Personalised information retrieval: survey and classification

M. Rami Ghorab, Dong Zhou, Alexander O'Connor, and Vincent Wade
Centre for Next Generation Localisation Knowledge & Data Engineering Group,
Trinity College Dublin, Dublin 2, Ireland.
{ghorabm, dong.zhou, alex.oconnor, vincent.wade}@scss.tcd.ie

Abstract. Information Retrieval (IR) systems assist users in finding information from the myriad of information resources available on the Web. A traditional characteristic of IR systems is that if different users submit the same query, the system would yield the same list of results, regardless of the user. Personalised Information Retrieval (PIR) systems take a step further to better satisfy the user's specific information needs by providing search results that are not only of relevance to the query but are also of particular relevance to the user who submitted the query. PIR has thereby attracted increasing research and commercial attention as information portals aim at achieving user loyalty by improving their performance in terms of effectiveness and user satisfaction. In order to provide a personalised service, a PIR system maintains information about the users and the history of their interactions with the system. This information is then used to adapt the users' queries or the results so that information that is more relevant to the users is retrieved and presented. This survey paper features a critical review of PIR systems, with a focus on personalised search. The survey provides an insight into the stages involved in building and evaluating PIR systems, namely: information gathering, information representation, personalisation execution, and system evaluation. Moreover, the survey provides an analysis of PIR systems with respect to the scope of personalisation addressed. The survey proposes a classification of PIR systems into three scopes: individualised systems, community-based systems, and aggregate-level systems. Based on the conducted survey, the paper concludes by highlighting challenges and future research directions in the field of PIR.

Keywords: personalisation, user modelling, user interests, information retrieval, multilingual information retrieval, adaptive hypermedia, search history, query adaptation, result adaptation, evaluation, survey.

1 Introduction

With the enormously increasing amount of information on the Web, there is a growing need for systems that offer personalised services to Web users, where information is adapted to the user's needs in terms of content and presentation (Brusilovsky, 2001, Jameson, 2008). Modelling user and usage information, whether on an individual user scope or community scope, is an essential process in personalised systems. Much research is being carried out concerning how to gather, represent, and make use of such information for providing personalised services on the Web (Gauch et al., 2007, Micarelli et al., 2007, Brusilovsky and Tasso, 2004, Brajnik et al., 1987).

Over the past decades, the area of Personalised Information Retrieval (PIR) has gained much attention in the literature (Micarelli et al., 2007, Agichtein et al., 2006a, Sugiyama et al., 2004, Brusilovsky and Tasso, 2004). Providing a personalised service to Web search users significantly helps them in satisfying their everyday information needs (Agichtein et al., 2006a, Speretta and Gauch, 2005, Teevan et al., 2005). Textual Information Retrieval systems have become wide-spread across the Web community, being used in search engines, online libraries, or local search facilities provided on numerous websites. A typical search process would involve users submitting queries, often in the form of a set of terms, to a retrieval system and receiving a ranked list of results in return. A natural characteristic of Information Retrieval (IR) systems is that if different users submit the same query, the system would yield the same list of results, regardless of the user. PIR systems, on the other hand, include the user in the equation (Brusilovsky and Tasso, 2004, Silvestri, 2010). In other words, a PIR system does not retrieve documents[1] that are just relevant to the query but ones that are also relevant to the user's interests; thus, different users may actually receive different results for the same query. This can be done by keeping track of the user's personal information and interests and then using this information to personalise the query or the presented results.

For example, assume a certain user is interested in critical reviews of works of literature (e.g. novels or plays) and submits the query "A Tale of Two Cities" to a search engine. The retrieval system will then attempt to retrieve all documents that are relevant to the query terms from the document collection. This will return many diverse documents as results for this search, such as: text excerpts from the body of the novel, information about the film that was created based on the novel, websites that offer to sell the novel or the film, critical reviews of the novel, information about the author Charles Dickens, and perhaps a number of irrelevant documents or documents that are related to another article or object that shares the same name. Therefore, users who are specifically interested in critical reviews or analysis of the literature will have to respond by either of two actions. Either they will have to sift through the many results that are not relevant to their information needs in order to find the ones that are relevant to them, or they will have to reformulate the query in order to specify their intent (e.g. submit a new query: "A Tale of Two Cities Analytic Review"). Now if the system had "known" that a particular user was interested in reviews of works of art, then it could have adapted the result list with respect to such interests. Results that represent analytic reviews about the novel would therefore be moved to the top of the ranked list where the user could more easily locate them. Furthermore, the system could adapt the original query itself, perhaps by automatically adding some terms to it, such as "critique", "criticism", "analysis", "analytic", or "review", so that more specific results could be retrieved in the first place.

A key feature of PIR systems is keeping track of the information needs of their users in order to personalise the service. Therefore, PIR systems should have a mechanism to learn about their users' search interests. The recorded search interests can then be used to tailor the users' future searches according to their inferred needs. For a PIR system to obtain user information, it could either request that users explicitly supply this information or it could implicitly gather this information in an unobtrusive manner from the users' search history[2]. Furthermore, a major concern in PIR systems is how to store and represent the gathered usage information. Some systems store this information in an individualised user model (Zhang et al., 2007, Speretta and Gauch, 2005, Pretschner and Gauch, 1999, Psarras and Jose, 2006), while other systems maintain an aggregate view of usage information (Agichtein et al., 2006b, Smyth and Balfe, 2006).

PIR systems generally pass through three stages in order to provide their personalised service (Gauch et al., 2007). The first stage is *information gathering*, where different tools and approaches are used to collect information about the users. The second stage is *information representation*, where different modelling approaches and data structures are used to represent the information that was gathered about the user. The third stage is the *implementation and execution of personalisation*, where different approaches are used to adapt the user's query or the results. This survey features a state-of-the-art review of these stages and classifies existing systems in the literature according to the various approaches exhibited in each stage. Furthermore, as evaluation is an important matter when considering functional systems, this survey also provides a review of the different methods used to evaluate PIR systems in the literature.

[1] The terms *document* and *result* are used interchangeably in this paper to denote any object in the result list retrieved in response to a query

[2] Search history entails objects that are exhibited in search logs; including queries, clickthrough data, and document snippets.

Personalised systems have been demonstrated in several areas in the literature, such as Web search (Stamou and Ntoulas, 2009, Teevan et al., 2009, Agichtein et al., 2006a), eLearning (Conlan et al., 2003, De Bra et al., 2003), and news dissemination (Katakis et al., 2009, Billsus and Pazzani, 2007). Information Filtering (IF) (Belkin and Croft, 1992, Oard, 1997) is an area that is closely associated with IR, with a number of common aspects under investigation in both. However, the key characteristic that distinguishes the two areas is that IF focuses on the continuous analysis of a stream of information (e.g. news, tweets, etc.) and the filtering of that stream based on the user's interests. In such systems the user is not required to issue a query, while IR and PIR focus on enhancing the search process which starts with the submission of a query to a search system with the aim of satisfying an information need at hand (Hanani et al., 2001). In this survey we focus on personalised search systems (which involve a query action from the user's side and then a system's response with results in return).

In personalised search systems, different sources can be exploited to obtain user information. This survey mainly focuses on systems that exploit logs of user interaction history as their main source of information (Silvestri, 2010, Jansen et al., 2008); nevertheless, for the purpose of completeness, a number of systems which exploit other sources of information in addition to logs will also be included in the survey. Furthermore, several personalisation dimensions are targeted by different personalised systems in the literature. For example, personalisation could be employed on the dimension of user's prior knowledge (Brusilovsky and Henze, 2007, Conlan et al., 2002), user's interests (Micarelli et al., 2007, Gauch et al., 2007), or user's context (Cool and Spink, 2002, Quiroga and Mostafa, 2002). This survey focuses on personalisation according to the user's search interests, which is the most common approach exhibited in PIR literature (Stamou and Ntoulas, 2009, Teevan et al., 2005, Speretta and Gauch, 2005). Inferring the users' interests from their past searches allows for personalising their future searches by providing them with more relevant search results.

A number of studies in the literature have aimed to review and classify different approaches to personalised search and the use of user models for personalisation in different application areas. In (Micarelli et al., 2007) the authors reviewed different personalisation techniques in the Web search domain. Furthermore, different types of personalised search systems were discussed. However, there was no clearly structured separation of personalisation stages. The authors in (Gauch et al., 2007) classified user models into keyword-based, semantic network-based, and concept-based user models. Moreover, they discussed implicit and explicit approaches to information gathering. However, the authors only focused on systems that made use of an individualised user model and they did not cover other systems that execute personalisation based on aggregate usage information (collective view of information from search logs). In (Kelly and Teevan, 2003), the authors provided an analytic review of implicit feedback approaches in PIR. However, the authors only covered the first stage of personalised systems, which is the information gathering stage.

The review presented in this paper mainly focuses on the area of personalised search systems and extends existing literature in the following ways:

1. A novel classification of PIR systems is presented in this paper where systems are categorised with respect to the scope on which personalisation is performed into three categories: *individualised personalisation*, *community-based personalisation*, and *aggregate-level personalisation*. Individualised personalisation is when the system's adaptive decisions are taken according to the information about each individual user as exhibited in his/her user model (Stamou and Ntoulas, 2009, Speretta and Gauch, 2005). Community-based personalisation takes a step further from individualised personalisation where the system's adaptive decisions are done in a collaborative manner (Teevan et al., 2009, Sugiyama et al., 2004). This involves systems in which a model is also constructed on a per-user basis, but where sharing of information between models can take place. Aggregate-level personalisation refers to the notion of a system that does not explicitly make use of a user model to represent each individual user; at which case personalisation is guided by aggregate usage data as exhibited in search logs (Agichtein et al., 2006b, Sun et al., 2005). This may be considered as a special case (a wider scope) of the community-based scope, but the difference is that no user model exists per se. An in-depth discussion of this classification is provided in Section 4.2.2.
2. In addition to monolingual PIR systems, this survey also covers multilingual PIR systems in the literature.
3. The survey features a review of the various sources and approaches of obtaining user and usage information, including social data which is an emergent approach in the literature.
4. A classification of user models according to their underlying data structure and the nature of their content is presented in Section 3.2.2.
5. An extensive discussion of query adaptation and result adaptation techniques in PIR is provided in Section 4. Furthermore, a novel classification of query adaptation techniques is presented in Section 4.2.3 where the techniques are divided into user-focused vs. non-user-focused (i.e. personalised techniques that involve user information in the process and non-personalised techniques that only involve information from query and document collection) and implicit vs. explicit (i.e. techniques that don't require user intervention and ones that require a specific user action).

6. This survey features a dedicated section for reviewing evaluation approaches in PIR systems (Section 6), where various evaluation techniques from the fields of Information Retrieval and Adaptive Hypermedia are discussed.

The rest of this paper is organised as follows. Section 2 discusses the information gathering stage of PIR systems where various techniques and sources are used to acquire the necessary information on which personalisation is based. Section 3 discusses the information representation stage where different data structures are used to maintain user and usage information in PIR systems. Section 4 discusses the personalisation implementation and execution stage where a variety of techniques are used for search personalisation. Section 5 discusses the different evaluation approaches and metrics used to evaluate PIR systems in terms of effectiveness and usability. Section 6 provides a general discussion and highlights challenges and future research directions in the field of PIR. Finally, conclusions are presented in Section 7.

2 Information gathering

2.1 Overview

This section of the survey is concerned with the first stage of personalisation, which is information gathering. A discussion is provided regarding the different sources and types of information on which personalisation can be based, and also regarding the different approaches of obtaining this information. The importance of discussing the information gathering stage stems from the idea that the nature of information available for a personalised system determines the way that the system can implement personalisation at later stages. The analysis is carried out over three criteria: the information gathering approach, the type of information, and the source of information. An overview of these three criteria is given below.

- **Information gathering approach**: the first criterion is the approach to gathering the information. Information can be gathered in an implicit manner where it is obtained without any extra effort from the user or in an explicit manner where the users have to explicitly supply information to the system.
- **Type of information:** the second criterion is the type of information gathered about the users and their usage behaviour when interacting with the system. User information is information collected about users themselves, such as their personal information, demographic information, or search interests. Usage information, on the other hand, is information recorded about the users' interactions with systems on the Web; for example, in the scope of Web search, this includes submitted queries, browsed pages, annotated content, bookmarked pages, tags, and personal profiles. User information is traced back to to a certain user, whereas usage information may be aggregated across many users.
- **Source of information:** the third criterion considered in this part of the analysis is the information source. Usage information can be gathered at the server-side or at the client-side. In addition, this criterion is also concerned with where the information is maintained, and highlights related privacy concerns.

The following section provides a detailed review and analysis of different systems in the literature. The analysis focuses on the information gathering stage of the surveyed systems, guided by the three criteria outlined above.

2.2 Review

2.2.1 Information gathering approach

Information can be gathered in an implicit or an explicit manner (Gauch et al., 2007). In the implicit method, information is gathered unobtrusively, without any effort from the user. This is typically the case when a system keeps track of the user's search history in terms of submitted queries and clicked results (Stamou and Ntoulas, 2009, Gao et al., 2007, Smyth and Balfe, 2006, Speretta and Gauch, 2005). This also includes processing any stored user documents or items (e.g. emails, calendar items, etc.) (Chirita et al., 2007, Agichtein et al., 2006a, Teevan et al., 2005), or harvesting information from the user's interactions with social applications (e.g. social networks, social tagging applications, blogs, etc.) (Zhou et al., 2012, Vallet et al., 2010, Carman et al., 2008). The implicit method attempts to automatically infer user's interests or context from the processed logs or user items.

In the explicit method, the users themselves have to explicitly supply information to the system, whether positive or negative. This can take the form of a user providing the system with an initial specification of interests or "non-interests" (Micarelli and Sciarrone, 2004), providing positive or negative relevance feedback about retrieved documents (Chen and Sycara, 1998, Asnicar and Tasso, 1997, Harman, 1992b), or scrutinising (inspecting and modifying) the information that the system has learnt about the user so far (Psarras and Jose,

2006, Micarelli and Sciarrone, 2004, Pitkow et al., 2002). Concerns regarding the explicit method are that users may not wish to exert the extra time or effort to supply the information to the system and that users may sometimes input inconsistent or incorrect information (Budzik and Hammond, 2000, Carroll and Rosson, 1987).

A good example of systems that depend on the explicit approach for gathering various information is the *WIFS* system (Micarelli and Sciarrone, 2004), which is a PIR system that operates on the domain of computer science literature. In WIFS, the system initially learns the user's interests through an interview form that is used when the user first registers with the system. The form allows the user to explicitly specify his/her degree of interest in different computer science topics on a scale from -10 to +10 (i.e. *very irrelevant* to *very relevant*). Moreover, the user can provide explicit relevance feedback about viewed documents on the same scale. Upon the user's feedback, the terms in the rated document are processed by the system which affects the user model by the alteration of interest weights, the removal of interests, or the insertion of new interests in the user model. Finally, the user model can be scrutinised where the user is allowed to inspect and modify the inferred interests and their weights. This facility helps in keeping the user model accurate and up to date.

It is necessary, at this point in the survey, to highlight the notion of what is deemed an implicit method and what is deemed an explicit method. To a certain extent an implicit method partially entails an action on the user's behalf, such as clicking on a result's link (in the case of learning from clickthrough history). However, this type of gathering user information is deemed as implicit because it does not require that users perform any *additional* activities other than the ones they would normally carry out during a search session. Likewise, an explicit method partially entails some form of automatic processing either before or after the user's action. For example, if the user is asked to provide explicit relevance feedback to the system by marking one of the results as relevant, then the next step would be that the system automatically processes the document to extract its keywords and append them to the user's interests. Nevertheless, this would still be deemed as an explicit method because it involved some extra activity by the user that is specifically carried out for obtaining information that would assist in the personalisation process.

Some systems gather user and usage information in a mixed approach of implicit and explicit methods. The *Outride* system, presented in (Pitkow et al., 2002), is an example of such PIR systems. In addition to implicitly gathering usage information from search logs, the Outride system also allows users to scrutinise the information that the system has learnt about them.

2.2.2 Type of information

The type of information gathered by a system, whether user or usage information, influences how that system can personalise its service. User information is usually in the form of personal or demographic information such as the user's name, age, language, or country. User information may also include the user's job title, job description, or competency. Usage information exists in many forms, including queries that the user submitted to the search system, clicked results and their snippets (titles and summaries of documents), full browsing activity[3], and dwell time[4] on clicked documents. User and usage information also include information that can be obtained from external resources (i.e. from resources other than the search system itself), such as tags, bookmarks, shared items on social networks, the user's emails, calendar items, and stored desktop documents on the user's machine.

A number of systems in the literature only keep track of clickthrough behaviour, which comprises submitted queries and clicked documents (Smyth and Balfe, 2006, Stamou and Ntoulas, 2009, Cui et al., 2003, Qiu and Cho, 2006). Other systems extend this information by also logging the text from the snippets of clicked results (Yin et al., 2009, Psarras and Jose, 2006, Ruvini, 2003, Shen et al., 2005). Snippets are regarded by several studies in the literature as query-focused summaries of documents and are therefore used to extract interest terms that are relevant to the context of the query. For instance, the *MiSearch* system (Speretta and Gauch, 2005) maintains snippet information with the aim of comparing the effectiveness of a user model where terms are obtained only from submitted queries to one where terms are obtained from snippets of clicked documents.

The majority of PIR systems in the literature maintain monolingual search logs, and relatively fewer ones operate on a multilingual level. An example of multilingual PIR systems is the cross-language search system described in (Gao et al., 2007) where the authors extend the logged information by keeping track of queries submitted in different languages. They are motivated by the idea that in the same period of time, many users from different language backgrounds will share similar information needs. Thus, similar queries in different languages will exist in the logs, which can be exploited for personalised cross-language search.

A number of systems in the literature gather a richer set of information about usage behaviour. For example, in the *OBIWAN* system (Pretschner and Gauch, 1999) cached Web pages on the user's machine and their estimated dwell time are analysed in order to determine the user's interests. Another example is (Teevan et al.,

[3] Browsing activity comprises URLs clicked from the result list and any pages followed afterwards, along with other browsing-related information

[4] Dwell time is the estimated time that the user spent viewing a document

2005) where the authors gather information about the user's queries, visited Web pages, emails, calendar items, and stored desktop documents. They state that the more the user information the better the personalisation.

An emergent approach to maintaining an extended set of usage information is to analyse social data and user-generated content on the Web (Vallet et al., 2010, Carmel et al., 2009, Xu et al., 2008). This approach builds on the advent of Web 2.0 standards where websites are designed to maximise user participation and to facilitate authoring and adding content to them. User-generated content includes a wide variety of online information such as users' interactions with social networks, tags and annotations in social bookmarking systems, and posts in blogs and forums. The advantage of exploiting this type of information is that it enables personalised systems to gain rich knowledge about their users' interests and preferences due to the wealth of information that is available nowadays on social websites. As discussed earlier, such rich user and usage information can be used to enhance personalisation. On the other hand, privacy can be a concern when systems are harvesting such information from the Web. Although, much of the information shared on social websites is shared on a public basis, users do not always realise that this information can be used for different purposes (e.g. to inform marketing applications or other commercial activities). Another challenge in the use of social data is ensuring accurate interpretation of the harvested information (e.g. correct sentiment analysis) as well as ensuring that this information derives appropriate personalisation.

2.2.3 Source of information

The amount of information available for PIR systems varies depending on the sources or repositories from which information is obtained. Moreover, *where* the gathered or processed information is maintained also has an effect on the personalisation process in terms of when and where the information is available to be exploited for personalisation. Furthermore, privacy concerns are raised concerning where the information will be stored and how it will be used.

Usage information can be obtained from the server-side where the user's interactions with the system are logged. Several research studies (Yin et al., 2009, Qiu and Cho, 2006, Psarras and Jose, 2006, Speretta and Gauch, 2005, Liu et al., 2004) and numerous live systems on the Web, such as *Google*[5], *Yahoo*[6], *Bing*[7], *Facebook*[8], del.icio.us[9], and *StumbleUpon*[10], maintain and process the history of users' interactions with the system at the server-side. One drawback of this approach, however, is that it may sometimes raise privacy concerns for the users. A number of privacy issues are discussed in Section 2.3.

A number of studies in the literature, while maintaining information at the server-side, took into consideration the privacy aspect. For example, in the *I-SPY* system (Smyth and Balfe, 2006), the authors argue that no user identification or personal details should be logged among the data at the server in order to preserve the anonymity of the user. This is believed to provide a certain comforting degree of privacy to the users of the system. The authors call this kind of personalisation: *anonymous personalisation*. However, the problem with this anonymous approach is that it limits the possibilities of individualised personalisation, as it has to be performed at the aggregate level of behaviour of the search users (i.e. in a collective manner).

Another example of systems with enhanced privacy features is presented in (Xu et al., 2007). The authors demonstrate a technique for partitioning a user model both horizontally (i.e. removing terms at a specified level of detail) and vertically (i.e. removing sensitive categories of terms). This permits the system to provide higher level information to combine with search result rankings, depending on the needs of the user in each category.

Usage information can also be gathered at the client-side. The advantage of gathering information at the client-side, compared to server-side logging, is that it allows for a richer set of information to be collected about user interactions and behaviour. For example, the exploration of information at the client-side gives opportunity for analysing the full browsing activity of the user which extends to pages that the user viewed after abandoning the search interface. This is done by accessing the browser's cache or by using software tools that are installed on the client's machine (e.g. browser plug-ins). Examples of such systems are (Stamou and Ntoulas, 2009, Chirita et al., 2007, Teevan et al., 2005, Shen et al., 2005, Pretschner and Gauch, 1999).

Another advantage of systems that maintain information at the client-side is that they offer a certain degree of privacy to their users by guaranteeing that user information will not be submitted to a remote server. However, some client-side systems lack this advantage as they submit the collected information to the server for further processing. Examples of such systems are presented in (Agichtein et al., 2006a, Sugiyama et al., 2004, Stefani and Strapparava, 1999).

[5] http://www.google.com
[6] http://www.yahoo.com
[7] http://www.bing.com
[8] http://www.facebook.com
[9] http://www.delicious.com
[10] http://www.stumbleupon.com

2.3 Summary and discussion

This section reviewed a number of approaches in the literature with respect to the information gathering stage of PIR systems. Table 1 offers a summarised view of these approaches along with examples from the literature.

Table 1: summary of information gathering approaches

Information Gathering Approach	Type of Information	Source of Information	Example Publications
Implicit	Queries, clicked documents, or snippets of clicked documents	Server-side	Yin et al. 2009, Smyth and Balfe 2006, Qiu and Cho 2006, Speretta and Gauch 2005, Cui et al. 2003, Ruvini 2003
Implicit	Queries in different languages and clicked documents	Server-side	Gao et al. 2007
Implicit	Queries, clicked documents, or snippets of clicked documents	Client-side	Stamou and Ntoulas 2009, Shen et al. 2005
Implicit	Queries, clicked and cached web pages, dwell time on pages, desktop documents, emails, or calendar items	Client-side	Chirita et al. 2007, Teevan et al. 2005, Pretschner and Gauch 1999
Implicit	Queries, clicked and cached web pages, dwell time on pages, or other usage features	Client-side (information submitted to server)	Agichtein et al. 2006, Sugiyama et al. 2004, Stefani and Strapparava 1999
Implicit	Tags and Bookmarks on online social applications	Server-side	Vallet et al. 2010, Carmel et al. 2009, Xu et al. 2008, Carman et al. 2008
Implicit & Explicit	Queries, clicked documents, and user supplied information (e.g. user can scrutinise model or specify categories)	Server-side and user intervention	Psarras and Jose 2006, Liu et al. 2004, Pitkow et al. 2002
Explicit	User's categorical interests, and user supplied information (e.g. user can provide explicit relevance feedback or scrutinise the model)	Client-side and user intervention	Micarelli and Sciarrone 2004, Chen and Sycara 1998

It is noted that there is a high tendency in more recent literature towards the use of implicit methods for information gathering. Three reasons may be given for this tendency. The first is that users have shown to be generally reluctant to providing explicit feedback to systems (Gauch et al., 2007, Carroll and Rosson, 1987). In other words, it has been shown that users dislike the idea of having to exert the extra time or effort required to explicitly supply information to a system; they would prefer to see that the system is correctly "guessing" what kind of information they need instead of them having to specify their needs or clarify their intentions to the system explicitly (Budzik and Hammond, 2000). The second reason is that some studies, such as (White et al., 2002), have shown that personalised systems can equally benefit from implicitly gathered information as from explicitly gathered information. The third reason is that implicit feedback generates masses of data, far more than could be gathered by explicit feedback. All this has encouraged many systems to exploit the benefits of search history for IR personalisation. Nevertheless, as discussed in this section, some systems opted to harness the benefit of both approaches, for example by implicitly constructing user models yet allowing the users to scrutinise them if they wish. This helps to maintain an up to date and accurate user model that truly reflects the interests of the users, and also gives them the chance to become aware of what the system has inferred about them at any given time. (Cook and Kay, 1994, Micarelli and Sciarrone, 2004).

The main objective of personalisation is to adapt aspects of a service or system to the needs of a "person", and therefore, collecting a richer set of information about that person may lead to improved personalisation. This was concluded by (Teevan et al., 2005) based on experiments that involved the use of different combinations of information sets for PIR. The information comprised queries, visited Web pages, emails, calendar items, and stored desktop documents. The results showed that a system operating on the full set of available information performed better than a system that operated on information drawn from Web pages only, followed by a system that operated on information drawn from queries only.

The controversial decision of whether systems should collect and maintain information at the server-side or at the client side has two dimensions: the functional dimension and the privacy dimension. With respect to the functional dimension, the advantages of client-side monitoring, excluding systems which submit the information to a server, would be: (1) the availability of a richer set of information that is beyond the reach of a server-side system; and (2) part of the system's burden of processing information (computing resources) is taken away from the server. However, the drawbacks of client-side systems are: (1) they usually require the installation of a

certain application or plug-in at the client's machine, either to monitor or to process data, which some users may reject; (2) logged information is not available or not complete if the user uses the system from multiple machines; and (3) it would not be possible for the system to perform any collaborative or collective processing over all the user models and usage information, which is the kind of processing that many search engines need to do in order to draw conclusions about popular and high quality pages that receive many hits (views).

If a system is keen on the privacy dimension, then the benefit of maintaining user information at the client-side is that it contributes to a higher degree of privacy for the user, where the user's information is not shared or transmitted to any other machine. Data protection and privacy are issues which strongly affect user confidence, and their likelihood to engage truthfully and fully with adaptive systems (Kobsa, 2007). Users can be reassured by a variety of means, such as privacy policies and explicit explanations. However, there is some experiential evidence to suggest that while users are often highly concerned with privacy, it can depend on their familiarity with the systems in question. For example, teenagers have been observed to have a clear model of risk-taking and privacy on social networks (Livingstone, 2008).

The challenge of maintaining privacy in a personalised system depends on two factors: ensuring that the user feels in control of their information, and ensuring the integrity of that information. Information integrity and privacy are increasingly areas of interest not just for users, enterprises and researchers, but also for legislative and regulatory authorities. For example, the legislative and legal environment described in (Volokh, 2000) notes that contract law is the main mechanism for ensuring information integrity. The article is applicable only to US law. Despite that, because many service providers situate themselves in the US, it is necessary for users who wish to inform themselves of their protections to be conscious of national borders and regulatory differences. The European Union goes significantly further in attempting to provide for "Data Protection" (Guarda and Zannone, 2009, Acquisti and Gross, 2006), and there is a clear need for privacy-aware systems to adapt and tailor their provisions and to maintain traceability based on the different applicable legislation in different areas (García-Barrios et al., 2009).

In the commercial context, there is substantial evidence that personalised user information is being used by enterprises on a mass scale. For example, Google has been personalising search results for signed-in and anonymous users since 2009[11]. Moreover, Microsoft Bing makes extensive use of user action in, for example, learning to correct common spelling errors in queries (Sun et al., 2010). Facebook has created a substantial repository of information in its *Graph*[12] (user information along with connections to items, such as images, videos, etc.), which can be leveraged by third parties either through instant personalisation or through *Facebook Connect*[13]. This permits third parties to access not only the users' information, but also the information of their "friends".

There is an explicit attempt to reflect the type of access that applications are given in the dialog that Facebook presents. There is evidence to suggest that user behaviour will change as individuals change roles in life, but that the use of the network itself is not subject to drastic change (Lampe et al., 2008). This suggests that privacy policies can be established over time and tested effectively, but that they must be cognisant of the fact that users' needs will also change over time. One piece of recent legislation is the EU Privacy Directive, which restricts how third party cookies can be used[14]. While such legislation may be comforting for European internet users, the practical implications of this legislation for real world system design are, as yet, unclear.

3 Information representation

3.1 Overview

This section of the survey is concerned with the second stage of PIR systems, which is the storage and representation of the information gathered in the previous stage. In many systems, a user model is constructed in order to represent the user's interests in an individualised manner. However, some personalised retrieval systems maintain an aggregate representation of users' preferences and general usage behaviour. This kind of collective information is exploited for personalisation across the cohort of aggregated users. In this survey, both kinds of systems are covered, with a more in-depth analysis of the former systems (i.e. the ones involving an individualised user model). Moreover, this section also discusses systems where a thesaurus or a knowledge source was used to organise the representation of the gathered information. Finally, this section also discusses the different mechanisms that are used to update the information maintained by PIR systems.

[11] The Official Google Blog - Personalized Search for Everyone: http://googleblog.blogspot.com/2009/12/personalized-search-for-everyone.html
[12] http://developers.facebook.com/docs/reference/api
[13] http://developers.facebook.com/blog/post/108
[14] The Wall Street Journal, Tech Europe - U.K. Publishes EU Cookie Directive Guidelines: http://blogs.wsj.com/tech-europe/2011/05/09/u-k-publishes-e-u-cookie-directive-guidelines

Gaining an insight into the information representation stage is important because it explores the nature and structure of user models that are a core part of many personalised systems. Furthermore, it gives way to understanding how query and result adaptation are performed, as both are closely dependent on the type of information maintained by the system (details of query and result adaptation will be discussed in Section 4).

The analysis presented in this section is carried out over the following three criteria:

- **Existence of an individualised user model and scope of interests:** the first criterion examined in this part of the analysis is concerned with systems which make use of an individualised user model and the scope of user interests maintained by the model (i.e. short-term or long term interest). Hereafter, the term *individualised user model* will be used to refer to the notion of the explicit existence of a user model in a system, regardless of the approach by which the model's information was gathered (be it explicitly or implicitly).
- **Usage information / user model representation:** the second criterion, which can be regarded as the most important criterion of this section, is concerned with how user and usage information is represented. This involves both, systems which make use of individualised user models and systems that represent information on an aggregate level.
- **Dynamism of user model and information update scheme:** the third criterion over which systems are discussed in this section is the degree of dynamism of the information stored in the user model and the mechanisms in place for updating this information. The information stored in the user model could be static, such as personal characteristics or demographic user information (which are rather permanent), or dynamic, such as the user's interests (which usually evolve with time).

The following section surveys different systems in the literature. The analysis goes through the information representation stage of the surveyed systems according to the three criteria outlined above.

3.2 Review

3.2.1 Existence of an individualised user model and scope of interests

A key component in many PIR systems is the user model which maintains the user's information on an individualised level, especially the terms that represent the user's search interests. These interests could be long-term or short-term interests.

In the context of IR systems, long-term interests are persistent interests that can be exhibited in the user's search history on the long run. Inferring these interests from past searches and exploiting them can help in enhancing similar future searches (Qiu and Cho, 2006, Speretta and Gauch, 2005, Psarras and Jose, 2006, Liu et al., 2004, Pretschner and Gauch, 1999). This is done by analysing the text of the user's queries and the clicked documents (or their snippets) and extracting key terms from them, for example by selecting the most frequently appearing terms. Interest terms are then used for adapting future queries or their results so that documents that are more relevant to the user are retrieved and displayed to the user at higher ranks. Besides harvesting interest terms from queries and clicked documents, some systems infer the users' long-term interests from their desktop documents, emails, or calendar items (Chirita et al., 2007, Teevan et al., 2005). Later sections in this survey discuss how interest terms are used for query and result adaptation.

Short-term interests are ephemeral interests that are usually satisfied by a few ad-hoc searches in a relatively shorter period of time (typically, one search session). Short-term interests are usually harvested from submitted queries and retrieved documents in a search session and used to personalise the search immediately within that search session (Ruvini, 2003, Shen et al., 2005).

Some systems in the literature perform personalisation based on both, long-term and short-term interests (Stamou and Ntoulas, 2009, Sugiyama et al., 2004). A good example is (Sugiyama et al., 2004) where the user's full browsing activity is monitored at the client-side in a live manner. This enables the system to deduce both, short-term and long-term user interests from terms available in the browsed Web pages. The two scopes of interests are stored separately in the user model. The TF.IDF[15] weighting scheme (Baeza-Yates and Ribeiro-Neto, 2011) is used to assign weights to the terms in order to depict different degrees of user's interest. The short-term interests are implicitly updated whenever the user clicks on a document and are thus immediately exploited for personalisation in the current search session. Long-term interests, and their weights, are also updated when the user clicks on documents, but the difference is that long-term interests are subject to a periodic weight-decaying mechanism that reduces term weights over time. This leads, on the long-run, to preserving only persistent interests that frequently appear in the user's browsing history. Mechanisms for updating user models are discussed in more details in Section 3.2.3.

[15] Term frequency multiplied by inverse document frequency

3.2.2 Usage information / user model representation

Several techniques and data structures can be used to represent user and usage information in PIR systems. This section starts by providing a discussion of the techniques used in systems that made use of an individualised user model. The discussion also involves knowledge sources that are sometimes used as the basis for representing users' interests. The section then moves on to discussing systems where usage information was represented at an aggregate level (i.e. without the use of an individualised user model).

In this survey, user models which represent the user's interests are classified with respect to two dimensions: *data structure* and *content*. The data structure dimension is concerned with the underlying storage mechanism used to represent interest terms in the model. This can either be a vector-based model or a semantic network-based model. The content dimension is concerned with the nature of the terms maintained in the user model. The terms can either be words that are freely mined from user/usage information or conceptual (categorical) terms that are drawn from some sort of knowledge source. The following review discusses the details of each of these types of user models and a summarised classification is shown in Table 2. This classification is an extension to the classification reported in (Gauch et al., 2007).

Table 2: classification of user models

	Terms	**Conceptual Terms**
Vector-based	models where user's interests are maintained in a vector of weighted keywords	models where user's interests are maintained in a vector of weighted concepts
Semantic network-based	models where user's interests are maintained in a network structure of terms and related terms	Models where user's interests are maintained in a network structure of concepts and related concepts

A vector-based user model is made up of a feature vector, which is a vector of terms and associated weights. The weights can be determined, for example, using a term weighting scheme such as TF, TF.IDF, or BM25 (a.k.a. Okapi BM25) (Baeza-Yates and Ribeiro-Neto, 2011, Robertson et al., 1995). One way to represent the terms in the model is by using words or phrases that are freely mined from user or usage information.

Vector-based user models may be composed of one or more vectors. For example, in (Shen et al., 2005) and (Ruvini, 2003) only one vector was used to store the user's short-term interests. In (Sugiyama et al., 2004), two vectors were used; one for short-term and one for long-term interests.

Gathering interest terms together in one vector may be more appropriate for maintaining short-term interests, but perhaps not for long-term interests. This is because a short-term vector would naturally comprise much fewer terms than a long-term vector as it is usually created for one search session. This is in contrast to a long-term vector where terms are continuously accumulated with every new search that the user performs. This may eventually lead to a *noisy* ocean of terms (i.e. a single vector that contains a wide variety of un-clustered terms) that may be harder to exploit for personalisation. To this effect, a number of systems in the literature, such as the systems described in (Psarras and Jose, 2006) and (Chen and Sycara, 1998), represented the user's long-term interests using multiple vectors; one vector per interest cluster. In such case, terms are usually grouped together under un-labelled clusters using unsupervised text clustering techniques (Witten et al., 2011).

An example of how words or phrases are harvested from search history and how they are used to populate a vector-based user model was illustrated in (Chen and Sycara, 1998). The system builds a vector-based user model which comprises multiple vectors of interest. The terms in the vectors are weighted using the TF.IDF scheme. Interest terms are extracted from documents which the user has explicitly marked as relevant, where each vector in the model corresponds to important keywords obtained from a single document. The full text of the document is not actually used for term extraction, rather only terms which are in the query's context[16]. The system has a threshold concerning the number of vectors to be maintained in the model (i.e. a maximum of N clusters of interest). Furthermore, the system also has a threshold for the number of terms stored in a vector (i.e. a maximum of M interest terms per cluster). If the threshold of N vectors was reached, and a new vector comes in, then all the vectors in the user model, in addition to the incoming vector, are textually compared to each other using cosine similarity (Baeza-Yates and Ribeiro-Neto, 2011). The two most similar vectors are then combined together in one vector. This is done by grouping together the terms from the two vectors, sorted in descending order of weights, and then keeping only the top M terms. The benefit of this approach is that, over time, terms which commonly appear in topics that were repeatedly searched for by the user will tend to cluster together in the model.

[16] Terms that surround the query terms in the document, extending for example to five words before and five words after each query term

Another way to represent the terms in a user model is by using conceptual terms. When conceptual terms are used in a vector-based user model it is commonly known as a concept-based user model. In this kind of model, the user's interests are represented by categorical terms that are drawn from some sort of knowledge source. Knowledge sources could be domain models that are developed by human domain-experts (such as databases of domain-specific terminologies), general knowledge repositories developed by human contributors (such as *Wikipedia*[17]), Web taxonomies or concept hierarchies (such as *ODP*[18]), or rich ontologies (such as *SUMO*[19]). *WordNet*[20], though may not be regarded as a knowledge source in the formal meaning, is considered as a rich source of linguistic and semantic language knowledge, and is therefore sometimes used to organise the terms in the model. The use of conceptual or categorical terms in concept-based user models serves to organise and accurately describe the user's interests with respect to the common terms used in a domain. The combination of a knowledge source with a user model is also known as an overlay model (Brusilovsky and Millán, 2007).

In MiSearch (Speretta and Gauch, 2005) two alternative vector-based user models are proposed to represent the user's long-term interests. The first comprises concepts extracted from the user's queries, and the second comprises concepts extracted from the snippets of clicked documents. Each user model is made up of multiple vectors; one per interest category. Both user models represent their categories and concepts based on the ODP hierarchy[21] (3 levels deep). Searching is not restricted to Web pages that are classified under ODP as it is performed on open corpora over the Web, using a wrapper of Google search. Thus, a need was raised concerning mapping retrieved documents to ODP categories. The authors handled this issue by training a Vector-space text classifier on 30 documents from each category, then using the classifier to classify the text of queries or snippets under a certain category. The authors concluded that both user models were equally capable of modelling the user's interests. Many systems in the literature also represented their vector-based user models using concepts from the ODP hierarchy (Qiu and Cho, 2006, Liu et al., 2004, Pitkow et al., 2002).

Another example of concept hierarchies that were exploited for constructing user models was the *Magellan*[22] concept hierarchy. In the OBIWAN system (Pretschner and Gauch, 1999), the user's long-term interests were modelled according to the Magellan hierarchy (4,400 nodes). The TF.IDF weighting scheme was used to weight the concepts stored in the user model and the document terms from which the concepts were extracted. Similarities between documents and the user model were computed using cosine similarity.

Concept hierarchies like ODP and Magellan may be regarded as simple ontologies with one straightforward relation between their nodes. This kind of relation is said to be an *"is-a"* or *"has-a"* relation between a parent node and a child node (Gauch et al., 2007). A number of studies in the literature prefer the use of a richer kind of ontology. The term *richer* refers to the notion of using a wider variety of relationships between nodes, such as *"related-to"*, *"associated-with"*, *"used-in"* or *"pre-requisite-of"*. Ontology representation languages, such as *OWL*[23], are used to represent such relationships. The details of ontologies are outside the scope of this survey. However, one example of PIR systems that make use of a rich ontology is (Stamou and Ntoulas, 2009), which uses manual and automatic techniques to create a merged ontology. The merged ontology basically depends on ODP for categorising Web content, and then the hierarchy is augmented with concepts from SUMO, WordNet, and *MultiWordNet Domains*[24]. More information about ontology-based models can be found in (Ye et al., 2007, Razmerita et al., 2003).

User models can be represented using a semantic network structure. In this case the model is made up of nodes and associated nodes that capture terms and their semantically-related or co-occurring terms respectively. Weights can be assigned to the nodes, their associated nodes, and the links between them. The advantage of semantic network-based models over vector-based models is that they can model the relationship between a key term or concept and its associated terms (e.g. synonymous terms or co-occurring terms in a document collection). The mapping between terms and related terms can be achieved using a thesaurus like WordNet. For example, the authors in (Stefani and Strapparava, 1999) state that the drawback of vector-based models is that the computation of term weights is only based on term frequencies (bag of words), which does not pay attention to *word sense disambiguation*. Therefore, in their SiteIF system (Stefani and Strapparava, 1998), they use WordNet to obtain semantic similarity between words (e.g. synonyms). In SiteIF, a main node holds a weighted term that represents a user's interest. The terms come from usage information, more specifically from clicked

[17] http://www.wikipedia.org/
[18] Open Directory Project: http://www.dmoz.org
[19] Suggested Upper Merged Ontology: http://www.ontologyportal.org/
[20] http://wordnet.princeton.edu/
[21] The ODP hierarchy is a Web taxonomy that was manually created, and is still maintained, by participation from over 92,000 human editors. It has 16 main categories which branch to approximately 1 million subcategories. It is formed as a tree with symbolic links between parent nodes and child nodes. A node may be linked to multiple parents, which makes ODP categories appear as a lattice rather than a strict hierarchy. The categories can be used as classification metadata. Approximately 4.9 million Web pages are classified under ODP, however, this only accounts for approximately 0.05% of the estimated number of Web pages indexed by Google (Chirita et al., 2005).
[22] Magellan was a project associated with the *Excite* search engine. According to (Gauch et al., 2007), when Magellan ceased to exist, the authors of OBIWAN switched to ODP.
[23] http://www.w3.org/TR/owl-overview/
[24] http://multiwordnet.fbk.eu

documents by the user. Moreover, semantically related terms to the main term are obtained from WordNet and stored into associated nodes. The associated nodes are connected to the main nodes using weighted links. The link weights represent the frequency of appearance of the associated terms with the main term in a document.

Another example is the user model used in the WIFS system (Micarelli and Sciarrone, 2004) which is, in part, based on a semantic network representation. In WIFS, the user explicitly specifies an initial set of weighted interests upon registering with the system, which are used to associate the user with a number of stereotypes. The stereotypes are a form of domain model that covers knowledge in the field of computer science, and were defined by human experts in the field. They are used as a way of default reasoning about the user (i.e. for bootstrapping a fresh model). *Active stereotypes* are the stereotypes that fit the current user's description and which are loaded in the user model. The WIFS user model maintains various pieces of information about the user: personal information, list of active stereotypes, and a set of semantic networks that represent the user's interests (as explicitly entered and as inferred from documents that the user marked as relevant). Each semantic network has a main node that holds a weighted topic term. This node is called a *planet*. Furthermore, when the user explicitly marks a document as relevant, the system harvests terms that co-occur with the topic term in the document. These co-occurring terms are then stored in nodes that are called the *satellite* nodes, which are connected to the planet node by weighted *arcs* (links).

A number of studies in the literature performed search personalisation on an aggregate level. Aggregate personalisation involves the exploitation of usage information in a collective manner where the search process is adapted to the needs of the many rather than the specific needs of the individual. In these studies, no user model is used for storing interests; rather, a general representation of usage information is used. For example, the I-SPY system (Smyth and Balfe, 2006) keeps track of all users' queries and their clicked documents in a matrix called the *hit matrix*. The rows of the matrix represent the queries and the columns represent the documents (document identifiers). A cell in the matrix holds the number of times that the designated document was clicked for the corresponding query. This representation can be thought of as storage of query-document pairs along with their click frequency (hits). Click frequencies are exploited by the system for assigning higher ranks for frequently clicked documents in the list of retrieved results to a common query.

Similarly, in (Agichtein et al., 2006a) aggregate usage information was maintained in the form of query-document pairs in a model called the *Implicit Feedback Model*. However, the model stored a much richer set of information about each pair. The model, which is represented as a feature vector, comprised a wide range of query-document aspects, such as *clickthrough information* (e.g. tracking the document's click frequency in relation to the click frequency of other documents that appeared higher or lower in the ranked result list), *browsing information* (e.g. average page dwell time), and *textual information* (e.g. overlap between the query terms and the terms of the document's URL, title, and snippet). The many diversified features in the model helped to accurately interpret and exploit aggregate implicit feedback.

The authors in (Gao et al., 2007) also process pairs of queries and documents but with the aim of deducing the degree of similarity between queries of different languages. The goal is to improve cross-language search by keeping records of queries and candidate similar queries from other languages that could be used for cross-lingual query suggestions.

The work reported in (Yin et al., 2009) is also motivated by the idea that query logs reflect the wisdom of the crowds, where users may seek the same information using different queries. Queries and clicked documents are represented using a Query-URL graph, on which a graph-based machine learning algorithm is applied. The Query-URL graph is a bipartite graph where the first set of vertices represents the queries and the second set represents the documents. The edges connecting the vertices of the two sets represent clickthrough information. The random walk algorithm that is applied on the graph generates probabilities between queries, where higher probabilities reflect higher query similarities. These similarities are then exploited to improve future searches by query adaptation. A limitation that was addressed in the study was that the random walk algorithm does not work well with queries for which no results were clicked. This challenge was overcome by textually comparing such queries with all other queries in the logs (using cosine similarity) to find ones that can be deemed similar.

Another example of using machine learning algorithms to analyse clickthrough information for improving search results is (Sun et al., 2005). Search engines usually generate a huge amount of clickthrough information involving different types of items, mainly: user identifiers, queries, and clicked Web pages. Over time, the relationships between these items can become complicated and the information can become highly sparse as each user only submits a relatively small number of queries, and then only a very small set of documents are clicked for them. Therefore, to capture latent features of the three-way relationship between the items, the authors propose a model called: *CubeSVD*, which is partially based on Latent Semantic Indexing techniques (Manning et al., 2008). In CubeSVD, clickthrough information was represented as a 3-order tensor, and was then analysed using the Higher-Order Singular Value Decomposition algorithm (Lathauwer et al., 2000). The analysis produced quadruplets of (user, query, page, weight), where for each quadruplet, the weight represents the likeliness that a certain user will visit the designated page upon submitting the designated query. This information is then used to adapt result lists to users in their future searches.

3.2.3 Dynamism of user model and information update scheme

Some user models are static, while others are dynamic in nature (Golemati et al., 2007, Hothi and Hall, 1998, Rich, 1983). Static user models are ones that maintain user information that is less likely to change over time and are therefore not subject to continuous updates. Examples of static information are personal characteristics, background knowledge, and demographic information. Maintaining static information allows PIR systems to group users into stereotypes and make high-level personalisation decision (e.g. localise the system's GUI based on the user's language, or adapt some of the services based on the user's geographic location). Dynamic user models, on the other hand, are ones that keep track of information that evolves over time. For example, models that maintain short-term user interests are usually created on-the-fly and are updated frequently over the span of the user's search session. Long-term interests can be considered as dynamic information as well if the system has a revision or update mechanism for them in place (e.g. increasing or decreasing the weights of the interests on a periodic basis, or adding new interests). More user-focused personalisation decisions can be made when the system maintains dynamic information; decisions that cater for the current user's context and interests.

Many PIR systems implement updating mechanisms in order to ensure that they maintain accurate and up to date information about the user. This mechanism can be triggered upon a certain user action, such as a click on a document or the provision of explicit relevance feedback about a document (Shen et al., 2005, Sugiyama et al., 2004, Ruvini, 2003, Chen and Sycara, 1998). At which point, information can be updated on-the-fly and the newly available information may be immediately exploited for personalisation in the current search session. Updating procedures can also be invoked by configuring the system to periodically revise the weights of learnt interests; a mechanism known as *decaying* or *aging* (Stefani and Strapparava, 1998, Asnicar and Tasso, 1997). Furthermore, manual updates can also take place by allowing the user to scrutinise the information that the system has inferred about them (Psarras and Jose, 2006, Micarelli and Sciarrone, 2004, Pitkow et al., 2002).

In systems where personalisation is based on short-term interests, the updating mechanism may be invoked several times within the same search session. This is to ensure that any new piece of information that becomes available, following a user action, would reflect in the model and would be immediately exploited for personalisation. For example, in (Ruvini, 2003) the model only keeps track of the user's current interests for a given query and the results browsed for it. That is, for every new query submitted by the user a new user model is created and is continuously updated as the user clicks on results. An insight into this updating mechanism can be gained by having a closer look at the system; the personalised search system is wrapped around the Google search engine and is mainly intended for use on limited-display devices (e.g. mobile phones) where only a small number of results can be displayed per page. A supervised machine learning approach is used to construct and update the model where a text classifier (Support Vector Machines) is trained on features extracted from the snippets of clicked result. The classifier operated under the assumption that clicked results are positive examples (of what is relevant to the user) and unclicked results are considered negative examples. When the user clicks on a result from the displayed page of results, the positive and negative examples are passed to the classifier to form a model of user's interests. Then, behind the scenes, the same query is re-submitted to Google and the top-N retrieved results are passed to the classifier to be labelled. Two groups of results are then formed; one for relevant result and one for non-relevant results. The ranking of Google is preserved for the results within each group. The user then actually avails of personalised results when he/she clicks to view the next result page. On the new page, a set of previously unseen results is displayed, where relevant results are displayed above non-relevant ones.

A good example of systems that maintained a model of both, static and dynamic user information is WIFS (Micarelli and Sciarrone, 2004). As partially mentioned earlier, the user model in WIFS contains various pieces of information about the user: personal information, search interests, and a list of active stereotypes that the user belongs to. In WIFS, multiple update schemes are implemented to ensure the dynamism of the user model. First, the user model can be updated manually where the user is allowed to fully scrutinise the model. This is done by: (1) adding or removing main topical terms or their associated co-occurring terms; (2) editing the weights of the terms or the links between them; or (3) manipulating the list of stereotypes. Second, the model is also updated by a weight decaying mechanism called the *renting* mechanism. However, this decaying mechanism is not performed on periodic basis. In the renting mechanism, the weights of a topic and its links are decreased if the topic does not appear in a document that the user has provided explicit relevance feedback for. Because of the different possibilities of updates, the authors devised a mechanism for maintaining the user model's consistency. The mechanism depends on a belief revision technique that is called *Justification-based Truth Maintenance System (JTMS)*. For example, if a user removes one of the active stereotypes, then all assertions that were justified by that stereotype (i.e. depended on it) are removed accordingly.

An example of updating schemes implemented in systems where usage information was maintained at an aggregate level can be found in the I-SPY system (Smyth and Balfe, 2006). In I-SPY, where a matrix was used to represent hits with respect to query-document pairs, two update issues were discussed by the authors: (1) documents that were indexed by the system at earlier times will tend to have higher click frequencies than more

recent ones, which may cause their rank to be higher even if more recent documents are more relevant to a given query; and (2) documents could be removed from the Web, thus leaving erroneous entries in the documents index. These issues were addressed by implementing two update schemes: first, the hit values (click frequencies) are reduced over time, and second, a garbage collection mechanism is run periodically to verify that indexed documents still exist on the Web.

3.3 Summary and discussion

This section provided a review and analysis of approaches exhibited in the literature concerning how information is represented in different PIR systems. Table 3 provides a condensed view of the approaches discussed in this section.

Table 3: summary of information representation approaches

Existence of User Model and Scope of Interests	Usage Information /User Model Representation	Use of Thesaurus or Knowledge Source	Dynamism of User Model and Information Update Scheme	Example Publications
Yes, short-term interests	Vector-based user model (keywords)	No	Dynamic: immediate update when the user clicks on a document	Shen et al. 2005, Ruvini 2003
Yes, long-term and short-term interests	Vector-based user model (keywords)	No	Dynamic: updated upon every new Web page that is browsed by the user. & Periodic decaying of interests.	Sugiyama et al. 2004
Yes, long-term interests	Vector-based user model (keywords)	No	Static+Dynamic: updated upon user's explicit relevance feedback for a document, user can scrutinise model, or model construction process can be repeated when needed	Chirita et al. 2007, Psarras and Jose 2006, , Teevan et al. 2005, Chen and Sycara 1998
Yes, long-term interests	Vector-based user model (conceptual terms)	Yes (ODP, Magellan, etc.)	Static+Dynamic: model construction process can be repeated when needed or user can scrutinise model	Qiu and Cho 2006, Speretta and Gauch 2005, Liu et al. 2004, Pitkow et al. 2002, Pretschner and Gauch 1999
Yes, long-term and short-term interests	Vector-based user model (conceptual terms)	Yes (Hybrid ontology of ODP, WordNet, MultiWordNet, SUMO)	Static+Dynamic: model construction process can be repeated when needed	Stamou and Ntoulas 2009
Yes, long-term interests	Semantic-network-based user model (weighted nodes of keywords and weighted links connecting co-occurring words)	No	Dynamic: updated upon user's explicit relevance feedback for a document. & Periodic decay of weights of nodes and links	Asnicar and Tasso 1997
Yes, long-term interests	Semantic-network-based user model (weighted nodes of keywords and weighted links connecting semantically related or co-occurring words)	Yes (WordNet or domain stereotypes built by human experts, etc.)	Static+Dynamic: updated upon user's explicit relevance feedback for a document, user can scrutinise model, or periodic reconsideration of weights of nodes and links	Micarelli and Sciarrone 2004, Stefani and Strapparava 1999
No	Multiple history matrices (one per community) to keep track of queries and frequency of clicked documents	No	Dynamic: click frequencies are decayed over time. & Periodic garbage collection mechanism (to check that indexed pages still exist on the Web)	Smyth and Balfe 2006
No	Statistical/Probabilistic information involving the relations between users, queries, clicked documents, or other features	No	Static+Dynamic: machine learning process can be repeated when needed	Yin et al. 2009, Gao et al. 2007, Agichtein et al. 2006, Sun et al. 2005

It is noted that relatively few studies in PIR literature performed personalisation based only on short-term user interests. The benefit of keeping track of the user's short-term interests is that it accounts for the user's ad-hoc information needs and allows systems to perform personalisation on-the-fly (Ruvini, 2003, Shen et al., 2005). To this end, the authors in (Shen et al., 2005) argue that the majority of the user's searches come from ad-hoc information needs which are usually satisfied by a small number of searches. Thus, they conclude that personalisation should target the scope of short-term user interests.

However, a concern that is associated with performing personalisation based only on short-term user interests (i.e. operating only on information obtained from the current search session) is that very little information is available to base the personalisation decisions on. For example, the analysis carried out in (Jansen et al., 2000) gives an idea of how limited the information from one search session could be. The analysis shows that the average number of queries per session is 1.6 queries and that the average number of results clicked in a session is approximately 2.4 results. Following on this, the large-scale PIR study conducted by (Teevan et al., 2005) shows that the amount of information available does affect the degree to which personalisation can be effective.

Therefore, with the apparent contradiction in the literature, perhaps it is wise to combine evidence from both scopes of interests to personalise the user's different searches. This may be achieved by partially basing personalisation decisions on short-term interests, yet relying on long-term interests when it makes sense to do so. This combined approach was shown to be useful in a number of studies such as (Stamou and Ntoulas, 2009) and (Sugiyama et al., 2004).

The advantage of using a knowledge source or a domain model for representing information in PIR systems is that it allows the system to better realise commonalities between search results and the user's interests. This is because a common domain vocabulary is used for mapping both, the user's interests and the text of the documents into conceptual terms. Furthermore, it facilitates the clustering of terms harvested from queries or documents into well-defined categories. These advantages, together with the capability of semantic network-based user models to adequately represent weighted relations between concepts, can help towards the construction of user models that accurately depict the user's interests. Furthermore, the use of powerful machine learning techniques, such as Bayesian Belief Networks, to discover these relations has shown to improve PIR effectiveness (Zhang and Koren, 2007, Pinheiro de Cristo et al., 2003, Pazzani and Billsus, 2007).

It is noted that the use of individualised user models in PIR literature was mostly investigated for monolingual PIR systems (especially for English, perhaps due to its inherent popularity). In a multilingual world, information that is relevant to the user's information need may exist in languages other than the language that the user used to query the system. With the advent in machine translation techniques, users can access documents that are beyond their native language. Furthermore, a proportion of the users may very well be familiar with multiple languages and are able to comprehend documents in those languages. Therefore, a viable direction for future research would be to take the multilinguality dimension into consideration in PIR systems, especially the investigation of how to construct user models that depict the aspects and interests of a multilingual search user.

4 Personalisation implementation and execution

4.1 Overview

This section of the survey focuses on the implementation and execution of the personalisation process. Personalisation in PIR systems is generally performed by adapting the query and/or the results. Adaptation can either target specific individualised user needs, or target common needs of groups of users. This section also discusses the types of services provided by the reviewed personalised systems.

As this section explores the details of how personalisation is implemented and executed in different systems, it can thus be regarded as the core part of the survey. The analysis presented in this section is carried out over three criteria: the system's type, the personalisation scope, and the personalisation approach. The following is an overview of these three criteria.

- **Type of service:** the first criterion is concerned with the domain or the type of IR service that a system offers, such as monolingual Web search, multilingual Web search, personalised news, eLearning, etc.
- **Personalisation scope:** the second criterion is the scope on which personalisation is performed. In this paper, systems are classified into three categories: individualised scope, community scope, and aggregate scope.
- **Personalisation approach:** the third, and most important, criterion in this part of the analysis is how personalisation is performed. This can be by query adaptation, result adaptation, or both.

The following section provides a detailed review of the literature. The analysis goes through the personalisation implementation stage of the surveyed systems according to the three criteria presented above.

4.2 Review

4.2.1 Type of Service

This section discusses the types of services provided by different PIR systems, and shows how the aspects of personalisation offered by these systems differ based on the services they provide. Although this survey mainly focuses on search systems, other systems from closely related areas are also discussed.

Textual search is perhaps the most prominent application of IR. Many systems presented in academic literature and ones which are currently deployed on the Web offer search services (Vallet et al., 2010, Stamou and Ntoulas, 2009, Yin et al., 2009, Speretta and Gauch, 2005). Some systems extend this to cross-language search, where a translation mechanism is used to translate the query or the document in order to allow the retrieval of documents that are not necessarily in the same language of the query (Oard, 2010, Gao et al., 2007, Ambati and Uppuluri, 2006, Cao et al., 2007). In personalised search, personalisation is often implemented by adapting the query (e.g. automatically or semi-automatically modifying the query terms to obtain a better description of the user's information need), adapting the results (e.g. re-ranking the list of results so that more relevant results are displayed higher in the list), or both. Section 4.2.3 provides a detailed discussion of these approaches.

Some systems study the provision of a personalised search service on hand-held devices (e.g. mobile phones). In such case, the study considers several GUI (Graphical User Interface) factors in the adaptation process. For example, in addition to investigating how to adapt the results with respect to the user, the authors in (Ruvini, 2003), also investigate the adaptation of result lists with respect to the limited display offered by mobile devices.

In most search systems, results are typically presented to the user in the form of a ranked list of results. To this effect, if result adaptation takes place, it mainly involves altering the ranks of these results in the list. However, the authors in (Steichen et al., 2009) present a different approach to result adaptation and presentation. The authors propose a search system that operates in an eLearning environment, where instead of displaying a typical ranked list of results, the content of the results is dynamically re-composed to generate a tailored hypertext presentation. The search is performed over a closed corpus of domain-specific Web pages that were harvested from the Web. Furthermore, the harvested Web pages were manually annotated to indicate the level and nature of the content presented in them (e.g. *introductory level information*, *advanced level information*, *theoretical/conceptual content*, *technical illustration/example*, etc.). The user model contained information about the user's prior knowledge with respect to the learning domain and the level of the user's experience in that domain. This information, together with the document annotations, provided adequate information to an adaptive engine so that it can re-structure the content of documents and display it in a presentation-style format that suits the user. Moreover, the authors extend the work and evaluate its application in the customer support domain (Steichen et al., 2011). They propose a search system that is intended to assist users who are searching for solutions to technical problems concerning a certain product. The system performs adaptive composition of a personalised hypertext presentation based on technical support content from heterogeneous data sources (open corpora, closed corpora, social networking, etc.). This content comprises technical information obtained from the product's manuals as well as user-generated content related to that product (e.g. discussion forums on the Web). Personalisation is based on the user's level of expertise with respect to individual product features and the user's query intent (e.g. "find out about product features" or "get at how-to"). In order to compose the result presentation, multiple versions of the query are submitted to the retrieval component. The queries are expanded using different terms and meta-data information obtained from the domain and content models[25]. The results retrieved for these queries are then re-composed according to the user information and query intent.

A number of systems offer personalised news services. In such systems, personalisation is concerned with "guessing" which pieces of news would be of interest to a particular user. For example, the *PersoNews* system described in (Katakis et al., 2009) disseminates RSS-feed news items to users based on their interests. Machine learning techniques are used to learn the users' interests based on the kind of news feeds that they subscribe to and the news items that they explicitly mark as relevant. The learnt interests are used to filter out news that is not relevant to the user. Another example is the *WebMate* system (Chen and Sycara, 1998) which also operates on the news domain. WebMate offers two services to its users: searching on news corpora (retrieval of information) and filtering news items according to the user's interests (filtering of information).

The area of Information Filtering (IF) (Belkin and Croft, 1992, Oard, 1997) is closely related to IR. Yet, there are a number of differences that distinguish between the two areas (Hanani et al., 2001, Belkin and Croft, 1992, Brusilovsky and Tasso, 2004). First, IR systems are generally intended for ad-hoc information needs, while IF systems are intended for persistent information needs that are exhibited on the long-run. Second, in IR, information needs are represented as queries, while in IF, the user models themselves can be considered as the

[25] The domain and content models were generated a priori using automatic meta-data generation techniques

representation of the user's information need. Third, the purpose of IR systems is to locate information, while the general purpose of IF systems is to disseminate information.

It may be deduced from the aforementioned argument that PIR based on user's long-term interests is essentially IF. Yet, with respect to the analysis and scope of this survey, the thin line that separates the two is how information is sought. The analysis in this PIR survey is concerned with search systems where the initial action in the information seeking process is the user submitting a query to the system with the aim of satisfying an information need (be it ephemeral or persistent). On the other hand, IF systems are hereby regarded as systems where there is a dynamic flow of unsolicited information that needs to be disseminated to users. The initial action in the IF process is thus the arrival of an incoming document. This distinctive feature was stated in (Belkin and Croft, 1992, Hanani et al., 2001) as one of the features which generally differentiate between IR and IF. However, given the gray area between PIR (with long-term interests) and IF, a number of systems in the literature, such as (Chen and Sycara, 1998) were able to provide a combination of both services in a unified interface.

The WIFS system (Micarelli and Sciarrone, 2004) is another example of systems which offered both, a search service and a filtering service. WIFS operated on domain-specific corpora, where the system allowed users to search for academic publications in the field of computer science, as well as recommend publications to them based on their exhibited interests in the user model.

4.2.2 Personalisation Scope

Different approaches to personalisation are exhibited in PIR literature. One of the main aspects that distinguish between these approaches is the level of information detail on which they operate. In that sense, some systems may operate on aggregate usage information as exhibited in search logs, while other systems may take a more fine-grained approach by operating on the scope of individual user information.

In this survey, we classify systems with respect to the scope on which personalisation is performed into three categories: *individualised*, *community-based*, and *aggregate-level*. This section provides a discussion concerning these three scopes, and highlights the rationale behind this classification.

Individualised personalisation is when the system's adaptive decisions are taken according to the information about each individual user as exhibited in his/her user model (Steichen et al., 2011, Stamou and Ntoulas, 2009, Speretta and Gauch, 2005, Shen et al., 2005). The advantage of this approach is that the system becomes truly personalised as it addresses the needs of a specific user; taking into consideration this user's personal interests, goals, prior knowledge, context, language, or country. This approach may lead to higher satisfaction degrees for the user. Yet, as discussed earlier, one of the disadvantages of the individualised approach is the fresh start (a.k.a. cold start) problem where a new user has just registered with the system and there is very little or no information available about him/her to work with at that point. Moreover, another disadvantage of this approach is that users may feel that their privacy is compromised when systems maintain, and in some cases share, their personal information.

Among the challenges facing individualised personalisation, and perhaps personalisation in general, are *the effect of getting it wrong* and *risk vs. reward* (Wade, 2009, Vassiliou et al., 2003, Espinoza and Höök, 1995, de La Passardiere and Dufresne, 1992). As personalisation may sometimes "go astray", PIR systems have to take into consideration that delivering an inaccurate personalisation service can have a profound negative effect on the user's perception of the system. In other words, inferences made by personalised system about their users are essentially a "guess"; the harm of getting it wrong can be greater than the benefit of getting it right. Moreover, some personalised systems may attempt to perform a limited form of personalisation based on a limited, yet reliable, set of attributes and information available about the user. In spite of such limitation, the reward of such cautious form of personalisation may be considered sufficient –to a certain degree- to satisfy the users of the personalised service. Performing a more aggressive form of personalisation entails a higher degree of risk, yet it might not produce huge transformations in the personalised service; at which point, the reward may not be worth the risk. Thus, it is important for future PIR systems to investigate successful tradeoffs for delivering the right amount of personalisation in a careful manner. It is also important for PIR systems to bear in mind the effects that personalisation introduces to the interface of the system; users should not be surprised or disoriented by the changes incurred by the adaptive service. Therefore, designers of PIR systems should provide adequate balance between the usability of the interface and the potential effectiveness of the system.

Community-based personalisation takes a step further from individualised personalisation as information can be shared between the user models (Teevan et al., 2009, Sugiyama et al., 2004, Mei and Church, 2008). The system's adaptive decisions are then based on a wider scope of users, and not just a single user. This may be the case when a system groups the users into stereotypes (Brajnik et al., 1987) according to certain similarity criteria between their user models; at which point the system can judge the relevance of a certain document or item to a user based on the information of other users who belong to the same group in a collaborative manner. It can also

be the case when information from some user models is used to determine or alter the weights of interests in other user models.

The main consideration in community-based systems is how users are grouped together. This can be done in the following ways: (1) by manually pre-defining labelled groups in which users can join when they sign up with the system. These groups can be related to topics of interests (e.g. music, sports, etc.); (2) by using machine learning techniques (e.g. clustering techniques) to automatically form clusters of users based on similarity features between their user models (e.g. textual similarity of interest terms); (3) by including content information, in addition to user information, when processing user models for similarity. This is, for example, the case with content-based recommendation systems (Pazzani and Billsus, 2007); (4) by grouping users based on their demographic information (e.g. language, location, line of work, etc.).

An argument in favour of community-based systems is presented in (Mei and Church, 2008). The authors argue that too much personalisation may sometimes degrade retrieval effectiveness just as severely as no personalisation at all. The authors suggest that personalisation should sometimes "back off"[26] to a larger number of users, rather than a single user, when not enough individual user information is available. To this effect, the authors in (Teevan et al., 2009), investigated how a user's model can be augmented with information from groups of similar users with the aim of improving retrieval effectiveness. Different ways to form groups of users were investigated, including demographic information. The authors called their approach "groupisation" (as opposed to personalisation).

A good example that demonstrates the advantage of community-based PIR is presented by (Sugiyama et al., 2004) where two search systems are proposed. The first system operates on an individualised scope while the second operates on a community-based scope where interest weights are assigned in a collaborative manner. In a typical collaborative system, items that are rated as relevant by some users may be recommended to other users who share similar interests. However, in this system, collaboration is actually applied on the interests stored in the user models, where terms and weights from the models of some users can be assigned to the model of another user. The weight of a "borrowed" term is then computed by obtaining an average from the closest neighbouring (similar) models. The authors showed that the community-based system outperformed the individualised system in terms of retrieval accuracy and therefore they recommended performing personalisation in a collaborative manner.

Community-based personalisation has also been widely implemented in the area of recommender systems. As this survey paper is mainly concerned with personalised search systems, more information about recommender systems can be found in (Schafer et al., 2007).

Aggregate-level personalisation refers to the notion of a system that does not explicitly make use of a "per-user" model to represent users; at which case personalisation is guided by collective usage data as exhibited in search logs (Agichtein et al., 2006b, Gao et al., 2007, Sun et al., 2005, Smyth and Balfe, 2006). For example, this is the case when a system ranks documents based on the number of times a document was visited by users.

It may be argued that systems which perform personalisation at an aggregate level should not be regarded as "personalised" systems, since they do not make use of a user model and thus do not tailor their service to a specific "person". However, when considering search personalisation in a broader sense, the objective is retrieving documents that satisfy users' information needs; this may indeed start at the higher level of adapting to the needs of the majority of users. Adapting to the needs of the majority can give some kind of guidance as to what an individual user may need. For example, a common information need can be inferred if at some point in time a large number of users issued the same query and clicked the same results for it. Therefore, exploiting this inferred common need may serve in adapting similar future searches. Yet, the success of aggregate level systems rely heavily on their capability of accurately analysing and interpreting aggregate usage information so that they could deduce the true needs of the majority.

The classification of PIR systems in this manner can be regarded as a way to identify the scope on which each system operates, rather than an attempt to define completely distinct categories. In this sense, the three introduced scopes may be regarded as special (or more generalised) cases of each other, where the individualised scope indicates that personalisation is performed per "only one user", the community-based scope indicates "more than one user", and the aggregate-level scope indicates "all users treated as one".

4.2.3 Personalisation Approach

Personalisation in PIR systems can be achieved by query adaptation, result adaptation, or both. In other words, adaptation can be performed over the information that users send out or the information that they receive. In systems that offer a multilingual service to the users, the adaptation process may also include query and result translation (Oard, 2010, Oard and Diekema, 1998).

[26] (Mei and Church, 2008) used the term "back off" to refer to the notion of broadening the scope of the number of users on which personalisation decisions are based.

4.2.3.1 Query Adaptation

Studies, such as (Furnas et al., 1987), show that users may not always be successful in using representative vocabulary when locating objects in a system. Therefore, query adaptation attempts to expand the terms of the user's query with other terms, with the aim of retrieving more relevant results (Manning et al., 2008). In some cases, source query terms may be completely replaced by other terms. Query adaptation also involves altering the weights (significance) of the query terms when submitting them to the retrieval component of the system.

Six techniques are mainly used for obtaining terms for query expansion, which can be classified in terms of whether they are user-focused or not and whether they are implicit or explicit: (1) *processing the user model*: which involves the implicit selection of terms from the user model (Zhou et al., 2012, Chirita et al., 2007, Psarras and Jose, 2006, Shen et al., 2005); (2) *processing aggregate usage information*: which involves implicitly obtaining terms from the query logs and/or their associated clicked documents under the assumption that the majority of user clicks would be on documents that are relevant to the queries they submitted (Yin et al., 2009, Gao et al., 2007, Cui et al., 2003, Billerbeck et al., 2003); (3) *pseudo-relevance feedback (local analysis)*: which involves performing an initial retrieval round (that takes place behind the scenes) using the source query and then implicitly selecting expansion terms from the top N retrieved documents (or their snippets) under the assumption that most of them would be relevant to the source query (Leveling and Jones, 2010, Ogilvie et al., 2009, Cao et al., 2008, De Luca and Nürnberger, 2006); (4) *global analysis*: which involves the implicit selection of expansion terms from a thesaurus (e.g. WordNet), a knowledge source (e.g. Wikipedia), or a large corpus (based on co-occurrence statistics in this corpus) (Callan et al., 1995, Xu and Croft, 1996, Nguyen et al., 2008); (5) *(explicit) relevance feedback*: which requires that the user explicitly provide relevance feedback about a number of documents from an initial set of retrieved results where documents marked as relevant are processed to obtain expansion terms (Ruthven and Lalmas, 2003, Harman, 1992b, Salton and Buckley, 1990); (6) *interactive query expansion*: which involves user interface that allows the user to explicitly select expansion terms from a candidate list of terms suggested by the system (Bast et al., 2007, Ruthven, 2003, Efthimiadis, 2000, Harman, 1988). Table 4 shows a summarised classification of query expansion techniques. Furthermore, details of these techniques are analysed across a number of example systems below.

Table 4: classification of query expansion techniques

	User-focused		**Not user-focused**
	Individualised	**Aggregate**	
Implicit	User Model	Usage Information (Search Logs)	Pseudo-relevance Feedback (Local Analysis) & Global Analysis
Explicit	Relevance Feedback & Interactive QE		

Processing the user model. The work reported in (Chirita et al., 2007) is an example of systems where query expansion terms are obtained from the user model. The user's interests are inferred from his/her *Personal Information Repository*, which is the collection of their desktop documents, emails and cached Web pages. The first step towards the selection of terms for expansion involves identifying documents in the user's repository which contain the source query terms. Second, these documents are sorted in descending order, with respect to the source query terms, based on a modified term frequency (TF) weighting scheme. Third, query-focused summaries of the top K documents are produced. Fourth, all the terms of the summaries are extracted and are sorted according to document frequency (DF) weighting based on the number of summaries they appeared in. Finally, the top four terms are used as expansion terms for the source query. The authors also conducted a set of experiments to determine the adequate number of terms to use for expansion. They suggested that the decision should be dynamically based on query features such as query length (number of terms in the query), query scope (IDF score of the query), or query clarity (query ambiguity). The use of such features in dynamic decisions for query expansion is an emergent approach in the literature that is known as s*elective query expansion*.

A number of systems in recent literature demonstrated the use of user defined tags (a.k.a. folksonomy) in PIR, some of which were based on query expansion. For example, the authors in (Zhou et al., 2012) propose a system where an individualised user model is constructed based on terms extracted from the user's tags and bookmarks on del.icio.us. A statistical tag-topic model is created to deduce latent topics from the user's tags and tagged documents. This model is then used to identify the most relevant terms in the user model to the user's query and then use those terms to expand the query.

In some systems, query adaptation is performed by re-writing the whole query based on a set of rules maintained in the user model. For example, the authors in (Koutrika and Ioannidis, 2004) propose a rule-based query re-writing process for personalising structured search across a database of movie information. The system

substitutes the submitted query with multiple queries using a set of rules that govern the process. These rules are based on the user's individual movie preferences. The queries are connected together in a disjunctive manner using the "OR" operator. For example, if a certain user, who is known to prefer movies of type comedy, enters a source query that requests a list of movies in a certain year, then the system will replace the source query with a query that seeks a list of movies of type comedy in that year.

Processing aggregate usage information. The study carried out by (Yin et al., 2009) is an example of performing query adaptation based on aggregate usage information as exhibited in search logs (submitted queries and snippets of clicked results). As briefly discussed earlier, the authors are motivated by the idea that users may seek the same information but using different queries. The authors use machine learning techniques to learn the similarities between queries in the logs, and exploit these similarities for query adaptation. They argue that traditional pseudo-relevance feedback has two drawbacks: (1) processing the full text of feedback documents (as opposed to processing only the snippets) obtained in the initial retrieval round is considered an overhead to the system; (2) not all feedback documents are guaranteed to be relevant, thus, some bad terms might be extracted from them (i.e. terms that may be harmful to retrieval effectiveness). The authors address these two issues by: (1) using the text of snippets instead of documents, which is further supported by the idea that, before clicking on results, users actually examine the result snippets in order to get a hint of how far a document is relevant to their information need; and (2) only selecting snippets that exceed a certain score threshold, where scores are assigned to snippets based on their rank and their similarity with the source query and similar target queries in the logs.

Query adaptation based on usage information is also investigated in (Gao et al., 2007). Furthermore, the authors extended into the multilingual dimension. Given a source query in a certain language, the system obtains related queries from other languages by exploiting multilingual search logs. This technique is also known as *Cross-Lingual Query Suggestion (CLQS)*. CLQS can be viewed as a technique that combines query translation and adaptation into a single process, where the formulation of the source query is expanded (or replaced) with common formulations of similar queries exhibited in the multilingual logs. The authors use machine learning algorithms in order to learn a cross-lingual similarity function that determines the degree of similarity between a query in the source language and another query in the target language. The process of determining cross-lingual similarity between two queries involves several features of monolingual similarity between the first query and the translation of the second query.

Pseudo-Relevance Feedback (PRF). Query expansion using PRF techniques (a.k.a. local analysis and blind relevance feedback) was subject to wide research in IR literature. The main issue with PRF is that the process is prone to noise caused by the fraction of feedback documents that are not relevant to the query, which may degrade retrieval effectiveness. This issue was addressed by a number of studies in the literature in a non-user-focused manner. For example, the research reported in (Cao et al., 2008) and (Leveling and Jones, 2010) investigate how automatic classification techniques can be used to identify good and bad terms for query expansion. Several features were used for the classification process, such as *term distribution* (the frequency of terms appearing in the feedback documents), *term specificity* (the number of documents in which the term appears in the entire collection), *term co-occurrence* (co-occurrence of query terms with candidate expansion terms in the collection or in a thesaurus), *term proximity* (the number of terms separating co-occurring terms), and *term string distance* (the Levenstein distance between terms, which may detect terms that are morphological variants of each other).

In multilingual PIR literature, the PRF approach is often used to expand the query in two ways, namely: *pre-translation* query expansion and *post-translation* query expansion. Pre-translation expansion involves expanding the source query (in its source language) by terms obtained from a retrieval round performed over documents of the source language. Afterwards, the source query is translated into one or more target languages using a translation mechanism (e.g. bilingual dictionaries or machine translation systems). Post-translation expansion is then applied to expand the translated query (in its target language) by terms obtained from another retrieval round that involves documents of the target language. The authors in (McNamee and Mayfield, 2002) discussed this process and mentioned that pre-translation expansion helps in improving translation by increasing the terms that are used as input to the translation module. This helps in overcoming any limitations in the translation method or limitations caused by Out of Vocabulary (OOV) terms[27]. The authors also mentioned that post-translation expansion helps in overcoming any output errors that may be exhibited in the terms produced by the translation module. Moreover, a comparison between the two approaches was carried out and the authors concluded that combining both approaches significantly improved retrieval effectiveness more than using any one of them alone. It was also concluded that pre-translation expansion contributed more than post-translation expansion towards the observed retrieval improvement.

The work reported in (Cao et al., 2007) is another example of studies which performed non-user-focused query expansion in a multilingual fashion. In the study, Markov Chains was used to combine query translation

[27] Out of vocabulary terms are new emerging terms that existing translation system may be unaware of.

and query expansion. Similar to the abovementioned CLQS work of (Gao et al., 2007), the process involved expanding the source query with semantically related terms in a different language. However, this was based on global analysis rather than local analysis.

Some other systems in the literature, such as the system presented in (Ambati and Uppuluri, 2006), investigated improving cross-lingual IR by exploiting search logs but with a different focus; instead of search personalisation, the main objective of the system was to improve translation methods.

Global analysis. An example of systems where expansion terms were implicitly obtained using global analysis techniques is the INQUERY system (Callan et al., 1995). In INQUERY, query terms can be expanded with other semantically related terms. This is achieved by grouping all terms in the collection into noun groups, where each noun group consists of a phrase (up to three adjacent terms), along with all the terms that co-occur with that phrase in a pre-define window size (e.g. within the distance of three sentences). TF and IDF are then used to weight the importance of the terms in the noun groups. Whenever a query is submitted to the system, the query terms are used to identify the appropriate noun groups. Then, related terms that exceed a certain weight threshold are selected from those noun groups and are used for expanding the source query.

Local and global analysis techniques are implicit (automatic) techniques, but are not user-focused. Opposite to those two techniques are relevance feedback and interactive query expansion, which are explicit feedback techniques that are user-focused (since the user is involved in the process).

Relevance feedback. In the relevance feedback approach to query expansion (a.k.a. explicit relevance feedback), users are asked to provide feedback about the relevance of result documents to their information need (Ruthven and Lalmas, 2003, Harman, 1992a, Salton and Buckley, 1990). This feedback can either be positive or negative, for example by marking documents on a binary scale of relevant vs. irrelevant. The system then analyses the feedback documents and modifies the source query accordingly. The new query is then used to retrieve documents that are similar to the positive examples, or filter out documents that are similar to the negative examples.

Relevance feedback is an iterative process, where users can keep providing feedback for every new result list provided to them. The process may eventually converge after a number of iterations (i.e. no more significant enhancements in the precision of the retrieved result list).

Although in relevance feedback there is no user model created (in the formal sense), the process can be considered personalised because the user is involved in specifying what is relevant and irrelevant to him/her. Furthermore, in a search session, the adapted query itself can be roughly regarded as a representation of the user's short-term (ad-hoc) interests with respect to the current information need.

Interactive Query Expansion (IQE). The IQE approach encompasses more involvement for the user (Bast et al., 2007, Ruthven, 2003, Efthimiadis, 2000, Harman, 1988). In IQE, the system suggests a set of terms, from which the user can select the ones to be used for expanding the query. An important initial step for IQE is that the system automatically produces a ranked list of candidate terms, a subset of which is presented to the user. These terms can be obtained from documents which have been marked relevant by the user or from a thesaurus, where terms that are semantically related to the query terms are identified.

Several studies conducted comparative evaluations between interactive and automatic query expansion (i.e. IQE vs. explicit relevance feedback) (Ruthven, 2003, Magennis and van Rijsbergen, 1997). It was shown that interactive techniques can sometimes be more effective than automatic techniques. However, it was also concluded that this is not always the case because IQE depends on other human factors like the degree of user's prior knowledge of the domain and the GUI of the application used to present the terms to the user.

4.2.3.2 Result Adaptation

The other common approach to search personalisation is result adaptation. Adaptation of result lists can be performed by result scoring, result re-ranking, or result filtering. Result re-ranking takes place after an initial set of documents have been retrieved by the system, where an additional ranking round is performed to re-order documents based on certain adaptation aspects (e.g. displaying certain documents at higher ranks in the result list based on the user's interests). Result filtering can be considered as a special case (or a step further) of result re-ranking, where after the result list is sorted in descending order of relevance scores, results that fall below a certain threshold are not displayed to the user. Result scoring involves incorporating adaptation features directly in the primary scoring function of the retrieval component of the system

Result re-ranking and result filtering. The result re-ranking approach is commonly used in many PIR systems. A good example is the MiSearch system (Speretta and Gauch, 2005), which is wrapped around Google search. Following a user's search, the results and snippets[28] retrieved from Google are passed to the result re-ranking component. The snippets are then analysed using text classification techniques. This is performed in

[28] Snippets, which are a form of summary or surrogate of a document, are regarded by several studies in the literature as query-focused document space representations. Thus, several studies opt to use the text of snippets instead of the full text of documents when processing result lists.

order to deduce their conceptual content so that they can be assigned under appropriate ODP categories. After the concepts of the snippets have been deduced, they are compared to the concepts in the user model using cosine similarity. The results are then re-ranked in descending order of the conceptual similarity score. Several modes of result re-ranking were tested in MiSearch, where the conceptual similarity ranking was combined with the original ranking of Google. An *alpha* factor was used to specify a certain weight for the conceptual ranking in relation to the original ranking. The value of alpha ranged between zero and one, where a value of zero led to completely ignoring the conceptual ranking (i.e. no adaptation applied), and a value of one led to completely ignoring the original ranking. Experiments with different values of alpha showed that a value of one achieved the highest improvements for retrieval effectiveness. Therefore, it was concluded that result re-ranking can be an adequate tool for adapting to the user's information needs.

Several systems in the literature, in which the retrieval components were wrapped around well-known search engines, do not apply the result re-ranking process on the full set of results retrieved from the search engine (which could be hundreds or thousands of documents). In fact, the process is often limited to the top N documents from the result list. For example, the authors in (Speretta and Gauch, 2005) decided to limit the re-ranking process to the top ten retrieved documents. This decision was based on some experiments that they carried out which involved a number of users using a non-personalised search system. The results of those experiments showed that 94% of users' clicks were on the top three results in the result list. To this end, the authors further investigated the effect of the *position bias phenomenon*[29]. The phenomenon was investigated by randomising the ranks of the top ten results retrieved from Google before displaying them to the user. The results of the investigation showed that the top three results of Google search only received 46% of the users' clicks when they were presented in a randomised order within the list of ten displayed results. The authors concluded that users are affected by the presentation order of the results and thus continued to randomise the top ten results retrieved from Google in their baseline system.

The authors in (Stamou and Ntoulas, 2009) also propose a system where personalisation is performed by re-ranking results that are retrieved from Google. However, as briefly discussed earlier, a notable aspect about their re-ranking process is that the weights of user interests are not only based on historical evidence (long-term interests) but also on evidence from the current search at hand (short-term interests). The authors implement this through a number of steps. First, the user's past conceptual preferences are identified by examining past queries and their corresponding clicked documents and then mapping them to concepts based on a merged ontology (see Section 3.2.2). Second, the user's current conceptual preference is identified by examining the current query. In other words, the system attempts to determine the user's current information need given a new query that has just been submitted to the system (where no documents have been clicked for it yet). If the same query was found in the logs, then the conceptual preferences that were determined for it in the first step are used. Otherwise, the system attempts to determine the similarity between the query and the documents listed under each of the ontology concepts (pre-classified). It might make sense to perform the re-ranking process only according to the identified conceptual interests of the current query (since it is the given evidence of the current information need); however, the evidence from the current query is weak evidence to some extent because it was supported only by a few terms in a single query. To this matter, the authors determine the degree of user's interest in conceptual topics by computing a combined value of historical evidence and current evidence. In this third step, an *alpha* value is used to explicitly specify weight for historical evidence in relation to current evidence. Lower values of alpha indicate a *conservative* approach that favours historical evidence (from past queries), while greater values of alpha indicate an *aggressive* approach that favours current evidence (from the current query at hand). Fourth, the retrieval process takes place, where the current query is submitted to Google and corresponding results are retrieved. Fifth, the conceptual topics present in the documents are determined with respect to the ontology, and are assigned weights. Finally, for each document, a relevance score is computed. The computed score involves the value obtained from the third step (user's conceptual interests) and the value obtained from the fifth step (documents' conceptual weights). The results are then re-ranked in descending order of the computed score.

Social data was also used for result re-ranking in PIR. For example, the authors in (Vallet et al., 2010) investigated how the ranking of search engine results can be improved with respect to users if the users' social information is taken into consideration. This was achieved by re-ranking results retrieved from Yahoo search engine based on a user model comprising tags extracted from the user's participation on the *del.icio.us* social bookmarking website. Users and documents were both represented by associated tags, where the tag distribution across 2000 users and about 160,000 documents were considered. A similar approach was also explored in (Noll and Meinel, 2007) where the system performed re-ranking of Google search results based on social bookmarks and tags harvested from del.icio.us. An advantage argued in the two studies is that the approach is independent

[29] Position bias phenomenon (a.k.a. trust bias) is the tendency of users to "trust" the ranking of a search engine and thereby click on the higher ranked documents even though more relevant documents may exist at lower ranks.

of a specific search engine, and thus any search engine can be used. However, the data sparsity problem poses a challenge to this approach as not all Web pages returned by search engines are tagged in the del.icio.us dataset.

Result filtering can be considered as an additional step that takes place after re-ranking the results with respect to the user's interests. An example of systems which employ result filtering is the WIFS system (Micarelli and Sciarrone, 2004). WIFS offers two services to its users: Web search and Web filtering. The filtering service autonomously retrieves Web pages and filters them according to the user's interests. The pages are first sorted in descending order of relevance scores and then pages that fall below a certain threshold are discarded.

The aforementioned result adaptation systems operated on an individualised scope. Opposed to this, are other approaches in the literature where result adaptation is performed on an aggregate-level scope. For example, in the I-SPY system (Smyth and Balfe, 2006), personalisation is collectively based on the deduced interests of the majority of users as exhibited in search history. As briefly discussed earlier, usage information in I-SPY is represented in a matrix that keeps track of each query and the number of times a corresponding document was clicked for that query. In order to re-rank results for a new search, the current query is checked for similarity against all the past queries recorded in the matrix. The similarity between queries is computed using term-based similarity measures which determine the degree of textual similarity between them. The outcome of this procedure is a list of candidate queries (ones which passed a certain similarity threshold). The click frequencies of the documents that were associated with the candidate queries are obtained from the matrix. For each document, the multiple frequencies that come from considering the multiple candidate queries are combined using a normalised weighted relevance metric which combines relevance scores for document-query pairs[30]. The new relevance scores are then used to re-rank the documents for the current query at hand.

An innate characteristic of the result re-ranking process is that two rounds of computation take place. In the first round a function is used to score the relevance of the documents with respect to the query in a pure IR manner. In the second round, another function is used to score the documents (or the top N documents) with respect to the user. Research studies which depend on an external retrieval component (i.e. where a search engine other than their own is used, such as Google, Bing[31], or Yahoo), they are obliged to work with the extra round of re-ranking, since they have no control over the factors of the first scoring function.

Result scoring. A number of other systems in the literature for which a retrieval component was implemented (i.e. ones that did not depend on one of the existing search engines), followed another approach for result adaptation which is result *scoring*. In result scoring, only one round of scoring is performed. The adaptive factors (variables) that are used to score the documents according to the users' needs are combined together with the IR factors in the original scoring function. For example, in (Agichtein et al., 2006a) result re-ranking and result scoring were both implemented and compared to each other. The two approaches operated on the scope of aggregate usage data. In the first approach, the one based on result re-ranking, the authors used machine learning algorithms to learn a function for relevance weighting based on implicit feedback features from the search logs. However, the rank orders that were obtained from the original scoring round were not totally ignored as they were combined with the ranks produced by the new learnt function. In other words, the first approach "honoured" the original scoring method by using an *additional* re-ranking function that combined the rank orders obtained from both, the original method and the new method. Furthermore, a factor was used to specify a certain weight for the ranks obtained from the new implicit feedback method in relation to the original method. This allowed control over the degree of bias towards the new method. In the second approach, which is based on result scoring, the authors included the implicit feedback features together with the original features in the main scoring function of the retrieval component. This allowed avoiding the extra scoring round. The experiments carried out by the authors showed that the second approach was more effective, thus they recommended performing personalisation by result scoring, rather than by result re-ranking.

Another technique for result scoring is the topic-sensitive PageRank algorithm (Haveliwala, 2002). In this algorithm, the system assigns multiple PageRank scores (Brin and Page, 1998) to each document, where each score is calculated with respect to one of ODP categories. In other words, each document is given multiple scores with respect to its popularity and its similarity with each ODP category. This information comes into play when a query is submitted, where the query's topic is used to identify which category score for a document will be used when ranking the documents. This work was extended by (Qiu and Cho, 2006) where the authors incorporated individual user interests into the process. This was done by exploiting clickthrough information to construct user models that are based on concepts from ODP. The evidence from these user models (i.e. the conceptual interests) was then factored into the equation, together with evidence from the deduced query's topic, to select the most appropriate document score to use in the ranking process.

4.2.3.3 Query Adaptation & Result Adaptation

[30] The relevance scores were combined by calculating the weighted sum of each relevance score, and then obtaining the average. The weighing was based on the degree of similarity between the source query and candidate queries.

[31] http://www.bing.com

A number of systems in the literature employ both, query adaptation and result adaptation. For example, in the *UCAIR* system (User Centred Adaptive Information Retrieval) presented in (Shen et al., 2005), the authors argue that the two main aspects of personalised search are: the user's interests and the search context (i.e. query disambiguation). The authors focus on modelling the user's short-term interests, in an approach called eager implicit feedback. In this approach, the current query's context is deduced using evidence from the immediate previous query (within the search session) and the results clicked for it. To determine if two successive queries are related, the system performs two searches; one with the previous query and one with the current query. The retrieved result lists for the two queries (50 results for each) are then compared to each other by checking how many terms are common between the titles and snippets of the two lists. If the two queries are related (based on a textual similarity threshold), the current query is expanded using terms from the short-term user model created for the previous query. Following the submission of the adapted query to the retrieval component, the retrieved result list is re-ranked based on the user model. The user model is updated in a live manner whenever the user clicks on a result from the displayed list. Based on the updated model, further result re-ranking takes place if the user clicks on the *next* link (i.e. live re-ranking is performed when the user requests to see the next results page).

Another example is the Outride system (Pitkow et al., 2002) where query and result adaptation are performed in an individualised manner. Query adaptation is performed by expanding the source query with the most relevant concepts from the user model. For adapting the results, the documents are re-ranked by comparing the documents' titles and metadata with the concepts in the user model using Vector Space methods.

A rather different approach for query and result adaptation was presented in (Liu et al., 2004). In one of the proposed systems, a vector-based user model of conceptual terms was maintained. The conceptual interests were based on *Google Directory*[32], which is a Web taxonomy that is based on ODP. Google Directory provides a facility to specify the category to which a query is to be submitted. Query adaptation was not performed by expanding the query terms, rather, by specifying the category of the query (e.g. Health, Arts, etc.). In other words, the system attempts to infer the concepts related to the submitted query and then use these concepts to provide context information to the retrieval system when submitting the query. An automatic and a semi-automatic approach were used to deduce candidate conceptual categories to which the query may belong. In the automatic approach the query terms were mapped into candidate concepts and then these concepts were scored against the concepts in the user model. The top N categories (up to three) related to highly scored concepts are then identified and specified when the query is submitted. In the semi-automatic approach, an additional step takes place, which is that candidate categories are shown to the user (three at a time). The user is then allowed to select the appropriate categories related to the query. After the categories are identified, the query is actually submitted multiple times for retrieval; once without specifying any categories, and one time for each of the identified categories. This leads to the retrieval of multiple result lists. The system then performs result adaptation by ranking and merging the results into a single list. A weighted voting based algorithm was used, where results that appeared on more than one list were favoured.

The advantage of this technique, besides catering for the user's long-term interests, is that it also accounts for the possibility of ad-hoc queries. This is because the fact that multiple result lists are sought by the system, one of which is based on the non-adapted version of the query, allows for some diversification in the kind of results presented to the user. This is opposed to other systems in the literature where only one result list is sought based on an inferred user interest; an approach where if the system's guess about the query's topic is wrong, the result list might be dominated by results that are irrelevant to the user's current information need. This approach is related to an approach in IR literature known as *result diversification*, where retrieval systems deliberately diversify the set of results presented to the user, especially on the first page of results (Santos et al., 2010, Minack et al., 2009, Gollapudi and Sharma, 2009). The rationale behind this approach is to guarantee that users with random or different intents will find at least one relevant document to their information need in the result list. Furthermore, this approach encourages users with explorative behaviour to learn more about diverse topics, which they may have not learnt about otherwise.

4.3 Summary and discussion

In summary, this section provided an analysis of the personalisation approaches exhibited in several systems in the literature. The analysis focused on the core process of executing personalisation using different techniques for query adaptation and result adaptation. Table 5 presents a summary of the analysis carried out in this section, along with some example systems.

[32] http://www.google.com/dirhp

Table 5: summary of personalisation approaches

Personalisation Approach	Personalisation Scope	System Type	Example Publications
Query Adaptation (query expansion using terms from user model)	Individualised	Web search	Zhou et al. 2012, Chirita et al. 2007, Psarras and Jose 2006
Query Adaptation (query rewriting)	Individualised	Structured search on movies database	Koutrika and Ioannidis 2004
Query Adaptation (query expansion using terms from query logs or generated thesaurus)	Aggregate-level	Web search	Yin et al. 2009, Cui et al. 2003
Query Adaptation (query suggestions using similar queries from query logs in other languages)	Aggregate-level	Cross-language Web search	Gao et al. 2007, Ambati and Uppuluri 2006
Result Adaptation (result re-ranking)	Individualised	Web search	Vallet et al. 2010, Noll and Meinel 2007, Speretta and Gauch 2005, Stamou and Ntoulas 2009, Teevan et al. 2005, Ruvini 2003, Pretschner and Gauch 1999
Result Adaptation (result re-ranking)	Individualised	Search and recommendations on computer science literature	Micarelli and Sciarrone 2004
Result Adaptation (result filtering and re-ranking)	Individualised	News	Katakis et al. 2009, Chen and Sycara 1998
Result Adaptation (result re-ranking)	Community-based	Web search	Teevan et al. 2009, Sugiyama et al. 2004
Result Adaptation (result re-ranking)	Aggregate-level	Web search	Smyth and Balfe 2006, Sun et al. 2005
Result Adaptation (result scoring)	Individualised	Web search	Qiu and Cho 2006
Result Adaptation (result scoring)	Individualised	Web search and document recommendations	Stefani and Strapparava 1999
Result Adaptation ((1)result scoring & (2)result re-ranking)	Aggregate-level	Web search	Agichtein et al. 2006
Result Adaptation (re-structuring and tailoring content of results into a hypertext presentation)	Individualised	eLearning (search on domain-specific corpora for education purpose)	Steichen et al. 2009
Query & Result Adaptation	Individualised	Web search	Shen et al. 2005, Liu et al. 2004, Pitkow et al. 2002
Query & Result Adaptation	Individualised	Customer support (search on domain-specific corpora for technical support)	Steichen et al. 2011

It is noted that individualised user models in PIR systems were mostly used for result adaptation compared to a relatively fewer number of systems where such models were used for query adaptation. Personalised query adaptation was often based on aggregate usage information as exhibited in search logs. A broader consideration of IR literature reveals that the majority of studies which investigated query expansion were based on approaches that are not user-focused, mainly PRF.

The authors in (Cui et al., 2003) compared query expansion based on search logs to query expansion based on PRF and showed that the former leads to higher retrieval effectiveness. A rather mixed approach of the two was used in (Shen et al., 2005) where the system inferred the user's current information need in a search session and expanded the query based on the inferred short-term interest. This was done by examining the snippets of the top result retrieved in an initial retrieval round using the given query (as in typical PRF), as well as examining the immediately preceding query in the same session and snippets of the clicked results associated with it (recent search history).

When considering the systems in multilingual PIR literature, it is noted that query adaptation is often employed in a non-user-focused manner. PRF is the most common technique used for expanding the query before and after translation. A challenge for future multilingual PIR systems would be to investigate the construction of individualised user models that cater for the needs and interests of a user with respect to multilingual or cross-lingual search.

5 Evaluation approaches

5.1 Overview

This section of the survey sheds light on evaluation of PIR systems. Although evaluation is not literally a stage in the personalisation process itself, it was nonetheless important to include it as a separate section in this survey. This is because it shows the effectiveness and the efficiency of the different personalisation approaches and mechanisms discussed in the previous three sections.

Four criteria are used to derive the discussion in this section: the aspect of evaluation targeted by the system, the evaluation metric or instrument used for evaluation, the datasets used in the experiments, and the experimental setting for evaluation. An overview of these four classification criteria is given below.

- **Aspect of evaluation:** the first criterion in this section is concerned with what is being evaluated in the system. Different aspects of a PIR system are subject to evaluation:
  1. *System performance*, which is usually concerned with measuring retrieval effectiveness (Yin et al., 2009, Chirita et al., 2007, Smyth and Balfe, 2006, Teevan et al., 2005). The advantage of evaluating PIR systems in terms of retrieval effectiveness is that it stands out as a well-defined quantitative comparison across different systems. However, a considerable drawback of such evaluation is that it does not truly reflect the personalisation aspect of the system (i.e. more system-focused than user-focused).
  2. *Usability*, which is concerned with the user's perception of, and satisfaction with, the system. Evaluating the usability of a system may be required as a pre-condition to further experimentation in order to make sure that poor subjective scores are not the result of poor system design. This involves evaluating the look-and-feel and ease-of-use of the baseline system with respect to the users and according to design standards (Tullis and Albert, 2008, Krug, 2005). Furthermore, in the area of personalised systems, usability evaluation extends beyond the notion of GUI evaluation to studying the impact of personalisation on the user experience. This involves investigating how far the introduction of personalised features into the system is successful in improving the user's perception of the service and in assisting the user in fulfilling the intended tasks (Steichen et al., 2009, Conlan and Wade, 2004, Micarelli and Sciarrone, 2004, Pitkow et al., 2002). As personalisation is concerned with adapting to the user's needs, the benefit of evaluating usability of a personalised system is that it pays attention to these needs and measures the degree of user satisfaction with respect to the adaptive service. A concern about this type of evaluation, however, is that it is hard to standardise across different systems and that it is subject to user bias.
  3. *User model accuracy*, which is concerned with how well the system is able to implicitly infer and represent user information, such as the user's interests (Pretschner and Gauch, 1999). The advantage of this type of evaluation is that it is user-focused. However, this approach is not very common in PIR literature, perhaps because many PIR systems implicitly gather a large amount of information about the user and it would be a tedious task to get the user to examine it.
- **Evaluation metric or instrument:** the second criterion is concerned with the different quantitative and qualitative metrics or instruments used for evaluation. Respectively following on the aspects of evaluation discussed in the previous criterion, the metrics can be as follows:
  1. Retrieval effectiveness can be quantitatively measured in a number of ways using well-known metrics in the IR community (Baeza-Yates and Ribeiro-Neto, 2011, Manning et al., 2008): (1) *Precision*, which is the number of retrieved relevant documents over the total number of retrieved documents; (2) *Recall*, which is the number of relevant documents that are retrieved over the total number of known relevant documents in the document collection; (3) *Precision at K*, which measures the fraction of retrieved relevant documents within the top K retrieved documents; (4) *Recall at K*, which measures the fraction of retrieved relevant documents within the top K documents over the total number of relevant documents in the document collection; (5) *Mean Average Precision (MAP)*, which is a single-valued metric that serves as an overall figure for directly comparing different retrieval systems. It is the average Precision at K values computed after each relevant document has been retrieved for a query, where the mean of all these averages is calculated across all the test queries; (6) *Normalised discounted Cumulative Gain (NDCG)*, which is s a precision metric that is designed for experiments where documents are judged using a non-binary relevance scale (e.g. highly relevant, relevant, or not relevant). It gives higher scores for more relevant documents being ranked higher in the ranked list of results; (7) *R-precision*, which measures precision with respect to a given number of documents that are known to be relevant; (8) *11-point Precision*, which measures the precision of retrieved results at 11 fixed values of recall; (9) *F-Measure*, which is the weighted harmonic mean of precision and recall; and (10) *Break-even Point*, which is the determining the point at which precision equals recall;

2. Usability can be qualitatively evaluated using usability questionnaires (Brooke, 1996, Harper et al., 1997) or quantitatively evaluated by measuring the user's performance in fulfilling certain tasks using the system, for example by keeping track of the time and number of actions needed to complete the task.
3. User model accuracy can be qualitatively evaluated by interviews or questionnaires that cover how accurate the users think the user model was able to depict their interests and the weights of those interests.

- **Datasets:** the third criterion is concerned with the datasets used in the experiments. In PIR, two kinds of datasets are used: document collections and search logs. Document collections (corpora) are datasets that comprise a large number of documents in one or more languages. Examples of these are the collections provided by TREC[1], CLEF[2], and NTCIR[3], which are widely used in the IR community. These collections, together with a set of manually selected information needs[4], are used as a test-bed for comparing retrieval and adaptation algorithms developed by researchers in the community. Not all experiments in PIR are conducted on standard test collections; several experiments were conducted on open Web corpora using retrieval components that are wrapped around live Web search engines. The advantage of this approach, over the use of standard test collections, is that the experiments are not over-fitted on the domain or characteristics of a specific test collection. However, the disadvantage of this approach is that it becomes hard to perform apples-to-apples comparisons between the results of different studies in the literature. Search logs, as discussed earlier, are datasets that comprise the history of user interactions with a system over a period of time. Search logs serve a very important role in PIR experiments since they hold usage information (aggregate or per user) which is a crucial element in search personalisation. When this information is analysed and represented in user models it becomes the basis of user-focused adaptation algorithms. Larger datasets of search logs can contribute towards more reliable results.
- **Experimental setting:** the fourth criterion in this section is concerned with the experimental setup put in place for evaluation. Some studies conduct experiments in a controlled setting that involves a small number of users and tasks (Stamou and Ntoulas, 2009, Steichen et al., 2009, Speretta and Gauch, 2005). The advantage of such setting is that it allows establishing control groups and conducting a richer evaluation of usability aspects. On the other hand, other studies base their evaluation on a large amount of data drawn from a realistic setting (e.g. well-known Web search engines) (Yin et al., 2009, Gao et al., 2007, Agichtein et al., 2006a). Large-scale experimental settings have the advantage of result reliability.

The following section presents a detailed review of the evaluation carried out by several systems in the literature.

5.2 Review

The discussion in this section starts with systems where experiments targeted the evaluation of the system's performance in terms of retrieval effectiveness (i.e. IR-style evaluation of retrieval precision or recall). The discussion then moves on to systems where experiments targeted the evaluation of other aspects of the system such as usability or user model accuracy (i.e. AH-style evaluation, which is more user-focused).

In the area of IR research, a common quantitative evaluation approach is to compare the effectiveness of a proposed search system to a baseline search system. For example, for the I-SPY system (Smyth and Balfe, 2006), the authors evaluated the precision and recall of their experimental PIR system against a non-personalised version of the system. The underlying retrieval component in both systems comprised a meta-search engine that performed search over open Web corpora by collating results from several well-known Web search engines. The experiments were conducted in an in-lab experimental setting that involved 92 users. The users were divided into two groups; one for training (45 users) and one for testing (47 users). In the first group (training group), each user was assigned 25 information needs to satisfy using a Web search interface (live meta search engine). The users were free to formulate any number of queries that described the given information need. The interactions of the first group with the baseline Web search system were logged and used for training I-SPY. The logs contained the submitted queries and the clicked results. The logged information was then used in two ways: (1) to create ground-truth relevance judgements, where the relevance of clicked documents with respect to the queries was manually assessed on a binary scale (i.e. relevant vs. irrelevant); and (2) to generate the hit matrix (i.e. to train the personalised system based on click frequencies on result documents). Users in the second group (the test group) used the personalised Web search system and were also given 25 information needs to fulfil using any number of queries. It is to be noted here that the approach of dividing the users into two

[1] Text REtrieval Conference: http://trec.nist.gov/
[2] Cross-Language Evaluation Forum: http://www.clef-campaign.org/
[3] NII Test Collection for IR Systems: http://research.nii.ac.jp/ntcir
[4] Test queries that are associated with each collection

groups was applicable because the baseline system did not perform personalisation in an individualised manner, but rather in an aggregate manner based on general usage history. Thus, it was not a must that the same group of users be subject to both systems. Several IR metrics were used for retrieval evaluation based on the ground-truth relevance judgements generated earlier. The results show that the I-SPY system achieved significant improvements between 117% and 266% over the baseline system using the Precision at K metric, where K varied between 5 and 30. The results also show improvements between 138% and 280% using the Recall at K metric with the same range of values for K. Moreover, the F-measure metric was also used to evaluate the personalised system where the results showed improvements up to 380% over the baseline system for K = 30. It is worth mentioning here that majority of studies in PIR literature report precision or recall improvements between 10% and 50% over a baseline. It is not very common to achieve improvements over 100% such as in this study. Besides the possibility that their proposed system was a very successful one, another viable possibility is that the baseline system used in the comparisons was a weak one.

Different personalisation approaches were evaluated in (Sugiyama et al., 2004), where three personalised systems were compared to each other: (1) an existing baseline system, which adapted results using a short-term user model that is based on explicit relevance feedback from users; (2) an experimental system which adapted results based on implicit construction of short-term and long-term user models; (3) a second system which extended the first experimental system by allowing long-term user models to borrow weighted interests from each other in a community-based manner. All systems used a retrieval component that performed search over open Web corpora using a Google search wrapper. The experiments were performed in an in-lab setting that involved 20 users using the baseline Web search system over a period of 30 days, where each user was assigned 50 queries. The queries were prepared by the authors based on 50 topics (information needs) that came from *TREC*[1] evaluation campaigns. The users were asked to provide relevance judgements for the top 30 results retrieved for each query. These relevance judgements were used as the ground-truth judgement for comparing the three systems. The user models for the two experimental systems were generated based on the search history accumulated from using the baseline system. The two systems were then run as an automated simulation using the same queries. The R-precision metric was used to evaluate the retrieval effectiveness of the three systems. The results showed that first proposed system (based on individual approach) outperformed the baseline system with a 28% improvement and that the second proposed system (based on collaborative approach) outperformed the baseline with 37% improvement; thus showing that community-based personalisation can be more effective than individualised personalisation.

A similar in-lab setting was also used in (Teevan et al., 2005) to compare a personalised system to a baseline system. The retrieval component was wrapped around MSN Search. Relevance judgements were performed in a non-binary manner, where documents were judged on a three-level scale: highly relevant, relevant, or not relevant. The results showed that their personalised system significantly outperformed the baseline system with an improvement of 24% in NDCG.

As discussed earlier, systems which implicitly infer the user's search interests can harvest terms from the queries that the users submitted, the documents that they clicked on, or the snippets of the clicked documents. With respect to these different sources, an interesting study was reported in (Speretta and Gauch, 2005) in which a system where terms are extracted from queries was compared to a system where terms are extracted from snippets of clicked documents. The retrieval effectiveness of the experimental systems was evaluated against a non-personalised system. All systems used a retrieval component that was wrapped around Google. The experiments involved six users who used the baseline system for their own daily searches (i.e. users' own information needs) over a period of six months. The baseline system randomised the top ten Google results before displaying them to the user. All the users' interactions with the system were logged. From the logs, 47 queries per user were extracted, where 40 queries were used for training the personalised systems (i.e. for constructing the user model either from the text of the queries or the text of the snippets), 5 queries were used for testing a number of parameters of the system (fine tuning), and 2 queries per user were used for validating the system. A notable difference between this study and other studies is that relevance judgments were not based on manual assessments. Rather, an implicit approach was used where documents that were clicked by users while using the baseline system were deemed as relevant. However, this approach only produced a very small number of judged documents with respect to test queries. A simple rank scoring measure was used to evaluate retrieval where each system was evaluated according to the rank it assigned to the relevant documents of the query (i.e. in which position in the list did the system place the few documents that were implicitly judged as relevant). The results showed that both proposed systems were equally capable of improving retrieval over the baseline system, with a slightly higher improvement for the snippet-based system (34%) compared to the query-based system (33%).

A number of studies in the literature, especially ones that were carried out by research teams who are affiliated to major search engine companies, conducted their experiments on large-scale datasets. This is

[1] Text Retrieval Conference: http://trec.nist.gov/

compared to the relatively much smaller datasets that are generated by in-lab experimental settings. For example, in (Agichtein et al., 2006a) a realistic experimental setting was arranged, where a large-scale dataset of usage data was obtained from a well-known search engine[1]. The dataset comprised search logs recorded for user interactions with the search engine over a period of eight weeks, which contained over 1.2 million unique queries and over 12 million user interactions (post-search actions, including clicking on results). A random sample of 3,000 queries was drawn from the dataset and was used for the experiments. For each of the queries, 30 result documents on average were manually judged for relevance. The authors noted that one of the characteristics of a realistic experimental setting is that implicit feedback can be noisy (e.g. inconsistent or incomplete). Nevertheless, they argued that this characteristic actually counts towards the reliability of the experimental results. Several personalised systems, in addition to the baseline system, were tested against each other. Personalisation was performed on an aggregate usage level where the systems made use of part or the entire evidence of implicit feedback. The systems mainly involved two personalisation approaches: result scoring and result re-ranking. Three metrics were used for retrieval evaluation: Precision at K, NDCG, and MAP. The experiments showed that: (1) the exploitation of implicit feedback information is useful in realistic Web search environments, despite the existence of noise in the recorded logs; (2) result scoring, where implicit feedback features are incorporated into one scoring function together with other existing scoring features, is more effective than result re-ranking; (3) using several implicit feedback features leads to better results than just using clickthrough features (i.e. it is recommended to make use of additional pieces evidence of implicit feedback such as dwell time on a page).

The experiments carried out by (Gao et al., 2007) were also conducted in a large-scale setting. As discussed earlier, the authors proposed a Cross-Lingual Query Suggestion (CLQS) system. Given a source query in a certain language, the system obtained related queries from other languages by exploiting multilingual search logs. The proposed CLQS method was intended to be used as a method that combines query expansion with query translation instead of the typical use of a translation component in CLIR. The experiments were conducted on large datasets of English and French search logs. The first dataset included 7 million unique English queries, obtained from MSN Search logs over a period of one month. The second dataset included 5000 randomly selected French queries out of 3 million queries from a French query log. The TREC-6 CLIR document collection and its 25 information needs were used in the experiments. The cross-lingual retrieval effectiveness of the proposed CLQS system was evaluated using the 11-point Precision metric against three systems: a monolingual system, a system that used Google French to English machine translation, and a dictionary-based query translation system using co-occurrence statistics for translation disambiguation. The proposed CLQS system achieved 7.4% improvement over the machine-translation-based system and 25% improvement over the dictionary-based system. It was able to achieve 88% of the monolingual system performance. A rather similar setting was also used in (Yin et al., 2009) where the experiments were conducted on a dataset of search logs obtained from Microsoft Live Search over a period of ten months. The dataset contained 12 million unique queries.

Evaluation of personalised systems which are based on social data also commonly used retrieval effectiveness as a key measure for evaluation. For example, the personalised system reported in (Zhou et al., 2012) evaluated the retrieval precision of a proposed query adaptation algorithm. Adaptation was based on terms obtained from a user model consisting of terms extracted from the user's tags and bookmarks on the del.icio.us website. The experiments were conducted on a large-scale dataset harvested from del.icio.us, involving the data of about 6,000 users, 1 million documents, and 280,000 tags. The intuition behind the evaluation process is that, if a user u bookmarked a document and tagged the document with a tag t, then it may be assumed that the tag t is considered relevant to the document by the user. Based on this idea, the system issues a query consisting of the keyword t on behalf of u, and then checks whether those documents tagged with t by u are ranked high in the returned result list. The MAP metric was used to evaluate retrieval effectiveness. The proposed personalised system was able to achieve a statistically significant improvement of 61% over a non-personalised baseline.

Evaluation in the area of Adaptive Hypermedia (AH), especially in the educational domain, has often focused on the effectiveness of the adaptive service within the given domain (Conlan and Wade, 2004, De Bra et al., 2003, Brusilovsky and Peylo, 2003). This type of evaluation reflects the two-fold challenge of evaluating adaptive systems: how to uniformly test a system which changes in response to the user and how to evaluate a complex user experience with an unbiased measure. This gives rise to the use of measures such as task time completion and user satisfaction as a basis for testing the adaptive experience.

For example, the authors in (Conlan and Wade, 2004) proposed an adaptive eLearning system based on the content of an undergraduate-level SQL (Structured Query Language) online course. The course was divided into two parts, a database theory part (given as face-to-face lectures) and a practical part (online) concerning the learning of SQL. Only the SQL part was presented via an adaptive eLearning course and was evaluated in large-scale experiments. The experiments involved a total of over 500 students, spanning a period of four years. The

[1] The study was conducted at Microsoft Research, but the name of the underlying search engine was not specified

experiments aimed at evaluating the effectiveness of the course provided by the adaptive system by examining the students' performance over a number of years in exams specifically related to SQL topics. This included exam scores over the period of the evaluation (four years) and also the two preceding educational years when an online non-adaptive version of the course was used. Evaluation was concerned with comparing how the students performed using the non-adaptive online course (before the introduction of the adaptive one) to how they performed using the proposed adaptive system. The results of the experiments demonstrated the success of the proposed adaptive system where an average of 13% increase in students' exam scores was reported for the adaptive system over the non-adaptive one. Furthermore, analysis of differences in student capabilities across the years was performed to ensure no natural bias between years.

The authors in (Steichen et al., 2009) carried out an assessment of the knowledge gain of the students in a domain-specific eLearning environment. The knowledge gain was assessed by comparing the students' initial knowledge, measured in a pre-test, with their answers to task-based questions in the adaptive system. The experimental setting involved 12 students who were asked to complete 3 learning tasks that were randomly selected from a pool of 6 tasks. The knowledge gain was calculated by scoring the students' answers in the pre-test on a scale from 0 to 5 (where 0 indicated that the student had no prior knowledge of the task area, and 5 indicated that the student had the knowledge needed to carry out the task) and by assessing the students' answers to the given tasks on the same scale (where 0 represented complete failure to solve the task and 5 represented complete success). The average knowledge gain of the 12 students using the system was 4.25, which reflected the educational impact of the proposed adaptive system. Moreover, the students were also asked to fill questionnaires to evaluate the usefulness and the usability of the system. The results suggested that students were satisfied with the relevance of the presented content to their information needs and that they liked the presentation of results in the form of adapted hypertext presentations (dynamic composition of results and eLearning content).

Relatively few personalised search studies in the literature attempted such user-focused evaluation that is common in the AH field. Among those few studies is a study conducted by (Pretschner and Gauch, 1999) where the accuracy of the user model was qualitatively evaluated by comparing the inferred interests in the user model to actual user interests. This was done by a questionnaire that asked users to indicate how well the top 20 inferred interests in the user model reflected their actual interests. The results indicated that, on average, the users found that more than half of the inferred interests truly reflected their interests. In addition to the user model accuracy evaluation, the retrieval effectiveness of the Web search system was evaluated using 11-point Precision. The experiments were conducted in an in-lab setting, and the retrieval component was wrapped around the *ProFusion*[1] search engine. The results showed an 8% improvement in precision for the personalised system over the baseline search engine.

Task-based evaluation was carried out in (Pitkow et al., 2002) where the system recorded the time and number of actions that the users needed in order to successfully complete a number of given search tasks. The experiments were carried out in an in-lab setting that involved 48 users using two systems: the experimental personalised search system (which was wrapped around Google) and any of the following well-known search engines: AOL[2], Excite[3], Yahoo, or Google. Each user was given 12 search tasks and a maximum time of 3 minutes to complete it. The results showed that the proposed personalised system enabled users to complete their tasks in less time and a smaller number of actions compared to the use of one of the search engines. It should be noted that the proposed system offered a rich user interface that comprised a number of special additional features that are not present in other search engines. Thus, the authors argue that a bias towards their system in the experimental results may be observed because some of the tasks were tailored to make use of those special features which enabled users to use them and finish their tasks faster and with fewer actions. A notable drawback in the evaluation process was that, due to experimental limitations, a default user model (i.e. the same user model) was used for all the users. The default user model contained information mined from browsing history of documents that are related to the search tasks. The users were given the chance to view the content of the user model prior to the experiment. Such use of a default user model is not a common approach in PIR studies and may render the experimental results doubtful.

5.3 Summary and discussion

A variety of evaluation mechanisms and approaches were discussed in this section. Table 6 presents a brief summary of these approaches and gives examples by a number of systems in the literature.

[1] The ProFusion meta search engine is now obsolete.
[2] http://www.aol.com/
[3] http://www.excite.com/

Table 6: summary of evaluation techniques

Scope of Evaluation	Evaluation Metric & Instrument	Datasets	Experimental Setting	Example Publications
System Performance (retrieval effectiveness)	Quantitative (P@K, Recall@K, F-measure, Break-even point, NDCG, R-precision)	Documents: open Web corpora. Logs: in-lab generated logs.	In-lab setting (6 to 47 users)	Smyth and Balfe 2006, Teevan et al. 2005, Speretta and Gauch 2005, Sugiyama et al. 2004
System Performance (retrieval effectiveness)	Quantitative (MAP, 11-Point Precision)	Documents: TREC collections. Logs: search engine query logs.	Large-scale setting (large number of live user interactions with a Web search engine: 3 to 12 million unique queries)	Yin et al. 2009, Gao et al. 2007
System Performance (retrieval effectiveness)	Quantitative (P@K, NDCG, MAP)	Documents: open Web corpora. Logs: search engine logs.	Large-scale setting (Large-scale setting (large number of live user interactions with a Web search engine: 1.2 million queries)	Agichtein et al. 2006
System Performance (retrieval effectiveness)	Quantitative (MAP, P@K, Recall@K)	Documents: subset of annotated documents from del.icio.us website. Logs: user tags from del.icio.us	Large-scale setting (200 users)	Zhou et al. 2012
User Evaluation (task-based)	Quantitative (time and number of actions needed to complete search tasks)	Documents: open Web corpora. Logs: in-lab generated logs.	In-lab setting (48 users)	Pitkow et al. 2002
System Performance and Usability	Quantitative & Qualitative (11-point precision & questionnaires about usability or user model accuracy)	Documents: open Web corpora. Logs: in-lab generated logs.	In-lab setting (16 to 24 users)	Micarelli and Sciarrone 2004, Pretschner and Gauch 1999
User Evaluation and System Usability	Quantitative & Qualitative (task score & usability questionnaires)	Corpora: domain-specific corpora, harvested from the Web	In-lab setting (12 users)	Steichen et al. 2009
User Evaluation and System Usability	Quantitative & Qualitative (exam scores & usability questionnaires)	Corpora: domain-specific eLearning corpus	Large-scale setting (500 users)	Conlan and Wade 2004

A key challenge that faces researchers in the field of PIR, is obtaining realistic search logs that can be used to infer users' behavioural patterns and search interests. Major search engines do not prefer to release their search logs to the public or even to the academic community. This may be attributed to two reasons: privacy concerns and competitive business or technological advantage. Thus, the alternative for researchers becomes in-lab-style experiments. Although an in-lab experiment would not yield a relatively large dataset of search logs, it has a number of advantages. First, more focused user studies and usability evaluations can be conducted by providing questionnaires to the users or by directly interviewing them. Second, the experiments can be repeated with different settings using the same test group of users, and therefore comparisons can be conducted between different experimental runs. In general, an important matter that researchers in the IR and PIR community should perhaps consider is to make their datasets available for other researchers to use. This would enrich the number and variety of datasets in the community which may lead in turn to enriching the quantity and quality of research conducted in the area. Furthermore, making the datasets available will enable researchers to replicate each others' experiments and therefore be able to perform apples-to-apples comparisons between their different proposed systems.

An allied approach has been to divide the challenge of evaluating a whole adaptive system into an evaluation of the user modelling component, and a separate evaluation of the adaptive decision-making component (Brusilovsky et al., 2004). This manner of summative evaluation might be useful for PIR systems, where the different components that make up the system can be evaluated in isolation and in a cooperative manner.

6 General discussion and challenges

The previous sections of this paper provided an analysis of the different approaches and techniques exhibited in PIR literature. In this section, a broader discussion is provided regarding a number of issues concerning PIR and IR in general. This section also discusses current and emerging challenges facing research in the field of PIR and highlights future research directions.

The majority of studies in the literature investigated personalisation in monolingual IR systems, and relatively fewer studies extended to multilingual IR. Furthermore, with respect to the use of an individualised user model for PIR it is noted that no studies attempted to investigate the construction of user models that would specifically represent and cater for the needs of a multilingual search user. Multilingual Information Retrieval (MIR) systems may greatly benefit from the use of individualised user models, for example by including information about the user's country and native or preferred language. This kind of cultural or linguistic information may be exploited for adapting both, the query and the results in MIR. Moreover, PMIR systems may also benefit from the creation of language or country stereotype models based on users' aggregate behavioural patterns as exhibited in MIR search logs (Ghorab et al., 2010, Jansen and Spink, 2003). Research in this area might reveal a need to alter the way that individualised user models are represented, so that they can accommodate the multilingual dimension of MIR. Such change of user model representation may in turn have a profound effect on query adaptation and result adaptation techniques in PMIR.

Very few studies in PIR literature addressed the area of personalised query adaptation based on information from the user model. More specifically, there is an exhibited gap in PMIR literature with respect to performing pre-translation and post-translation query expansion based on terms obtained from the user's search interests. It is perhaps worthwhile for future studies to investigate if result-list precision in personalised search can be improved by performing query adaptation that is based on user model information. In other words, research in this area can benefit of studies that test the hypothesis of whether higher degrees of search personalisation can be achieved if the query adaptation process is more user-centred. One of the main challenges facing this area of research is how to determine which terms in the user model are most related to a given query so that they can be selected for expansion.

The selective personalisation approach (see selective query expansion in Section 4.2.3.1), which involves systems dynamically making decisions about different aspects of the personalisation process at runtime, is gaining attention in recent literature. For example some systems such as (Dou et al., 2009, Teevan et al., 2008, Amati et al., 2004) dynamically decide whether or not to apply search personalisation based on certain query features, while other systems, such as (Ogilvie et al., 2009, Chirita et al., 2007) use query features to dynamically decide about how many terms to use when expanding a query. This kind of studying of query features is known by several names in the literature, such as: query ambiguity, query clarity, and query performance. This is an interesting area of research, specially that personalisation is centred around the idea that content and services are dynamically adapted to users at runtime, and it would therefore be interesting to see how such dynamic nature of personalisation can be stretched even further to include dynamic decisions of when and how personalisation should be performed.

Various techniques for results ranking and presentation have been explored in the literature, some of which were well studied in the context of PIR, while others may still require more attention and comparative evaluation regarding how they can be incorporated with PIR. For example, a characteristic of the result diversification technique is that it aims at displaying diverse results within the first set of results presented to the user (Santos et al., 2010, Minack et al., 2009). This notion can be considered as opposed to personalisation techniques, where the aim is to display many results from the topic that is inferred to be of relevance to the user. To this end, there may be scope for investigating how these two complementary techniques can be brought together under one roof. An example of this kind of research is the work reported in (Gollapudi and Sharma, 2009) where the system attempts to deduce a result scoring function that makes a trade-off between the two techniques.

Another result presentation method that is exhibited in recent literature and is currently used by some of the well-known search engines is the aggregation of search results from different *verticals* (Arguello et al., 2011, Diaz and Arguello, 2009). This refers to the notion of incorporating different kinds of multimedia items in the result list, such as images, videos, and text. It also refers to incorporating results from a variety of genres, such as news, blogs, company profiles and people profiles. Two research directions are being investigated in this topic: (1) how to select the most appropriate vertical for a given query; and (2) how to rank items from different verticals when displaying them in the result list. It would be interesting to see more studies that focus on how user models can represent the user's preferences with respect to different verticals in association with different search domains, and how this information can be exploited for PIR. Following on this, there may be even more room for research on search results' presentation techniques that move away from the traditional ranked list paradigm, where not only the "list" of results is adapted, but also the "content" of the results is re-structured and tailored to meet the user's knowledge and needs (Levacher et al., 2011, Hearst 2009).

A certain degree of controversy was exhibited in the literature regarding the use of long-term interests vs. short-term interests for search personalisation. Since short-term interests are incidental interests that emerge from ad-hoc information needs, they might not benefit much from what has been learnt about the user on the long run. Yet, there might be scope for investigating how long-term interests may be used for personalising ad-hoc searches. This area of research may perhaps be aligned with the approach of selective personalisation where a PIR system can put into effect certain thresholds concerning when to make use of long-term interests to adapt the user's search depending on how similar the current query is to the interests exhibited in the user model. Furthermore, this topic might also benefit from putting result diversification techniques into practice as these techniques can help amend situations where the system's inference about the user's information need has gone wrong.

It is noted that the investigation of implicit information gathering approaches for user modelling gained more attention over explicit approaches in more recent literature. The use of implicit methods has shown the ability to improve PIR, especially when a large quantity of historical information is available about the users' interactions with the system. Such history of interactions is a useful resource for inferring the user's interests and preferences. However, the process of interpreting implicit feedback and inferring the user's interests from it may not always be very accurate. For example, it is not always the case that clicked results are relevant to the user's search, and furthermore, the user may sometimes even spend some time reading a document before realising it is not relevant to their information need. On the other hand, explicit feedback methods may be considered as a more reliable source of information about what the users like and dislike, and can therefore be used to revise the system's inferences about its users in order to ensure that the user model does not go astray. Yet, the challenge with explicit feedback is getting the users to actually provide it. To this end, several new systems on the Web exhibit intelligent and unobtrusive ways to gather "explicit-like" feedback from users by providing features that are integrated as part of the system's service; ones that do not require the users to deviate from the natural flow of system usage.

For example, an early attempt to provide such a feature was carried out by the *Excite*[1] search engine. A feature called "more like this" was provided, which was basically a link displayed beside each result in the result list (Jansen et al., 2000). By clicking this link, users indicated that the result was relevant to their information need and that they wanted to get more similar results. A number of other examples are now provided by many new systems on the Web. Google search was recently observed to sometimes display a "similar" link beside each result, which can be used to obtain other related results. The *Yippy*[2] search engine groups search results under labelled clusters; when users click on a cluster to specify the category to which their query belongs, then this can be considered as a reliable hint of relevance provided by the user to the system. Moreover, with respect to interactive query adaptation techniques, several search engines currently offer a query suggestion feature, where an *auto-complete* mechanism is used to instantly display various expanded forms of the query while the user is typing in the search field. Besides search engines, numerous social and recommender websites, such as Facebook and StumbleUpon, are examples where similar feedback features are used. The "Like" or "I like it" buttons in those websites are considered a very good source of information about what the user is interested in. In an indirect way, such buttons gather explicit-like feedback from the users, but in an intelligent manner that does not give the users the feeling that they are explicitly providing feedback to the system about their preferences.

The presence and wide use of this form of rich-feedback features in current commercial systems on the Web give a new meaning to the notions implicit and explicit forms of feedback. Users may have shown to be reluctant to the idea of having to provide explicit feedback to the system, but this seems to only be the case when the main flow of their system usage is interrupted. It would be interesting to see how future systems continue to intelligently provide feedback-gathering features that are "disguised" as application features.

In addition to feedback features, many online social applications almost completely revolve around the idea of user participation. Examples are Facebook, del.icio.us, Dogear, Twitter, and numerous other new websites that conform to Web 2.0 standards. The large amount of user-generated content on those websites is certainly a rich source of information for personalisation. The exploitation of such content for user modelling has gained attention in recent studies in the literature. It would be interesting to see more research in this direction in a way that would maximise the utilisation of these new and rich sources of information in PIR and other related areas.

A higher tendency is noted in PIR literature towards evaluating systems based on retrieval effectiveness, compared to evaluating systems based on other aspects like the accuracy of the user model or the usability of the adaptive service. This tendency can be attributed to the wide usage of precision-based evaluation metrics in the field of IR where retrieval effectiveness is the focus of evaluation. On the other hand, many systems in AH literature, where personalisation is implemented in other application areas such as eLearning, it can be noted that evaluation focused on aspects like the system's usability, the performance of users in given tasks, the user

[1] http://www.excite.com
[2] http://search.yippy.com/

model, or other user related aspects. Since the area of PIR can be recognised as an area where there is a hybrid fusion of techniques from both, IR and AH, then perhaps there is scope for "borrowing" more evaluation techniques from the AH field; ones which pay more attention to the user factor in the equation. Nonetheless, IR evaluation metrics should not be disregarded as they are standardised and allow bench mark testing across many systems.

Many PIR systems operate over the top results obtained from popular search engines like Google, Bing, or Yahoo. It may therefore be argued that the improvements achieved by such PIR systems partially owe to the quality of the underlying IR components. In other words, it is true to a certain extent that the effectiveness of the result adaptation component of a PIR system depends on the ability of the retrieval component to retrieve highly relevant results to the query in the first place, before further adapting those results to the user. This suggests that in order for personalisation to be effective it must operate on the best available content, and therefore, researchers in the field of PIR need to make sure that they are using the best available IR techniques before layering their personalisation techniques on top of them. This calls for evaluation frameworks that facilitate and standardise the evaluation of the different components of PIR in isolation, as well as the evaluation of the overall effectiveness of the combined parts.

7 Conclusion

This survey paper presented a critical review and novel classification of State-of-the-Art approaches in the field of Personalised Information Retrieval. The analysis was carried out over four stages: (1) information gathering, which was concerned with approaches for collecting information about system users; (2) information representation, which focused on different approaches of maintaining and modelling usage and user information; (3) personalisation implementation and execution, which presented an in-depth analysis of approaches to search personalisation; and (4) system evaluation, which provided a review of the experimental settings and evaluation mechanisms involved in the evaluation of PIR systems.

Furthermore, this paper presented a classification of PIR systems into three categories according to the scope of personalisation addressed, namely: individualised personalisation, community-based personalisation, and aggregate-level personalisation. The paper also presented a classification of query adaptation techniques from a personalisation perspective. This classification featured two attributes: (1) user-focused vs. non-user-focused techniques; and (2) implicit vs. explicit techniques.

Moreover, the survey provided a discussion of general issues related to information retrieval, personalisation, user modelling, and adaptive hypermedia. The survey also highlighted challenges and research directions that can be addressed by future studies in the field of PIR.

In conclusion, we argue that the Web community has moved to a situation where global multilinguality is becoming an ever more important aspect of the users' daily interaction with information on the Web. Yet, research in the area of Personalised Multilingual Information Retrieval is still in an early stage. Research in this area should enable users to achieve maximum benefit of information on the Web, beyond the barriers of language and country. Therefore, researchers should be looking at how PIR can be tailored with two things in mind: a multilingual Web and a multilingual user. The consideration of this natural multilinguality characteristic may have a profound effect on the way personalised systems gather, model, and exploit user information for the delivery of a service that not only adapts to the user's knowledge and interests, but also to the user's cultural and linguistic background.

We also argue that the rapid development of how information is presented on the Web and how users interact with different personalised systems should be taken into account when designing future PIR systems. This should affect the way the user interface is designed so that it allows the system to learn more about its users by gathering as much information as possible about them in an unobtrusive, yet reliable, manner. This should also affect the way machine learning algorithms are used to accurately interpret the true interests and needs of the users. Such accurate modelling of the users will certainly help towards providing better and more effective PIR systems on the Web.

Acknowledgements

This research is supported by the Science Foundation of Ireland (grant 07/CE/I1142) as part of the Centre for Next Generation Localisation (www.cngl.ie) at Trinity College, Dublin. The authors would like to thank the reviewers for their valuable comments to the earlier versions of this paper. The authors would also like to thank Helen Ashman for the fruitful discussions.

References

ACQUISTI, A. & GROSS, R. 2006. Imagined Communities: Awareness, Information Sharing, and Privacy on the Facebook. *6th Workshop on Privacy Enhancing Technologies (PET 2006) in Lecture Notes in Computer Science.* Cambridge, UK: Springer Berlin / Heidelberg, pp. 36-58.

AGICHTEIN, E., BRILL, E. & DUMAIS, S. 2006a. Improving Web Search Ranking by Incorporating User Behavior Information. *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006).* Seattle, Washington, USA: ACM, pp. 19-26.

AGICHTEIN, E., BRILL, E., DUMAIS, S. & RAGNO, R. 2006b. Learning User Interaction Models for Predicting Web Search Result Preferences. *29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2006).* Seattle, Washington, USA: ACM, pp. 3-10.

AMATI, G., CARPINETO, C. & ROMANO, G. 2004. Query Difficulty, Robustness, and Selective Application of Query Expansion. *Lecture Notes in Computer Science. The 26th European Conference on Information Retrieval (ECIR 2004).* Sunderland, U.K.: Springer, pp. 127-137.

AMBATI, V. & UPPULURI, R. 2006. Using Monolingual Clickthrough Data to Build Cross-lingual Search Systems. *New Directions in Multilingual Information Access Workshop of SIGIR 2006.* Seattle, Washington, USA: ACM

ARGUELLO, J., DIAZ, F., CALLAN, J. & CARTERETTE, B. 2011. A Methodology for Evaluating Aggregated Search Results. *33rd European Conference on Information Retrieval (ECIR 2011).* Dublin, Ireland, pp. 141-152.

ASNICAR, F. A. & TASSO, C. 1997. ifWeb - a Prototype of User Model-Based Intelligent Agent for Document Filtering and Navigation in the World Wide Web. *Adaptive Systems and User Modeling on the World Wide Web.* Chia Laguna, Sardinia.

BAEZA-YATES, R. & RIBEIRO-NETO, B. 2011. *Modern Information Retrieval: The Concepts and Technology behind Search (2nd Edition)*, Addison-Wesley.

BAST, H., MAJUMDAR, D. & WEBER, I. 2007. Efficient Interactive Query Expansion with Complete Search. *16th ACM Conference on Information and Knowledge Management (CIKM 2007).* Lisbon, Portugal: ACM, pp. 857-860.

BELKIN, N. J. & CROFT, W. B. 1992. Information Filtering and Information Retrieval: Two Sides of the Same Coin? *Communications of the ACM,* 35**,** 29-38.

BILLERBECK, B., SCHOLER, F., WILLIAMS, H. E. & ZOBEL, J. 2003. Query Expansion using Associated Queries. *12th International Conference on Information and Knowledge Management (CIKM 2003).* New Orleans, LA, USA: ACM, pp. 2-9.

BILLSUS, D. & PAZZANI, M. 2007. Adaptive News Access. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 550-570.

BRAJNIK, G., GUIDA, G. & TASSO, C. 1987. User Modeling in Intelligent Information Retrieval. *Information Processing & Management,* 23**,** 305-320.

BRIN, S. & PAGE, L. 1998. The Anatomy of a Large-scale Hypertextual Web Search Engine. *7th International World Wide Web Conference (WWW1998).* Brisbane, Australia.

BROOKE, J. 1996. SUS-A quick and dirty usability scale. *In:* JORDAN, P. W., THOMAS, B., WEERDMEESTER, B. A. & MCCLELLAND, A. L. (eds.) *Usability Evaluation in Industry*, pp. 189-194.

BRUSILOVSKY, P. 2001. Adaptive Hypermedia. *User Modeling and User-Adapted Interaction,* 11**,** 87-110.

BRUSILOVSKY, P. & HENZE, N. 2007. Open Corpus Adaptive Educational Hypermedia. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 671-696.

BRUSILOVSKY, P., KARAGIANNIDIS, C. & SAMPSON , D. 2004. Layered evaluation of adaptive learning systems. *International Journal of Continuing Engineering Education and Lifelong Learning,* 14**,** 402-421.

BRUSILOVSKY, P. & MILLÁN, E. 2007. User Models for Adaptive Hypermedia and Adaptive Educational Systems. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 3-53.

BRUSILOVSKY, P. & PEYLO, C. 2003. Adaptive and Intelligent Web-based Educational Systems. *International Journal of Artificial Intelligence in Education,* 13**,** 157-299.

BRUSILOVSKY, P. & TASSO, C. 2004. Preface to Special Issue on User Modeling for Web Information Retrieval. *User Modeling and User-Adapted Interaction,* 14**,** 147-157.

BUDZIK, J. & HAMMOND, K. J. 2000. User Interactions With Everyday Applications as Context for Just-in-time Information Access. *5th International Conference on Intelligent User Interfaces (IUI 2000).* New Orleans, Louisiana, USA: ACM, pp. 44-51.

CALLAN, J. P., CROFT, W. B. & BROGLIO, J. 1995. TREC and TIPSTER Experiments with INQUERY. *Information Processing & Management,* 31**,** 327-343.

CAO, G., GAO, J., NIE, J.-Y. & BAI, J. 2007. Extending Query Translation to Cross-Language Query Expansion with Markov Chain Models. *14th ACM International Conference on Information and Knowledge Management (CIKM 2007).* Lisbon, Portugal: ACM, 351-360.

CAO, G., NIE, J.-Y., GAO, J. & ROBERTSON, S. 2008. Selecting Good Expansion Terms for Pseudo-Relevance Feedback. *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008).* Singapore, Singapore: ACM, pp. 243-250.

CARMAN, M. J., BAILLIE, M. & CRESTANI, F. 2008. Tag Data and Personalized Information Retrieval. *Workshop on Search in Social Media (SSM at CIKM 2008).* Napa Valley, California, USA: ACM, pp. 27-34.

CARMEL, D., ZWERDLING, N., GUY, I., OFEK-KOIFMAN, S., HAR'EL, N., RONEN, I., UZIEL, E., YOGEV, S. & CHERNOV, S. 2009. Personalized Social Search based on the User's Social Network. *18th ACM Conference on Information and Knowledge Management (CIKM 2009).* Hong Kong, China: ACM, pp. 1227-1236.

CARROLL, J. M. & ROSSON, M. B. 1987. The Paradox of the Active User. *In:* CARROLL, J. M. (ed.) *Interfacing Thought: Cognitive Aspects of Human-Computer Interaction.* Cambridge, MA: MIT Press, pp. 80-111.

CHEN, L. & SYCARA, K. 1998. WebMate: A Personal Agent for Browsing and Searching. *2nd International Conference on Autonomous Agents.* Minneapolis, Minnesota, United States: ACM, pp. 132-139.

CHIRITA, P.-A., FIRAN, C., S. & NEJDL, W. 2007. Personalized Query Expansion for the Web. *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).* Amsterdam, The Netherlands: ACM, pp. 7-14.

CONLAN, O., HOCKEMEYER, C., WADE, V. & ALBERT, D. 2003. Metadata Driven Approaches to Facilitate Adaptivity in Personalized eLearning Systems. *Journal of the Japanese Society for Information and Systems in Education*, 1, 38-45.

CONLAN, O. & WADE, V. 2004. Evaluation of APeLS – An Adaptive eLearning Service Based on the Multi-model, Metadata-Driven Approach. *In:* DE BRA, P. & NEJDL, W. (eds.) *Lecture Notes in Computer Science. 3rd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2004).* Eindhoven, The Netherlands: Springer Berlin / Heidelberg, pp. 504-518.

CONLAN, O., WADE, V., BRUEN, C. & GARGAN, M. 2002. Multi-Model, Metadata Driven Approach to Adaptive Hypermedia Services for Personalized eLearning. *Lecture Notes in Computer Science. 2nd International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2002).* Malaga, Spain: Springer, pp. 100-111.

COOK, R. & KAY, J. 1994. The Justified User Model: a Viewable, Explained User Model. *4th International Conference on User Modeling (UM 1994).* Hyannis, Massachusetts, USA, pp. 145-150.

COOL, C. & SPINK, A. 2002. Issues of Context in Information Retrieval (IR): an Introduction to the Special Issue. *Information Processing & Management,* 38**,** 605-611.

CUI, H., WEN, J.-R., NIE, J.-Y. & MA, W.-Y. 2003. Query Expansion by Mining User Logs. *IEEE Transactions on Knowledge and Data Engineering,* 15**,** 829-839.

DE BRA, P., AERTS, A., BERDEN, B., DE LANGE, B., ROUSSEAU, B., SANTIC, T., SMITS, D. & STASH, N. 2003. AHA! The Adaptive Hypermedia Architecture. *14th ACM Conference on Hypertext and Hypermedia (Hypertext 2003).* Nottingham, UK: ACM.

DE LA PASSARDIERE, B. & DUFRESNE, A. 1992. Adaptive Navigational Tools for Educational Hypermedia. *In:* TOMEK, I. (ed.) *Computer Assisted Learning, Lecture Notes in Computer Science.* Springer Berlin / Heidelberg, pp. 555-567.

DE LUCA, E. W. & NÜRNBERGER, A. 2006. Adaptive Support for Cross-Language Text Retrieval. *Lecture Notes in Computer Science. 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006).* Dublin, Ireland: Springer, pp. 425-429

DIAZ, F. & ARGUELLO, J. 2009. Adaptation of Offline Vertical Selection Predictions in the Presence of User Feedback. *32nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2009).* Boston, MA, USA: ACM, pp. 323-330.

DOU, Z., SONG, R., WEN, J.-R. & YUAN, X. 2009. Evaluating the Effectiveness of Personalized Web Search. *IEEE Transactions on Knowledge and Data Engineering,* 21**,** 1178-1190.

EFTHIMIADIS, E. N. 2000. Interactive Query Expansion: A User-based Evaluation in a Relevance Feedback Environment. *Journal of the American Society for Information Science,* 51**,** 989-1003.

ESPINOZA, F. & HÖÖK, K. 1995. An Interactive interface to an Adaptive Information System. *User Modelling for Information Filtering on the World Wide Web Workshop.* Hawaii, USA.

FURNAS, G. W., LANDAUER, T. K., GOMEZ, L. M. & DUMAIS, S. T. 1987. The Vocabulary Problem in Human-System Communication. *Communications of the ACM,* 30**,** 964-971.

GAO, W., NIU, C., NIE, J.-Y., ZHOU, D., HU, J., WONG, K.-F. & HON, H.-W. 2007. Cross-Lingual Query Suggestion Using Query Logs of Different Languages. *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).* Amsterdam, The Netherlands: ACM, pp. 463-470.

GARCÍA-BARRIOS, V. M., HEMMELMAYR, A. & LEITNER, H. 2009. Personalized Systems Need Adaptable Privacy Statements! How to Make Privacy-related Legal Aspects Usable and Retraceable. *2nd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2009).* Porto, Portugal, pp. 91-96.

GAUCH, S., SPERETTA, M., CHANDRAMOULI, A. & MICARELLI, A. 2007. User Profiles for Personalized Information Access. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* 1 ed.: Springer, pp. 54-89.

GHORAB, M. R., LEVELING, J., ZHOU, D., JONES, G. J. F. & WADE, V. 2010. Identifying Common User Behaviour in Multilingual Search Logs. *In:* PETERS, C., DI NUNZIO, G., KURIMO, M., MANDL, T., MOSTEFA, D., PEÑAS, A. & RODA, G. (eds.) *Lecture Notes in Computer Science (6241/2010), Multilingual Information Access Evaluation I. Text Retrieval Experiments.* Springer, pp. 518-528.

GOLEMATI, M., KATIFORI, A., VASSILAKIS, C., LEPOURAS, G. & HALATSIS, C. 2007. Creating an Ontology for the User Profile: Method and Applications. *Research Challenges in Information Science (RCIS 2007).* Ouarzazate, Morocco, pp. 407-412.

GOLLAPUDI, S. & SHARMA, A. 2009. An Axiomatic Approach for Result Diversification. *18th International Conference on World Wide Web (WWW 2009).* Madrid, Spain: ACM, pp. 381-390.

GUARDA, P. & ZANNONE, N. 2009. Towards the development of privacy-aware systems. *Information and Software Technology,* 51**,** 337-350.

HANANI, U., SHAPIRA, B. & SHOVAL, P. 2001. Information Filtering: Overview of Issues, Research and Systems. *User Modeling and User-Adapted Interaction,* 11**,** 203-259.

HARMAN, D. 1988. Towards Interactive Query Expansion. *11th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1988).* Grenoble, France: ACM, pp. 321-331.

HARMAN, D. 1992a. Relevance Feedback and Other Query Modification Techniques. *In:* FRAKES, W. B. & BAEZA-YATES, R. (eds.) *Information Retrieval.* Prentice-Hall, Inc, pp. 241-263.

HARMAN, D. 1992b. Relevance Feedback Revisited. *15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1992).* Copenhagen, Denmark: ACM, pp. 1-10.

HARPER, B. D., SLAUGHTER, L. A. & NORMAN, K. L. 1997. Questionnaire Administration Via the WWW: A Validation & Reliability Study for a User Satisfaction Questionnaire. *World Conference on the WWW and Internet.* Toronto, Canada.

HAVELIWALA, T. H. 2002. Topic-sensitive PageRank. *11th International Conference on World Wide Web (WWW 2002).* Honolulu, Hawaii, USA: ACM, pp. 517-526.

HEARST , M. A. 2009. *Search User Interfaces,* New York, NY, Cambridge University Press.

HOTHI, J. & HALL, W. 1998. An Evaluation of Adapted Hypermedia Techniques using Static User Modelling. *2nd Workshop on Adaptive Hypertext and Hypermedia.* Pittsburgh, USA.

JAMESON, A. 2008. Adaptive Interfaces and Agents. *In:* SEARS, A. & JACKO, J. A. (eds.) *The Human-Computer Interaction Handbook: Fundamentals Evolving Technologies and Emerging Applications.* 2nd ed.: CRC Press.

JANSEN, B. J. & SPINK, A. 2003. An Analysis of Web Searching by European AlltheWeb.com Users. *Information Processing & Management,* 41**,** 361-381.

JANSEN, B. J., SPINK, A. & SARACEVIC, T. 2000. Real life, real users, and real needs: A study and analysis of user queries on the web. *Information Processing and Management,* 36**,** 207-227.

JANSEN, B. J., SPINK, A. & TAKSA, I. (eds.) 2008. *Handbook of Research on Web Log Analysis*: Information Science Reference.

KATAKIS, I., TSOUMAKAS, G., BANOS, E., BASSILIADES, N. & VLAHAVAS, I. 2009. An adaptive personalized news dissemination system. *Journal of Intelligent Information Systems,* 32**,** 191-212.

KELLY, D. & TEEVAN, J. 2003. Implicit Feedback for Inferring User Preference: A Bibliography. *SIGIR Forum,* 37**,** 18-28.

KOBSA, A. 2007. Privacy-Enhanced Web Personalization. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 628-670.

KOUTRIKA, G. & IOANNIDIS, Y. 2004. Rule-based Query Personalization in Digital Libraries. *International Journal on Digital Libraries,* 4**,** 60-63.

KRUG, S. 2005. *Don't Make Me Think!: A Common Sense Approach to Web Usability*, New Riders.

LAMPE, C., ELLISON, N. B. & STEINFIELD, C. 2008. Changes in Use and Perception of Facebook. *ACM Conference on Computer Supported Cooperative Work (CSCW 2008).* San Diego, CA, USA: ACM, pp. 721-730.

LATHAUWER, L. D., MOOR, B. D. & VANDEWALL, J. 2000. A Multilinear Singular Value Decomposition. *SIAM Journal on Matrix Analysis and Applications,* 21**,** 1253–1278.

LEVACHER, K., LAWLESS, S. & WADE, V. 2011. A Proposal for the Evaluation of Adaptive Content Retrieval, Modification and Delivery. *Workshop on Personalised Multilingual Hypertext Retrieval (PMHR 2011).* Eindhoven, Netherlands: ACM, pp. 18-25.

LEVELING, J. & JONES, G. J. F. 2010. Classifying and Filtering Blind Feedback Terms to Improve Information Retrieval Effectiveness. *Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO 2010).* Paris, France: LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, pp. 156-163.

LIU, F., YU, C. & MENG, W. 2004. Personalized Web Search for Improving Retrieval Effectiveness. *IEEE Transactions on Knowledge and Data Engineering,* 16**,** 28-40.

LIVINGSTONE, S. 2008. Taking risky opportunities in youthful content creation: teenagers' use of social networking sites for intimacy, privacy and self-expression. *New Media & Society,* 10**,** 393-411.

MAGENNIS, M. & VAN RIJSBERGEN, C. J. 1997. The Potential and Actual Effectiveness of Interactive Query Expansion. *20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997).* Philadelphia, Pennsylvania, United States: ACM, pp. 324-332.

MANNING, C. D., RAGHAVAN, P. & SCHUTZE, H. 2008. *Introduction to Information Retrieval*, Cambridge University Press.

MCNAMEE, P. & MAYFIELD, J. 2002. Comparing Cross-Language Query Expansion Techniques by Degrading Translation Resources. *25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002).* Tampere, Finland: ACM, pp. 159-166.

MEI, Q. & CHURCH, K. 2008. Entropy of Search Logs: How Hard is Search? With Personalization? With Backoff? *International Conference on Web Search and Web Data Mining (WSDM 2008).* Palo Alto, California, USA: ACM, pp. 45-54.

MICARELLI, A., GASPARETTI, F., SCIARRONE, F. & GAUCH, S. 2007. Personalized Search on the World Wide Web. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* 1 ed.: Springer, pp. 195-230.

MICARELLI, A. & SCIARRONE, F. 2004. Anatomy and Empirical Evaluation of an Adaptive Web-Based Information Filtering System. *User Modeling and User-Adapted Interaction,* 14**,** 159-200.

MINACK, E., DEMARTINI, G. & NEJDL, W. 2009. Current Approaches to Search Result Diversification. *First International Workshop on Living Web: Making Web Diversity a True Asset.* Washington DC., USA.

NGUYEN, D., OVERWIJK, A., HAUFF, C., TRIESCHNIGG, D., HIEMSTRA, D. & DE JONG, F. 2008. WikiTranslate: Query Translation for Cross-Lingual Information Retrieval Using Only Wikipedia. *Lecture Notes in Computer Science. Cross-Language Evaluation Forum (CLEF 2008).* Aarhus, Denmark: Springer, pp. 58-65.

NOLL, M. & MEINEL, C. 2007. Web Search Personalization Via Social Bookmarking and Tagging. *In:* ABERER, K., CHOI, K.-S., NOY, N., ALLEMANG, D., LEE, K.-I., NIXON, L., GOLBECK, J., MIKA, P., MAYNARD, D., MIZOGUCHI, R., SCHREIBER, G. & CUDRÉ-MAUROUX, P. (eds.) *6th International Semantic Web Conference and 2nd Asian Semantic Web Conference (ISWC/ASWC 2007).* South Korea: Springer Berlin / Heidelberg, pp. 367-380.

OARD, D. 1997. The State of the Art in Text Filtering. *User Modeling and User-Adapted Interaction,* 7**,** 141-178.

OARD, D. W. 2010. Multilingual Information Access. *Encyclopedia of Library and Information Sciences, 3rd Edition***,** 3682 - 3687.

OARD, D. W. & DIEKEMA, A. R. 1998. Cross-Language Information Retrieval. *In:* WILLIAMS, M. (ed.) *Annual Review of Information Science (ARIST)*, pp. 472-483.

OGILVIE, P., VOORHEES, E. & CALLAN, J. 2009. On the Number of Terms Used in Automatic Query Expansion. *Information Retrieval,* 12**,** 666-679.

PAZZANI, M. & BILLSUS, D. 2007. Content-Based Recommendation Systems. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 325-341.

PINHEIRO DE CRISTO, M. A., CALADO, P. P., DE LOURDES DA SILVEIRA, M., SILVA, I., MUNTZ, R. & RIBEIRO-NETO, B. 2003. Bayesian Belief Networks for IR. *International Journal of Approximate Reasoning,* 34**,** 163-179.

PITKOW, J., SCHUTZE, H., CASS, T., COOLEY, R., TURNBULL, D., EDMONDS, A., ADAR, E. & BREUEL, T. 2002. Personalized Search. *Communications of the ACM,* 45**,** 50-55.

PRETSCHNER, A. & GAUCH, S. 1999. Ontology Based Personalized Search. *11th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 1999).* Chicago, Illinois, USA: IEEE, pp. 391-398.

PSARRAS, I. & JOSE, J. 2006. A System for Adaptive Information Retrieval. *Lecture Notes in Computer Science. 4th International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH 2006).* Dublin, Ireland: Springer Heidelberg, pp. 313-317.

QIU, F. & CHO, J. 2006. Automatic Identification of User Interest for Personalized Search. *15th International Conference on World Wide Web (WWW 2006).* Edinburgh, Scotland: ACM, pp. 727-736.

QUIROGA, L. M. & MOSTAFA, J. 2002. An Experiment in Building Profiles in Information Filtering: The Role of Context of User Relevance Feedback. *Information Processing & Management,* 38**,** 671-694.

RAZMERITA, L., ANGEHRN, A. & MAEDCHE, A. 2003. Ontology-Based User Modeling for Knowledge Management Systems. *Lecture Notes in Computer Science. 9th International Conference on User Modeling (UM 2003).* Johnstown, Pennsylvania, USA: Springer Berlin / Heidelberg, pp. 213-217.

RICH, E. 1983. Users are Individuals: Individualizing User Models. *International Journal of Man-Machine Studies,* 18**,** 199-214.

ROBERTSON, S. E., WALKER, S., JONES, S., M.HANCOCK-BEAULIEU, M. & GATFORD, M. 1995. Okapi at TREC-3. *3rd Text REtrieval Conference (TREC-3)*, pp. 109-126.

RUTHVEN, I. 2003. Re-examining the Potential Effectiveness of Interactive Query Expansion. *26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2003).* Toronto, Canada: ACM, pp. 213-220.

RUTHVEN, I. & LALMAS, M. 2003. A survey on the use of relevance feedback for information access systems. *The Knowledge Engineering Review,* 18**,** 95-145.

RUVINI, J.-D. 2003. Adapting to the User's Internet Search Strategy. *Lecture Notes in Computer Science. 9th International Conference on User Modeling (UM 2003).* Johnstown, Pennsylvania, USA: Springer, pp. 55-64.

SALTON, G. & BUCKLEY, C. 1990. Improving Retrieval Performance by Relevance Feedback. *Journal of the American Society for Information Science,* 41**,** 288-297.

SANTOS, R. L. T., MACDONALD, C. & OUNIS, I. 2010. Exploiting Query Reformulations for Web Search Result Diversification. *19th International Conference on World Wide Web (WWW 2010).* Raleigh, North Carolina, USA: ACM, pp. 881-890.

SCHAFER, J. B., FRANKOWSKI, D., HERLOCKER, J. & SEN, S. 2007. Collaborative Filtering Recommender Systems. *In:* BRUSILOVSKY, P., KOBSA, A. & NEJDL, W. (eds.) *The Adaptive Web.* Springer, pp. 291-324.

SHEN, X., TAN, B. & ZHAI, C. 2005. Implicit User Modeling for Personalized Search. *14th ACM International Conference on Information and Knowledge Management (CIKM 2005).* Bremen, Germany: ACM. Pp. 824-831.

SILVESTRI, F. 2010. Mining Query Logs: Turning Search Usage Data into Knowledge. *Foundations and Trends in Information Retrieval,* 4**,** 1-174.

SMYTH, B. & BALFE, E. 2006. Anonymous Personalization in Collaborative Web Search. *Information Retrieval,* 9**,** 165-190.

SPERETTA, M. & GAUCH, S. 2005. Personalized Search based on User Search Histories. *IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005).* Compiegne University of Technology, France, pp. 622-628.

STAMOU, S. & NTOULAS, A. 2009. Search Personalization Through Query and Page Topical Analysis. *User Modeling and User-Adapted Interaction,* 19**,** 5-33.

STEFANI, A. & STRAPPARAVA, C. 1998. Personalizing Access to Web Sites: The SiteIF Project. *2nd Workshop on Adaptive Hypertext and Hypermedia* Pittsburgh, Pennsylvania, USA.

STEFANI, A. & STRAPPARAVA, C. 1999. Exploiting NLP Techniques to Build User Model for Web Sites: the Use of WordNet in SiteIF Project. *2nd Workshop on Adaptive Systems and User Modeling on the World Wide Web.* Toronto, Canada.

STEICHEN, B., LAWLESS, S., O'CONNOR, A. & WADE, V. 2009. Dynamic Hypertext Generation for Reusing Open Corpus Content. *20th ACM Conference on Hypertext and Hypermedia (Hypertext 2009).* Torino, Italy: ACM, pp. 119-128.

STEICHEN, B., O'CONNOR, A. & WADE, V. 2011. Personalisation in the Wild: Providing Personalisation Across Semantic, Social and Open-Web Resources. *22nd ACM Conference on Hypertext and Hypermedia (Hypertext 2011).* Eindhoven, The Netherlands: ACM, pp. 73-82.

SUGIYAMA, K., HATANO, K. & YOSHIKAWA, M. 2004. Adaptive Web Search Based on User Profile Constructed without Any Effort from Users. *13th International Conference on World Wide Web (WWW 2004).* New York, USA: ACM, pp. 675-684.

SUN, J.-T., ZENG, H.-J., LIU, H., LU, Y. & CHEN, Z. 2005. CubeSVD: A Novel Approach to Personalized Web Search. *14th International Conference on World Wide Web (WWW 2005).* Chiba, Japan: ACM, pp. 382-390.

SUN, X., GAO, J., MICOL, D. & QUIRK, C. 2010. Learning Phrase-Based Spelling Error Models from Clickthrough Data. *48th Annual Meeting of the Association for Computational Linguistics.* Uppsala, Sweden: Association for Computational Linguistics, pp. 266-274.

TEEVAN, J., DUMAIS, S. T. & HORVITZ, E. 2005. Personalizing Search via Automated Analysis of Interests and Activities. *28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005).* Salvador, Brazil: ACM, pp. 449-456.

TEEVAN, J., DUMAIS, S. T. & LIEBLING, D. J. 2008. To Personalize or Not to Personalize: Modeling Queries with Variation in User Intent. *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008).* Singapore, Singapore: ACM, pp. 163-170.

TEEVAN, J., MORRIS, M. R. & BUSH, S. 2009. Discovering and Using Groups to Improve Personalized Search. *2nd ACM International Conference on Web Search and Data Mining (WSDM 2009).* Barcelona, Spain: ACM, pp. 15-24.

TULLIS, T. & ALBERT, W. 2008. *Measuring the User Experience: Collecting, Analyzing, and Presenting Usability Metrics*, Morgan Kaufmann.

VALLET, D., CANTADOR, I. & JOSE, J. 2010. Personalizing Web Search with Folksonomy-Based User and Document Profiles. *In:* GURRIN, C., HE, Y., KAZAI, G., KRUSCHWITZ, U., LITTLE, S., ROELLEKE, T., RÜGER, S. & VAN RIJSBERGEN, K. (eds.) *32nd European Conference on Information Retrieval (ECIR 2010).* Milton Keynes, UK: Springer Berlin / Heidelberg, pp. 420-431.

VASSILIOU, C., STAMOULIS, D., SPILIOTOPOULOS, A. & MARTAKOS, D. 2003. Creating adaptive web sites using personalization techniques: a unified, integrated approach and the role of evaluation. *In:* PATEL, N. V. (ed.) *Adaptive evolutionary information systems.* IGI Publishing, pp. 261-285.

VOLOKH, E. 2000. Personalization and Privacy. *Communications of the ACM,* 43**,** 84-88.

WADE, V. 2009. Challenges for the Multi-dimensional Personalised Web. *In:* HOUBEN, G.-J., MCCALLA, G., PIANESI, F. & ZANCANARO, M. (eds.) *Proceedings of User Modeling, Adaptation, and Personalization Conference (UMAP 2009). Lecture Notes in Computer Science.* Springer Berlin / Heidelberg, pp. 3-3.

WHITE, R. W., RUTHVEN, I. & JOSE, J. M. 2002. The Use of Implicit Evidence for Relevance Feedback in Web Retrieval. *Lecture Notes in Computer Science. 4th BCS-IRSG European Colloquium on IR Research (ECIR 2002).* Glasgow, UK: Springer, pp. 449-479.

WITTEN, I. H., FRANK, E. & HALL, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques (3rd Edition)*, Morgan Kaufmann.

XU, J. & CROFT, W. B. 1996. Query Expansion Using Local and Global Document Analysis. *19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1996).* Zurich, Switzerland: ACM, pp. 4-11.

XU, S., BAO, S., FEI, B., SU, Z. & YU, Y. 2008. Exploring Folksonomy for Personalized Search. *31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008).* Singapore, Singapore: ACM, pp. 155-162.

XU, Y., ZHANG, B., CHEN, Z. & WANG, K. 2007. Privacy-Enhancing Personalized Web Search. *16th International Conference on World Wide Web (WWW 2007).* Banff, Alberta, Canada: ACM, pp. 591-600.

YE, J., COYLE, L., DOBSON, S. & NIXON, P. 2007. Ontology-based models in pervasive computing systems. *The Knowledge Engineering Review,* 22**,** 315-347.

YIN, Z., SHOKOUHI, M. & CRASWELL, N. 2009. Query Expansion Using External Evidence. *Lecture Notes In Computer Science. 31st European Conference on Information Retrieval (ECIR 2009).* Toulouse, France: Springer, pp. 362-374.

ZHANG, H., SONG, Y. & SONG, H.-T. 2007. Construction of Ontology-Based User Model for Web Personalization. *Lecture Notes in Computer Science. 11th International Conference on User Modeling (UM 2007).* Corfu, Greece, pp. 67-76.

ZHANG, Y. & KOREN, J. 2007. Efficient Bayesian Hierarchical User Modeling for Recommendation System. *30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2007).* Amsterdam, The Netherlands: ACM, pp. 47-54.

ZHOU, D., LAWLESS, S. & WADE, V. 2012. Improving search via personalized query expansion using social media. *Information Retrieval***,** 1-25.

Author Biographies

Mr. M. Rami Ghorab, School of Computer Science & Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland.
Rami is a PhD student in Computer Science at Trinity College. He received his BSc in Computer Science from the Modern Academy, Egypt, in 2001. He completed a post-graduate Diploma at the Information Technology Institute, Egypt, and received his MSc in Information Technology from the University of Nottingham, UK, in 2003. His main research interests are in the fields of Multilingual Information Retrieval, Web Search, Personalisation, and User Modelling. Previous research interests included Peer-to-Peer Networks.

Dr. Dong Zhou, School of Computer Science & Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland.
Dong currently holds the position of research fellow in the Knowledge & Data Engineering Group (KDEG), the School of Computer Science & Statistics, Trinity College Dublin. He received his MSc degree in Information Processing from the University of York, UK, and his PhD in Computer Science from the University of Nottingham, UK. His primary research interests include Cross-language/Multilingual Information Retrieval, Personalisation, Natural Language Processing, Machine Learning, Hypertext and Hypermedia Systems, and Adaptive Applications for the Web. His current research focuses on developing novel models/systems for Personalised Multilingual Information Access.

Dr. Alexander O'Connor, School of Computer Science & Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland.
Alexander is a Research Fellow at the Knowledge & Data Engineering Group (KDEG), School of Computer Science & Statistics, Trinity College Dublin. He received his, BAI, MSc and PhD from Trinity College Dublin. His main research interests are in Ubiquitous Computing, Context Systems, and Adaptive Systems. His current position is a a post-doctoral researcher with the Centre for Next Generation Localisation (CNGL), where he is part of the Digital Content Management research track, focusing on Adaptivity, Web Systems, Knowledge Representation and Customer Care.

Prof. Vincent Wade, School of Computer Science & Statistics, Trinity College Dublin, College Green, Dublin 2, Ireland.
Vincent is a Professor of Computer Science in the School of Computer Science & Statistics, Trinity College Dublin. He holds a BSc (Hons) in Computer Science from University College Dublin, an MSc and PhD from Trinity College Dublin (TCD). Vincent is the head of the Intelligent Systems Discipline in TCD (2007-2012), which addresses research in the areas of Knowledge and Data Engineering, Graphics & Computer Vision, Computational Linguistics and Artifical Intelligence. He has been the Deputy Director of the Centre for Next Generation Localisation (CNGL) (2007-2012), a world leading multi-institutional research centre focusing on Multilingual, Multi modal Globalisation of Digital Content. Vincent has authored over 200 papers and has edited several books. In 2002, he was awarded the Fellowship of Trinity College (FTCD) for his contribution to research in the areas of Knowledge Management, Web-based Personalisation and Adaptive Content and Service Technologies.