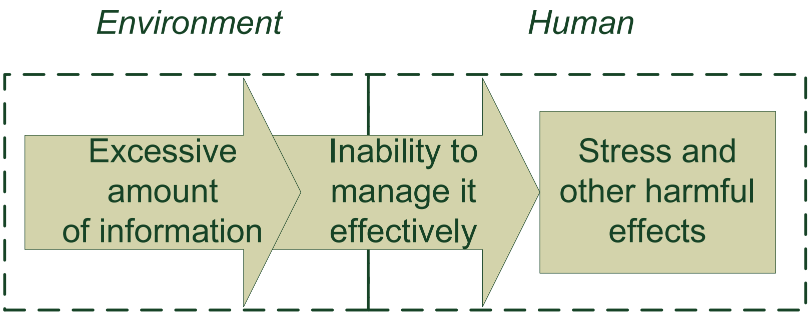

Environment
Human
Excessive amount of information
Inability to manage it effectively
Stress and other harmful effects

Figure 1.1 Information overload in process

Clearly, to alleviate the information overload for an individual, we can reduce the amount of information and/or increase our processing capabilities. Considering the fact that people with high organisation skills and information literacy have less perceived information overload and usually require better tools to process information, and people with constantly perceived information overload requires better training as to how to manage it (Janssen and de Poot, 2006), probably the first step to alleviate the problem is providing information literacy and organisation instructions prior to providing the tools. After such countermeasures become ineffective due to the overwhelming amount of information, filtering, summarising, organising and other tools have to be applied. Certainly, there is no need for a separation of the approaches and normally they are used together, however the presence of the first one is a must.

In this study we focus on the technological approach considering each and every individual's interests, preferences and expertise in order to provide selective information retrieval and access, thus expediting the acquisition of desired and relevant information. Section 1.3 will clarify the research questions and objectives, and give a further outline of the approach.

1.2 Growing Role of Human in Information Creation, Assessment and Sharing

In addition to the fact that information overload is a subjective phenomenon and it is a human who is affected by it and has to copy with it, it is easy to see that the phenomenon itself is largely caused by a human and his activities. It started to be particularly tangible with popularisation of *user-generated content* (user-generated media, or user-created content) which, in turn, was enabled by new technologies, such as weblogging (or blogging), wikis, podcasting, photo and video sharing on the Web. Wikipedia defines user-generated content as

"*… various kinds of media content, publicly available, that are produced by end-users*" (Wikipedia).

The motivations for people to share their time and knowledge are, as discussed by Nov (2007) for the case of Wikipedia, 1) altruistic contribution for others' good, 2) increasing or sustaining one's social relationships with people considered important for oneself, 3) exercising one's skills, knowledge and abilities, 4) expected benefits in terms of one's career, 5) addressing one's own personal problems, 6) contributing to one's own enhancement[3], 7) fun and 8) ideological concerns, such as freedom of information. Although not exhaustive, investigation on blogging by Nardi et al. (2004) shows quite similar reasons for people to blog.

According to Nielsen//NetRatings (2006), in July 2006 "user-generated content sites, platforms for photo sharing, video sharing and blogging, comprised five out of the top 10 fastest growing Web brands". Among them were ImageShack, Flickr, MySpace and Wikipedia – the brands that are also well-known nowadays to any more or less literate Web user. Such user-generated content sites continue growing by attracting new users of various ages and social groups. For instance, MySpace is reported to attract 230,000 new users per day (Sellers, 2006) and more than half of its users are of age 35 and older. The Interactive Advertising Bureau estimates the number of user-generated content consumers is 100.8 million (52% of Internet users) and the number of content creators is 83 million (42.8% of Internet users). And the number of user-generated content consumers will reach 130.1 million (60% of Internet users) and the number of content creators – 108.5 million (50% of Internet users) in year 2012 (IAB, 2008).

With the emergence of user-generated content (UGC) concept, an individual's role as a creator and active evaluator of the shared Web information has become central, and perhaps will become critical in future. With increase of human activities on the Web, the percentage of information related to such activities grows; hence, it is becoming more and more *user-centric*. Such centricity becomes a cause of

[3] These categories are closely related to the concept of *self-extension* we have outlined within social networking services (Shtykh et al, 2008)

creation of excessive, sometimes useless, amounts of information, but, on the other hand, also helps people to overcome information overload problem with the wisdom of crowds (Surowiecki, 2005). People use the power of user-generated content to make decisions on their daily activities, whether it be work or leisure, and researches are investigation how to leverage it in order to benefit from it in a great number of work tasks. JupiterResearch (2008) has found that 42 percent of online travelers using user-generated content trust the choices of other travelers and such UGC is very influential on their accommodation decisions. Exchange of user-generated content facilitates an enrichment of our life by creating new social ties and promoting interaction within communities, as, for instance, discussed in the study of enhancing a local community with IPTV platform to exchange user-generated audio-visual content conducted by Obrist et al. (2008). However, along with the virtues, such user-centricity of UGC brings new problems of trust, and quality and credibility (e.g., Flanagin and Metzger, 2008) of volunteered content that are transformed to adjust the UCG context. As an example, trust becomes a metric for identifying useful content and can be defined as "belief that an information producer will create useful information, plus a willingness to commit some time to reading and processing it" (Golbeck, 2008).

It should be noted that in our research we do not focus particularly on user-generated content, but, as everyone's Web experiences can show, the number of such content is great and its significance cannot be neglected. Although UGC has its specific problems, such as above-mentioned credibility and trust, to be solved, it shows the growing importance of every individual and proves the power of experience of online users taken altogether, which is an important pillar of our research. Generated by human, user-generated content is rapidly growing and influencing many aspects of human life. In other words, it can be named as a mechanism of indirect societal regulation by human, and this regulation is done by not a group of limited number of specialists, but by all interested people willing to participate. So the role of each and every individual in the modern society is growing and becomes more important than ever. Moreover, in the situation of information overload such an engagement is even essential to overcome the problems of excessive information that are, strictly speaking, created by the participants themselves. To reformulate this, nowadays we have to benefit from each other's expertise and this has to be enabled by appropriate technological solutions, which in turn ought to become as human-centric as possible to understand requirements to them in particular work task settings and employ all power of human expertise.

1.3 Research Objectives

Let us consider the research objectives of this work on two levels – macro and micro. Macro level will give us explanation of the objectives from the perspective of the presented concepts of information overload and user-centeredness of information creation, assessment and sharing on the Web. Micro level will help to outline the research questions and objectives we are working on in a closer perspective and domain of information retrieval (IR).

Alleviating Information Overload (macro level)

In this work we tackle the problem of information overload primarily from technical perspective within which a consideration of situational and subjective nature of the problem is done. In other words, although we propose a technological solution for the problem, we attempt to consider it as a problem lying also in a subjective dimension. We believe that no solution can be effective enough without considering a person's processing capabilities and information needs which are very individual, as we discussed above, and situational respectively.

Better Understanding and Satisfying Human Information Needs (micro level)

IR is an important research and application area in the era of digital technology. Today information retrieval tools are essential for information acquisition. However, with information overload becoming more tangible every day, such tools reach their limits of providing information pertinent to users' information needs. This is a reason for revival of interest of scientists and enterprises to information filtering and personalisation today. In order to perform effectively, an IR system has to understand a user's information needs in a particular situation, context, work task and settings, and only after such knowledge about the user is available (through inference or other methods) the search has to be done. The understanding of situational and contextual nature of seeking and endeavors to harness it for more effective seeking process stimulated the research of the cognitive aspects of IR, known today as cognitive information retrieval (CIR) (refer to Ingwersen and Jarvelin, 2005; Spink and Cole, 2005). Inferring the user's interests and determining his/her preferences is one of the useful techniques not only for CIR, but also for personalised IR (PIR). Since the difference between the two may be not clear-cut, we consider PIR as, though often considering the user's search context and situation, not making special focus on cognitive aspects of information seeking.

In our research we propose a collaborative information search and sharing framework called BESS (BEtter Search and Sharing) in attempt to incorporate the discussed user-centeredness into information seeking tasks. We present a holistic approach as to how to harnesses relevance feedback from users in order to estimate their interests, construct user profiles reflecting those interests and apply them for information acquisition in online collaborative information seeking context. The thesis proposes the notions of *subjective* and *objective* index in IR system, and demonstrates the methods for user interest inference, dynamic multi-layered profile construction changing with change of interests, evaluation of shared information with regard to each user's expertise, and *subjective concept-directed vertical search*.

1.4 Organisation of the Thesis

First of all, in Chapter 2 we discuss human-centric solutions for information seeking and exploration with main focus on personalisation, its advances in academy and business, and its problems and speculate on other personalisation-related issues. The survey and discussion on user profiles as the main approach in personalisation and adaptation to services in general follow. Further, we give an extensive overview of user interest inference, its methods, design and use, as an important part of personalisation process.

In Chapter 3 we discuss BESS collaborative information search and sharing framework, present its conceptual basis and present its model and architecture. Chapter 4 narrates about our original interest-change-driven modelling of user interests, discusses its role and position within the framework and compare with other profile construction approaches. Generally speaking, our proposed user modelling approach does not presuppose a method of user interest inference, and any online incremental method can be used, but the online incremental clustering method described in Chapter 5 can be preferred in the environment settings similar to ours. The method is straightforward for implementation and shows reasonable classification results and good relevance feedback processing speed. Chapter 6 discusses shared information assessment and search in the framework. A demonstration of a search scenario is given to better reveal the concepts and information seeking strengths of BESS.

Finally, Chapter 7 concludes the thesis with the summary of the presented research and outlines future research issues.

2 Enhancing Information Seeking and Exploration: User-Centric Approach

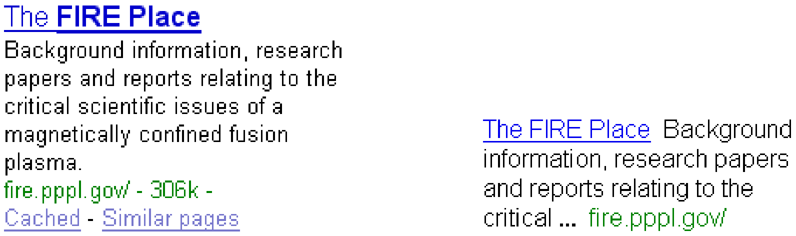
Information overload problems have made a human reconsider information retrieval process and IR tools that seemed to be effective to a certain point. It has become clear that the success of retrieval does not only consist in improving search algorithms, IR models and computational power of IR frameworks – new approaches to make information seeking closer to the end-user are needed. Such approaches include research in user interfaces better adapted to the user's operational environments, systems understanding the user's needs and whose intelligence spreads beyond an algorithmic query-document match seen in conventional "Laboratory Model" of IR discussed in (Ingwersen and Jarvelin, 2005). This resulted, for instance, in the emergence of interactive TREC track and raise of great interest in user-centred and cognitive IR research. IR systems are seeking to incorporate the human factor in order to improve the quality of their results. Information seeking today is getting considered in dynamic context and situation rather than static settings, and a human is its essential and central part actively processing (receiving and interpreting) and even contributing information. Contextual information of the user is obtained from his/her behaviours collected by the system the user interacts with, organised and stored in user profiles or other user modelling structures, and applied to provide personalised information seeking experience.

In this chapter we introduce endeavours to improving Web IR by means of user interface improvements and support of exploration activities, and focus on personalisation as the most wide-spread approach to user-centric IR. We discuss user profile (UP) as the core element of most personalisation techniques, show its structural variety and construction methods.

2.1 Improving Web Information Retrieval

It is well known that alongside with search engine performance improvements and functionality enhancements one of the determinant factors of user acceptance of any search service is the interface. To build a true user-centric information seeking system, this factor must not be underestimated. Here we will show its importance considering mobile Web search, as the need for improvements are particularly tangible due to small screen limitations of handheld devices most of us possess today.

Landay and Kaufmann (1993) in 1993 noted that “researchers continue to focus on transferring their workstation environments to these machines (portable computers) rather than studying what tasks more typical users wish to perform”. In spite of all the advances of mobile devices, probably the same can be said about mobile Web search judging from its state today. Search today is poorly adapted to mobile context – often, it is a simplistic modification of search results from PC-oriented search services. For instance, many commercial mobile Web services, like those of Yahoo!, provide search results that consist of titles, summaries and URLs only. However, although all redundant information like advertisements is removed to facilitate search on handheld devices, users may still experience enormous scrolling due to long summaries. To improve the experience some services, like Google, reduce the size of summary snippets. However, this can hardly lead to the improvements and, quite the contrary, can thwart the search. As shown in Figure 2.1, a mobile user searching for “fireplace” cannot know that the result page is about plasma and does not match his/her needs, and has to load the page to find it out. According to Sweeney and Crestani (2006)’s investigation on the effects of screen size upon presentation of retrieval results, it is best to show the summary of the same length, regardless whether it is displayed on laptops, PDAs or smartphones.



The **FIRE Place**
Background information, research papers and reports relating to the critical scientific issues of a magnetically confined fusion plasma.
fire.pppl.gov/ - 306k -
Cached - Similar pages

The FIRE Place Background information, research papers and reports relating to the critical **...** fire.pppl.gov/

Figure 2.1 The same search result item for PC-oriented Web search (left) and mobile Web search (right)

Improvements to mobile Web search done in academia go further. For example, De Luca and Nürnberger (2005) implement search result categorisation to improve the retrieval performance and present the information in three separate screens: screen for search and presentation of the results in a tree, screen to show search results and bookmarks’ screen. Church et al. (2005) substitute summary snippets, which are coming with each result item, with the related queries of like-minded individuals – queries leading to the selection of a particular Web page in the search result list. The researchers argue that such queries can be as informative as summary snippets and using this approach they provide more search results per one screen.

In contrast to the existing approaches, Shtykh et al. (2008) (see also Shtykh and Jin (2008a)) do not make any modifications to the search results, but propose an interface to handle the results provided by any conventional search service. The approach abolishes fatigue-inducing scrolling while preserving "quality" summaries of PC-oriented Web search. The proposed interface, called slide-film interface (SFI), is a kindred of "paging" technique. Unlike most mobile Web search services that truncate summary snippets of the search result items to reduce the amount of scroll and in this way facilitate easier navigation through search results that often can lead to difficulties in understanding of the content of a particular result, (owing to the availability of one slide of a screen size for one search result) our approach has an advantage to provide the greater part of one slide screen to place the full summary without any fear to make the search tiresome. SFI was compared with the conventional method of mobile Web search and the experimental results showed that, though there was no statistically significant difference in search speed when the two interfaces are used, SFI was highly evaluated for its viewability of search results and ease to remember the interface from the first interaction.

Although such approaches to improve the search with focus on the user, his/her usability are very important and user-oriented, they treat the user regardless of his/her contextual and situational information. As we already mentioned and will discuss more in Chapter 3, information need and human behaviour are very contextual. Therefore peculiarities of information behaviour, proclivities, preferences and everything that can give a better conception of the user, his/her behavioural patterns and needs must be considered in order to be able to provide a truly personalised information seeking experience. Although in the thesis we focus on information seeking specifically, the application area of personalisation spreads far beyond it. It is applied to Web recommendations and information filtering, user adaptation of Smart Home and wireless devices, etc.

Through our research we were particularly interested in personalising and facilitating a human's interactions with various Web services. And search is not the only activity in Web information space users are engaged in. As empirical studies show (McKenzie and Cockburn, 2001), most of time users rediscover things they used to find in the past, and often they browse without any specific purpose discovering information space around them or with a particular purpose, such as learning miscellaneous information. To support such a discovery, we designed an exploratory information space (Shtykh and Jin, 2006) that makes use of human-centred power of bookmarking for information selection. The information space is built as a result of a search for something a user intends to discover, and serves as a place for rediscoveries of personal findings, socialisation and exploration inside discovery chains of other participants of the system.

2.2 Personalisation

Today personalisation is the term we often relate to Web search personalisation, such as in Google's iGoogle, recommendation system of Amazon.com, or contextual advertisements on Web sites. It is also about Decentralised-Me (O'Brien, 2007) of emerging Web 3.0 or is an essential part of Mitra (2007)'s formula of Web 3.0 – *Web 3.0 = (4C + P + VS)*, where *4C* is Content, Commerce, Community, and Context, *P* is personalisation, and *VS* is vertical search. However, the notion of personalisation is much more diverse than that. It differs with regard to its application area and is being transformed over time and advances in its research. It is sometimes synonymous to *customisation* and often to *adaptation*. It concurs with *information filtering* and *recommendation*.

In 1999 Hansen et al. (1999) outlined two knowledge management strategies for business – *codification*, i.e., impersonalised storing knowledge in databases and its reuse, and *personalisation*, which focuses on dialogue helping people to communicate knowledge. The authors claim that emphasising the wrong strategy or pursing the both at the same time can undermine a business. However, today, in the situation of information overload, the both strategies often complement each other. Greer and Murtaza (2003) define personalisation as "a technique used to generate individualized content for each customer" and investigate the factors that influence the acceptance of personalisation on an organisation's Web sites. The research finds that ease of use, compatibility with an individual's value and his/her intents and expectations, and trialability ("the degree to which personalization can be used on a trial basis") are the key factors for personalisation adoption. Monk and Blom (2007) in their earlier works define personalisation as "a process that changes the functionality, interface, information content, or distinctiveness of a system to increase its personal relevance to an individual", and Fan and Poole (2006) extends this definition to "a process that changes the functionality, interface, information access and content, or distinctiveness of a system to increase its personal relevance to an individual or a category of individuals" which serves as the working definition for the thesis.

Such a great diversity in understanding of what personalisation is results in difficulties to produce a holistic view on personalisation, hurdles for sharing findings for researches of different fields and difficulties to compare approaches. And this is one of the conceivable reasons why the current approaches focus on "how to do personalisation" rather than "how can personalisation can be done well", as Fan and Poole (2006) has noted. In attempt to bring different approaches together, the researchers present a high-level multi-dimensional framework of personalisation perspectives and personalisation implementation classification scheme in their multi-paradigm review. Personalisation perspectives are distilled from found literature on personalisation – namely,

- architectural perspective associated with environmental psychology, urban planning and architecture and concerned about creation of an environment and personal experiences according to a user's needs;
- instrumental perspective that refers to utilisation of IS for enhancement in "efficiency and personal productivity by providing, enabling, and delivering useful, usable, user-friendly tools in a way that meet the user's situated needs" (Fan and Poole, 2006, p. 192);
- relational perspective that refers to "mediation of personal relationships and utilization of relational resources to facilitate social interactions by providing a convenient platform for people to interact with others in a way that is compatible with the individual's desired level of communality and privacy" (Fan and Poole, 2006, p. 193);
- and commercial perspective that reflects one of the most important human activities – consumption.

Implementation classification scheme is another high-level abstraction of the framework that consists of three dimensions:

- personalised object/media (content, user interface, information channel/access, and functionality);
- personalisation target (an individual or group);

- personalisation doer (a system or human).

Such framework gives a better understanding of what personalisation is, how it might be defined and what can be its expected result.

User-Initiated Personalisation

Most personalisation approaches on the Web are *system-initiated*, i.e., considering *adaptivity* which is the ability to adapt to a user automatically based on some knowledge or assumptions about the user. But another concept – of *adaptability*, which is a *user-initiated* (or *explicit* by Fan and Pool (2006)) approach to modify the system's parameters in order to adapt its functionalities to his/her particular contexts, – is also important when considering personalisation. Monk and Blom (2007) emphasised that people always personalise their surroundings, and their Web environment is not an exception, and presented their theory of user-initiated personalisation of appearance. In the user study conducted in 2001 with users having personal Web pages the authors identified dispositions and effects of personalisation, and assessed the correlation between individuals' dispositions to personalise (frequency of use, knowledge of personalisation and social emotional context of use) and the degree of personalisation evident in their pages. The study's results have produced several interesting interpretation of disposition–personalisation and personalisation–effects correlations, such as the following disposition–personalisation–effects causative relationships:

- dispositions leads to personalisation that leads to the effects (ease of use, fun, positive associations, attachment to the system, feeling in control and feeling of ownership, to name a few);
- probable feedback from effects to the user's dispositions;
- unavailability of causative link between disposition and personalisation. In this case, personalisation's positive effects cause a greater disposition;
- unavailability of causative link between personalisation and the effects;
- dispositions, personalisation and effects are caused by some unspecified factors.

The research still leaves many possibilities for further explorations.

Another study on user-initiated personalisation by Oulasvirta and Blom (2008) examines the psychological aspects of why users are willing (or not) to make efforts to personalise. The authors argue that there is no special need for personalisation and "personalisation behaviour is caused by motivational states that are not rooted to behaviour itself". They apply modern theories of motivation to explain a user's personalisation behaviour – specifically, the *Self-Determination Theory*. According to the theory, *self-determination* is the need to have a choice in behaviour, and humans are motivated to perform an action for two reasons:

1) to maintain the optimal level of stimulation, and
2) to achieve competence and personal causation, or self-determination.

The study shows that personalisation promotes autonomy (willingness to engage in an activity), support the user's competence (ability to influence and benefit from the environment) and relatedness (becoming an emotionally close part of something, for instance, community) which are the basic categories of self-determination, and their analysis is important part of user-initiated personalisation process. As the authors say, "... users are willing to expend effort when the product involves and nurtures their psychological needs of autonomy, competence, and relatedness, taps into and extends their interests and preferences, and makes it possible for a user to transform a company-supplied one-size-fits-all technology so that it becomes a personalised, personally useful, and enjoyable tool that can be used to improve and enjoy life and work".

Personalisation has a lot of advantages over impersonalised approaches, some of which are obvious and some of which are hidden and have to be empirically proven. For instance, Guida and Tardieu (2005) prove that personalisation, similarly to long-term working memory, helps to overcome working memory limitations, expanding storage and processing capabilities of human-beings. Although the discussed personalisation is considered as a creation of the situation of individual expertise that is generally not exactly what modern personalisation systems can provide, such approach indicates the need in better considering context and situation in order to fully employ its merits.

However, along with the merits for organisations and individual users, it brings new concerns (described in Section 2.2.2) we have to face and challenge. Furthermore, in no way should personalisation become a purpose of personalisation or a guise for something different, as shown in Hartley (2007)'s study which reveals personalisation of education in England as inchoate re-branding guise.

2.2.1 Application Areas

As discussed, personalisation is applied to a variety of systems in various areas. In this section we discuss its major application fields confined to information systems (IS), such as Web search, digital libraries, e-learning environments, Web pages customisation and ubiquitous services.

Web Search

The most ebullient area of personalisation on the Web is search. Its niche in personalisation is very huge and variety of personalisation solutions is great. Today anyone can easily find a number of "My-" or "YourSomething" services, like, for instance, myAOL personalised home page, or "Your Fluther" feature of Q&A site paired with recommendation component feeding a user with questions on topics found in the user's profile. Personalisation in context of Web search engines is often "the ability of the Web site [(Web search engine)] to match retrieved information content to a user's profile" (Khopkar et al., 2003), but the notion is changing and has a tendency today to better consider the user's context. There is also a tendency for social media to become an auxiliary part of Web search personalisation, as can be exemplified with Wikio, "a personalizable [customisable] news page featuring a news search engine that searches media sites, blogs and the contributions of Wikio members". Note that in a number of cases personalisation refers to adaptability of the service, notion described earlier in this chapter, which is of less interest to the thesis. So forth we will focus on the notion of adaptivity of services.

Pitkow et al. (2002) outlined two general approaches to Web search personalisation – query augmentation and re-ranking search results based on information about the user. Till today the number of solutions and approaches to Web search personalisation has grown. They include complex user modelling frameworks, agent-based frameworks and personalisation by social bookmarking and tagging. As an example, Noll and Meinel (2008) proposed a document re-ranking personalisation solution based on social bookmarking and tagging. In the approach bookmarking is used to learn about a user and pages he/she is searching for and personal bookmarks are the indicators of interests stored in user profile which is a vector of *m* tags with values denoting tag counts in the user's bookmark collection. In addition, document profile is created as another vector of tags. When a search is done, the system estimates user-document similarity and re-ranks the list of search results returned by a Web search engine accordingly. Teevan et al. (2005) leverage implicit information about a user found from visited Web pages and other documents stored locally on the user's desktop in order to personalise Web search by re-ranking. Ligon et al. (2006) propose a multi-layered agent architecture for Web search in which agents learn from users' explicit and implicit feedback.

These are only several salient examples of personalisation for Web search taking different approaches we give here for illustrative purposes. More examples of such systems are encountered through the chapter.

Digital Libraries

Digital libraries is another area personalisation is extensively used. Personalisation and recommendation offer a great potential to solve information-overload-related problems in digital libraries, contributing to their proactiveness. It is closely related to Web search, since on higher abstraction level search on the Web and in digital libraries often have the same aims and issues to solve, and use common techniques for this, but can be distinguished by the emphasis done on information sharing within a community and high recognition of heterogeneity of library resources (Smeaton and Callan, 2005).

E-learning

Students at school have different level of knowledge and skills, different characters and proclivities that has to be carefully considered in the teaching and learning process. Personalisation, one the intrinsic tasks of which is to develop the understanding of and to structure each individual's characteristics for further adaption in individual-system interactions, is likely to be a good enhancement reducing workload of instructors and facilitating learning efforts of students.

Santally and Alain (2005) claim that one of the main problems of e-learning systems is the lack of personalisation and propose their e-learning adaptation framework. The framework implements an algorithm for personalised instruction which decide on what is the most appropriate learning object by selecting one with the highest confidence factor with regard to the student profile. Stiubiener et al. (2007) propose a personalising e-learning system that provides didactic materials and learning activities matching pedagogical goals with the learner's skills and preferences. The system is implemented in a LMS (Learning Management System) defining Personalisation Learning Policy (PLP), which is a configurable set of rules for didactic content and learning activity adaptation lead by the Orientation Layer containing such adaptation criteria. Orientation Layer's constituents are level of knowledge, psychological model, learning style, media format, etc. Another essential element of personalisation in the system is Learner Profile (LPROF). LPROF is a set of observed student's behavioural characteristics like "media preferences", "learning style" and "participation in chats" which are dynamically updated. Together with PLP LPROF is used to estimate the student's preferences and provide relevant didactic material and learning activities. Although such personalisation is deemed to organically assist e-learning process, the authors do not provide results and discussions on the effect of the proposed personalised approach on learning.

Web Usage Mining

Web usage mining (WUM) is a family of personalisation techniques applied to users' behaviours on the Web in order to find their usage patterns. By analysing server access and proxy logs, Web browser caches, etc., WUM systems extract users' navigational behaviours, model user behaviours and apply the models for personalisation of Web pages.

Baraglia and Silvestri (2007) propose the WUM system called SUGGEST for dynamic personalisation of Web sites without user intervention. Personalisation in SUGGEST consists of three phases first two of which are done offline:

- Preprocessing for knowledge base construction from access logs;
- Pattern Discovery for evaluating useful data patterns in the knowledge base using data mining techniques;
- Pattern Analysis for finding interesting patterns after which recommendations to the user are done.

Personalisation of Mobile and Other Ubiquitous Services

The research on volume control personalisation by Ypma et al. (2006) is a good example how ubiquitous and important to all aspects of human life personalisation is. The work presents a learning method "to absorb user adjustments to the volume control of a hearing aid in the parameters of the volume control algorithm". Other endeavours to service ubiquitous personalisation are ontology-based framework for mobile services by Jorstad and Do van Thanh (2007) and Personalised Context Ontology for location modelling by Niu and Kay (2008).

2.2.2 Personalisation-related Issues

There are a number of Web personalisation-related issues like privacy, trust, robustness and scalability. Anand and Mobasher (2005) outline ten issues and discuss available solutions. Here we list up the issues and add several other novel solutions not discussed in the original paper.

- The "cold start" and latency problem.

  Cold start issue arises when a new user with zero UP joins the personalised system and at that point the system cannot provide useful personalised information to the user. But it can be successfully coped with, for instance, hierarchical Bayesian network proposed by Zigoris and Zhang (2006) which incorporates the prior learnt from other users. Latency issue occurs when there are not enough ratings for a new item introduced to the system;
- data sparseness, which arises because of a huge amount of items sparsely rated by users;
- scalability, intrinsic to memory-based approaches which suffer when a number of users in the system grows;
- privacy (Kobsa, 2007);
- recommendation list diversity, which is an important factor to increase users' satisfaction in recommendations;
- adapting to a user's context, which is usually inferred from user-system interactions (e.g., (Limbu et al., 2006)).

  It is an important direction for personalisation for near future;

- using domain knowledge, often derived from ontologies or various semantic structures with information about the user and items of the system;
- managing the dynamics in user interests, which is one of the main focuses of this thesis (Chapter 4);
- robustness, with regard to false ratings and other abuse;
- trust, as a method to improve personalisation in collaborative settings and alleviate abuse threats.

In the further discussion we will focus on privacy issue as the most acute for a user today.

Privacy

Any Web personalisation system requires storing and analyzing private information. This is seen to be a problem by privacy advocates (e.g., Privacy International) since there is never enough security to protect personal information. And such worries are not groundless at all. Relevance feedback is the most widely used technique to collect information about users and largely contribute to personalisation. But despite its virtues, there is the other side of the coin with all chances for a user whose information is being monitored and collected to be jeopardised. Claypool et al. (2001) points to the great potential of privacy abuse brought by implicit relevance feedback. As the authors note, protecting the privacy of user profiles is always important, but empowered by implicit interest indicators (feedback) profiles become more accurate and therefore valuable not only to personalisation services but also to potential privacy abusers. Therefore privacy has to be one of the main concerns in personalisation research.

The problem of collecting and reusing personal data is becoming a big concern not only for research in academia. It is faced by the modern technological society, and therefore legislators in many countries are working on the laws to secure personal data by setting rules for collecting and processing personal information by businesses (e.g., Canada’s “The Personal Information Protection and Electronic Documents Act” (Office of the Privacy Commissioner of Canada)), which is an important normative approach reassuring consumers and thus facilitating promotion of businesses on the Web.

Individual privacy concerns and stringent legislation may have a negative impact on personalisation, as the less information about a user is available the higher the risks for personalisation mechanisms to produce less reliable results. In order to reconcile the tensions between personalisation and privacy, Kobsa (2007) proposes an interdisciplinary approach of *privacy-enhanced personalisation* that strives “for best possible personalization within the boundaries set by privacy”. The researcher also shows that personalisation is generally valued by users, but personalisation systems have to do some efforts for the users to increase their use. Awareness of and control over the use of personal information, trust and normatively-supported policies are the key factors for a user to show more willingness to disclose personal information and, as a result, get better personalised experiences. Chellappa and Sin (2005) point to the trade-off between personalisation and privacy users have to face and develop a model to predict consumers’ use of online personalisation as a result of such concern. And again the findings say that personalisation success greatly influenced by trust of consumers in the service, and service providers’ understanding of consumers’ values and efforts to build trust are crucial to the success of personalised services.

Despite of the above-mentioned concerns, personalisation is still of great value to users disposing them to a number of benefits worth of afore-mentioned efforts. Interestingly, along with conceivable threats to jeopardise a user, Web personalisation can be used to protect him. Aknine et al. (2005) propose a multi-agent method to protect users from racist discourse. The approach is realised with the help of a Web crawler searching for offensive documents and providing the list of their sites to self-protective programs. Because the offensive nature of such discourse is often hidden by racist authors, a multi-agent system is proposed to combine different textual analysis techniques.

2.3 Modelling User Interests

In order to be user-centric, a service has to know each user it interacts with. This is the task personalisation attempts to fulfil with a variety of methods in various work task and environmental settings. Personalisation systems extract the user’s interests, infer his/her preferences, update and rely on knowledge about the user accumulated and structured in user profiles that differ by the data used for their definition, their structure and complexity, and construction approaches.

At this point we have to note that in modelling user interests we do not make a distinction between Web search personalisation, recommendation or information filtering because the differences in their methods and goals are very subtle. All such approaches utilise a certain scheme to know the user's preferences to adapt to his/her future interactions with the system and information it provides, and constructing user profiles (or user modelling) is the most popular method. It has been extensively used from days of first information filtering systems, for instance as a user-specified profile or a bag-of-words extracted from the documents accessed by the user, and today it takes many richer and diverse forms to meet the requirements of the variety of information systems.

2.3.1 Relevance Feedback as a Modelling Material

As the reader can see from the above discussions, use of relevance feedback for personalisation is very important and widely utilised.

Feedback Types

Relevance feedback is extensively used in Web IR for efficient collection of user behavioural data for further user behaviour analysis and modelling. Relevance feedback can be *explicit* (provided explicitly by the user) or *implicit* (observed during user-system interaction). The first form of relevance feedback is high-cost in terms of user efforts and the latter one is low-cost but requires a thorough analysis to reduce the noise it normally contains. Implicit relevance feedback in IR systems consists of a number of elements, such as a query history, a clickthrough history, time spent on a certain page or a domain, and others, that can be considered in general as a collection of implicit behaviours of users interacting with the information retrieval system. It is conducted without interruption of user activities, unlike explicit one that requires direct user interferences, that is why many are showing keen interest in it. Interested readers are referred to (Ruthven and Lalmas, 2003) for survey on the use of classic relevance feedback methods and (Kelly and Teevan, 2003) for extensive bibliography of papers on implicit feedback, or any modern information retrieval (IR) textbook for the detailed introduction of relevance feedback.

Let us give several examples of how relevance feedback is used and what kind of data is considered feedback. Russian Internet advertisement company Begun states that behavioural advertisement based on watching user actions on certain Web sites and search portals brings about the same amount of profit as semantic advertisement. Begun creates user profiles that are later updated according to user actions on specified Web sites. The main source of information for profile formation is user search queries leading to the advertised sites, user routes from one site or page to another, and history of interactions with certain advertisements. Moreover, users are organised into groups and a part of prediction algorithm uses profile similarities to fill the gaps in other users' profiles and do correct predictions. A similar research is being conducted by Microsoft AdCenter Labs. Taking the probabilistic approach, they are building user profiles based on page views, searches, and other online behaviours for targeted advertisement. Further, such profiles are clustered and segments of customers with similar interests are created. In addition to implicit feedback found the Web, Teevan et al. (2005) use documents found on the user's desktop, such as Web pages, e-mails, calendar items and documents.

With emergence of social network, new types of feedback become available. Thus, social bookmarking and tagging, as described in (Noll and Meinel, 2008), are sui generis mixture of both implicit and explicit relevance feedback. On one hand, bookmarking is an explicit action done by a user and not monitored for by the system, on the other hand, in contrast to explicit feedbacks, it is normally not a burden for the user. We would classify such a feedback as *motivated explicit feedback*, since it is motivation that removes burdens from the explicit nature of the feedback.

Another emerging type of relevance feedback that is worth mentioning is *contextual relevance feedback* which shows again an increasing attention to context for personalisation. As a matter of fact, it is often of no difference from many other approaches based on user profiles. Thus, in (Harper and Kelly, 2006)'s approach contextual relevance feedback is a feedback to a search result list to filter it based on user-collected document piles. Another example is contextual relevance feedback architecture by Limbu et al. (2006) which, in addition to profiles, utilises ontologies and lexical databases.

Types of Data for Relevance Feedback

As to the types of data used for profile construction, their choice depends on the application domain of the system to be personalised. For IR systems, relevance feedback is normally documents, queries, network session duration and everything related to information search process on the Web and beyond. For instance, Teevan et al. (2005) extend the conventional relevance feedback model to include the information “outside of the Web corpus” – implicit feedback data is derived from not only search histories but also from documents, emails and other information resources found in the user's PC. With the change of the application domain the type of data differs. For instance, mobile device features and location can be considered for profile construction in nomadic systems (Carrillo-Ramos et al., 2007), and user interests can be learnt from TV watching habits, as in (Wang et al., 2008). Naturally, any user behaviour can be considered as a source for inference of his/her interests and further user profiling, and there are as many selection decisions in regard to use of a particular feedback type as there are systems that utilise them. Fu (2007) proposes to examine a variety of behavioural evidences in Web searches to find those that can be captured in a natural search settings and reliably indicate users' interests.

2.3.2 Modelling Methods

With the afore-mentioned data, user interests can be inferred and user profiles (models) can be created in a number of ways and various methods. It is hard to clearly classify all of them, since some of them are very domain-data-dependent and thus their methods are very specific. To give an idea about their diversity we present several different modelling methods[1] distinguished on model representation approaches.

Vector-Space Modelling

In vector-space modelling approach a user’s interests are represented as vectors of *n* features $X = \{x_1, ..., x_n\}$. In (Çetintemel et al., 2000) use profile is a set of profile vectors

$$p_i = ((t_{i1}, w_{i1}), ..., (t_{im}, w_{im})), i = 1, ..., n \qquad (2.1)$$

where t is a term and w is its weight. Relevance feedback (documents) is used as a material for vector

[1] The classification here is not a complete list of existing approaches (for instance, we do not give an example of approaches based neural networks or rule sets), rather its purpose is to indicate some of the major methods

construction. All documents are clustered with an incremental clustering method, and each cluster holds its representative vector, namely p_i.

Many methods of this group (e.g., Semeraro et al., 2005a) extend the Rocchio algorithm, one of the most popular classification methods. Generally, as in the previous example, user profile consists of a set of vectors, each related to a specific category of a user's interests

$$\vec{x}_i = \{w_{i1}, \ldots, w_{im}\} . \tag{2.2}$$

Each representative vector is computed as

$$w_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{w_{kj}}{|\, POS_i \,|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{w_{kj}}{|\, NEG_i \,|} . \tag{2.3}$$

Probabilistic Modelling

Probabilistic modelling is very diverse including methods relying on language modelling, Bayesian networks, and so forth.

In (Zigoris and Zhang, 2006) user model of user *u* is built using explicit and implicit relevance feedback. It is a function that takes information about a document x^u and returns its ratings y^u as to the user's interest in the document: $f^u : x^u \rightarrow y^u$. When the user starts with the system for the first time, a prior belief about the user model is borrowed from existing users via hierarchical Bayesian framework and written as

$$f^u \sim P(f \mid \theta) . \tag{2.4}$$

As the user uses the system and rates items, the system updates its belief about the user model based on feedback data and get posterior distribution over user models

$$P(f^u \mid \theta, D_u) = \frac{P(D_u \mid f^u, \theta) P(f^u \mid \theta)}{P(D_u \mid \theta)} \tag{2.5}$$

where D_u is a set of document-rating pairs.

As the user uses the system more, the prior learnt from other users become less important.

User profile presented in (Barbu and Simina, 2003) is learned with PLSA (Probabilistic Latent Semantic Analysis) and positioned in the "Time – Probabilistic Latent Semantic Space" which besides weights dimensions *P(z|d)* and *P(z|Q)* (*z* – aspect, or hidden value, *d* – document, *Q* – query) introduces a temporal dimension to reflect raising and decaying interests as time goes.

Graph-based Modelling

In Baraglia and Silvestri (2007)'s recommendation approach Web usage information is represented as an undirected graph whose nodes are associated with the accessed pages, and each edge is associated to a measure of the correlation between nodes. The graph is partitioned into "session clusters" that lead to the resultant list of suggestions. The approach does not consider a page's content to measure the user's interest in the page, it considers the occurrences of the page in a session. The weight of each edge is calculated as

$$W_{ij} = N_{ij} / \max(N_i, N_j) \tag{2.6}$$

where N_{ij} is the number of sessions containing both pages *i* and *j*, N_i and N_j are the number of sessions containing only *i* or *j* respectively. In this way the authors obviate pages that can be considered as non-informative, such as those the user usually starts Web browsing and search from.

Domain-specific Modelling

Often user interest modelling is done specifically for the system it is applied to with regard to its application domain and based on the specific data that can be obtained from user-system interactions of this particular system. Consequently, modelling methods for user interests will be constrained to that that type of systems, in contrast to other generic modelling approaches.

For instance, the personalised peer-to-peer television system by Wang et al. (2008) is interested in user interests inferred from TV watching habits. For user u_k the interest in program i_m is calculated as

$$x_k^m = \frac{WatchedLength(m,k)}{OnAirLength(m) \cdot freq(m)} \tag{2.7}$$

where *WatchedLength(m,k)* is the duration of program i_m in seconds watched by user u_k, *OnAirLength(m)* is the full duration of program i_m, and *freq(m)* denotes the number of times its has been broadcast. Models in e-learning, in addition to interests, often consider learning styles and performance, cognitive aspects of a learner, etc. They are complex and require explicit directives and assessments of an instructor. For instance, student profile in (Santally and Alain, 2005) consists of four components: 1) cognitive style, 2) cognitive controls, 3) learning style and 4) performance. It is created by a student registering to the course and complemented by the instructor's and psychological experts' surveys on the user's cognitive and learning styles. It is updated with the student's feedback, monitored performance and the instructor's decisions based on the user's learning history.

2.3.3 Structural Components

There is a great variety of profile structure types. The simplest and most widespread one is to represent user interests learnt from relevance feedback with document term vectors for each interest's category (e.g., Potamias et al., 2005; Semeraro et al., 2005a; U and Varma, 2007; Karpouzis et al., 2004). Shapira et al. (1997) enhance such vectors with sociological data (profession, position, status). Profiles in Sobecki (2004) are *attribute-value* tuples, where the attributes characterise usage such as visited pages or past purchases, or demographic data such as name, sex, occupation, etc. In Ligon et al. (2006)'s agent-based approach user profiles are a combination of information categories and a preference database containing search histories related to the categories.

User profiles become more elaborate and complex trying to reflect the dynamics of constantly changing user context and interests. For instance, Bahrami et al. (2007) distinguish static and dynamic user interests for profile construction in their information retrieval framework. Barbu and Simina (2003) distinguish *Recent* and *Long-Term* continuously learnt user profiles and apply them to information filtering tasks. Further, information systems utilised by mobile devices often extend the notion of user profile in conventional IR systems bringing specific contextual information into it. For instance, in (Sendin et al., 2003) *spatial awareness* is proposed as an important property of mobile personalisation services. It is an important part of environment recognition which is done by automatic extraction of the objects of interest according to a user's profile, state and current situation with regard to his/her position. Carrillo-Ramos et al. (2007), in attempt to adapt information to a nomadic user by taking *context of use* into consideration, introduce Contextual User Profile which consists of user preferences and current context (location, mobile device features, access rights, user activities) of use. Ferscha et al. (2006) propose context-aware profile description language (PPDL) expressing mobile peers' preferences with respect to a particular situation. Finally, some attempts to provide more holistic approaches to profile structuring, such as Gargi (2005)'s Information Navigation Profile (INP) defining attributes for characterising IR interfaces, interaction and presentation modes, are made resulting in complex profiles that consist of multiple search criteria.

2.3.4 On User Contexts

As we already noted, personalisation with better focus on user contexts and situations is the topic to be better investigated in the near future. As personalisation depends much of the intents of and results expected by a user, it is essential to accurately assess his/her contextual characteristics.

In spite the fact that a number of personalisation approaches today use the notion of context, such 'context' is usually derived from queries and retrieved documents and inferred from user actions. They are not likely to accurately capture the situation and the context which includes far more factors than taken in such approaches[2]. Furthermore, the definition differs from one solution to another. And, naturally, the diversity grows in mobile and ubiquitous personalisation approaches because of context peculiarities. While context of a user is learnt, for instance, from documents and ontologies (Sieg et al., 2004), multiple context attributes like environmental and other properties (time, location, temperature, space, speed, etc.) are considered in (Ferscha et al., 2006) to define context-aware profiles. And probably because of such differences related to application domain, there is very little exchange of verified practices among researchers working on personalisation in different areas and, despite available similarities in various domains, the one-sided views on context are not rare. There are endeavours to utilise context and situation in a holistic fashion (e.g., Ingwersen and Jarvelin, 2005), however they are on the level of theory. We believe that accurately and timely estimated contextual information will greatly contribute the field of personalisation, therefore further endeavours to characterise, methods to capture and systematise knowledge about it should be continued and deepened and corroborated with empirical studies.

[2] Recognising the fact that our approach does not fully capture user contexts, the notion of "fragmentary context" is used

3 User-Centric Information Search and Sharing with BESS

3.1 Being User-Centric by Knowing User's Preferences through Contexts

One of the main driving forces of human information behaviour is information need that is recognition of one's knowledge inadequacy to satisfy a particular goal (Case, 2002), or "consciously identified gap" in one's knowledge (Ingwersen and Jarvelin, 2005). Therefore its understanding is crucial for systems that are supposed to facilitate information acquisition. However, in many cases capturing and correctly applying individual information needs is extremely difficult, even impossible. For instance, in IR systems a user's input cannot usually be considered as a correct expression of his/her information needs – that results in invalidity of many traditional relevance measures (Kagolovsky and Moehr, 2003). And this happens not only in IR, but in any system when context, in which an information need was developed, is lost.

Then, the following question arises. In "Introduction" we defined user-centric system as a system that "understands" (is able to capture) the user's information need in order to satisfy it effectively. But

how can the system be user-centric and satisfy sufficiently the user's information need without being able to capture it?

Information need emerges in one's individual context, and both context and information need are evolving over time. Information behaviours happening to satisfy the information need and leading to an information object selection also take place in the same particular context (Figure 3.1). Therefore, although knowing particular contexts does not give us the full understanding of a particular user's information needs, such knowledge can give us some conception (or a hint) of conceivable information a user tries to obtain in a particular context, i.e., lead us to the potentially correct object selection. As shown in Figure 3.1, particular information need in a particular context leads to information behaviours which, in their turn, result in object selections from two groups of similar objects. Knowing information behaviour patterns (and their contexts) resulting in particular object selections, in our research we try to induce a user's current preferences for a particular object without clear knowledge of current information need. Such knowledge gives a chance for a service to identify user contexts during user-service interaction and help with correct information object selection. Further, by matching context information of one particular user with contexts of other users that utilise the same service, we can try to foresee a situation new to the user (an unknown context) and facilitate his/her information behaviour.

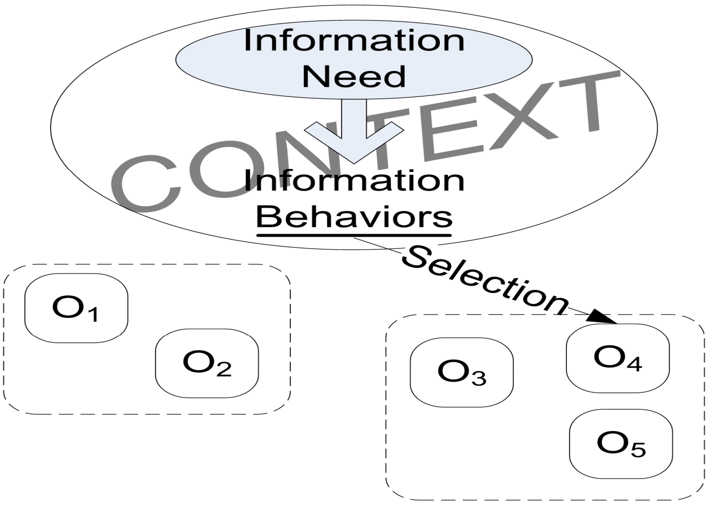


Figure 3.1 Information object selection in context

Essentially, context can be considered as a formation of many constituents – an individual's geographical location, educational background, emotions, work tasks and situations, etc. With the advances of spatial data technologies, ubiquitous technologies and kansei engineering we are likely be able to collect a large part of them in the near future, but this task is still very challenging. Even more challenging is the task to effectively utilise all these constituents in various user-centric services. Moreover, the need in some particular constituent of the whole context depends on the task the system is trying to facilitate.

In information seeking tasks we are studying, as in most tasks that support information activities today, it is impossible to collect all contextual information, so the contexts considered here have a *fragmentary nature* – basically consisting of information behaviours obtained from users' explicit and implicit relevance feedback (Shtykh and Jin, 2008b). Generally, it is a feedback of textual, temporal or behavioural information with regard to the resources a user interacts with.

3.2 User-Centrism in BESS: Main Concepts of the Proposed Approach

In the proposed approach we attempt to utilise acquired user contexts as much as possible to make the services of BESS user-centric and consequently help users with effective acquisition of information pertinent to their particular contextual and situational information needs. The main concepts for achieving such user-centeredness after having appropriate contextual information are

1) concept;
2) multi-layer user profile;
3) interest-change-driven profile construction mechanism;
4) subjective index creation and its collaborative assessment;
5) subjective concept-directed vertical search.

Determining and Organising Personal Interests

Information seeking, as any information behaviour, is done in a context determined by situation, interest, person's task, its phase and other factors. In the process, some user interests tend to change often influenced with temporal work tasks and personal interests, and some tend to persist. Capturing them gives us a fragmentary understanding about current user contexts and can be used to induce a general understanding about a user. In our research such interests are inferred from relevance feedback information provided by the user and are a set of conceivably semantically-adjacent terms. Therefore they are called *concepts*.

However, such concepts are not much of interest when they are not grouped by some criterion that helps an IR system to understand their tendency to emerge and change. In order to organise user interests and have the whole contextual picture, we chose user profile construction based on temporal criterion. As a result, user profiles in BESS are *multi-layered* – each of layers reflecting user interests temporally, corresponding to long-lasting, short-term and volatile interests. Furthermore, they are generated with *interest-change-driven profile construction mechanism* which relies entirely on dynamics of interest change in process of profile construction and determination of current user interests (see Chapter 4).

Obviously, in order to infer interests we have to handle user's relevance feedback separately from all information resources available at the system. Therefore, each user has its own *subjective index* data which is collected from his/her relevance feedback. It distinguishes from index data of conventional search engines we call *objective index* by its *social nature* – it is created based on the information found valuable in the context of a specific information need and submitted by users, in contrast to objective index which is collected by crawlers or specialists without any particular consideration of context, situation or information need. Collecting such personal information pieces gives us access only to highly selective information tied to a specific context – without such a relation preserved, this information is not much different from that stored in conventional search systems.

From I-Centric to We-Centric Information Search and Sharing

Determining and organising a user's personal interests is very helpful to further facilitate user-system interactions in general, and information seeking tasks in particular. However, would such facilitation be fully user-centric without collaboration of all members of the system? Probably, it would be. But, as we discussed in Introduction, such an approach would not benefit from "wisdom of crowds" (Surowiecki, 2005) of other users and loose much predictive power it could draw upon other users' experiences. In addition, personalisation oriented on one individual will lead to different experiences among community of users that can increase problems of transparency and interpretation (Smeaton and Callan, 2005), and sharing information with others creates new possibilities for discovery and reinterpretations. Recognising this and following our endeavours on collaboration (Shtykh et al., 2005), BESS is designed as a highly collaborative information search and sharing system. It harnessed collective knowledge of its users who share their personal experiences and benefit from experiences of others. In other words, this is *We-Centric* part of the system, in contrast to *I-Centric* one harnessing solely personal experiences.

To emphasise the collaborative nature of relevance feedback submitted by users explicitly, it is called a *contribution* in our research. Although explicit feedback can disrupt search user activities, it is important for subjective index creation, and explicit measures in information retrieval tasks are found to be more accurate than implicit ones (Nichols, 1997). Together with implicit feedback it forms subjective index of each user which in turn is used for concept creation. As we already mentioned, concepts correspond to user interests, and, placed into user profiles, they are used to assess each user's expertise with regard to a concept of the relevance feedback the user contributes. These assessments are an important mechanism to estimate the value of a particular piece of information based on the contributor's expertise, which is induced from dynamically changing user profiles, and help to find relevant information to people with similar interests and work tasks through *subjective concept-directed vertical search*, which is discussed in detail in Chapter 6.

To summarise, the search experience we are trying to provide can be characterised as collaborative and personalised. Users' searches and contributions have a personalised (I-Centric) nature, and information pieces found valuable by every user in context of his/her current information needs are shared among all users (We-Centric-ness).

3.3 Position of BESS among Modern Web Personalisation Systems

Reconsidering information retrieval in the context of each person is essential to continue searching effectively and efficiently. That is why so much attention is paid to this problem and consequently a number of approaches to Web search personalisation have arisen recently. Nowadays we are experiencing the much anticipated breakthrough in personalised search efficiency by "actively adapting the computational environment - for each and every user - at each point of computation" (Pitkow et al., 2002).

To show the peculiarities of existing Web search personalisation systems and the position of BESS inside Web search personalisation approaches we classify them as *vertical* and *horizontal*, *individual-oriented* and *community-oriented* based on breadth of search focus and degree of collaborativeness they possess (see Figure 3.2; arrows denote current trends in search personalisation).

Outride (Pitkow et al., 2002) and similar systems take a contextual computing approach trying to understand the information consumption patterns of each user and then provide better search results through query augmentation. Sugiyama et al. (2004) experiments with a collaborative approach constructing user profiles based on collaborative filtering to adapt search results according to each user's information need. Almeida et al. (2004) harnesses the power of community to devise a novel ranking technique by combining content-based and community-based evidences using Bayesian Belief Networks. The approach shows good results outperforming conventional content-based ranking techniques. Systems like Swicki, Rollyo, and Google Custom Search Engine correspond to vertical and mostly community-oriented approach of search personalisation. They provide community-oriented personalised Web search by allowing communities to create personalised search engines around specific community interests. Unlike horizontal (or broad-based) search systems mentioned above, such systems are considered personalised in the sense that available document collections are selected by a group of people with similar interests and the systems can be collaboratively modified to change the focus of search. Although not Web-based, we take tools like Google Desktop Search as an example of individual-oriented vertical search systems. They search contents of files, such as e-mails, text documents, audio and video files, etc., inside a personal computer. The absence (to the best of our knowledge) of salient Web-based systems of this kind can be explained by the increasing popularity of services on the Web benefiting from community collaboration and favouring fast transition of each person's activities from passive browsing to active participation.

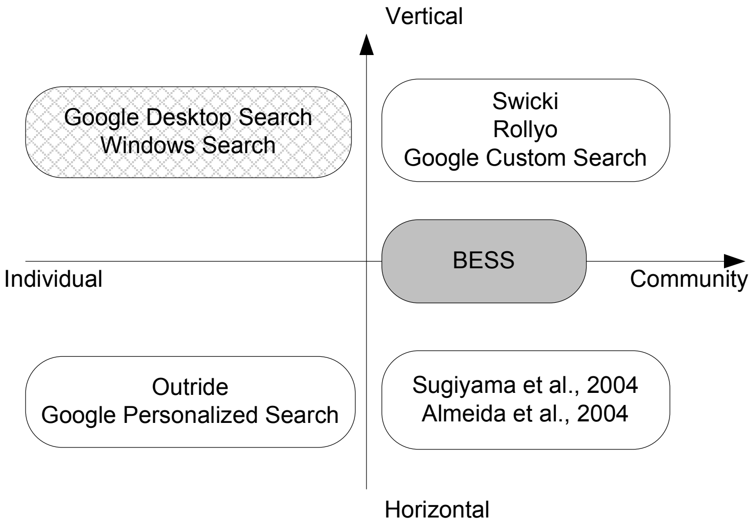


Figure 3.2 Search personalisation services and BESS

As it is shown in Figure 3.2, BESS is a community-oriented system having the features of both horizontal and vertical search system. It performs search on information assets of both horizontal (objective index) and vertical (subjective index) nature. The notion of *subjective index* in our research is similar to 'social search' of vertical community-oriented systems presented above, but differ in higher degree of personalisation for every user, high granularity of vertical search model (see *subjective concept-directed* *vertical search* in Chapter 6) and, finally, the way of collecting and (re-)evaluating information pieces. Groups of users are created dynamically without a user's interference based on match of interests/expertise, and the role of community is indispensable for search quality improvement and the system's evolution in general.

3.4 Architecture and System Overview

BESS is a complex system consisting of several components for relevance feedback collection, analysis and evaluation, online incremental clustering, user profile generation, indexing and a few elements realising several search functionalities.

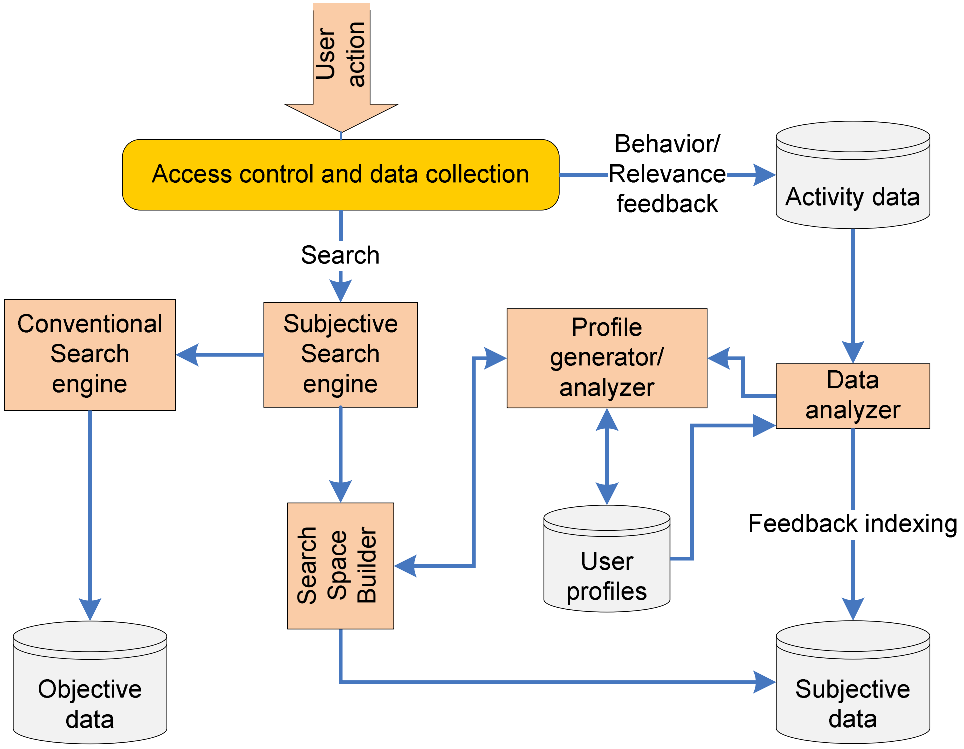


Figure 3.3 General system architecture

As we have already discussed, the main purpose of BESS is to realise collaborative personalised search. And to achieve its assigned tasks, first of all, our collaborative search and sharing system has to be capable of distinguishing users, and collecting and analyzing their personal feedback. "Access control and data collection" module of BESS is responsible for this. A user is authenticated when accessing the system, so we know whom it is used by. After that, his/her interactions with the system are logged. To have an understanding of the user's interests we are primarily interested with contributions (explicit feedback), done through the contribution widget of a Web browser, and implicit feedback, collected by monitoring the user's clickthrough. All the interaction data is stored in "Activity data" database, as shown in Figure 3.3. Then, this 'raw' data is processed and clusters (concepts) reflecting the user's interests are created by "Data analyzer". Existing concepts are incrementally updated. At this moment the interests are inferred and known, but are of little interest because they say nothing about their temporal characteristics. As a result, some concepts can be outdated, others can be recent and topical.

In order to organise the concepts, "Profile generator/analyzer" generates a user profile using interest-change-driven profile construction mechanism, as described in Chapter 4, and they are stored. We have to note that, as it is also discussed in the next chapter, user profile is very central for the system functioning in general. As it is shown in Figure 3.3, user expertise, together with expertise of other users, with regard to a particular topic (concept) is used for assessing his/her feedback, which is then indexed and stored in the "Subjective data" repository for further retrieval. This personal and 'collectively evaluated' feedback becomes a piece of the user's subjective index data.

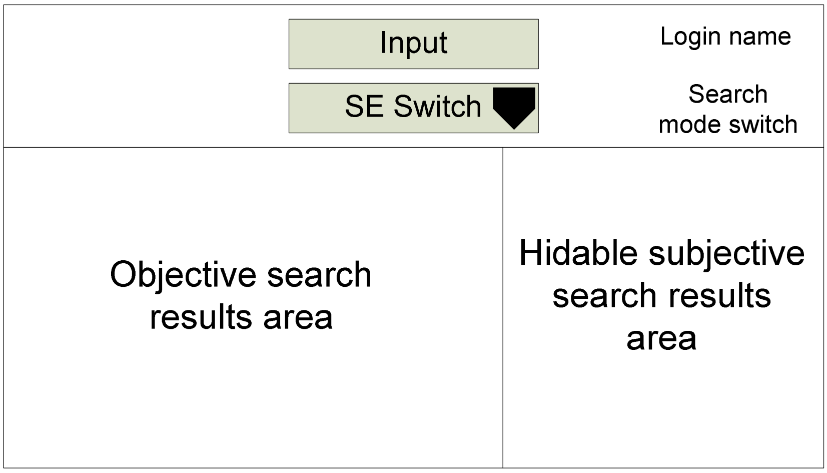

Input
Login name
SE Switch
Search mode switch
Objective search results area
Hidable subjective search results area

Figure 3.4 User interface schematically

Now, when we have data to be searched on, let us consider search.

Here the user has an opportunity to search both with conventional search engines and the search engine provided with BESS. Essentially, both are used when a search request is issued. The results of the conventional one are shown in “Objective search results area” and the results of the one provided by BESS are shown in “Hidable subjective search results area”. The user can select his/her favourite Web search service from “SE Switch” and hide “Hidable subjective search results area” if there is not enough subjective contributions for the topic in concern, or he/she is simply not interested in collaboration temporarily and want to concentrate on objective search only. In any case, the user is enriching his/her personal subjective index, and consequently all shared subjective index.

Search on the subjective index data is normally done in all-shared mode, when the subjective index of all users is searched on. In this case, query-document matching is performed, and all matched documents are retrieved and listed according to the ranking algorithm. However, the user has another option – to search on the subjective index data of the users whose user profiles are conceptually close to his/her current user profile by switching with “Search mode switch”. This is what we mentioned as *subjective concept-directed vertical search* already[1].

3.5 Notes on Implementation Technologies

In order to realise all the described functionalities, BESS employs a number of technologies, such as online incremental clustering, indexing and search. Indexing and search is done with help of customised Apache Lucene. User profile construction and online incremental clustering are the modules implemented according to the methods described in the following chapters of the thesis. All implementation is done with Java, using JSP (Java Server Pages), Java Servlet, Spring Framework and other Java technologies. For contribution submission Firefox component development, we used AJAX (Asynchronous JavaScript and XML) and XUL (XML User Interface Language).

[1] Detailed discussion of the ranking algorithm and subjective *concept-directed vertical search* is given in Chapter 6.

4 Constructing Interest-Change-Driven User Profile

As we have shown in Chapter 2, there are many different ways to construct and organise a user's interests using user profiles (UP). The organisation structure usually depends on what characteristics of the user and his/her interests a UP is designed to capture. User profiles in BESS are designed to timely and effectively capture the user's interests, to update his/her profile in regard with its temporal, and transitively interest-involvement-degree, characteristics, and to be used for collaborative contribution evaluation and information retrieval. UPs are composed from concepts which serve as representatives of the user's interests. They are *multi-layered* with layers reflecting temporal characteristics of user contexts. Furthermore, they are dynamically updated to precisely reflect changes in interests using *interest-change-driven profile construction mechanism* presented further in this chapter.

4.1 The Role and Position of User Profile

User profiles play a key role in our BESS information retrieval framework. The framework is developed in attempt to capture information needs and information seeking contexts of every individual, and better facilitate information seeking activities by identifying and providing information resources pertinent to every individual's needs. This is achieved by modelling a user's changing interests from relevance feedback (explicit feedback, called contributions, and observed user behaviour, such as clickthrough information) over time and using the models

- to evaluate the feedback by considering the contributor's expertise and his/her past experiences with the concept the user feedback belongs to, and
- to change the focus of search, similarly to what occurs in vertical search engines, but automatically, detecting users with similar contexts and using their concepts.

These steps ensure the search is done on highly selective documents evaluated by the users with similar interests taking into account their expertise, or the degree of their involvement into a particular topic.

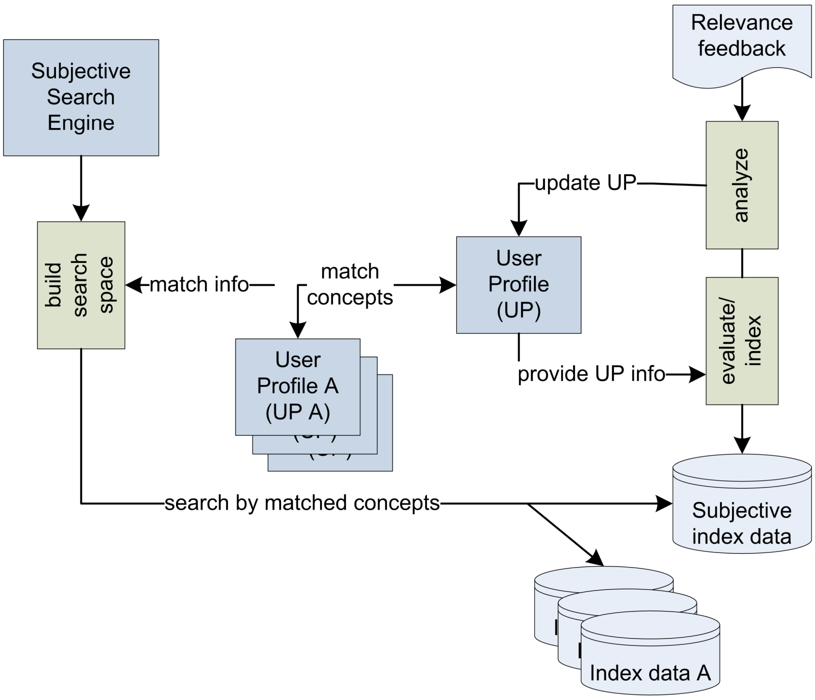


Figure 4.1 User profile inside system services of BESS

Figure 4.1 is a schematic fragment of the system architecture describing the position and the role that user profiles have inside the system. First, the analyzed relevance feedback is used to update a user's profile with a newly created or updated concept. Then, the updated profile and its concepts' peculiarities are used for the evaluation of the same relevance feedback item[1]. And finally, the feedback is indexed in every individual's subjective index repository which is shared among all users of the system. When a user searches, he/she can search on the multiple sets of information assets evaluated according to each user's expertise or narrow his/her search to the resources of those users whose interests (concepts in user profiles) are similar to his/her own.

As it can be seen from this short description, the position of user profile in the system operations is central and the quality of the profile is of vital importance not only to information seeking experiences of one user but to the experiences of all users of the system. Therefore, in this paper we pay the particular attention to the profile construction and to the quality of the concepts, which are the constituents of user profiles and indicators of user interests, in particular. In Chapter 5 we propose the clustering method for

[1] see Chapter 6 for the details of the evaluation mechanism

the extraction of concepts from relevance feedback in information seeking tasks and give its extensive explanation and evaluation.

4.2 Concept as a Principal Profile Component

Relevance feedback is an essential element of any information filtering system and a significant part of the proposed system. It is exhaustively researched in its various forms. Explicit feedback often disrupts normal user activities; therefore another form of feedback that can be collected with no extra cost to the user – implicit – is used widely. Sometimes these two forms are combined to get better insight about a user's peculiarities. Kelly et al. (2003) gives a good classification and overview of works on implicit feedback. Generally, user behaviour is considered to be an implicit feedback, and its analysis is done for improving information retrieval by predicting user preferences, re-ranking Web search results and disambiguating queries (Agichtein et al., 2006a; Agichtein et al., 2006b; Shen et al., 2005a).

Often relevance feedback is used in attempt to find out user behavioural patterns and generate individual user profiles reflecting current user interests. There exist many approaches for profile modelling. Nanas et al. (2003) discusses profiles made from concept hierarchies that are generated from user specified documents and applied for information filtering. Profiles are divided into three layers by heuristic threshold each of which determines the topic, subtopics and subvocabulary for the specified topic. Term weighting approach (Relative Document Frequency) is extensively used for hierarchy construction. Semeraro et al. (2005b) uses a different approach for profile construction. As in (Nanas et al., 2003), profiles consist of concepts, but the approach employs ontologies where semantic user profiles are built with the use of content-based algorithms extended using WordNet (Fellbaum, 1998). Such an approach is proved to help infer more accurate user profiles. User profile defined (in Koutrika et al., 2005) "treats query disambiguation and personalization as a uniform term rewriting process". Profile, a directed graph representing term relations, dictates modifications of search queries.

BESS makes extensive use of both explicit and implicit relevance feedback for construction of personal information assets and user profile. Unlike profiles in the above-mentioned approaches, profiles in BESS are constructed with the main focus on users' interest change when searching, and concepts are loosely coupled and dynamic.

User profile in BESS is a structured representation of user contexts which are in turn consist of preferences and interests of a user. It consists of concepts (semantic clusters), and each concept is the system's piece of 'knowledge' about what the user is interested in. Each concept is modelled as a cluster c_i of n document vectors $X = (x_1, ..., x_n)$ from the individual document set grouped by a specific 'knowledge' criteria. Concepts are extracted from minimal user search and post-search behaviours (user-system interactions while searching, browsing and contributing Web pages) – relevance feedback captured by BESS. The system is configured to capture the following data with the Web proxy:

- user ID used for authentication;
- search query terms;
- URL of the page user is interacting with;
- type: query, click or feedback;
- timestamp;
- session ID.

Prior to concept extraction, documents from individual document collections are linearised by removing HTML and script tag data, non-content-bearing 'stopwords' are deleted and document vectors are normalised. Then, a classification method is used to extract concepts from the document vectors. Virtually, any method can be applied for this. In Chapter 5 we present an online incremental clustering algorithm H2S2D which is suitable for incremental unsupervised concept extraction and tailored specifically for uniform relevance feedback data. In (Shtykh and Jin, 2008b) we experimented with *Intelligent K-Means* algorithm, or *iK-Means* (Mirkin, 2005). iK-Means uses *Anomalous pattern* method as a procedure to meaningfully determine the number of initial seeds for K-Means that performs a disjoint partitioning of document vectors and computes a centroid for each partition. Normalised centroids contain valuable semantic information about the partitions (concept), therefore they can be treated as representative components (vectors) of concepts.

We applied iK-Means to data produced by users interacting with a Web social bookmaking service located in Japan. Social bookmarks were considered as explicit relevance feedbacks and used as a corpus for concept formation in the experiment. The followings are an example of the sample concepts obtained in the experiment and their representative terms.

Table 4.1 Concepts produced by simulation on Web social bookmarking service data (The original corpus is in Japanese. The terms are translated to English)

Concept A	**Concept B**	**Concept C**	**Concept D**	**Concept E**
Diary	Hyogo (region)	Text editor	Balance	Technique
Book	Brand	Code	Effect	Recruitment
Reader	Shop	Document	Cause	Mid-career
Category	Confectionery	Free	Female	Company
		Author	Nutrition	Annual income

4.3 User Profile Structure

Information seeking, as any information behaviour, is done in a context determined by situation, interest, person's task, its phase and other factors. In process of seeking information needs and their contexts are changing even within the same seeking task.

Recognising the fact, we introduce a temporal dimension to user profiles by splitting and combining (generalising) all concepts on a time line. For this, we make user profiles in BESS multi-layered – each layer reflects user interests within a certain period. It consists of four layers – static $pr^{(st)}$, session $pr^{(ss)}$, short-term $pr^{(sh)}$ and long-term $pr^{(ln)}$ (Figure 4.2). Thus, profile of user a can be defined as

$$\Pr_a = (pr_a^{(st)}, pr_a^{(ss)}, pr_a^{(sh)}, pr_a^{(\ln)}) \tag{4.1}$$

Each layer consists of concepts which are the components of profiles representing user contextual information by topics:

$$pr_a^{(l)} = (C_{a1},...,C_{ak}), \tag{4.2}$$

where *l* is a layer and *k* is a concept number.

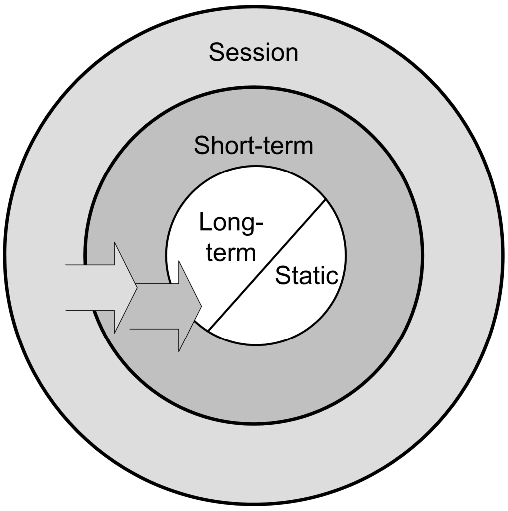


Figure 4.2 Layered user profile

Each layer has a pool of concepts that characterise best a user's seeking context during the layer's time span. The static layer is defined at the start of user-system interaction to solve so-called "cold start" problem when the system has no information about the user and cannot facilitate his/her activities or can even damage the whole interaction. Other three layers can be classified as dynamic layers, since they are dynamically constructed and changed along with changing user information needs and their contexts.

The session layer contains the fragmentary context of the current information behaviour of a particular user. It is a highly changeable layer and defined by a concept that best matches one of the concepts available in the short-term layer or a newly created concept. In other words, the session layer is the indicator of context switch at the lowest level. The short-term layer is a central layer of the whole system – it consists of concepts formed in all user-system interaction sessions within a specified period of time, and its generation itself serves as an important factor for collaborative feedback evaluation mechanism. And finally, the long-term layer is derived from the most frequent concepts of the short-term layer, as discussed in the profile construction section, and reflects general user context of interaction with the system. When there is enough information for its formation, it is created and gradually supersedes the static layer. The profile layer construction mechanism is further described in the next sub-section.

4.4 Dynamic Interest-change-driven Profile Construction

As we have already described, user profile plays an important and central role in BESS for collaboratively evaluating documents contributed to the community and for adjusting the focus of search. Therefore user profiles have to be precise and accurate, and this is achieved by correctly specifying their concepts by the online incremental clustering algorithm we present further in the next section. Moreover, profiles ought to timely reflect the changeability of user interests while maintaining the steadiness of persistent preferences. In our *interest-change-driven* model for dynamic user profile generation we proposed in (Shtykh and Jin, 2008b) we adopt *recency*, *frequency* and *persistency* as the three important criteria for profile construction and update.

Once we have concepts extracted from a user's feedback, we can detect the change of a user's context and set the latest one as the current context (*recency criterion*), which is the session layer in multi-layered user profiles. By observing concept creation dynamics we can set some to be the short-term layer according to the following (*frequency* and *recency*) rule:

For n concepts in the latest clustering output, choose newly-created and already existing concepts whose input item growth is high in a reverse order (newness) of the output sequence.

And finally, the long-term layer is formed from *n* most frequent concepts which have also been observed in the short-term layer.

Thus, concept extraction (clustering) method produces $C_a = \{C_{a1}, \ldots, C_{an}\}$ set of *n* concepts which are ordered by recency criterion, i.e., a concept that is newly created or most recently updated appears at the top of the recency list. C_{a1} is the most recent concept and considered to be the current context and the session layer of the profile of user *a*, i.e., $pr_a^{(ss)} = C_{a1}$.

The short-term layer consists of *m* most frequently updated and used concepts, which are, in their turn, chosen from *r* most recent (top) concepts in the concept recency list. In other words, these are the concepts that are frequently used and still of some interest for the user. Figure 4.3 explains how the short-term layer is created.

1 Fetch *r* most recent concepts from the recency list.

2 Sort them by update frequency in the descending order.

3 Take top *m* concepts to create the short-term layer $pr_a^{(sh)} = \{C_a^1, ..., C_a^m\}$, *m* <= *r*. The concepts are ranked by frequency.

Figure 4.3 Short-term layer creation procedure

The goal of the long-term profile layer is to find persistent user interests. Therefore its construction is based on *persistency* criterion and, indirectly, on frequency and recency considered for the short-term layer creation – the layer is derived from the concepts of the short-term layer which were most frequently observed as the layer's components. To determine the concepts matching the afore-mentioned criteria, in addition to concept update frequency $freq_c$, we introduce frequency measure $freq_s$ for the number of times the concept was a component of the short-term layer and find *m* concepts whose *persistency factor PF* is high. Persistency factor is a measure to infer the user's continuous interests by combining a concept's frequency count with its evidence of being a user's short-term layer's constituent.

$$PF_{C_{ai}} = \alpha \frac{freq_c(C_{ai})}{\max freq_c(C_a)} + (1 - \alpha) \frac{freq_s(C_{ai})}{\max freq_s(C_a)} \qquad (4.3)$$

where α is set experimentally. C_{ai} is a concept of the set of concepts C_a produced from relevance feedback of user *a*.

The concepts for the long-term layer are found by the procedure shown in Figure 4.4.

1 Calculate PF_{Cai} for each concept.

2 Take *m* concepts with the highest values to create the long-term layer $pr_a^{(\ln)} = \{C_a^1, ..., C_a^m\}$.

Figure 4.4 Long-term layer creation procedure

All the layers dynamically created at time *t* form concept-based *interest-change-driven model* of user *a*, and are the representation of the user's interests at *t*. A change of the concepts in terms of their ranking in the short-term profile layer signifies a change of user interests and emergence of a new model of user *a*. The model update is not constrained with the predefined parameters, such as fixed time period after which the update occurs, and driven by natural dynamics of changing user interests. This mechanism is used to find a user's *n* past profiles and their concepts to determine the areas of expertise of the user to be used in his/her feedback evaluation mechanism, as described in Chapter 6.

4.5 User Profile Construction: An Example

In order to demonstrate profile construction using the proposed online incremental clustering method and profile construction scheme and show the rationality of the chosen approach, we give an example of profile construction and discuss its peculiarities.

First, we implemented the profile construction system where every user relevance feedback was processed one by one and the extracted concepts were used to create user profiles according to the scheme described in Section 4.4. Then, we prepared relevance feedback of several users collected using the same method we used for data collection in the discussion about the assumption of our proposed clustering algorithm in Section 5.1, processed it with the system and produced user profiles. Overall, 20 concepts were created using H2S2D clustering with 0.1 threshold.

Here we show typical user profile construction results for one user. Since the session profile layer is simple – consisting of one currently used concept – and very frequently changed with the change of the user's current interests and needs, we skip it to illustrate the dynamics of short-term and long-term layers. Figure 4.5 shows how the user's short-term profile layer is being generated during concept extraction process. "Processed items" axis refers to the number of relevance feedback items processed by H2S2D method. So, for instance, label "288" indicates 288 items processed one by one and it is a point of change of user interests – literally, change of rank of concepts C1, C6, C4 and C8 in the short-term profile layer. "Rank" axis refers to the rank of concepts in the layer (explained in sub-section 3.3), where only top *m* items are considered being the layer's concepts and the others are given to show the change dynamics of concepts during the period the user supplies relevance feedback data when interacting with the system. As shown in the figure, the rank of the concepts, as it can be expected, tend to change often initially and become more stable when more feedback, and accordingly better-quality concepts, is available. In fact, we do not expect the scheme to produce an unchangeable layer, since this layer has to reflect short-term user interests and the change in concept rank indicates emergence of a new user model. This model is called *interest-change-driven model*, since a new profile generation/update is caused not by some predefined settings, such as days, hours, etc., but by the dynamics of model generation itself (concept rank change). For instance, if we choose *three* most frequent of *r* most recent concepts (*m* <= *r*) in Figure 4.5 to be in the short-term layer, we can see that concept rank change, and accordingly a new model definition, occurs after item 94, 131 and so on are processed. The most highly ranked concept C5 keeps its top position since the user keeps working on plugin implementation for Firefox browser. The third highly ranked concept changes often from interests in news to travel and conference-related topics. See Table 4.2 for simple explanation of concepts presented as a number of terms representative of the concept topics.

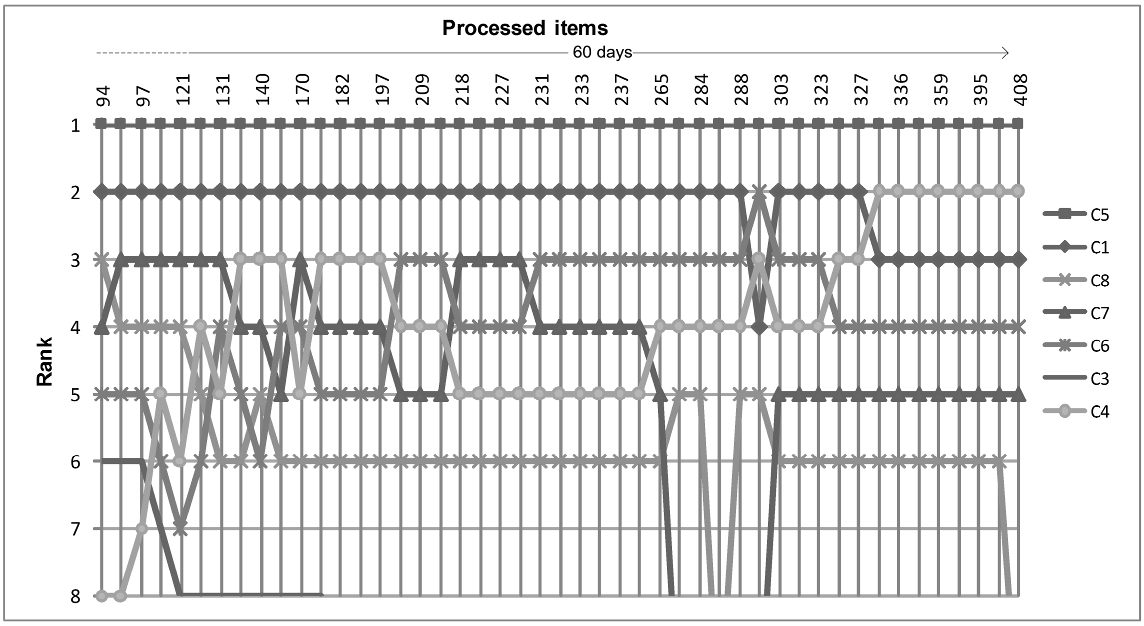


Figure 4.5 Short-term profile generation

Figure 4.6 shows the generation process of the long-term profile layer which was being constructed alongside with the short-term layer. Again *m* concepts with the highest *persistency factor PF* (with α=0.1) are chosen to be the long-term layer. But, in contrast to the short-layer concepts, they are not ranked and only the fact of their belonging or not belonging to the top *m* concepts is important for the definition of the layer. The concept ranks in Figure 4.6 are shown to indicate the change dynamics of the concepts' *PF* value and their less frequent changeability in comparison to the short-term layer's concepts. In other words, if we choose *three* concepts with the highest *PF* to be the long-term layer's concepts, the layer remains the same from the formation time (after item 217 is processed) – $Pr_a^{(ln)} = \{C6, C7, C4\}$. It changes only after concept *C1* gains *PF* value that is higher than the value of *C7* – $Pr_a^{(ln)} = \{C6, C4, C1\}$. *Persistency factor* ensures that only those concepts that have a tendency for being long-term interests gain higher value. For instance, concept *C1* is ranked highest in the short-term layer and there is no evidence that this is not a long-lasting temporary interest and the user will return to it in future, therefore its *PF* value is not one of the highest. However, if such evidence will be available in future when the user returns to interests reflected by *C1*, its *PF* value will increase and the chances to be a concept of the long-term profile layer will grow.

Table 4.2 Five discriminative keywords for concepts in Figure 4.5 and 4.6

C5	**C1**	**C8**	**C7**	**C6**	**C3**	**C4**
Firefox	real estate	software	Japanese news	conference	massage	travel
customise	square meters	hardware	company	workshop	Tokyo	tour
plugin	area	computer	business	mobile	stress	Malta
colored scrollbar	rent	forum	politics	multi-media	traditional	ticket
browser	floor	talks	finance	IR	nerves	hotel

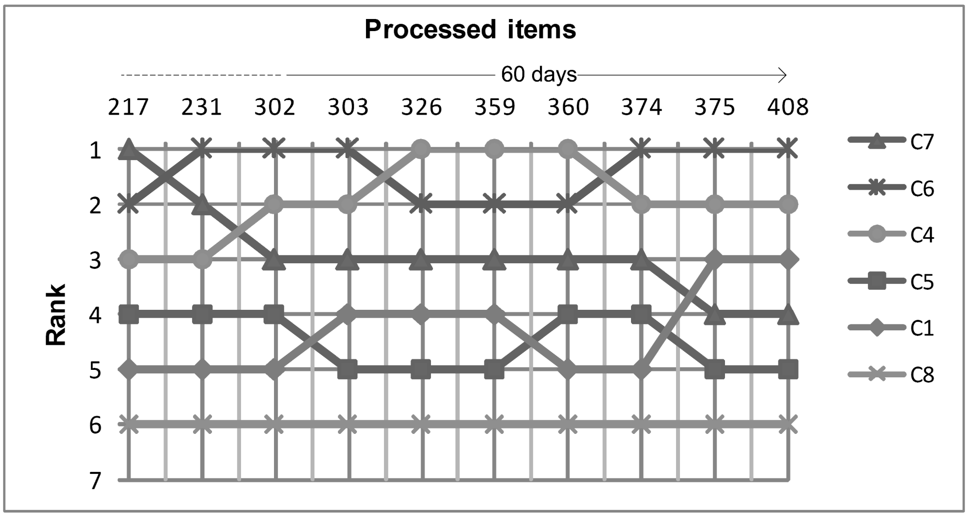


Figure 4.6 Long-term profile layer generation

As it can be seen from the figures and their explanations, the constructed layers meet our expectations and requirements to reflect a user's current interests for the session profile layer, to be a representation of both recent and frequent interests (i.e., the recent and vivid interests lasting for some time) for the short-term layer, and to collect persistent interests for the long-term layer.

5 Inferring User Interests with H2S2D Online Incremental Clustering

As we have already mentioned, concepts are the elementary parts from which user profile layers are constructed. Such concepts can be obtained, for instance, by grouping/classifying user relevance feedback items, such as the objects selected by a user during his/her interaction with a search system or explicitly shown user preferences. In information seeking systems the most widely used and easily obtained material for formation of concepts (or micro-models) is a feedback document.

Although for the extraction and formation of concepts (or user interest inference) virtually any classification method[1] can be used, its choice is determined by the properties of data set (e.g., type) and a priori knowledge (e.g., conceivable classes) the method has about the data set. To learn user interests from continuously incoming data items, such as relevance feedback during information seeking, an online incremental method has to be considered. In contrast to non-incremental clustering methods, which assume all data set is available before the clustering begins, online incremental methods are usually not aware of full data set and receive data items one by one or in portions. Further, the incoming data can be classified in supervised or semi-supervised fashion if there is some a priori knowledge about the data, otherwise unsupervised classification can be applied.

The algorithm we propose for concept extraction can be classified as online incremental and unsupervised, since we are dealing with the incoming data stream and have no knowledge of the data except its type. But in contrast to other methods of this group, it is tailored for concept extraction from uniform relevance feedback data in information seeking tasks and uses the peculiarities of a user's seeking behaviour for concept creation to achieve high effectiveness and accuracy. It is based on the assumptions described in the following section.

5.1 Assumptions for Relevance Feedback Clustering

The proposed user interest inference method is based on the following intuitive assumptions.

- When a user searches, he/she usually sees (clicks on, focuses the attention on, etc.) several

[1] Some of them are discussed in Chapter 2

documents (links or other objects) until the most relevant is found. Most of these documents are potentially inter-similar to some extent and can give a conception about a particular user interest.

- Even if some similar documents are not sequenced till the present moment, there are documents related to the persistent user interests and re-searches on these interests are likely to occur. In these cases a user either clicks on the links he/she found before or on the links leading to the documents highly similar to those found before.

Such assumptions allow us to think of user relevance feedback, such as search queries and clickthrough, as sequentially-incoming data *S* with subsequences of n (more than one) or more highly similar items $S = S_1S_2...S_n...$ which are linked through by a particular information need, and consider them as potentially new semantic clusters (concepts), if there are no such clusters created yet. In other words, an inference about a new user interest (concept) is done when n or more similar items that have no or very little similarity to others are observed sequentially during t_0 to t_n (see Figure 5.1). Those items that are not coming in high-similarity subsequences are still considered as potentially related to user interests, but since they are not much useful for profiles they are put into a candidate pool to be retrieved and used for concept formation later when a subsequence of similar feedback data items is observed (the second assumption). Otherwise, they are considered to be outliers and can be purged later.

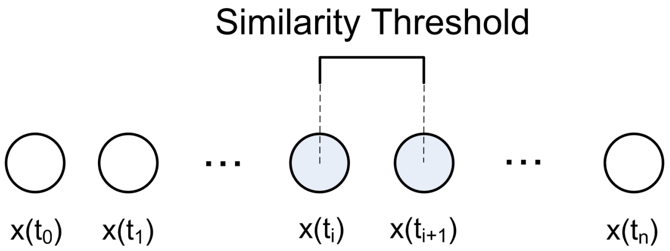


Figure 5.1 High-similarity items in data stream

In order to check the validity of our assumptions, we conducted a short user study. The information search records (clickthrough and Web browser's URL bar input) of 12 users in their mid-twenties to mid-forties were collected and analyzed. The study's duration was two weeks, and on average 320 records per a participant were collected. From the documents which were retrieved from URLs of the records we created multi-dimensional vectors using the classic TF-IDF (Term Frequency-Inverse Document Frequency) weighting scheme. To analyse the sequential characteristics of the data, we compared inter-similarity of items in the relevance feedback data sequence using the classic cosine similarity measure. Items were considered similar if their similarity were bigger than a particular similarity threshold: 0.05 for the first analysis, 0.1 for the second and 0.2 for the last one[2].

The analysis results have shown that 36% of the relevance feedback sequential data consisted of subsequences of documents which inter-similarity satisfies 0.05 similarity threshold (see Figure 5.2 (a)). 14% and 9% of all relevance feedback data (about 64% of all sequential data) are two-item and three-item subsequences respectively. The percentage of subsequence data decreased to 21% (Figure 5.2 (b)) for 0.1 and 7% for 0.2 thresholds. Due to the small scale of the study, its estimations should be considered rough, however its results are very indicative of the nature of the considered relevance feedback data sequence. From these results we can say that the first assumption about inter-similarity is satisfied, however the inter-similarity measure has to be set with care. The study results have also shown that the percentage of re-accessed and all inter-similar documents in the collected data is high (second assumption): about 83% for 0.05 threshold, and 74% and 57% for 0.1 and 0.2 thresholds.

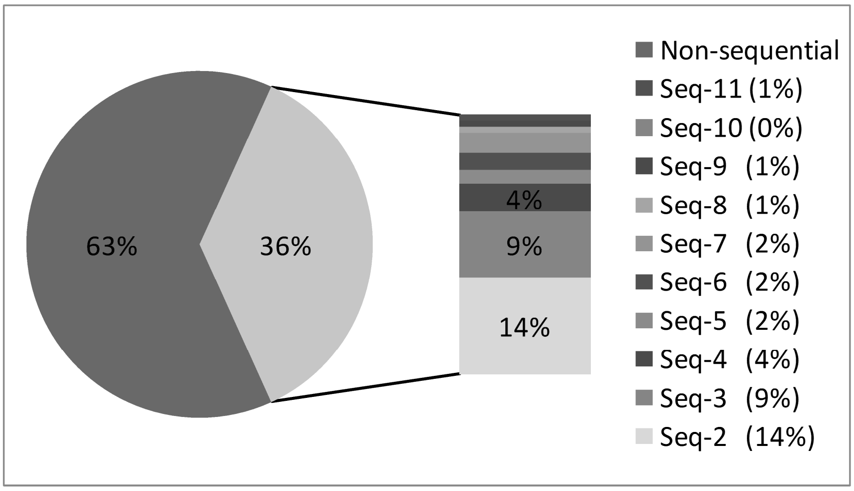


Figure 5.2 (a) Subsequence percentage for 0.05 threshold

[2] These same thresholds are used for the algorithm evaluation

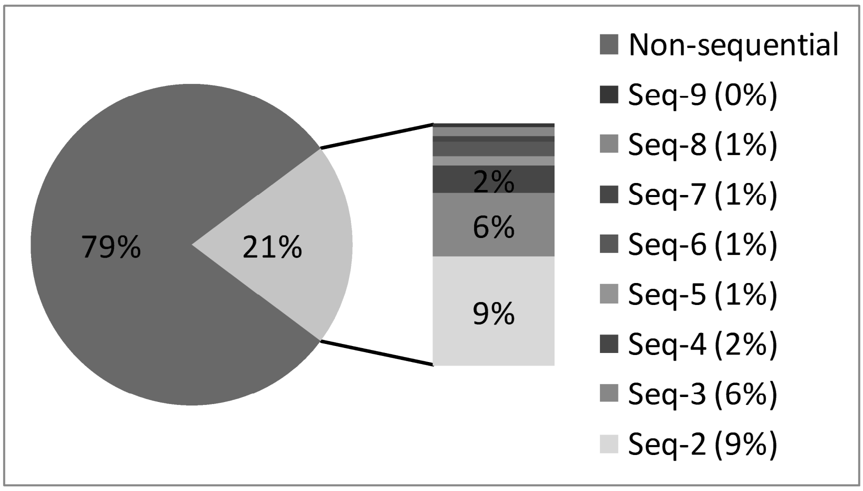


Figure 5.2 (b) Subsequence percentage for 0.1 threshold

5.2 Relevance Feedback Clustering with H2S2D Method

As we have already mentioned, the proposed clustering method is online and incremental. Charikar et al. (1997) define incremental clustering problem as follows – "for an update sequence of *n* points of *M*, maintain a collection of *k* clusters such that as each input point is presented, either it is assigned to one of the current *k* clusters, or it starts off a new cluster while two existing clusters are merged into one." Greedy Incremental Clustering and the Doubling Algorithm are the representatives of the approach (Charikar et al., 1997). However, setting fixed number clusters does not allow capturing new contexts with high precision when the number of contexts exceeds the number of clusters. Some methods, as SHC (similarity histogram-based clustering) (Hammouda and Kamel, 2003) or ECM (Evolving Clustering Method) (Kasabov, 2007), do not define the limit for the number of clusters, but introduce conditions when an incoming data should create a new cluster and when it is immersed into an existing cluster. Thus, SHC processes data according to clusters' coherence criteria. ECM judges how to treat an incoming item using the specified distance threshold.

```
Step 1:
Set similarity threshold $S_{th}$ for the
clustering task;
Step 2:
Receive item $x(t_i)$ from the input stream;
Step 3:
Compute its similarity $s$ to clusters
in cluster list $L$ and find $s_{max}$.
If $L$ is empty
   - add $x(t_i)$ to candidate pool $P$;
   - set the latest candidate ($k$)'s value to $x(t_i)$
   - go to Step 2.
Else
   - proceed to Step 4;
Step 4:
If $s_{max} < S_{th}/2$
   - check $P$ for $k$'s value;
   - go to Step 5;
Else
    If $s_{max} < S_{th}$
  - add $x(t_i)$ to $P$;
  - set $k$ to $x(t_i)$;
    Else
  - assign $x(t_i)$ to cluster with the highest $s$;
  - set $k$'s value to null;
  - go to Step 2.
Step 5:
If $k$ is null
   - add $x(t_i)$ to $P$;
   - set $k$ to $x(t_i)$;
Else
   - compare $x(t_i)$ with $k$ and find their $s$
   If $s < S_{th}$
      - add $x(t_i)$ to $P$;
      - set $k$ to $x(t_i)$;
   Else
      - create new cluster $C_n$ with $x(t_i)$
        as a reference point $Rp_n$ and add it to $L$;
      - add $k$ to $C_n$;
      - find items from $P$ whose $s > S_{th}$
        and assign to $C_n$;
      - set $k$ to null;
- go to Step 2.
```

Figure 5.3 H2S2D algorithm

Similarly to ECM algorithm, the proposed H2S2D algorithm (shown in Figure 5.3) uses similarity threshold S_{th}, which is the only parameter we have to set for clustering, to specify how similar data items should be to be considered similar by the algorithm. The key features for H2S2D are the followings:

- a new cluster definition relies upon sequential characteristics of relevance feedback;
- assignment of an incoming data item is delayed if there is no similar enough cluster, and performed when such a cluster is created.

In order to create a cluster (i.e., identify a particular user interest), relevance feedback data sequence is examined. When a subsequence of more than *n* highly similar data items (whose similarity in subsequence is larger than the specified threshold S_{th}) is observed, a new cluster is created using the data in the subsequence.

In contrast to other approaches (e.g., ECM) where the cluster centre's recalculation happens, H2S2D chooses a reference point *Rp* from the items of the newly-created cluster and it remains invariant. These items are considered as a rough representation of a well-separated user interest and the reference point groups all cluster items within S_{th}. However, for some information retrieval tasks a mass-centre computation for each cached cluster may be done separately. We define the operation as *delayed mass-centre definition* – a computation of every cluster's mass-centre is done as a pre-processing step before using the clusters in retrieval. As this is a completely separate process, it does not affect H2S2D clustering performance.

The algorithm does no reassignment of data items to other clusters – an assignment for every incoming data item $x(t_i)$ is done once and only if there is a cluster whose item–cluster similarity is bigger than S_{th}. The item–cluster similarity is measured between *Rp* of each cluster and the data item. If no suitable cluster is found, $x(t_i)$ is placed into the candidate pool and remains there until a new cluster the item can be assigned to is created (see Step 5 in Figure 5.3). Thus, we always have a number of well-defined and updated clusters at hand and ready for profiling tasks. While the items are in the candidate pool, they are of no interest and may only introduce inaccuracies in such tasks. As we will show in the evaluation section of the paper, the number of such items tends to reduce to a certain level as persistent user interests are being determined. If the big size of the candidate pool becomes a performance concern, an aging mechanism can be introduced to purge its items gradually.

5.3 Evaluation

5.3.1 Data Sets

For the evaluation of the proposed clustering algorithm, first we used Reuters-21578 collection and then “20 newsgroup” data set. The former collection contains manually categorised short financial news items. A document of the collection can belong to more than one category, and some are not categorised. For evaluation we extracted and used text items having a topic enclosed in <TOPICS> </TOPICS> tags and discarded the rest. Also we removed multi-topic documents leaving only those having one topic. As a result, we obtained 8,643 documents of 65 topics – in fact, eight topics had only one document, thus becoming members of the candidate pool by the algorithm’s definition. “20 newsgroup” data set is a collection of 20,000 messages from 20 different newsgroups (1,000 messages for each group).

The input for the algorithm was pre-processed in the following way:

- normalised term vectors (stop-words removed) were created to serve as input data items;
- data items were arranged so that the input contained at most five[3] subsequences of two items of each particular topic. All other subsequences were generated randomly.

Items are processed by the algorithm as they arrive in the input stream.

5.3.2 Evaluation Measures

To evaluate the algorithm we used standard relevance measure metrics for document classification. And since our goal is to determine basic user interests that can be used as building blocks of a user profile, *precision* and *accuracy* of clustering were of particular interest. During the experiment the following quantities were recorded.

- *TP* (True Positives) – number of class-relevant documents correctly identified as relevant;
- *FP* (False Positives) – number of class-irrelevant incorrectly identified as relevant;

[3] The number which is supposedly enough to represent a topic in the sequence

- *TN* (True Negatives) – number of class-irrelevant correctly identified as irrelevant;
- *FN* (False Negatives) – number of class-relevant documents incorrectly identified as irrelevant.

Accuracy is the fraction of correct classifications. It is calculated as

$$Acc = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.1}$$

Precision is the fraction of correctly classified data items, and calculated as

$$P = \frac{TP}{TP + FP} \tag{5.2}$$

To give better idea about the method, we also calculated *F-measure* which combines precision and recall:

$$F = \frac{2PR}{P + R} \tag{5.3}$$

where

$$R = \frac{TP}{TP + FN} \tag{5.4}$$

is *recall*.

Since candidate items are considered unclustered at moment t_i and some percentage of them will be clustered during further clustering iterations, they are not taken into account for the above measures.

5.3.3 Evaluation Results and Discussion

In (Shtykh and Jin, 2008c) we compared H2S2D clustering method with Center-Greedy algorithm (Charikar et al., 1997), which is a popular and simple method for incremental clustering with a preset number of clusters, against Reuters-21578 collection. We have found that our proposed clustering method performs much faster and often better then Center-Greedy approach. However, the two clustering methods are very different by their characteristics, therefore such a comparison, though indicative, does not reveal all the peculiarities of the proposed approach. Hence, we conducted comparison with ECM algorithm (Kasabov, 2007) we describe below. ECM method, as H2S2D, does not limit the number of clusters and the threshold value plays the role of 'vigilance' parameter for the clustering process determining the ranges within which items to be assigned to one particular cluster lies. However, we have to note that in ECM this 'vigilance' parameter is a distance, while in H2S2D it is a similarity, so in the experiments we introduce hereafter the distance thresholds for the algorithm is converted into similarity. ECM processed the same input data stream H2S2D did.

Table 5.1 shows values obtained by the afore-mentioned measures on Reuters-21578 collection for 500 to 8000 processed data items[4] with similarity thresholds 0.05, 0.1 and 0.2, which were also used in the user study (Section 2). H2S2D operates with similar accuracy and precision for $S_{th}=0.05$ and shows higher results for $S_{th}>0.1$ clustering. However, we have to note again that the items from the candidate pool are not considered for clustering with H2S2D, because the decision about their belonging to a particular cluster is delayed until an appropriate cluster is created. Until then, they are considered both as candidates for assignment and potential outliers. Therefore they are not considered in the above measures for the evaluation of H2S2D since we cannot be sure in the results until such assignment happens. And therefore the comparison with ECM should not be considered a full-fledged clustering comparison and the results in Table 5.1 are shown to indicate the quality of inferred concepts only, which is the primary concern of the research.

The number of clusters produced in the experiment (Table 5.2) and numeric results in Table 5.1 shows that our chosen strategy for cluster creation based on similarity subsequences in relevance feedback works well with S_{th}=0.2 and produces clusters with high precision and accuracy values. In contrast, ECM, which considers primarily the threshold value for cluster creation and clusters all incoming items, produces a large number of clusters (much larger than the number of topics in the data set), and as a result, performs worse in terms of precision and accuracy. As for the results with S_{th}=0.1, they are quite similar for the both methods. H2S2D produces more clusters than ECM with 0.05 threshold, but their number is justified with accuracy and precision results. Nevertheless, as we see from clustering results of Reuters-21578 collection, high similarity threshold is the better choice for the proposed algorithm.

[4] Values 3000 to 6000 are omitted, since they follow the general tendency observed from the values presented in the table

Table 5.1 Clustering results (Reuters-21578)

		H2S2D			ECM		
S$_{th}$	**Items**	***Acc***	***P***	***F***	***Acc***	***P***	***F***
0.05	500	0.78	0.56	0.62	0.80	0.50	0.56
	1000	0.83	0.50	0.58	0.81	0.47	0.49
	2000	0.86	0.52	0.52	0.82	0.50	0.52
	…	…	…	…	…	…	…
	7000	0.89	0.48	0.49	0.85	0.47	0.49
	8000	0.89	0.49	0.50	0.86	0.47	0.50
0.1	500	0.95	0.61	0.64	0.89	0.61	0.57
	1000	0.90	0.64	0.60	0.85	0.52	0.45
	2000	0.90	0.65	0.54	0.85	0.54	0.44
	…	…	…	…	…	…	…
	7000	0.95	0.56	0.56	0.85	0.58	0.39
	8000	0.95	0.57	0.57	0.86	0.60	0.38
0.2	500	0.99	0.95	0.94	0.84	0.76	0.30
	1000	0.91	0.93	0.89	0.85	0.73	0.30
	2000	0.91	0.82	0.79	0.85	0.71	0.26
	…	…	…	…	…	…	…
	7000	0.93	0.81	0.70	0.84	0.67	0.20
	8000	0.93	0.81	0.70	0.83	0.68	0.19

Table 5.2 Number of clusters (Reuters-21578)

Items	S_{th}=0.05		S_{th}=0.1		S_{th}=0.2	
	H2S2D	ECM	H2S2D	ECM	H2S2D	ECM
500	5	4	8	10	5	48
1000	6	4	11	13	10	61
2000	9	4	13	14	11	74
3000	9	4	14	16	12	84
4000	11	5	20	21	20	93
5000	13	6	22	22	21	102
6000	14	6	24	23	22	113
7000	14	6	25	25	24	117
8000	14	6	25	27	24	123

The results of clustering "20 newsgroup" data set are shown in Table 5.3. In comparison to the results of Reuters-21578's clustering, we observed a decrease of precision in the both clustering methods. This is explained by higher categorisation quality of Reuters-21578 collection which was manually classified. Also, as one can see from Table 5.4, after completing clustering of all "20 newsgroup" data set with S_{th}=0.2 we obtained 22 concepts, the number close to the true number of classes in the data set. However, in order to obtain it, all data had to be clustered. On the other hand, clustering with S_{th}=0.1 produced the reasonable number of clusters with reasonable precision and accuracy much earlier.

Table 5.3 Clustering results (20 newsgroup)

		H2S2D			ECM		
S_{th}	**Items**	***Acc***	***P***	***F***	***Acc***	***P***	***F***
0.05	500	0.89	0.31	0.39	0.81	0.25	0.35
	1000	0.91	0.37	0.46	0.83	0.19	0.28
	2000	0.92	0.31	0.35	0.87	0.22	0.29
	3000	0.92	0.27	0.29	0.88	0.18	0.24
	…	…	…	…	…	…	…
	7000	0.94	0.32	0.27	0.91	0.27	0.29
	8000	0.94	0.30	0.25	0.92	0.26	0.28
0.1	500	0.87	0.58	0.70	0.93	0.36	0.22
	1000	0.94	0.53	0.62	0.93	0.31	0.19
	2000	0.94	0.45	0.51	0.94	0.28	0.14
	3000	0.95	0.45	0.40	0.94	0.22	0.13
	…	…	…	…	…	…	…
	7000	0.95	0.46	0.32	0.94	0.27	0.13
	8000	0.95	0.44	0.31	0.94	0.26	0.12
0.2	500	0.5	0.5	0.67	0.94	0.62	0.15
	1000	0.5	0.5	0.67	0.94	0.53	0.11
	2000	0.94	0.63	0.77	0.95	0.48	0.09
	3000	0.90	0.72	0.63	0.95	0.47	0.07
	…	…	…	…	…	…	…
	7000	0.93	0.66	0.51	0.95	0.42	0.06
	8000	0.94	0.61	0.45	0.95	0.38	0.06

Table 5.4 Number of clusters (20 newsgroup)

Items	S_{th}=0.05		S_{th}=0.1		S_{th}=0.2	
	H2S2D	ECM	H2S2D	ECM	H2S2D	ECM
500	8	5	3	27	1	89
1000	9	6	7	34	1	110
2000	11	8	11	48	2	144
3000	15	9	20	52	7	169
4000	19	12	22	56	9	181
5000	20	14	26	63	13	193
6000	20	14	29	66	18	213
7000	24	15	30	68	20	225
8000	27	16	32	70	22	226

Since all data items whose clustering are delayed are placed into the candidate pool, there can be performance concerns in regard to the growing number of candidates. But, as the experiment shows, the ratio of items in the candidate pool to the number of processed items (“Items” in Tables 5.1 and 5.3) is gradually decreasing and reaches a certain level (Figure 5.4). Hence, if a user’s context does not change we can expect all clusters to reach a certain unchangeable state in terms of their numbers. However, if the growing number of candidate items when clustering with high similarity thresholds may cause performance problems, an aging algorithm can be introduced to remove a part of the items.

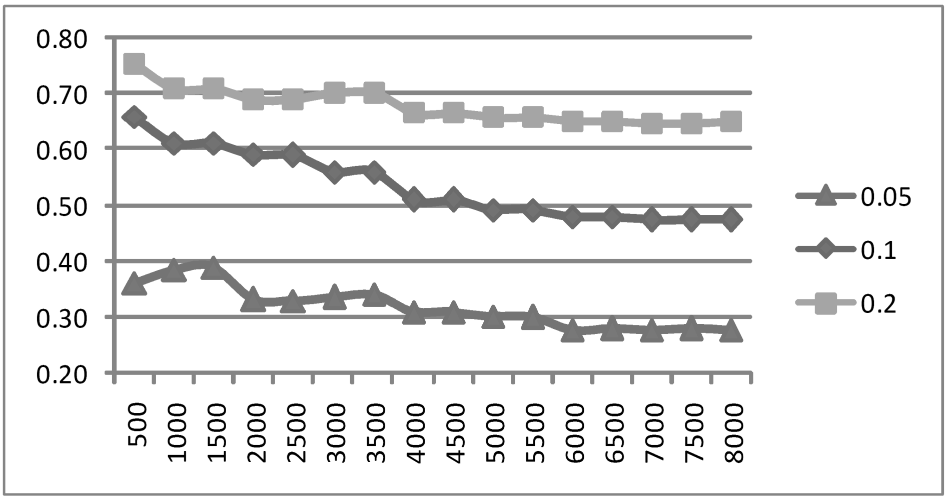


Figure 5.4 Ratio of candidate items to the number of processed items (20 newsgroup)

5.4 Summary

Although the proposed method suffers from some of the common problems of incremental clustering algorithms, such as sensitivity to the order of input data and information loss due to the characteristics of cluster abstraction model, and does not ensure convergence, it is a reasonable approach, as shown by the experiments, to infer user interests in online and incremental fashion. It requires a careful selection of ‘vigilance’ parameter – threshold for each particular data set. However, it does not perform cluster centre re-definition and thus performs fast. Furthermore, it ‘promotes’ clustering suspension of data that do not have enough evidence for cluster creation, which is, in its turn, is obtained from similarity assumptions of uniform relevance feedback, thus mitigating sensitivity to the order of input data of the method and keeping items that are potential candidates for concept creation out of clusters’ scope until such an evidence is available and they are recognised as a user’s interests.

The method is designed to be used in information seeking tasks, but it can be applied to other tasks where relevance feedback is used for user profiling and the assumptions described in Section 5.1 are satisfied. Relevance feedback can be taken in its wide sense as any uniform user activity in the information system that is considered to have some relevance degree to user interests, tasks and activities. Therefore, eventually, a number of feedback items are collected and many of them do not necessarily reflect a user’s interests. In order to distinguish the items representative of the user’s interests, the proposed method uses high similarity of sequential data from the input stream (e.g., clickthrough data in information seeking tasks) for creation of new clusters (concept inference) and updates them incrementally with newly incoming highly relevant feedback items.

6 On Sharing and Search

The user-centeredness and collaborative spirit of the proposed approach reveals through information sharing and search we present in this chapter. Here we show how the dynamics of interest-change-driven user profile, its contextual information and its construction process are employed for contribution co-evaluation and search. A reader will be able to see the central role of user profiles we described in Chapter 4 and understand how information search and sharing are done in BESS.

6.1 Contribution Co-evaluation

One of the main characteristics of the proposed framework is community-empowered user-centeredness realised through users' contributions. Together with implicit relevance feedback they are used for the users' interest inference and formation of dynamically changing user profiles (UPs). In contrast to implicit feedback, which does not give a clear understanding of a user's interests before it is analysed, contributions are clear indicators of the user's interests and can be used for promotion of contributed documents with regard to his/her expertise to all community members of the information sharing and search system. In BESS, contributions and the user's expertise drawn from his/her user profile are used for co-evaluation of information resources the search is done on.

In order to 'promote' documents contributed by a particular user *a*, first of all, his/her user profile, and specifically concepts in it, has to be known. Clearly, not all concepts can be used for various reasons – an involvement of too many concepts into the computation process is costly, or many concepts can be considered unsuitable in terms of the current user context and situation. Therefore, we consider concepts of the short-term layer of UP as a highly dynamic and, at the same time, rather consistent layer reflecting most of the recent and significant concepts of user *a*. As we discussed in Chapter 4, a change of the concepts in terms of their ranking in the short-term profile layer signifies a change of user interests and emergence of a new model of user *a*. The short-time profile layer generation process serves as an important factor for feedback value definition mechanism. In other words, its change-driven nature helps to determine the period of time for amount of data used to evaluate user contributions – to determine the concepts that will be used for contribution assessments.

The layer generation process is illustrated in Figure 6.1. Note that concept re-ranking is done according to the criteria described in Chapter 4.

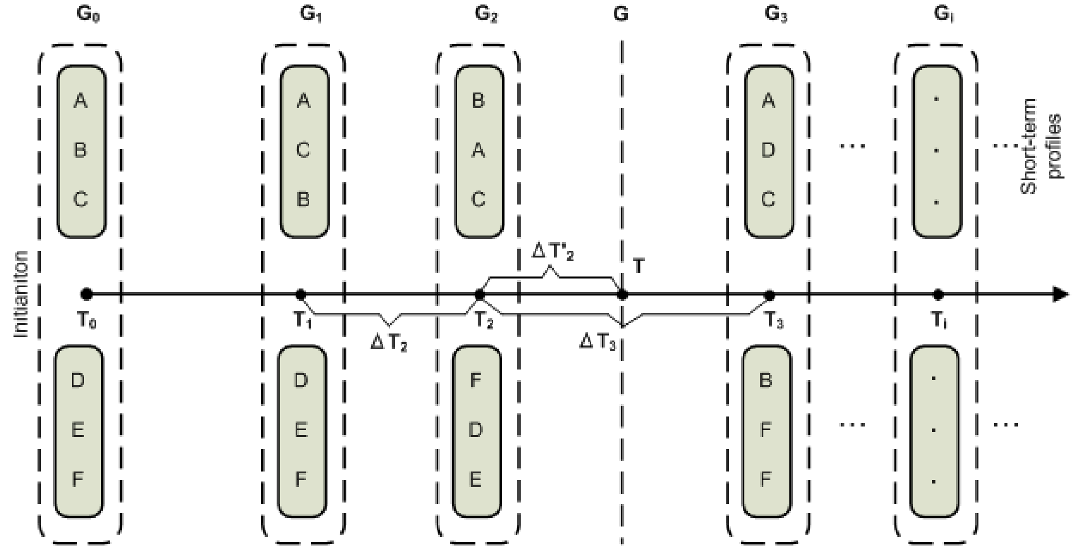


Figure 6.1 Short-term profile dynamics

i – number of interest changes. G_i – group of dynamic profile concepts. *A*, *B*, *C*, *D*, *E*, *F* – interest concepts.

$$\{A, B, ..., F\} \in G_i$$

T_i – moment of concept change in profiles. *G* and *T* – current concept group and moment respectively.

Then, the period to be used for contribution assessments can be defined as

$$P(T_s, \Delta T) = \begin{cases} P(T_{i-1}, T - T_{i-1}), \; G = G_i \\ P(T_i, T - T_i), \; G = G_{i+1} \neq G_i \end{cases} \tag{6.1}$$

where *i* is the number of interest changes, T_s – start time of evaluation period, ΔT – the interval to the current position on profile generation time line, and *G* and *T* are the current concept group and moment respectively.

For instance, we define the evaluation period with the last interest changing point at T_2 as

$$P(T_s, \Delta T) = \begin{cases} P(T_1, T - T_1), G = G_2 \\ P(T_2, T_3 - T_2), G = G_3 \neq G_2 \end{cases} = \begin{cases} P(T_1, \Delta T_2 + \Delta T_2'), G = G_2 \\ P(T_2, \Delta T_3), G = G_3 \neq G_2 \end{cases} \tag{6.2}$$

After the period is defined, a contribution can be evaluated according to the following criteria with regard to the available concepts of this particular user and all concept space of the system – concepts of all users interacting with the system:

- contribution activeness;
- contribution popularity.

Contribution activeness (*CA*) is defined as the ratio of the contribution number of user *a* to the number of the most active contributor with regard to concept *i*. To rephrase, it is a indicator of how active user *a* is in a particular area of expertise compared to the most active

$$CA_{aCi} = \frac{|C_{ai}|}{\max|C_{ui}|} \tag{6.3}$$

Contribution popularity (*CP*) is defined as the ratio of the sum of all contributions to concept *i* to all contributions done in the framework.

$$CP_{Ci} = \frac{\sum_{u=1}^{U} |C_{ui}|}{\sum_{i=1}^{I} |C_i|} \tag{6.4}$$

where $C_i \in C = \{C_1, ..., C_I\}$ and $C_{ui} \in C_u = \{C_{u1}, ..., C_{uI}\}$.

Using the two criteria, the contribution of user *a* to document *d* which belongs to concept *i* is estimated as

$$cntr_{aCi} = \alpha \cdot CA_{aCi} \cdot CP_{Ci} \tag{6.5}$$

where *α* is a constant regulating the power of different contribution types – for instance, if implicit feedback is considered as a contribution, its estimated value has to be weakened because it is less indicative of the user's interests compared to explicit feedback.

Since a document often includes several concepts, we have to consider them in the estimation as

$$cntr_{aCi} = \frac{\sum_{Cai} \alpha \cdot CA_{Cai} \cdot CP_{Ci}}{|C_d|} \qquad (6.6)$$

where $|C_d|$ is the number of (top n) concepts d belongs to.

Evaluating contributions in this way we give an assessment of a user's expertise by his/her involvement degree (*CA*) and potential value the user brings to the whole community by his/her contribution (*CP*). More specifically, these criteria take into account:

- contributions by user *a* to a particular concept;
- contributions by all users to a particular concept;
- potential value of contributions with regard to a particular concept;
- activeness of the whole community with regard to a particular concept.

As the same contribution can be done by multiple users, its score is re-assessed according to the following formula

$$cntr_d = \frac{\sum_{u=1} cntr_d}{|U_d|} \qquad (6.7)$$

and stored for document ranking in retrieval process described in the next section.

6.2 Collaborative Search

In Chapter 3 we mentioned that besides subjective search the framework provides objective search functionality – searching through conventional Web search services. Objective search is the functionality each of us uses daily and, therefore, not of much interest in our discussion. On the other hand, subjective search of BESS is collaborative and includes two modes using a custom rank function as shown in Table 6.1.

Table 6.1 Subjective search modes

Mode	Normal	Focused
Ranking	Custom Rank	Custom Rank

Both modes are labelled as collaborative. Search in normal mode is collaborative in the sense it is performed on the collection of contributed and shared documents and co-evaluations of the documents are considered for search result re-ranking. Focused search is even more collaborative since it detects search space from subjective document collections of users with similar contexts.

6.2.1 Normal Subjective Search

Normal subjective search mode is the basic mode to search on subjective index data of BESS. When using the mode, all index data becomes the search target. Furthermore, in contrast to objective search where query-document match is done regardless of a user's search context, BESS provides research results ranked with regard to the user's degree of involvement in the query context using conceptual information in user profiles.

When retrieving documents from subjective index repositories, first a simple impersonal query-document match, as in objective index search, is done to find documents objectively relevant to the user's query. Then the system attempts to personalise the retrieved results by considering the user's search contexts found in UP using the following formula and rank them with maximum sim_L values at the top.

$$sim_L = \sum_{k=1}^{K} \alpha_k \cdot sim(d, l_k) \qquad (6.8)$$

where L – user profile, l_k – layer k of UP, d – document, and α_k – ratio in the mixture which is set experimentally and $\sum_{k=1}^{K} \alpha_k = 1$. Normally, a_{ss} has to be considered as the minimum and a_{ln} the maximum values, since they represent the most volatile and persistent concepts respectively.

6.2.2 Subjective Concept-directed Vertical Search

Another method to search in BESS is concept-based *subjective concept-directed vertical search*, or *focused search*. It is realised by matching of concepts of the individual user profile with concepts of other users when retrieving through the subjective search engine. This operation forms the target search document space for the query by finding users with similar interests and reaching their subjective index repositories. After the search document space is determined relevant documents are retrieved by comparing query and document vector similarity.

By forming the search document space, we change the focus of search, similarly to what occurs in vertical search engines, but automatically detecting users with similar information seeking contexts. In this way we ensure the search will be done on highly selective documents evaluated by the users with similar interests taking into account their expertise, or the degree of their involvement into the topic. That is, by drawing upon community expertise and similarity of information needs, we perform *subjective concept-directed vertical search*.

Figure 6.2 can give a better picture on how the search is done in the system, presenting it on the level of concepts user profile consists of.

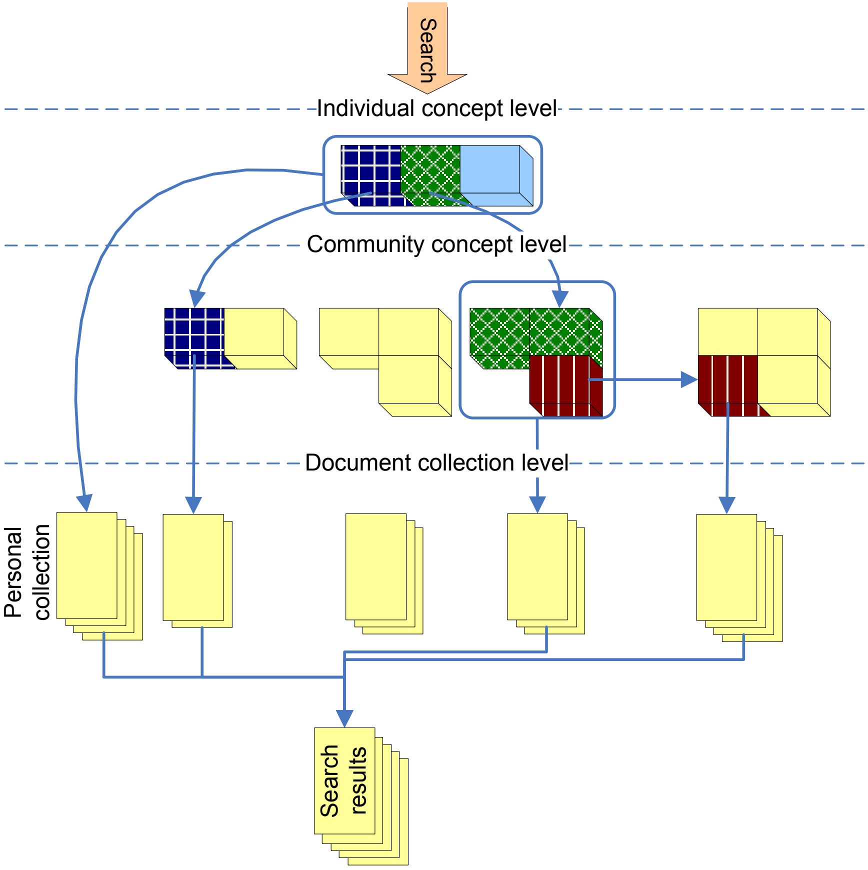


Figure 6.2 Search by concepts

Users with similar profiles, and hence their subjective index collections, are found by comparing reference points of concepts in each layer of UP one by one. The following formula is used for the estimation.

$$\sum_{k=1}^{K} sim_L(l_{ak}, l_{bk}) = \sum_{k=1}^{K} \frac{\sum_{j=1}^{J}\sum_{i=1}^{I} sim(Rp_j, Rp_i)}{JI} \alpha_k \tag{6.9}$$

When *n* top users with similar contexts are found, their collections form the search document space. Then, the search is done as in normal subjective search mode.

6.2.3 Re-ranking with Contribution Assessments

Regardless a search mode, re-ranking of retrieved document is done by listing them with maximum sim_L values at the top. But this is not the only value considered for re-ranking. Another value is a contribution assessment value $cntr_d$ described in Section 6.1. Re-ranking is done by sorting according to the following rules:

Table 6.2 Re-ranking example

Document	sim_L	$cntr_d$
d$_1$	0.69	0.65
d$_2$	0.81	0.59
d$_3$	0.56	0.35
d$_4$	0.42	0.21
d$_5$	0.3	0.52
d$_6$	0.3	0.51

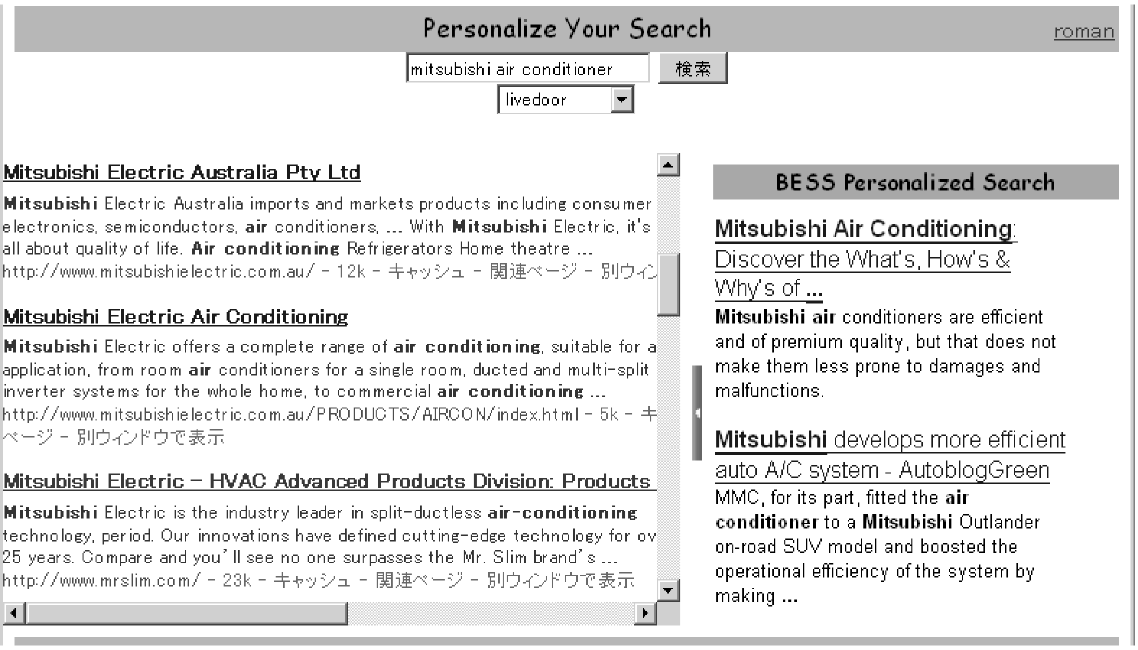
1) documents with $cntr_d$ values exceeding a certain contribution threshold and similarity threshold (0.5

and 0.6 respectively in Table 6.2) are considered as high priority in ranking and sorted by $cntr_d$ field;

2) all other documents are sorted by sim_L field.

6.3 Search Scenario

To give a better idea how, for instance, subjective concept-directed vertical search works, let us consider several users of the system having the following rather abstract concepts generated from their contributions and search and post-search behaviours.



Mitsubishi Electric Australia Pty Ltd
Mitsubishi Electric Australia imports and markets products including consumer electronics, semiconductors, **air** conditioners, ... With **Mitsubishi** Electric, it's all about quality of life. **Air conditioning** Refrigerators Home theatre ...
http://www.mitsubishielectric.com.au/ - 12k - キャッシュ - 関連ページ - 別ウィン

Mitsubishi Electric Air Conditioning
Mitsubishi Electric offers a complete range of **air conditioning**, suitable for a application, from room **air** conditioners for a single room, ducted and multi-split inverter systems for the whole home, to commercial **air conditioning** ...
http://www.mitsubishielectric.com.au/PRODUCTS/AIRCON/index.html - 5k - キ ページ - 別ウィンドウで表示

Mitsubishi Electric - HVAC Advanced Products Division: Products
Mitsubishi Electric is the industry leader in split-ductless **air-conditioning** technology, period. Our innovations have defined cutting-edge technology for ov 25 years. Compare and you' ll see no one surpasses the Mr. Slim brand' s ...
http://www.mrslim.com/ - 23k - キャッシュ - 関連ページ - 別ウィンドウで表示

Mitsubishi Air Conditioning: Discover the What's, How's & Why's of ...
Mitsubishi air conditioners are efficient and of premium quality, but that does not make them less prone to damages and malfunctions.

Mitsubishi develops more efficient auto A/C system - AutoblogGreen
MMC, for its part, fitted the **air conditioner** to a **Mitsubishi** Outlander on-road SUV model and boosted the operational efficiency of the system by making ...

Figure 6.3 Personalised Web search results

- User A: cars, Japan politics, golf;
- User B: action movies, Disney, blogging;
- User C: online shopping, cars, travel;
- User D: computer games, messenger.

Let us assume that *user A* searches for pages explaining air-conditioning in Mitsubishi cars and submits “Mitsubishi air conditioner”. The user does not know that Mitsubishi Electric produces air-conditioning systems and have Mitsubishi cars on his/her mind. As a result of the query, conventional search engine ranks high the documents covering room air-conditioning systems, whereas the proposed system will restrict results to document collections of users with similar interests and rank them according to user evaluations if available. In the case we are considering, user A’s interests match those of *user C* and the query retrieves documents explaining air-conditioning in Mitsubishi cars (Figure 6.3).

7 Conclusions

7.1 Summary

With the exponential growth of information in the Internet and, as a result, failures to manage and process it effectively and efficiently, solutions beyond conventional information organisation and filtering are being sought. Realising the subjective nature of information overload and witnessing the fast proliferation of user-generated media, academia and business turn to the human more and more often today – human problems are proposed to be solved directly by humans, rather than being mediated by information systems. Algorithms and systems to facilitate information access and acquisition are still very important and their role cannot be diminished, however they have to get better understanding whom they are used by in order to do their job with the expected efficacy and efficiency – that is to say, they have to be *user-centric*. Probably having the complete knowledge of each particular individual will suffice to call a system user-centric, but, in our understanding, such a definition of user-centrism is one-sided and lacks a concept of community as an important part of human context. An approach considering an individual user only will not benefit from "wisdom of crowds" (Surowiecki, 2005) of other users, loose much predictive power it can draw upon other users' experiences, and will not be able to collect all contextual information necessary for complete understanding of the user. Thus, an essential part of user-centrism is considering a user not only in his/her individual scope, but expanding it to the user's community participation quintessence.

Recognising this, we designed and implemented BESS (BEtter Search and Sharing) as a highly collaborative information search and sharing system. In addition to learning individual interests of each particular user, BESS harnesses collective knowledge of its users who share their personal experiences and benefit from experiences of others. We made an endeavour to develop a holistic approach from how to harnesses relevance feedback (both explicit and implicit) from users in order to estimate their interests, construct user profiles reflecting those interests to applying them for information acquisition in online collaborative information seeking context. To emphasise the collaborative nature of relevance feedback submitted by users explicitly, it is called a *contribution* in our research. In comparison to implicit feedback, contributions are better indicators of a user's interests and therefore have bigger values for shared information assessment.

The central part in the system is allotted to user profiles claimed to contain fragmentary user contexts and used for contribution evaluation and search. Each user profile consists of a set of concepts representing user interests and inferred with *High Similarity Sequence Data-Driven clustering* method thoroughly discussed in Chapter 5. The method is easy to implement, and, as the experimental results show, it produces concepts of reasonably good quality fast and in online and incremental fashion. To ensure that a user profile is always updated and reflect current, recent and long-term interests, it is designed multi-layered and changing with change of the user's interests as a result of estimations done with *interest-change-driven profile construction mechanism*. The mechanism adopts *recency*, *frequency* and *persistency* as the three important criteria for profile construction and update.

As mentioned, user profiles are important for both relevance feedback evaluation and search. And they are especially important to realise a unique mechanism of *subjective concept-directed vertical search*, or *focused search*. It is realised by matching of concepts of the individual user profile with concepts of other users when retrieving through the subjective search engine.

Furthermore, in order to distinguish between impersonal data of conventional search systems and individualised data of the members searching and sharing in the proposed system, we introduced the notions of *subjective* and *objective* index in IR system. Such distinction aims to enrich users' experiences with both conventional and personalised search results.

7.2 Future Research Directions

In order to achieve highly user-centric experience the discussed research approach puts user profiles in the centre of search and sharing. Profiles contain concept-based information about users inferred from relevance feedback and is designed to dynamically reflect user contexts. However, contexts are of fragmentary nature, reflecting only user interests inferred primarily from textual information of feedback. Recognising the importance of contextual knowledge about a user for personalisation, new methods to define, capture and organise contextual information, extension of the architecture and algorithms to include more information from a user and assessments of effects of such inclusion on personalised experience are one of the future directions of our research.

Another direction is a closer examination of motivational factor for relevance feedback in social bookmarking and tagging, and search for means to collect more precise information about a user without a burden.

Finally, measuring satisfaction from using the search interface is important to know whether it contributes to the approach and how users utilise it.

Bibliography

Agichtein, E., Brill, E. and Dumais, S., (2006a) "Improving Web Search Ranking by Incorporating User Behavior Information", *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 19-26.

Agichtein, E., Brill, E., Dumais, S. and Ragno, R., (2006b) "Learning User Interaction Models for Predicting Web Search Result Preferences", *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 3-10.

Aknine, S, Slodzian, A. and Quenum, G. (2005) "Web personalisation for users protection : A multi-agent method", *Intelligent Techniques for Web Personalization*, Lecture Notes in Computer Science, Vol. 3169, pp. 306-323.

Almeida, R. B. and Almeida V. A. F. (2004) "A Community-Aware Search Engine", *Proceedings of the 13th International Conference on World Wide Web*, ACM Press, pp. 413-421.

Anand, S. S. and Mobasher, B. (2005) "Intelligent Techniques for Web Personalization", *Intelligent Techniques for Web Personalization*, Lecture Notes in Computer Science, Vol. 3169, pp. 1-36.

Bahrami, A., Yuan, J., Smart, P. R. and Shadbolt, N. R. (2007) 'Context-Aware Information Retrieval for Enhanced Situation Awareness', *Military Communications Conference (MILCOM 2007)*, pp.1 – 6.

Baraglia, R. and Silvestri, F. (2007) "Dynamic personalization of web sites without user intervention", *Communications of the ACM*, Vol. 50, No. 2, pp. 63-67.

Barbu, C. and Simina, M. (2003) 'A probabilistic information filtering using the profile dynamics', *IEEE International Conference on Systems, Man and Cybernetics,* Vol. 5, pp.4595 – 4600.

Beaudoin, C. E. (2008) "Explaining the Relationship between Internet Use and Interpersonal Trust: Taking into Account Motivation and Information Overload", *Journal of Computer-Mediated Communication*, Volume 13 Issue 3, pp. 550-568.

Carrillo-Ramos, A., Villanova-Oliver, M., Gensel, J. and Martin, H. (2007) "Profiling Nomadic Users Considering Preferences and Context of Use", *On the Move to Meaningful Internet Systems 2007: OTM 2007 Workshops*, Lecture Notes in Computer Science, Vol. 4805, pp.457-466.

Case, D. O. (2002) *Looking for Information: A Survey of Research on Information Seeking, Needs, and Behavior*, Amsterdam: Academic Press.

Çetintemel, U., Franklin, M.J., and Giles, C.L. (2000) "Self-adaptive user profiles for large-scale data delivery", *Proceedings of the 16th International Conference on Data Engineering*, pp. 622-642.

Charikar, M., Chekuri, C., Feder, T. and Motwani R. (1997) "Incremental Clustering and Dynamic Information Retrieval", *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*, pp. 626-635.

Chellappa, R. K. and Sin, R. G. (2005) "Personalization versus Privacy: An Empirical Examination of the Online Consumer's Dilemma", *Information Technology and Management*, Vol. 6, No. 2-3, pp. 181-202.

Chen, Y. C., Shang, R. A., and Kao, C. Y. (2007) "The Effects of Information Overload on the Outcomes of On-Line Consumption Behavior", *Proceedings of International Conference on Wireless Communications, Networking and Mobile Computing, 2007 (WiCom 2007)*, pp. 3791-3794.

Church, K., Keane, M. T. and Smyth, B. (2005) "An Evaluation of Gisting in Mobile Search", D.E. Losada and J.M. Fern´andez-Luna (Eds.): *ECIR 2005*, LNCS 3408, pp. 546-548.

Claypool, M., Brown, D., Le, P., and Waseda, M. (2001) "Inferring User Interest", *IEEE Internet Computing*, Vol. 5, No. 6, pp. 32-39.

comScore (2006) "More than Half of MySpace Visitors are Now Age 35 or Older, as the Site's Demographic Composition Continues to Shift", *Press Release*, http://www.comscore.com/ press/release.asp?press=1019.

De Luca, E. W. and Nürnberger, A. (2005) "Supporting information retrieval on mobile devices", *Proceedings of the 7th international conference on Human computer interaction with mobile devices & services*, pp. 347-348.

eMarketer (2008) "Can User-Generated Content Generate Revenue?" http://www.emarketer.com/Article.aspx?id=1006190&src=article1_newsltr (*Last accessed on Oct. 11th 2008*).

Fellbaum, C. (1998) *WordNet: An Electronic Lexical Database*, MIT Press.

Ferscha, A., Hechinger, M., Riener, A., Schmitzberger, H., Franz, M., dos Santos Rocha, M. and Zeidler, A. (2006) 'Context-aware profiles', *International Conference on Autonomic and Autonomous Systems*, p.48.

Flanagin, A. J. and Metzger, M. J. (2008) "The credibility of volunteered geographic information", *GeoJournal (Springer)*, Vol. 72, No. 3-4, pp. 137-148.

Fu, X. (2007) "Evaluating Sources of Implicit Feedback in Web Searches", *Proceedings of ACM Recommender Systems 2007 (RecSys '07)*, pp.191-194.

Gargi, U. (2005) 'Information Navigation Profiles for Mediation and Adaptation', *Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05)*, pp.515-520.

Golbeck, J. (2008) "Weaving a Web of Trust", *Science*, Vol. 321, No. 5896, pp. 1640-1641.

Greer, T. H., and Murtaza, M. B. (2003) "Web Personalization: The Impact of Perceived Innovation Characteristics on the Intention to Use Personalization", *Journal of Computer Information Systems*, Vol. 43, No.3, pp. 50-55.

Grisé, M.-L. and Gallupe, R.B. (1999) "Information Overload: Addressing the Productivity Paradox in Face-to-Face

Electronic Meetings", *Journal of Management Information Systems*, Vol. 16, No. 3, pp. 157-185.

Guida, A. and Tardieu, H. (2005) "Is personalisation a way to operationalise long-term working memory?" *Current Psychology Letters*, Vol. 1, pp. 1-17.

Hall, A. and Walton, G. (2004) "Information overload within the health care system: a literature review", *Health Information & Libraries Journal*, Vol. 21, No. 2, pp. 102-108.

Hammouda, K. M. and Kamel, M. S. (2003) "Incremental Document Clustering Using Cluster Similarity Histograms", *2003 IEEE/WIC International Conference on Web Intelligence (WI'03)*, pp.597–601.

Hansen, M. T., Nohria, N. and Tierney, T. (1999) "What's Your Strategy for Managing Knowledge?" *Harvard Business Review*, pp. 106-116.

Harper, D. J. and Kelly, D. (2006) "Contextual Relevance Feedback", *Proceedings of the 1st International Conference on Information Interaction in Context*, pp. 129-137.

Hartley, D. (2007) "Personalisation: the emerging 'revised' code of education?" *Oxford Review of Education*, Vol. 33, No. 5, pp. 629-642.

Himma, K. E. (2007) "The concept of information overload: A preliminary step in understanding the nature of a harmful information-related condition", *Ethics and Information Technology*, Vol. 9, No. 4, pp. 259-272.

Hunter. G. L. (2004) "Information Overload: Guidance for Identifying When Information Becomes Detrimental to Sales Force Performance", *Journal of Personal Selling & Sales Management*, Vol. 24, No. 2, pp. 91-100.

Hunter, G. L. and Goebel, D. J. (2008) "Salespersons' Information Overload: Scale Development and Validation and its Relationship to Salesperson Job Satisfaction and Performance", *Journal of Personal Selling and Sales Management*, Vol. 28, No. 1, pp. 21-35.

Ingwersen, P. and Jarvelin, K. (2005) *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*, Springer-Verlag New York, Inc., Secaucus, NJ.

Jacoby, J., Speller, D. E. and Kohn, C. A. (1974) "Brand choice behavior as a function of information load", *Journal of Marketing Research*, pp. 63-69.

Janssen, R. and de Poot, H. (2006) "Information overload: Why some people seem to suffer more than others", *Proceedings of the 4th Nordic Conference on Human-Computer Interaction*, pp. 397-400.

Jorstad, I. and Do van Thanh (2007) "A Framework and Tool for Personalisation of Mobile Services Using Semantic Web", *Proceedings of 2007 International Conference on Mobile Data Management*, pp. 402-406.

JupiterResearch (2008) "US Online Travel Consumer Survey, 2008", http://www.jupiterresearch.com/.

Kagolovsky, Y. and Moehr, J. R. (2003) "Current Status of the Evaluation of Information Retrieval", *Journal of*

Medical Systems, Vol. 27, No. 5, pp. 409-424.

Karpouzis, K., Moschovitis, G., Ntalianis, K., Ioannou, S. and Kollias, S. (2004) 'Web Access to Large Audiovisual Assets Based on User Preferences', *Multimedia Tools and Applications*, Vol. 22, No. 3, pp.215-234.

Kasabov, N. (2007) *Evolving Connectionist Systems: The Knowledge Engineering Approach (Evolving Connectionist Systems)*, Springer London.

Kelly, D. and Teevan, J. (2003) "Implicit Feedback for Inferring User Preference: a Bibliography", *ACM SIGIR Forum*, Vol. 37, No. 2, pp.18–28.

Khopkar, Y., Spink, A., Giles, C. L., Shah, P. and Debnath, S. (2003) "Search engine personalization: An exploratory study", *First Monday* (*Peer-Reviewed Journal on the Interne)t*, http://www.firstmonday.org/issues/issue8_7/khopkar/ (*Last accessed on Oct. 22, 2008*).

Kim, K., Lustria, M. L. A., Burke. D., and Kwon, N. (2007) "Predictors of cancer information overload: Findings from a national survey", *Information Research*, Vol. 12, No. 4.

Klausegger, C., Sinkovics, R. R., and Zou, H. J. (2007) "Information overload: a cross-national investigation of influence factors and effects", *Marketing Intelligence & Planning*, Vol. 25, No. 7, pp. 691-718.

Kobsa (2007) "Privacy-Enhanced Personalization", *Communications of the ACM*, Vo. 50, No. 8, pp. 24-33.

Koutrika, G. and Ioannidis, Y. (2005) "A Unified User-Profile Framework for Query Disambiguation and Personalization", *Proceedings of Workshop on New Technologies for Personalized Information Access (PIA 2005)*, pp. 44-53.

Landay, J. A. and Kaufmann, T. R. (1993) "User Interface Issues in Mobile Computing", *Proceedings of the Fourth Workshop on Workstation Operating Systems*, pp. 40-47.

Levy D. M. (2005) "To Grow in Wisdom: Vannevar Bush, Information Overload, and the Life of Leisure", *Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital libraries*, pp. 281-286.

Ligon, G. L., Balachandran, M. B., and Sharma, D. (2006) "Personalisation of Web Search: An Agent Based Approach", *Knowledge-Based Intelligent Information and Engineering Systems*, Lecture Notes in Computer Science, Vol. 4253, pp. 1192-1200.

Limbu, D. K., Connor, A., Pears, R. and MacDonell, S. (2006) "Contextual Relevance Feedback in Web Information Retrieval", *Proceedings of the 1st International Conference on Information Interaction in Context*, pp. 138-143.

McKenzie B. and Cockburn, A. (2001) "An empirical analysis of web page revisitation", *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, pp. 128-137.

Miller, G. A. (1956), "The magical number seven, plus or minus two: some limits on our capacity for processing information", *Psychological Review*, Vol. 63, pp. 81-97.

Mirkin, B. (2005) *Clustering for Data Mining: A Data Recovery Approach*, Chapman & Hall/CRC.

Mitra, S. (2007) "Web 3.0 = (4C + P + VS)", http://www.sramanamitra.com/2007/02/14/web-30-4c-p-vs/ (*Last accessed on Oct. 22, 2008*).

Mulder, I., de Poot, H., Verwij, C., Janssen, R. and Bijlsma, M. (2006) "An information overload study: using design methods for understanding", *Proceedings of the 2006 Australasian Computer-Human Interaction Conference (OZCHI 2006)*, pp. 245-252.

Nichols, D. M. (1997) "Implicit ratings and filtering", *Proceedings of the 5th DELOS Workshop on Filtering and Collaborative Filtering*, pp. 31-36.

Nielsen//NetRatings (2006) "User-Generated Content Drives Half of U.S. Top 10 Fastest Growing Web Brands, According to Nielsen//NetRatings", http://www.nielsen-netratings.com/pr/PR_060810.PDF (*Last accessed on Oct. 11th 2008*).

Niu, W. T. and Kay, J. (2008) "Pervasive Personalisation of Location Information: Personalised Context Ontology", *Adaptive Hypermedia and Adaptive Web-Based Systems*, Lecture Notes in Computer Science, Vol. 5149, pp. 143-152.

Noll, M. G. and Meinel, C. (2008) "Web Search Personalization Via Social Bookmarking and Tagging", *The Semantic Web*, Lecture Notes in Computer Science, Vol. 4825, pp. 367-380.

Nov, O. (2007) "What Motivates Wikipedians?" *Communications of the ACM*, Vol. 50, No. 11, pp. 60-64.

O'Brien, R. (2007) "The next thing after 2.0!" http://www.outofrhythm.com/2007/04/14/the-next-thing-after-20/ (*Last accessed on Oct. 22, 2008*).

Obrist, M., Beck, E., Kepplinger, S., Bernhaupt, R., and Tscheligi, M. (2008) "Local Communities: Back to Life (Live) through IPTV", *Changing Television Environments*, Lecture Notes in Computer Science, Vol. 5066, pp.148-157.

Office of the Privacy Commissioner of Canada, "The Personal Information Protection and Electronic Documents Act", http://www.privcom.gc.ca/legislation/02_06_01_e.asp (*Last accessed on Oct. 23, 2008*).

Oulasvirta, A. and Blom, J. (2008) "Motivations in personalisation behaviour", *Interacting with Computers*, Vol. 20, No. 1, pp. 1-16.

Pennington, R. and Tuttle, B. (2007) "The Effects of Information Overload on Software Project Risk Assessment", *Decision Sciences*, Vol. 38, No. 3, pp. 489-526.

Pereira, R. and da Silva, S. R. P. (2008) "The Use of Cognitive Authority for Information Retrieval in Folksonomy

Based Systems", *Proceedings of the Eighth International Conference on Web Engineering (ICWE '08)*, pp. 325-331.

Pitkow, J., Schütze, H., Cass, T., Cooley, R., Turnbull, D., Edmonds, A., Adar, E. and Breuel T. (2002) "Personalized search", *Communications of the ACM*, Vol. 45, No. 9, pp. 50-55.

Potamias, G., Koumakis, L. and Moustakis, V. S. (2005) 'Enhancing Web Based Services by Coupling Document Classification with User Profile', *The International Conference on Computer as a Tool, EUROCON 2005,* pp.205-208.

Privacy International, http://www.privacyinternational.org/index.shtml (*Last accessed on Oct. 21, 2008*).

Reuters-21578 collection, http://www.daviddlewis.com/resources/testcollections/reuters21578.

Reuters, Ltd. (1996) "Dying for information: an investigation into the effects of information overload in the USA and worldwide", London: Reuters Limited.

Ruthven, I. and Lalmas, M (2003) "A survey on the use of relevance feedback for information access systems", *The Knowledge Engineering Review*, Vol. 18, No. 2, pp.95-145.

Santally, M. I. and Alain, S. (2006) "Personalisation in Web-Based Learning Environments", *International Journal of Distance Education Technologies*, Vol. 4, No. 4, pp. 15-35.

Sellers, P. (2006) "MySpace cowboys", http://money.cnn.com/magazines/fortune/fortune_archive/2006/09/04/8384727/index.htm.

Semeraro, G., Degemmis, M., Lops, P. and Palmisano I. (2005a) "WordNet-based User Profiles for Semantic Personalization", *Proceedings of Workshop on New Technologies for Personalized Information Access (PIA 2005)*, pp. 74-83.

Semeraro, G., Lops, P. and Degemmis, M. (2005b) 'Personalization for the Web: Learning User Preferences from Text', in Hemmje, M., Niederée, C. and Risse, T. (Eds.): *From Integrated Publication and Information Systems to Virtual Information and Knowledge Environments*, Lecture Notes in Computer Science, Vol. 3379, pp.162-172.

Sendín, M., Lorés, J., Montero, F., López, V. and González P. (2003) "User Interfaces: A Proposal for Automatic Adaptation", *Web Engineering*, Lecture Notes in Computer Science, Vol. 2722, pp. 263-266.

Shen, X., Tan, B. and Zhai C. X. (2005a) "Context-sensitive information retrieval using implicit feedback", *Proceedings of the ACM Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 43-50.

Shen, X., Tan, B. and Zhai C. X. (2005b) "UCAIR: Capturing and Exploiting Context for Personalized Search", *Proceedings of the ACM Conference on Research and Development in Information Retrieval - Information*

Retrieval in Context Workshop (IRiX'2005), Salvador, Brazil.

Shtykh, R. Y., Chen, J. and Jin, Q. (2008) "Slide-Film Interface: Overcoming Small Screen Limitations in Mobile Web Search", C. MacDonald et al. (Eds.): *ECIR 2008*, LNCS 4956, pp. 622-626.

Shtykh, R.Y. and Q. Jin (2006) "Design of Bookmark-Based Information Space to Support Exploration and Rediscovery", *Proc. CIT2006*, p.30.

Shtykh, R.Y. and Q. Jin (2008a) "Improving Mobile Web Search Experience with Slide-Film Interface", *Proc. SITIS2008/KARE2008 (First International Workshop on Knowledge Acquisition, Reuse and Evaluation, in conjunction with the Fourth IEEE International Conference on Signal-Image Technology & Internet Based Systems)*, to appear.

Shtykh, R. Y. and Jin, Q. (2008b) "Harnessing user contributions and dynamic profiling to better satisfy individual information search needs", *Int. J. Web and Grid Services*, Vol. 4, No. 1, pp. 63-79.

Shtykh, R. Y. and Jin, Q. (2008c) "Capturing User Contexts: Dynamic Profiling for Information Seeking Tasks", *Proceedings of I-CENTRIC 2008 (International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services)*.

Shtykh, R.Y., Jin, Q., Nakadate, S., Kandou, N., Hayata, T. and Ma, J. (2008) "Mobile SNS from the Perspective of Human Self-Extension", *Handbook of Research on Mobile Multimedia (Second Edition)*, Chapter XLVIII, IGI Global.

Shtykh, R., Zhang, G., and Jin, Q. (2005) "Peer-to-Peer Solution to Support Group Collaboration and Information Sharing", *International Journal of Pervasive Computing and Communications (Troubador Publishing)*, Vol. 1, No. 3, pp. 187-197.

Sieg, A., Mobasher, B., and Burke, R. (2004) "Inferring User's Information Context: Integrating User Profiles and Concept Hierarchies", *presented at the 2004 Meeting of the International Federation of Classification Societies*.

Sifry, D. (2007) "The State of the Live Web, April 2007", http://technorati.com/weblog/2007/04/328.html. (*Last accessed on 10/07/2008)*.

Smeaton, A. F. and Callan, J. (2005) "Personalisation and recommender systems in digital libraries", *International Journal on Digital Libraries*, Vol. 5, No. 4, pp. 299-308.

Spink A. (Ed.) and Cole C. (Ed.) (2005) *New Directions in Cognitive Information Retrieval*, Springer.

Stickel, C., Holzinger, A., and Ebner, M. (2008) "Useful Oblivion Versus Information Overload in e-Learning: Examples in the context of Wiki Systems", *Proceedings of the ITI 2008 30th International Conference on Information Technology Interfaces*, pp. 171-176.

Sugiyama, K., Hatano, K. and Yoshikawa M. (2004) “Adaptive web search based on user profile constructed without any effort from users”, *Proceedings of 13th international conference on World Wide Web*, ACM Press, pp. 675-684.

Surowiecki, J. (2005) *The Wisdom of Crowds*, Anchor; Reprint edition.

Sweeney, S. and Crestani, F. (2006) “Effective search results summary size and device screen size: is there a relationship?” *Information Processing and Management*, Vol. 42, Vol. 4, pp. 1056-1074.

U, R. and Varma, V. (2007) ‘A Novel Approach for Re-Ranking of Search results using Collaborative Filtering’, *International Conference on Computing: Theory and Applications*, pp.491-496.

Wang, J., Pouwelse, J., Fokker, J., de Vries, A. P. and Reinders, M. J. T. (2008) “Personalization on a peer-to-peer television system”, *Multimedia Tools and Applications*, Vol. 36, pp. 89-113.

Wikipedia, “User-generated content”, http://en.wikipedia.org/wiki/User-generated_content (*Last accessed on Oct. 11th, 2008*).

Ypma, A., de Vries, B. and Geurts, J. (2006) “Robust Volume Control Personalisation from on-Line Preference Feedback”, *Proceedings of the 2006 16th IEEE Signal Processing Society Workshop on Machine Learning for Signal Processing*, pp. 441-446.

Zigoris, P. and Zang Y. (2006) “Bayesian Adaptive User Profiling with Explicit and Implicit Feedback”, *Conference on Information and Knowledge Management (CIKM’06)*, pp.397-404.

Websites

Begun, http://begun.ru/ (*Last accessed on Oct. 23, 2008*).

Fluther, http://www.fluther.com/ (*Last accessed on Oct. 22, 2008*).

Google Custom Search Engine, http://google.com/coop/cse/ (*Last accessed on Oct. 15th, 2008*).

Google Desktop Search, http://desktop.google.com (*Last accessed on Oct. 15th, 2008*).

Java Servlet Technology, http://java.sun.com/products/servlet/ (*Last accessed on Oct. 15th, 2008*).

JavaServer Pages Technology, http://java.sun.com/products/jsp/ (*Last accessed on Oct. 15th, 2008*).

Microsoft AdCenter Labs, http://adlab.msn.com/ (*Last accessed on Oct. 23, 2008*).

Mitsubishi Electric, http://global.mitsubishielectric.com/

myAOL, http://my.aol.com/ (*Last accessed on Oct. 22, 2008*).

Rollyo, http://www.rollyo.com (*Last accessed on Oct. 15th, 2008*).

Spring Framework, http://www.springframework.org (*Last accessed on Oct. 15th, 2008*).

Swicki, http://www.eurekster.com (*Last accessed on Oct. 15th, 2008*).

Wikio, http://www.wikio.com/ (*Last accessed on Oct. 22, 2008*).

Windows Search, http://www.microsoft.com/windows/products/winfamily/desktopsearch/ (*Last accessed on Oct. 15th, 2008*).