Personalisation of Generic Library Search Results Using Student Enrolment Information

Marwah Alaofi
Taibah University
maofi@taibahu.edu.sa

Grace Rumantir
Monash University
grace.rumantir@monash.edu.au

This research explores the application of implicit personalisation techniques in information retrieval in the context of education. Motivated by the large and ever-growing volume of resources in digital libraries, coupled with students' limited experience in searching for these resources, particularly in translating their information needs into queries, this research investigates the potential of incorporating student enrolment information, that is, published information on the units/subjects they are enrolled in, to identify students' learning needs and produce personalised search results.

We propose, implement, and evaluate a personalisation approach that makes use of the collection of units a student is enrolled in to generate a student profile used to estimate the relevance of the library resources. To do this, we propose the use of a *Final Relevance Score* (*FRS*) measure, which assigns a relevance score for each query-dependent resource based on its similarity to both the student profile and the submitted query, with α parameter controlling the effect of both. To examine the effectiveness of this approach and whether it truly produces any improvement over the library generic approach, this approach was translated into an application called PersoLib and evaluated by a group of 16 students who were doing foundation units in the Masters of Information Technology course at Monash University.

The evaluation results show that the personalisation approach significantly outperforms the library generic approach. This shows the potential of incorporating student enrolment information to create a more effective search environment in which students' search results are not only driven by the submitted query, but also by the units they are enrolled in.

Keywords: Educational Digital Libraries, Information Retrieval for Education, Implicit Search Personalisation, Student Information Needs

1. INTRODUCTION

Inspired by the implicit personalisation technology that has been successfully employed to improve user experience in searching the World Wide Web, this paper explores the implementation of this technology to improve student experience in searching a university digital catalogue. This paper addresses the question of whether incorporating student enrolment information, that is, the information about the units[1] they are currently enrolled in, can effectively be used as a basis of identifying their learning needs and customising their library search results accordingly. To this end, we have designed, implemented, and evaluated a personalisation approach PersoLib (shown in Figure 1), which incorporates student enrolment information to improve the retrieval relevance of a university digital library search engine.

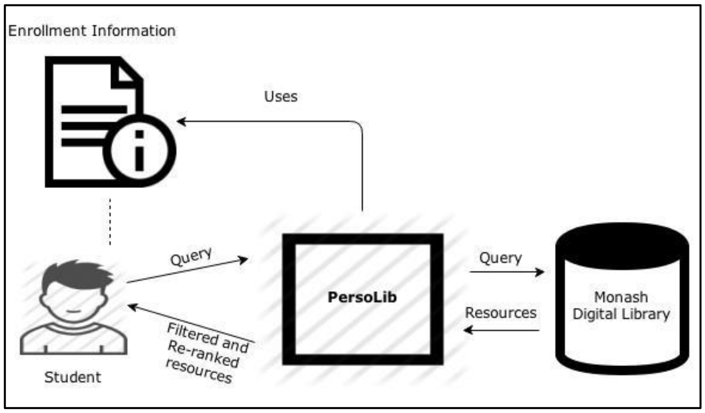


Figure 1: PersoLib Conceptual Framework

This paper is organized as follows: Section 2 and 3 present the motivation behind this research and the research background. The related work is presented in Section 4. Section 5 outlines the system design while Section 6 presents the experimental design including the method used for evaluation. Section 7 presents and discusses the results and the conclusions and future work are presented in Section 8 and 9 respectively.

2. RESEARCH MOTIVATION

Users are generally characterized as vague in specifying their queries (Carmel et al., 2009) and students are no different. In the university context, most students, especially those commencing in a new discipline, have limited knowledge of the subjects they are enrolled in, and thus, their search queries are expected to be consequently vague.

The number of resources, both physical and electronic, that students search through is large, and it continues to grow rapidly. This increase in the volume of library resources can be clearly seen by analysing the annual reports of Monash University Library, the library we are targeting in this study. In 2013—for which the latest report was published—Monash University Library had more than two million physical monographs (e.g. physical books), a number that has increased by approximately 15% over a period of 10 years. The increase in

[1] A ‘unit’ is a term conventionally used in Australia to refer to a unit of teaching within an academic program, such as ‘FIT5131 Programming Foundation’; often referred to as a course/subject in other parts of the world.

the library's electronic monograph collection (e.g. electronic books) has far exceeded even that of its physical monographs. In 2013, there were an estimated 662,980 electronic monographs, with an approximate increase of 589% over a period of 10 years (See Figure 2 and Figure 3). These two factors—the vagueness of student queries and the growing search space—could make library search an overwhelming experience which can potentially lead to poor utilisation of digital libraries.

The large investments libraries make in their collections would be best utilised if supported by effective search engines through which students could find the resources to satisfy their information needs. Unless digital libraries can offer efficient, customised and easy-to-use services, students may turn away from them. This would result in a significant waste of effort toward developing such digital libraries, reducing the discovery of resources in which the university has made large financial investments. More importantly, it may result in students not using high-quality resources, as they may turn to commercial products instead of the university library, the latter of which filters resources such that only those of high quality are provided.

We believe that using an automated process that could learn about students' academic interests and learning needs by utilising the available information about the units they are studying could provide a potential improvement. It could reduce the gap between their information needs and the available resources, bringing them the resources most similar to their academic interests, thus facilitating the learning process.

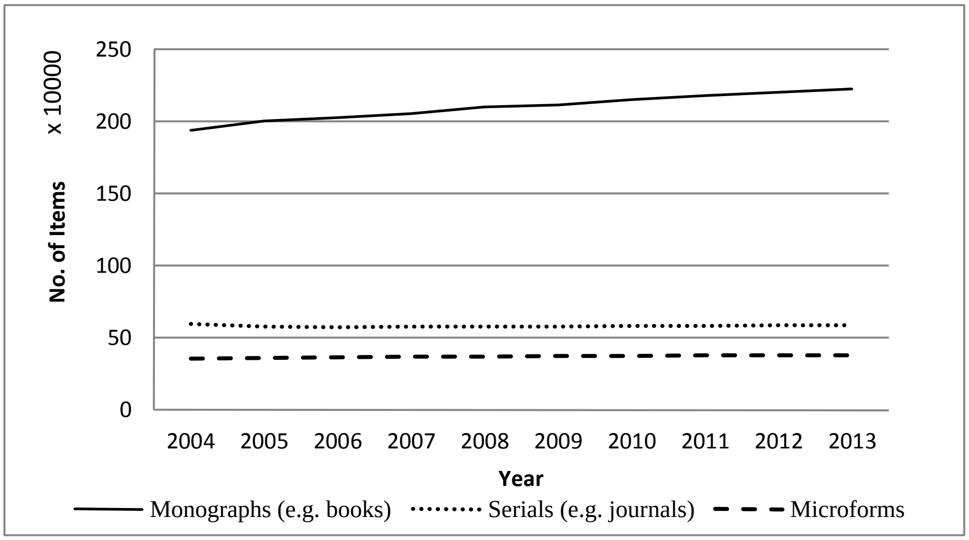


Figure 2: Growth of the physical collection at Monash Library over the last decade (Source: Monash Library Annual Reports, 2005-2014).

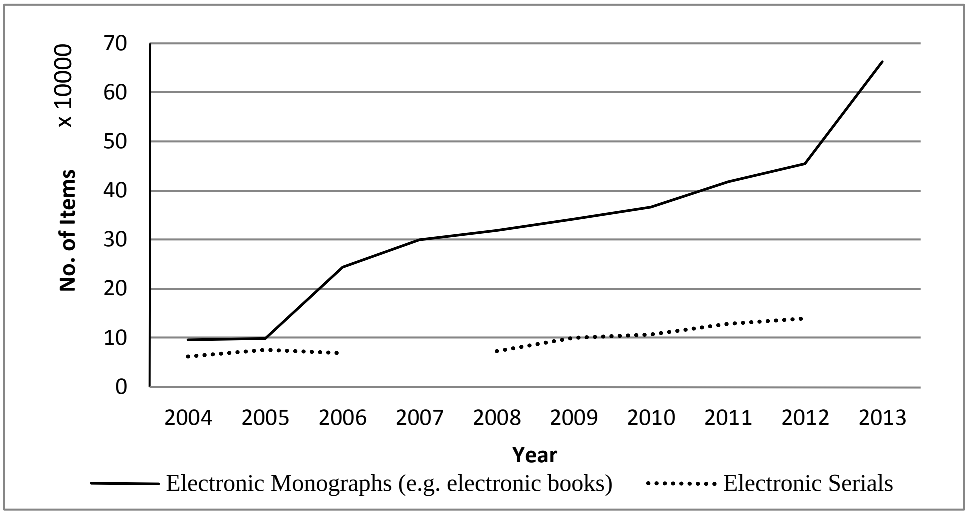


Figure 3: Growth of the electronic collection at Monash Library over the last decade (Source: Monash Library Annual Reports, 2005-2014).

3. RESEARCH BACKGROUND

This section highlights some of the main concepts addressed in this study, namely, Digital Libraries and the very specific need of personalisation within this area.

3.1. DIGITAL LIBRARIES (DLS)

According to Pomerantz, Choemprayong, and Eakin (2008), the first use of the term *Digital Library (DL) in print* seems to have been in Kahn and Cerf (1988), in which the latter propose a framework for a digital library system for a national information infrastructure.

Definitions of the notion of a DL that emerged from the computer and information science research community have evolved in both scope and content. The first textbook to address DLs, **PRACTICAL DIGITAL LIBRARIES** (Lesk, 1997), simply defines a DL as an organised collection of digital information. The working definition of DLs, which was set forth by The Digital Library Federation (DLF) to ensure a common understanding of the notion of DLs among their partners, is as follows:

> Digital Libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities. (Waters, 1998, p. 1)

The collections held in DLs may involve a wide range of information covering a large number of disciplines, such as science, literature, business and economics. This information can be in the form of digitised text, images, video and audio (Callan et al., 2003). The central contribution of DLs lies in the selection of quality-assured content and in providing supporting metadata for efficient structuring and easier content discovery. Advanced services for content access, use, and sharing are also offered to enhance users' experiences with DLs. Several types of DLs have emerged, such as those serving specific educational needs, academic institutions, organisations and cultural heritage (Callan et al., 2003).

Based on the context in which DLs are employed, the notion of DLs has several meanings, which may differ slightly from the aforementioned general definitions. For example, some DLs involve electronic resources and also provide access to searchable physical resources, all of which are catalogued and enriched with supporting metadata for better organisation and enhanced discovery. This extended view of DLs appears to be in play in universities' DLs such as the one we are targeting in this study. The DLs belonging to universities provide access to digital content, as well as create records for the physical collection available in the physical library to assist students and staff in searching the catalogue electronically and more easily.

3.2. THE NEED FOR PERSONALISATION IN DLS OF UNIVERSITIES

The first generations of DLs were built with the assumption that users were well informed and could accurately describe their information needs, which would then be easily matched with resources, as the size of the corpus was relatively small and had mostly homogeneous content (Callan et al., 2003). Due to advances in technology, social recognition and substantial funding, DLs have grown greatly and have become widely used, offering richer and more diverse content and services (Ashraf & Gulati, 2010).

The new generations of DLs are more heterogeneous in terms of the collections they hold and their expected users. As for collections, DLs can hold a high volume of resources (see Figure 2 and Figure 3) in different disciplines for different levels of comprehension. Users are also expected to be from a wide spectrum of expertise and experience, with different backgrounds, interests and skills, ranging from absolute beginners to experts in specific areas (Callan et al., 2003). This vast growth of both information diversity and volume, with users having different backgrounds, drives the need for more intelligent and effective library services. One of these services is based upon the concept of *personalisation*, in which users' backgrounds can be leveraged and taken into consideration to better understand their information needs and to tailor the content accordingly. DLs need to keep up with new technologies and take the initiative to develop more intelligent and adaptive services that tailor information to individuals and communities (Callan et al., 2003).

Personalisation has proven useful in many areas, mainly in the e-commerce domain, and it has shown successful results when reproduced in other areas, such as education (Klašnja-Milićević, Vesin, Ivanović, & Budimac, 2011). It has also been used in DLs and has produced better services with more satisfied users. However, the utilisation of this concept in this domain is still limited when compared to other areas of adoption. Nonetheless, the promising results obtained from previous efforts encourage more investigation, particularly on how to identify users' interests and map different emerging needs to offer more customised services, thereby narrowing the gap between the amount of offered content and the individual characteristics and particular preferences of each user.

Although tailoring the content to serve specific information needs is highly desirable in all types of DLs, it might be more valuable in those in which no specific domain is addressed and more diversity is observed, such as in the case of universities' DLs, like that on which this study is conducted. In this case, it is fair to say that these types of libraries would normally contain content that covers all disciplines within the university's major interests. They would also be dedicated to a large and diverse campus audience, such as students, lecturers and researchers, whose backgrounds can differ substantially.

4. RELATED WORK

Previous work on personalized services has mainly been featured in the context of implementation, the data source used to learn about users, and the technique by which a user is profiled. This section presents some of the studies that have been done on personalizing information retrieval services (i.e. search engines), with an emphasis on those designed for DLs. This section will also look at some of the research efforts that have been made to implement information access personalisation in its broader sense to benefit the educational domain while highlighting how the research efforts presented in this paper advance on the literature.

4.1. PERSONALIZED INFORMATION RETRIEVAL IN DLS

General web search personalisation is a mature research area, and it has been extensively studied. Several sources for interest identification have been leveraged to identify users' preferences. Some of these track user navigational and browsing history (Kim, Collins-Thompson, Bennett, & Dumais, 2012; Matthijs & Radlinski, 2011; Shen, Tan, & Zhai, 2005), while others utilise their desktop documents, emails, or social media (Carmel et al., 2009; Karweg, Huetter, & Bohm, 2011; Teevan, Dumais, & Horvitz, 2005; Zhou, Lawless, & Wade, 2012). The identified interests are then represented using a specific data structure, resulting in what is referred to as a 'user model' or 'profile'.

Not as many as those addressed the web, there have been some studies which attempted to personalise digital library search services. Potey, Pawar, and Sinha (2013) extend a traditional search engine to re-rank DL resources based on their relevance to individuals' interests. Other studies, however, addressed particular groups such as scholars in (Amini, Ibrahim, Othman, & Rastegari, 2011; Jomsri, Sanguansintukul, & Choochaiwattana, 2012) and domain experts as in (McKeown, Elhadad, & Hatzivassiloglou, 2003). Jomsri et al. (2012) attempted to incorporate the backgrounds of the scholars in personalising and re-ranking their digital library search results. They examine the feasibility of using academic social bookmarking systems to understand the interests of the researchers and customise their academic paper search results accordingly. User profiles are built using user-defined data, such as tags, abstracts, and titles of all the papers the user has posted in academic social bookmarking systems. When users search for academic papers, their profiles are used to measure the similarity between the returned articles and their profiles. With a more semantic representation of scholars' knowledge, Amini et al. (2011) follow a similar approach, in which they built conceptual profiles for scholars and personalised their digital library search results accordingly.

4.2. PERSONALISATION IN THE EDUCATIONAL DOMAIN

As discussed earlier, information overload characterizes DLs in general and the effect might be more apparent on students searching educational DLs, where more diversity in both content, and user level of expertise and interests might be observed. Ways in which the large space of information can be presented to students in those libraries is a major research question that has been addressed in the literature. With the massive amount of information available online, traditional information access methods have become undesirable and many alternative methods have been proposed in response to improve students' experience in finding relevant resources that fit under their academic interests and needs.

Brusilovsky, Farzan, and Jae-wook (2005) proposed Knowledge Sea II system, in which they explore comprehensive access to educational resources through multiple pathways, encompassing information visualization, information browsing, keyword-based search and instructor recommendations. A personalized information access by which students can locate resources that are relevant to their goals, knowledge, and interests is also explored in the study. To consider the relevance aspect of the educational resources in information visualization, social navigation support is used, which relies upon both traffic information and students' annotations. This information is used to reveal the resources that are frequently visited and those having positive annotations, which when combined can further assist students finding relevant resources.

Educational recommender systems have probably gained the major attention in providing personalized information access in educational DLs. Klašnja-Milićević et al. (2011) proposed a recommendation system which automatically recommends educational resources based on the learner's learning style, learning characteristics, interests, habits and knowledge levels. Tang, Winoto, and McCalla (2014) have advanced the concept of recommender systems to tailor the educational domain. They have realized the very specific need of the educational domain in which some contextual factors should be considered to assess the pedagogical value of resources. Some educational DLs–where users are not obligated to create accounts neither do they provide feedback–pose some challenges in regards to providing personalized information access as the latter often depends on attributes derived from user profiles. In response, Akbar, Shaffer, Weiguo, and Fox (2014) proposed a framework based on a Deduced Social Network (DSN) which analyses the sessions of anonymous users including their clicks, page views, times in pages, to proactively recommend educational resources.

While the literature presents some initiatives in attempting to personalize information access in DLs for students, none of which seems to go beyond utilising information mainly derived from students' interaction with the DL. To the best of our knowledge, there has been no research examining the usefulness of utilising enrollment information to personalize information retrieval services. There appears to be much room to explore ways in which better personalised services can be provided to support students in their information-seeking activities and reduce the gap between their information needs and the resources in DLs.

This research aims to connect the education and library sectors and examine the usefulness of some information retrieval principles to create customised search environments. We believe that library search engines should not be guided only by students' queries, but also by their academic scope, through which their search context can be identified and thus the relevance of the resources can be accurately quantified. The main goals are to investigate the impact of incorporating students' learning needs into the library search engine and to propose a framework that implicitly learns about students' needs through their enrolment information, as described in their unit guides, and ranks the query-dependent resources accordingly.

5. PersoLib Design

PersoLib is designed to help students find relevant resources while searching the Monash University Digital Library. The underlying concept is simple, and it is based primarily on incorporating students' enrolment information, i.e. the information about the units they are enrolled in, to quantify the relevance of the library resources obtained by the library search engine[2] and thus filter and re-rank them accordingly. In other words, PersoLib aims to personalise students' library search results based on the units they are studying, such that the obtained results are driven not only by the submitted query, but also by student enrolment information. The complete framework for PersoLib is shown in Figure 4.

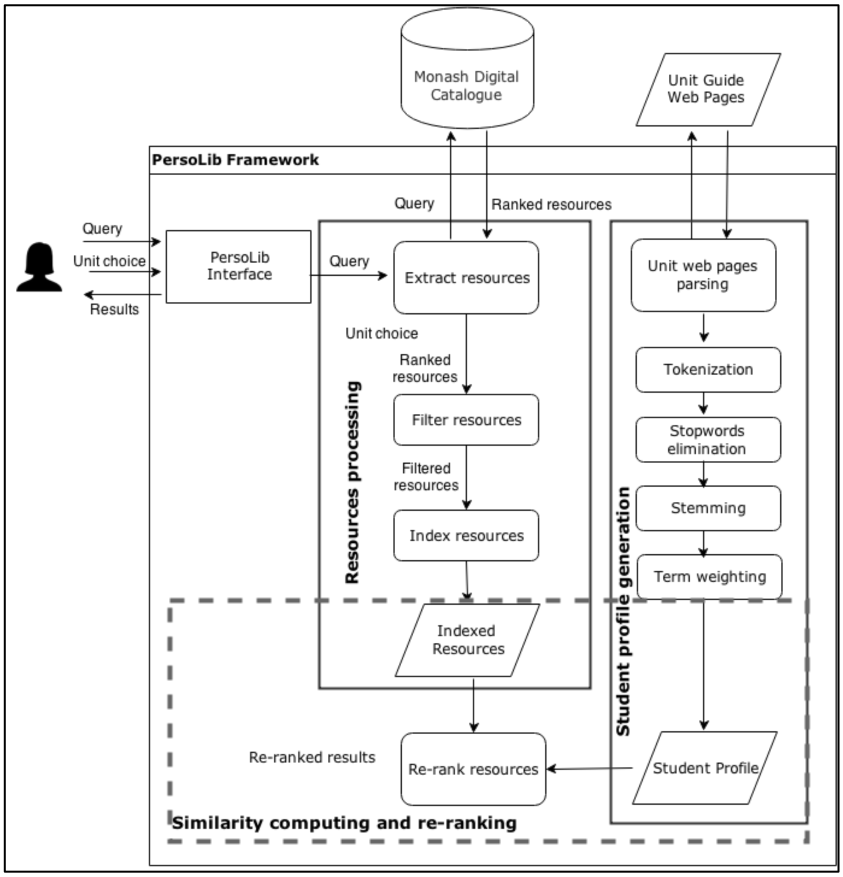


Figure 4: PersoLib Framework

[2] The library search engine can be found at: http:// lib.monash.edu

As seen in Figure 4, the personalisation functionality in PersoLib involves a number of tasks. The first task is parsing student unit guide web pages[3] of the units the students are enrolled in from the Monash University website to construct their profiles, which will serve as the basis upon which the degree of relevance of each resource is partially measured. Resources are extracted from the library website based on the queries submitted by the students and are then filtered and indexed for relevance computing. The similarity between each resource and the student profile is measured and then used to partially estimate the final rank of the resource. The major tasks of the system can be thus divided into three main tasks as follows: a) Student profile generation, b) Resources processing and c) Similarity computing and re-ranking. The following sections address all of the tasks involved and show the main technique(s) used in accomplishing each.

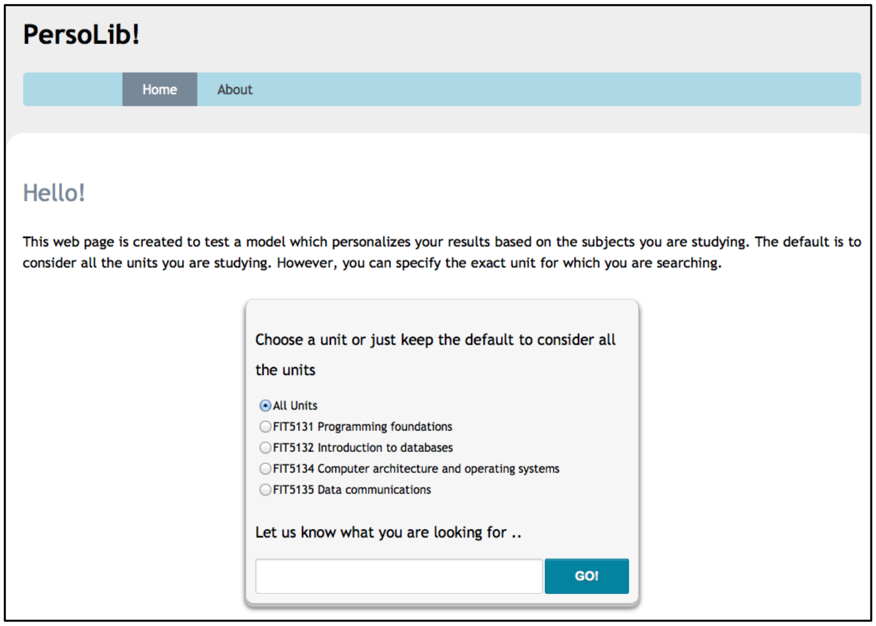
5.1. STUDENT PROFILE GENERATION

Student profiles essentially contain student learning needs as reflected in the units they are studying. The unit guide for each unit is available on the Monash University website and is considered to represent the main topics and learning outcomes of each unit. The learning needs of each student are fetched on the fly. That is, all unit guide web pages of the units in which the student is enrolled are parsed once the student starts searching the library. Parsing all units is the default, but students can further specify a particular unit under which their queries fall for better customisation.

In practice, students are presented with a web-based interface like that shown in Figure 5, where they can submit a query and choose to either accept the default personalisation or further specify a particular unit. *The unit parser* is then used to parse the information included in the unit guides. The unit guides consist of several parts, some of which do not represent the academic content of the unit, but rather focus on unit delivery and assessment. These latter parts do not represent the unit in terms of its academic content and are thus not extracted by the parser for student profiling. The included parts from each unit guide are as follows: a) Unit academic overview, b) Unit learning outcomes and c) Unit weekly topics.

Table 1 shows the main parts of FIT5131 Programming Foundation unit extracted by the unit parser.

[3] A unit guide, sometimes referred to as handbook, contains useful information about the unit, involving academic overview, learning outcomes and topics breakdown. An example of a unit guide web page can be found at: http://www.infotech.monash.edu.au/units/archive/2014/s2/fit5131.html



Hello!

This web page is created to test a model which personalizes your results based on the subjects you are studying. The default is to consider all the units you are studying. However, you can specify the exact unit for which you are searching.

Choose a unit or just keep the default to consider all the units

● All Units
○ FIT5131 Programming foundations
○ FIT5132 Introduction to databases
○ FIT5134 Computer architecture and operating systems
○ FIT5135 Data communications

Let us know what you are looking for ..

Figure 5: PersoLib web-based interface

Table 1: Text Extracted from FIT5131

Unit Guide Part	Content	
Unit academic overview	This unit aims to provide students with the basic concepts involved in the development of well structured software using a programming language. It concentrates on the development of problem solving skills applicable to all stages of the development process. Students gain experience with the translation of a problem specification into a program design, and the implementation of that design into a programming language. The subject introduces software engineering topics such as maintainability, readability, testing, documentation, modularisation, and reasoning about correctness of programs. Students are expected to read and understand existing code as well as develop new code.	
Learning Outcomes	design, construct, test and document small computer programs using Java; interpret and demonstrate software engineering principles of maintainability, readability, and modularisation; explain and apply the concepts of the "object-oriented" style of programming.	
Unit Schedule	Week 1	Introduction to FIT5131 and expectations; introduction to programming, basic OO concepts, objects, classes, attributes, behaviour, state and identity.

	Week 2	Class definition, fields, constructors, methods, parameter passing, variables, expressions, statements, assignment, primitive data types, arithmetic operators, strings,basic output.
	Week 3	Selection (if and switch statements), conditions, relational & logical operators, shorthand operators, ++ operator, precedence, scope and lifetime, basic input.
	Week 4	Object creation and interaction, abstraction, modularisation, class & object diagrams, object creation, primitive vs. object types, method calling, message passing, method signatures,method overloading.
	Week 5	Class libraries, importing classes, collections, ArrayLists, arrays, iteration,pre and post test loops.
	Week 6	Testing, unit testing, testing heuristics, regression testing,debugging.
	Week 7	Class documentation, Javadoc, identity vs. equality, more on strings, sets and maps,conditional operator.
	Week 8	Information hiding, encapsulation, access modifiers, scoping, class variables, class methods,constants.
	Week 9	Program design, design methods, responsibility-driven design, design documentation, testing a program,specifying a test strategy.
	Week 10	Programming errors, exception handling,file I/O.
	Week 11	Code quality, coupling, cohesion, refactoring,using the Java SDK.
	Week 12	Inheritance, superclasses, subclasses, subtypes, substitution, polymorphic variables, protected access, casting, wrapper classes,collection hierarchy.

Student profiles are represented as *weighted keyword profiles* in which each unit guide is represented as a set of weighted terms known as a term vector. Terms are automatically extracted from the extracted text from unit guide web pages, tokenized, filtered from stopwords, stemmed and associated with numerical weights representing their importance within the extracted text. A student profile P_s can therefore be viewed as a collection of term vectors, each of which represents a unit the student is enrolled in as follows:

$$P_s = \begin{pmatrix} tf(t_{1,1}) \\ tf(t_{2,1}) \\ tf(t_{3,1}) \\ \vdots \\ tf(t_{j,1}) \end{pmatrix}, \begin{pmatrix} tf(t_{1,2}) \\ tf(t_{2,2}) \\ tf(t_{3,2}) \\ \vdots \\ tf(t_{j,2}) \end{pmatrix}, \dots, \begin{pmatrix} tf(t_{1,i}) \\ tf(t_{2,i}) \\ tf(t_{3,i}) \\ \vdots \\ tf(t_{j,i}) \end{pmatrix} \quad (1)$$

where each vector is the frequency vector of a unit i, and $tf(t_{j,i})$is the number of occurrences of index term $t_{i.i}$.

The students targeted in this study are first-year students, all of whom are doing four units, as can be seen in Figure 5. Their complete profile, therefore, consists of four frequency vectors, each of which represents the number of occurrences of each index term in a given unit. Selected parts of one profile are shown in Table 2.

Table 2: Selected Parts of an MIT Student Profile

FIT5131 Vector		FIT5132 Vector		FIT5134 Vector		FIT5135 Vector	
Term	TF	Term	TF	Term	TF	Term	TF
:	:	:	:	:		:	:
arithmetic	1	sql	5	linux	5	lan	2
arraylist	1	subquery	1	management	7	layer	6
array	1	system	1	memory	5	layered	1
assignment	1	table	1	modern	1	link	2
attribute	1	technology	1	network	2	local	2
basic	4	theoretical	2	networking	1	method	3
behaviour	1	theory	1	operating	13	metropolitan	1
calling	1	transaction	1	performance	1	model	1
casting	1	trend	1	peripheral	1	multiplexing	1
class	9	trigger	1	personal	1	network	8
:	:	:	:	:	:	:	:

5.2. RESOURCES PROCESSING

When the student submits the query through the PersoLib web-based interface shown in Figure 5, PersoLib sends an http request to the Monash University Digital Library search engine asking for search results for the student's submitted query. The website responds to the request, and the search results are extracted accordingly. The extracted results are filtered and indexed in a later stage and are then fed to the next task for similarity computing and re-ranking. This task can be divided into three sub-tasks, as follows: a) Resource extraction, b) Resource filtering and c) Resource indexing.

The following three sections provide a detailed explanation of the mechanism used in each task.

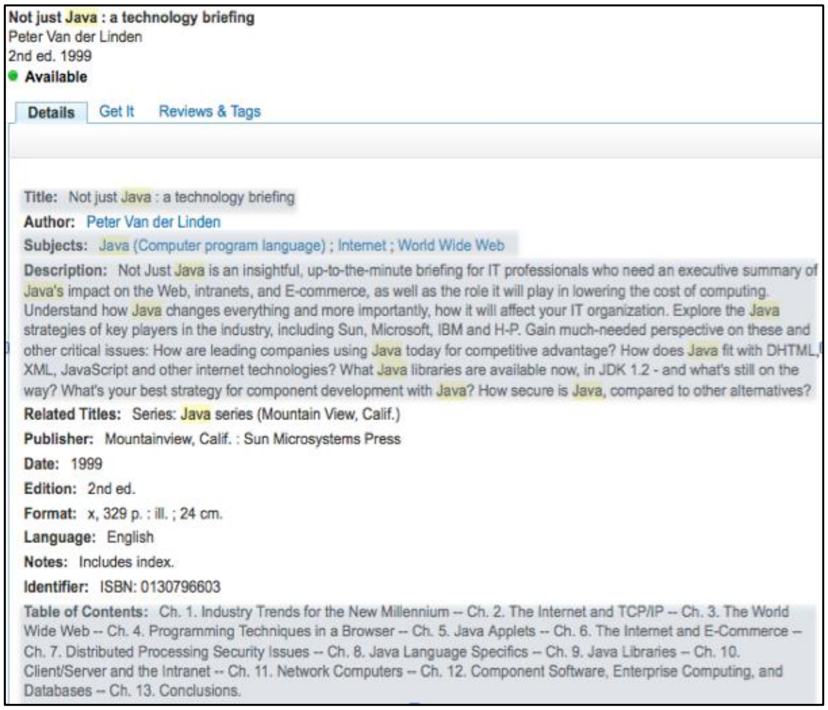
5.2.1. Resource Extraction

This is an essential task in which the initial set of search result candidates are obtained. The goal of this task is to extract the first 50 query-dependent results produced by the Monash University Digital Library search engine. The produced results are the result candidates from which the final set of filtered and re-ranked resources is formed. The choice to obtain the first

50 resources is based on the conclusions of some studies that have suggested that for vague queries, the chances are high that a search engine will be successful in getting search results that fall under different categories within this range (Matthijs & Radlinski, 2011; Teevan et al., 2005).

For each candidate resource we obtain its information through the resource details page as seen in Figure 6. The associated information for each resource may differ slightly according to the type of resource in hand. A book may have a table of contents, while a journal article most likely will not. Also, not all resources are guaranteed to be associated with the required information for library-specific reasons. Our approach, however, extracts all the available information that is assumed to provide useful information about the resource. Such information includes a resource's title, description, subjects and table of contents. Figure 6 shows the extracted information from a book entitled **NOT JUST JAVA: A TECHNOLOGY BRIEFING**.

The output of this task can be summarised as a set of 50 resources, each of which is described by its title and possibly enriched by its subjects, a description and its table of contents, if available.



Not just Java : a technology briefing
Peter Van der Linden
2nd ed. 1999
● **Available**

Title: Not just Java : a technology briefing

Author: Peter Van der Linden

Subjects: Java (Computer program language) ; Internet ; World Wide Web

Description: Not Just Java is an insightful, up-to-the-minute briefing for IT professionals who need an executive summary of Java's impact on the Web, intranets, and E-commerce, as well as the role it will play in lowering the cost of computing. Understand how Java changes everything and more importantly, how it will affect your IT organization. Explore the Java strategies of key players in the industry, including Sun, Microsoft, IBM and H-P. Gain much-needed perspective on these and other critical issues: How are leading companies using Java today for competitive advantage? How does Java fit with DHTML, XML, JavaScript and other internet technologies? What Java libraries are available now, in JDK 1.2 - and what's still on the way? What's your best strategy for component development with Java? How secure is Java, compared to other alternatives?

Related Titles: Series: Java series (Mountain View, Calif.)

Publisher: Mountainview, Calif. : Sun Microsystems Press

Date: 1999

Edition: 2nd ed.

Format: x, 329 p. : ill. ; 24 cm.

Language: English

Notes: Includes index.

Identifier: ISBN: 0130796603

Table of Contents: Ch. 1. Industry Trends for the New Millennium -- Ch. 2. The Internet and TCP/IP -- Ch. 3. The World Wide Web -- Ch. 4. Programming Techniques in a Browser -- Ch. 5. Java Applets -- Ch. 6. The Internet and E-Commerce -- Ch. 7. Distributed Processing Security Issues -- Ch. 8. Java Language Specifics -- Ch. 9. Java Libraries -- Ch. 10. Client/Server and the Intranet -- Ch. 11. Network Computers -- Ch. 12. Component Software, Enterprise Computing, and Databases -- Ch. 13. Conclusions.

Figure 6: Information extracted from a resource's details page

5.2.2. Resource Filtering

The search results produced by the Monash University Digital Library may include resources outside what a student is looking for, which is highly likely to be due to a low degree of clarity in the submitted queries. For example, a query like 'Java' would result in resources falling under many categories. While only the resources under 'Java Programming Language' are within the Master's of Information Technology student's interests when searching for information relevant to their studies; resources under other categories, such as 'Java Island Travel' or 'Java Coffee', might be included in the results list.

To exclude such resources, the text extracted from each resource, i.e. title, subjects, description and table of contents, is classified, and only those resources in the areas of students' enrolled units are considered. In particular, PersoLib makes use of an online Topic Classifier developed under the project uClassify[4] to estimate the degree to which a particular resource belongs to a number of topics, namely arts, business, computers, games, health, home, recreation, science, society and sports. The classifier is a Naïve Bayesian classifier, which is trained using one of the most comprehensive human-edited directories of the web, known as the Open Directory Project[5]. Naïve Bayesian classifier, which is a popular method for building text classifiers, uses Bayes' theorem and performs probabilistic classification on data with the assumption that all attributes are independent given the class (John & Langley, 1995).

Since the participants are MIT students, PersoLib uses the classifier to classify resources and to include only those that are estimated to be at least 50% under computer-related topics, which covers a wide range of areas, such as operating systems, data communication, artificial intelligence and computer and human interaction.

5.2.3. Resource Indexing

After obtaining the list of computer-related resources, all resources are indexed. In particular, the extracted text of each resource is transformed into a frequency vector using the same mechanism adopted in representing each unit guide. That is, the text associated with each resource, i.e. title, subjects, description and table of contents, is tokenised, filtered for stopwords, stemmed and then weighted using a TF weighing schema. We therefore define a term vector for each resource as follows:

$$\vec{R}_i = \begin{pmatrix} tf(t_1) \\ tf(t_2) \\ tf(t_3) \\ \vdots \\ tf(t_j) \end{pmatrix} \quad (2)$$

where $\vec{R}_i$ is the frequency vector of resource i, and $tf(t_j)$ is the number of occurrences of index term t_j in a resource R_i, which contains j index terms.

[4] http://www.uclassify.com

[5] http://www.dmoz.org

5.3. SIMILARITY COMPUTING AND RESOURCE RE-RANKING

This task can be viewed as the core of PersoLib, in which the obtained resources, i.e. 50 query-dependent search results, are re-ranked with respect to the student profile. We propose a *Unit Relevance Score* (*URS*) by which we rank the search results based on their relevance to the units the student is enrolled in. It assigns a score to each resource that quantifies the average similarity of the resource frequency vector to the student profile, represented by unit frequency vectors, as follows:

$$\mathrm{URS}(R_i) = \frac{\sum_{j=1}^{n} \mathrm{Sim}(\vec{R_i}, \vec{U_j})}{n} \qquad (3)$$

where R_i is the resource for which its score is measured, n is the number of units the student is enrolled in, and $\mathrm{Sim}(\vec{R_i}, \vec{U_j})$ is the cosine similarity between the resource and a given unit frequency vector. It is worth noting that both vectors will be made of the same dimension, which is equal to the size of the term vocabulary.

Though the URS is expected to rank the search results based on how relevant they are to the student, this may result in subjective search results, where the effect of the query is omitted. Therefore, we combine URS with the *Query Relevance Score* (*QRS*) given by the library search engine. For each resource, we assign a *Final Relevance Score* (*FRS*), which combines both the URS and the QRS, with α controlling the impact of each score on the *FRS*, as follows:

$$\mathrm{FRS}(R_i) = \alpha\, \mathrm{URS}(R_i) + (1 - \alpha)\mathrm{QRS}(R_i) \qquad (4)$$

The value assigned to α has an impact on the FRS and therefore influences the final ranking of PersoLib. In the evaluation of PersoLib, we experimented with three values of α as follows: $\alpha = 0.0$, where only QRS is considered, i.e. non-personalised (Library generic approach); $\alpha = 0.5$, where both QRS and URS are considered equally (PersoLib α = 0.5); and $\alpha = 1.0$, where only URS is considered (PersoLib α = 1.0).

The FRS is computed for each resource, and the resources are re-ranked accordingly, from highest to lowest.

6. EXPERIMENTAL DESIGN

An experiment was conducted for this study to evaluate the effectiveness of the proposed solution (PersoLib). This section describes the background of the participants and the method used to evaluate the proposed solution.

6.1. PARTICIPANTS BACKGROUNDS

With the assumption that new students would be more likely to benefit from the personalisation approach proposed in this research, as they are not familiar with the key terminologies of their discipline and they have limited experience in searching the library, we chose the sample participants from the population of first-year students in the Masters of Information Technology course, all of whom were studying IT foundation units. In order to recruit students, the research project was advertised at the end of tutorial sessions in students' foundation units. The research was briefly introduced to students, and a flyer was distributed to provide more details for those interested in participating.

6.2. PersoLib Evaluation Approach

Evaluating personalised searching is quite challenging, as it is highly subjective and can only be assessed by the searchers themselves. To assess whether PersoLib is producing an actual improvement in the retrieval relevance, it is important that it is evaluated by participants performing actual searches that represent real information needs. Therefore, participants were encouraged to choose queries they had already used or to formulate ones that reflected their current information needs.

The experiment took place in the middle of the semester. Participants submitted their queries through a web-based interface for PersoLib (see Figure 5) and were presented with three result sets, each of which represented the top 10 search results returned for each variation of α ($\alpha = 0.0$, where only QRS is considered, i.e. non-personalised; $\alpha = 0.5$, where both QRS and URS are considered equally and $\alpha = 1.0$, where only URS is considered). Participants were then asked to evaluate the relevance of each search result in each result set as being *very relevant*, *relevant*, or *not relevant*. The criterion to which each result set was ranked was hidden from participants so as not to bias them.

To compare the retrieval quality of the three result sets, we used the Normalized Discounted Cumulative Gain (NDCG) which is a measure commonly used in information retrieval. NDCG was proposed by Järvelin and Kekäläinen (2002) and has been used widely in the information retrieval literature as it introduces a graded relevance assessment and credits information retrieval systems that can retrieve highly relevant documents at the top of the ranking.

7. Results and Discussion

The quality of the search results is measured by the NDCG, averaged over 16 queries submitted by 16 participants, who evaluated the top 10 results produced by each of the three approaches (i.e. 480 relevance judgments in total). Three relevance levels were used to evaluate each search result, namely, very relevant, relevant and not relevant, for which we assign gain values of 2, 1 and 0, respectively. All of the reported statistical significance measures are obtained using a two-tailed t-test.

We compare the performance of our approach (PersoLib with two variations, $\alpha = 0.5$ and $\alpha = 1.0$) against the generic non-personalised Monash University Digital Library search engine. The results are summarised in Table 3 which shows the level of precision of the produced search results from the three approaches, as measured by NDCG@10 (at the 10th rank position), as well as the amount of improvement achieved by both personalisation approaches over the generic one.

Table 3: Summary of Evaluation Results

Approach	Average NDCG@10	Improvement (%)
Non-personalised	0.699	-
PersoLib (α= 0.5)	0.828	18.5%
PersoLib (α= 1.0)	0.876	25.4%

Two main observations emerge from these results. Not surprisingly, measured by NDCG@10, both personalisation approaches significantly outperform the library's non-personalised approach (PersoLib $\alpha = 0.5$, $p < 0.001$; PersoLib $\alpha = 1.0$, $p < 0.001$). Interestingly enough, PersoLib ($\alpha = 1.0$), the personalisation approach that does not take the original library ranking into account and relies purely on the relevance of the resources to student enrolment information, yields the highest average NDCG@10 score, with 25.4% improvement over the non-personalised library approach. This may indicate that, at least in some cases, some relevant resources are ranked low (down in the search results list) by the library search engine. In this light, considering the ranking returned by the library search engine lowers the FRS for such resources, and it makes it difficult to push them up into the top 10 of the re-ranked list. Ignoring the library original rank, as in PersoLib ($\alpha = 1.0$), therefore helps promote such resources so that they are included in the top 10 search results.

The average NDCG scores for the three approaches across all of the 10 rank positions are shown in Figure 7. It can be seen that both personalisation approaches manage to produce a superior performance over the non-personalised library approach across all rank positions.

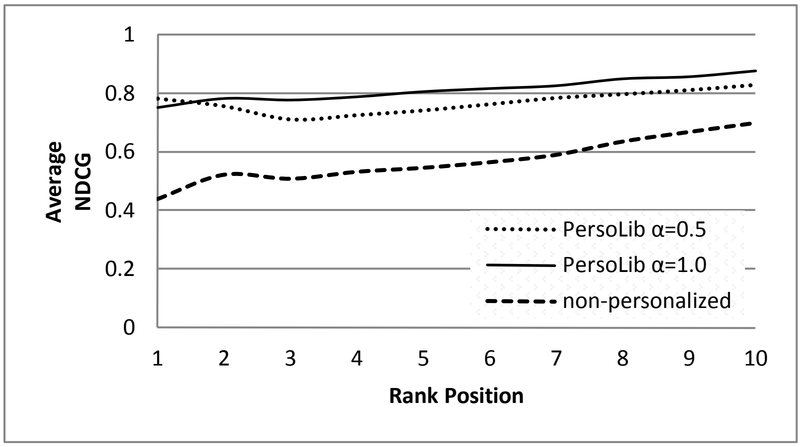


Figure 7: Average Normalized Discounted Cumulative Gain (NDCG) score obtained at each rank position from each of the three approaches.

A more detailed presentation of the results can be seen in Figure 8, in which we show all the submitted queries along with the NDCG@10 scores produced by each approach.

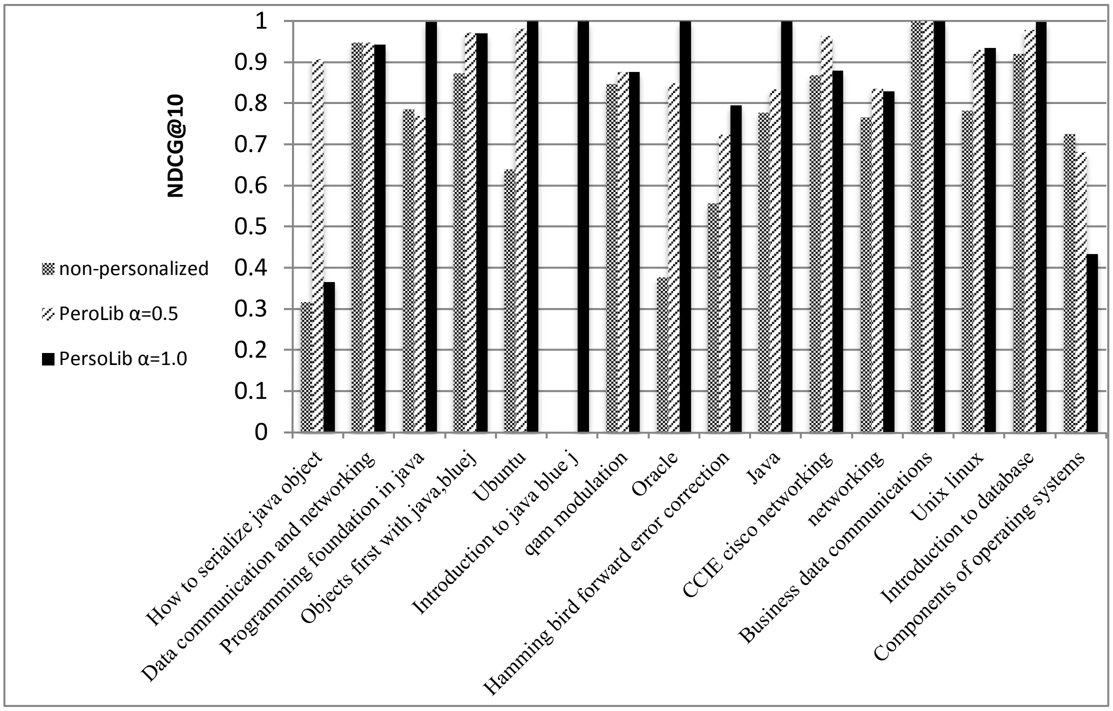


Figure 8: Comparison between the NDCG@10 scores produced by each approach for each query.

8. CONCLUSIONS

In this paper, we investigate the potential of incorporating student enrolment information into the process of identifying their learning needs to evaluate the relevance of the query-dependent resources returned from the Monash University Digital Library. In particular, we propose a framework that makes use of the collection of units a student is enrolled in to generate a student profile. The relevance of the query-dependent resources is, therefore, measured against the student profile. In doing so, we propose a Final Relevance Score (FRS) measure, which assigns a relevance score for each query-dependent resource based upon how similar the resource is to both the student profile and the submitted query. To examine the effectiveness of this framework and whether it truly produces any improvement over the library generic approach, this framework was implemented in an application called PersoLib and evaluated by a group of 16 Masters of Information Technology students doing foundation units at Monash University.

We have collected a total of 460 relevance judgments for the top 10 search results for the three approaches (library generic approach, PersoLib $\alpha = 0.5$, and PersoLib $\alpha = 1.0$). The evaluation results show that both personalisation approaches, the one that considers the library ranking (PersoLib $\alpha = 0.5$) and the one that does not (PersoLib $\alpha = 1.0$), significantly outperform the library generic approach. Although the framework we have implemented is not the most efficient in terms of the techniques employed, it shows the potential of incorporating student enrolment information in the creation of a better search environment in which students' search results are driven not only by the submitted query but also by the units in which they are enrolled.

9. Limitations and Future Research

The techniques we employ in implementing our personalisation approach are not the most efficient, and yet they produce promising results. The main purpose is to examine the potential of incorporating students' enrolment information; the techniques are not necessarily the most effective ones. Clearly, incorporating students' enrolment information has proven useful in creating customised search environments for students, and it would be interesting to explore more sophisticated techniques, particularly in indexing students' units as well as library resources. More semantic representation of the unit guide web pages and the resources' information could produce better results.

Obviously, the sample size is one of the limitations in this study and in order to draw more reliable conclusions, more subjects are needed. However, the initial results are significant in that they provide an initial insight into the usefulness of utilizing this technology using a novel manner, i.e. using textual information about students' units, and are hoped to encourage further investigation.

Student profiles are based solely on the information provided in their unit guide web pages. One of the research directions could investigate extending student profiles to incorporate a more complex representation of their learning needs and academic interests. As an example of extension, student profiles could incorporate the notion of time and represent students' current learning needs, which might be inferred from their posts and activities in the online learning system, e.g. Moodle or Blackboard, and probably the weekly topics listed in their unit guide web pages. Also, one of the research directions may investigate ways in which students' level of expertise in the areas they are studying, as well as the level of complexity of the library resources, can be measured and taken into account when ranking the library resources.

References

AKBAR, M., SHAFFER, C. A., WEIGUO, F., & FOX, E. A. (2014). *Recommendation based on Deduced Social Networks in an educational digital library.* Paper presented at the 2014 IEEE/ACM Joint Conference on Digital Libraries (JCDL).

AMINI, B., IBRAHIM, R., OTHMAN, M. S., & RASTEGARI, H. (2011). *Incorporating scholar's background knowledge into recommender system for digital libraries.* Paper presented at the 2011 5th Malaysian Conference in Software Engineering (MySEC), Johor Bahru, Malaysia.

ASHRAF, T., & GULATI, P. A. (2010). Digital Libraries: A Sustainable Approach. In T. Ashraf, J. Sharma & P. Gulati (Eds.), *Developing Sustainable Digital Libraries: Socio-Technical Perspectives* (pp. 1-18). Hershey, PA: IGI Global.

BRUSILOVSKY, P., FARZAN, R., & JAE-WOOK, A. (2005). *Comprehensive personalized information access in an educational digital library.* Paper presented at the Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, 2005. JCDL '05.

CALLAN, J., SMEATON, A., BEAULIEU, M., BRUSILOVSKY, P., CHALMERS, M., RIEDL, J., . . . TOMS, E. (2003). Personalisation and Recommender Systems in Digital Libraries: Joint NSF-EU DELOS Working Group Report.

CARMEL, D., ZWERDLING, N., GUY, I., OFEK-KOIFMAN, S., HAR'EL, N., RONEN, I., . . . CHERNOV, S. (2009). *Personalized social search based on the user's social network*. Paper presented at the the 18th ACM conference on Information and knowledge management, Hong Kong, China.

JÄRVELIN, K., & KEKÄLÄINEN, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst., 20*(4), 422-446. doi: 10.1145/582415.582418

JOHN, G. H., & LANGLEY, P. (1995). *Estimating continuous distributions in Bayesian classifiers*. Paper presented at the Proceedings of the Eleventh conference on Uncertainty in artificial intelligence.

JOMSRI, P., SANGUANSINTUKUL, S., & CHOOCHAIWATTANA, W. (2012). A Personalized Re-ranking Technique for Academic Paper Searching Based on User Profiles. *International Journal of Digital Content Technology and its Applications(JDCTA), 6*(16). doi: 10.4156/jdcta.vol6.issue16.62

KAHN, R. E., & CERF, V. G. (1988). *The Digital Library Project Volume I: The World of Knowbots, (DRAFT): An Open Architecture For a Digital Library System and a Plan For Its Development*. Reston, VA: Corporation for National Research Initiatives.

KARWEG, B., HUETTER, C., & BOHM, K. (2011). *Evolving social search based on bookmarks and status messages from social networks*. Paper presented at the the 20th ACM international conference on Information and knowledge management, Glasgow, Scotland, UK.

KIM, J. Y., COLLINS-THOMPSON, K., BENNETT, P. N., & DUMAIS, S. T. (2012). *Characterizing web content, user interests, and search behavior by reading level and topic*. Paper presented at the the fifth ACM international conference on Web search and data mining, Seattle, Washington.

KLAŠNJA-MILIĆEVIĆ, A., VESIN, B., IVANOVIĆ, M., & BUDIMAC, Z. (2011). E-Learning personalization based on hybrid recommendation strategy and learning style identification. *Computers & Education, 56*(3), 885-899. doi: 10.1016/j.compedu.2010.11.001

LESK, M. (1997). *Practical digital libraries: books, bytes, and bucks*. San Francisco, CA: Morgan Kaufmann Publishers Inc.

MATTHIJS, N., & RADLINSKI, F. (2011). *Personalizing web search using long term browsing history*. Paper presented at the the fourth ACM international conference on Web search and data mining, Hong Kong, China.

MCKEOWN, K. R., ELHADAD, N., & HATZIVASSILOGLOU, V. (2003). *Leveraging a common representation for personalized search and summarization in a medical digital library*. Paper presented at the the 3rd ACM/IEEE-CS joint conference on Digital libraries.

POMERANTZ, J., CHOEMPRAYONG, S., & EAKIN, L. (2008). The development and impact of digital library funding in the United States. *Advances in Librarianship, 31*, 37-92. doi: 10.1016/S0065-2830(08)31002-2

POTEY, M. A., PAWAR, S. P., & SINHA, P. K. (2013). *Re-ranking for personalization using concept hierarchy in DL environment.* Paper presented at the 15th International Conference on Advanced Computing Technologies (ICACT).

SHEN, X., TAN, B., & ZHAI, C. (2005). *Implicit user modeling for personalized search*. Paper presented at the the 14th ACM international conference on Information and knowledge management, Bremen, Germany.

TANG, T., WINOTO, P., & MCCALLA, G. (2014). Further Thoughts on Context-Aware Paper Recommendations for Education. In N. Manouselis, H. Drachsler, K. Verbert & O. C. Santos (Eds.), *Recommender Systems for Technology Enhanced Learning* (pp. 159-173): Springer New York.

TEEVAN, J., DUMAIS, S. T., & HORVITZ, E. (2005). *Personalizing search via automated analysis of interests and activities*. Paper presented at the the 28th annual international ACM SIGIR conference on Research and development in information retrieval, Salvador, Brazil.

WATERS, D. J. (1998). What are digital libraries. *CLIR (Council on Library and Information Resources) Issues, 1*(4), 5-6. Retrieved from http://www.clir.org/pubs/issues/issues04.html

ZHOU, D., LAWLESS, S., & WADE, V. (2012). Web Search Personalization Using Social Data. In P. Zaphiris, G. Buchanan, E. Rasmussen & F. Loizides (Eds.), Theory and Practice of Digital LibrariesLecture Notes in Computer Science (Vol. 7489, pp. 298-310): Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-642-33290-6_32. doi: 10.1007/978-3-642-33290-6_32