UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di Laurea Specialistica in Statistica e Informatica

AN EXPLORATIVE STUDY OF “THE EUROPEAN LIBRARY” PORTAL LOG-FILE TOWARD IMPLICIT FEEDBACK

Relatore: Chiar.mo Prof. Massimo Melucci

Tesi di laurea di:
Anna Munaro
Matricola n. 566342

ANNO ACCADEMICO 2007–2008

Ad Andrea

TABLE OF CONTENTS

RINGRAZIAMENTI

Grazie ai miei genitori, per avermi permesso di raggiungere questo traguardo, e per l'orgoglio che dimostrano nei miei confronti.
Grazie ad Andrea, per non aver mai dubitato che ce l'avrei fatta.
Grazie a Nicola per l'adsl e ad Alessandra per tutti i libri.
Grazie a Laura e Giorgio per tutte le stampe.
Grazie a Simona e Andrea, sempre pronti a chiarire ogni dubbio.
Un grazie speciale a Veronica, Serena, Mattia, Irene, Marta e Carlo, per le volte in cui mi hanno costretta a staccare la spina.

Ringrazio il professor Melucci per avermi guidato con infinita disponibilità in questo lavoro di tesi.
Ringrazio l'ing. Di Nunzio per avermi aiutata nell'elaborazione dei dati.
Infine, un ringraziamento al gruppo di ricerca in Sistemi di Gestione delle Informazioni del Dipartimento di Ingegneria dell'Informazione dell'Università di Padova per aver messo a disposizione il log file.

INTRODUZIONE

TELplus è un progetto e-Content Plus ECP-2006-DILI-510003 specifico "per le biblioteche digitali sostenuto dalla Conferenza delle Biblioteche Nazionali Europee (CENL). The European Library (TEL) è un servizio gestito dalla Biblioteca Nazionale dei Paesi Bassi a nome della CENL, che ha ricevuto sostegno alle varie fasi da parte della Commissione Europea."[1] TELplus, "iniziato nel settembre del 2007, è un altro elemento di Europeana, la biblioteca, archivio e museo di oggetti digitali europea, ed è volto a rafforzare, estendere e migliorare il servizio di The European Library. Questo obiettivo sarà raggiunto affrontando una serie di questioni fondamentali, compreso il miglioramento dell'accesso attraverso l'accordo con OAI, rendendo disponibili, mediante OCR, i contenuti digitali di più di 20 milioni di pagine dai documenti delle biblioteche nazionali europee, migliorando la ricerca e il reperimento multilingua e aggiungendo servizi per la manipolazione e l'uso di contenuti."[2]

L'Università di Padova è leader di *WP5 on User personalisation services – log file analysis and use of annotations.* Con questo *work package*, TELplus ha l'obiettivo di migliorare le funzionalità per l'interazione con il sistema, focalizzandosi sull'analisi delle esigenze dell'utente e sul disegno di funzionalità di ricerca innovative. Tra queste innovative funzionalità di ricerca, particolare attenzione è stata posta sulla personalizzazione dei servizi necessari per rendere la ricerca per l'utente finale più efficace rispetto alla corrente versione del portale TEL. Una questione fondamentale per la personalizzazione è il contesto, le sue dimensioni e i suoi fattori.

Durante il seminario *National Science Foundation (NSF) Information and Data Management Principal Investigator* del 2003, un gruppo di ricercatori si è riunito per discutere un maggiore uso del contesto per l'accesso all'informazione. I

[1] http://ec.europa.eu/information_society/activities/econtentplus/projects/cult/telplus/index_en.htm

[2] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

ricercatori intervenuti in questo seminario provengono da diverse comunità di ricerca; nonostante ciò, c'è un ampio accordo sul fatto che il contesto è importante e non sufficientemente sfruttato dagli attuali sistemi di informazione.

Al fine di migliorare l'efficacia della ricerca, è comunemente accettato nella comunità di ricerca dell'Information Retrieval che l'osservazione delle azioni degli utenti finali è un compito necessario, soprattutto se si devono progettare e attuare servizi di personalizzazione; infatti il bisogno di informazioni di un utente in un certo momento e luogo è diverso da quello della stessa persona in un'altra situazione, o dal bisogno di un altro utente in situazioni analoghe.

Per studiare le azioni degli utenti, un'utile e talvolta necessaria fonte di informazioni è costituita da un file di log che registra le azioni svolte da ogni singolo utente; se il file di log è ben progettato, da esso si possono ottenere gli indicatori di ciò che l'utente fa quando interagisce con il sistema; attraverso l'interazione l'utente fornisce implicitamente indizi circa la sua soddisfazione e circa l'utilità o la rilevanza dei documenti reperiti. TEL offre agli utenti l'opportunità di registrarsi: la registrazione permette agli utenti di beneficiare di servizi personalizzati, e consente ai ricercatori di disporre di ulteriori informazioni riguardo agli utenti stessi. Attualmente, il numero di registrati è relativamente piccolo rispetto al totale degli utenti, ed è stato notato che gran parte di essi ha utilizzato il portale TEL solo per un giorno o solo per una sessione di ricerca. Tuttavia, tale numero è sufficientemente elevato, in termini assoluti, da consentire l'analisi dei dati.

Per migliorare l'efficacia della ricerca si possono sfruttare le tecniche di reperimento dell'informazione basate sull'interazione tra utente e sistema; queste tecniche sono chiamate Relevance Feedback. L'idea è quella di prendere i documenti reperiti in risposta a una determinata interrogazione, utilizzare le informazioni fornite dagli utenti implicitamente o esplicitamente su quali siano i documenti rilevanti e quali non lo siano, estrarre parole chiave o altri dati provenienti dai documenti rilevanti, modificare l'interrogazione ed eseguire l'interrogazione modificata. Si possono pertanto distinguere due diversi tipi di Relevance Feedback: il primo esplicito e il secondo implicito, a seconda delle azioni effettuate dagli utenti per esprimere la rilevanza dei documenti.

Quando si applica l'Explicit Feedback, gli utenti sono tenuti a esprimere giudizi di rilevanza per indicare il loro interesse verso le pagine reperite. Nello studio riportato in questa tesi, si è deciso di non applicare l'Explicit Feedback perché, come documentato nella letteratura pertinente, l'utente è riluttante a segnalare i documenti rilevanti per il sovraccarico a cui è sottoposto. Inoltre, doversi fermare per fornire le valutazioni può modificare i normali comportamenti di navigazione e di lettura.

L'Implicit Relevance Feedback, invece, utilizza le azioni degli utenti come indicatori impliciti di interesse; questo dà alcuni vantaggi sia agli utenti che ai ricercatori: gli indicatori impliciti evitano agli utenti di dover valutare i documenti; potenzialmente, ogni interazione dell'utente può contribuire alla creazione di un indicatore implicito. Inoltre, essi possono essere raccolti gratuitamente e possono essere combinati con altri indicatori impliciti per una valutazione più accurata, o con indicatori espliciti per una migliore classificazione.

Gli indicatori basati sul comportamento più analizzati includono il tempo di visualizzazione, il salvataggio, la stampa, la selezione e il *bookmarking*. Un esempio di insieme di dati al quale si possono applicare le tecniche di Implicit Relevance Feedback è un file di log, cioè un file in cui un'applicazione di Information Retrieval registra le operazioni nell'ordine in cui sono state eseguite.

Vi sono sostanzialmente due modi per raccogliere i giudizi di rilevanza dagli utenti: (1) attraverso l'osservazione delle azioni di un gruppo di utenti invitati a svolgere alcune azioni fissate da un certo protocollo; (2) utilizzare i file di log di un sistema informatico che consente l'accesso ai propri file da parte degli utenti. Nel primo modo, sarebbe difficile raccogliere dati da un numero significativo di persone e i risultati ottenuti non sarebbero generalizzabili, cioè il modello ottenuto potrebbe non adattarsi bene alla popolazione generale composta da tutti gli utenti del sistema; inoltre, questi esperimenti sono difficili da replicare perché la loro ripetizione implica la selezione di soggetti con le stesse caratteristiche.

In questa tesi si riporta l'analisi dell'action log file del portale TEL che si riferisce al periodo tra il 1° novembre 2007 e il 28 febbraio 2008; il file contiene 498292 righe riferite a 63528 diverse sessioni. Il file di log contiene informazioni sui comportamenti di interazione degli utenti del portale TEL. In particolare, ogni azione che l'utente esegue durante la sua attività di ricerca è registrata nel file di log, e corrisponde a una riga del dataset. Ricerche come quella riportata in questa tesi sono ripetibili grazie alla disponibilità del file di log di TEL; per ripetere lo studio, non è richiesto il reclutamento di nessun soggetto con nessuna particolare caratteristica.

Il principale obiettivo di questa tesi è di fornire un'analisi esplorativa del file di log di TEL per individuare la più appropriata tecnica di Implicit Feedback.

In questo file di log, i giudizi di rilevanza non sono presenti. Tuttavia, vale la pena ricordare che il nostro dataset contiene alcuni indicatori che possono essere considerati come indizi di rilevanza. Tali indicatori sono definiti come "proxy", dove proxy si riferisce a un indicatore statistico che descrive un certo fenomeno non direttamente osservabile o non oggettivamente misurabile. In particolare, a nostro parere le azioni che forniscono informazioni circa l'interesse dell'utente, cioè i proxy, sono tutte quelle relative alle operazioni di conservazione, come ad esempio la stampa, il salvataggio e l'invio per e-mail di un record reperito.

Sulla base delle caratteristiche del dataset e degli utenti emerse dall'analisi del file di log, si è scelto di studiare una tecnica di Implicit Feedback la cui proprietà principale è quella di prendere in considerazione il contesto in cui ogni attività di ricerca è effettuata.

In particolare, l'analisi del file di log ha mostrato che gli utenti del portale TEL costituiscono un gruppo eterogeneo di soggetti, e che essi svolgono una serie di attività di ricerca, ciascuna diversa dalle altre, quindi è importante considerare una metodologia che distingua ogni attività di ricerca dalle altre attraverso la descrizione del contesto. Inoltre, si è constatato che molti indicatori impliciti di interesse possono essere ottenuti dal dataset. Gli indicatori impliciti di interesse permettono di delineare il contesto in cui ogni utente esegue le sue azioni di ricerca, quindi si considera il maggior numero di indicatori possibile. Se si riuscissero ad accumulare molteplici aspetti di interazione con l'utente, invece di sfruttare un solo indicatore, diventerebbero disponibili più elementi di prova sulla rilevanza, e potrebbero potenzialmente essere creati algoritmi di Implicit Relevance Feedback più robusti. Pertanto, è stato fatto uso di tutti gli indicatori impliciti deducibili dal file di log per definire il contesto in cui ogni utente esegue la sua attività di ricerca.

In sostanza, se si riuscisse a cogliere e sfruttare il contesto in cui è maturata l'esigenza informativa, allora sarebbe possibile reperire in risposta all'interrogazione tutti i documenti composti nello stesso contesto dell'interrogazione stessa e, di conseguenza, migliorare l'efficacia dei sistemi di reperimento dell'informazione.

Infine, si presenta un modello geometrico basato sugli spazi vettoriali che utilizza più indicatori impliciti di interesse per sviluppare un modello di Implicit Feedback personalizzato per ciascun utente. L'intuizione alla base di questa metodologia è che un vettore è generato da una base così come un oggetto informativo o un'esigenza informativa è generato all'interno di un contesto. Precedenti studi condotti in letteratura dimostrano l'efficacia della suddetta tecnica.

La tesi è strutturata come segue.

Il primo capitolo è dedicato all'Implicit Relevance Feedback: in primo luogo, si introduce questo concetto, poi si presenta una panoramica della letteratura sulle tecniche di Implicit Relevance Feedback. Si citano alcuni esperimenti sviluppati per identificare gli indicatori impliciti di interesse e si presentano alcuni studi sperimentali che illustrano algoritmi di IRF. In seguito, vengono fornite alcune considerazioni circa l'applicabilità delle tecniche descritte al caso TEL. Si illustrano le differenze tra il dataset utilizzato negli studi descritti e il file di log di TEL. Infine, vengono menzionati i risultati dei suddetti studi utili a suggerire una metodologia da applicare al caso TEL.

Nel secondo capitolo si fornisce un'analisi esplorativa del file di log di TEL. In primo luogo, si illustrano le caratteristiche rilevanti dei sistemi di automazione delle biblioteche e dei sistemi di biblioteca digitale che devono essere presi in considerazione per affrontare lo studio dei file di log nelle biblioteche digitali. Dopo di ciò, si mostra il contenuto informativo del file di log di TEL; in particolare, si fa capire che cosa succede nel file di log quando un utente accede al portale di The European Library.

Nel terzo capitolo si studia la metodologia da adattare al caso TEL. Si fornisce un'illustrazione della metodologia per la navigazione e la ricerca tenendo in considerazione il contesto, e poi si propone un modo per adattare la tecnica illustrata al file di log di TEL, fornendo un'applicazione utilizzando dati del file di log.

INTRODUCTION

TELplus has been an e-Content Plus ECP-2006-DILI-510003 targeted project "for digital libraries supported by The Conference of European National Librarians. The European Library (TEL) is a service managed by National Library of the Netherlands on behalf of CENL which has received support at various stages from the European Commission."[3] TELplus, "started in September 2007, is another building brick in the creation of Europeana, the European digital library, museum and archive, and is aimed to strengthen, extend and improve The European Library service. This will be achieved by addressing a number of key issues, including improving access through OAI compliancy, making more than 20 million pages from the European National Libraries' digital content available with OCR, improving multilingual search and retrieval and adding services for the manipulation and use of content."[4]

The University of Padua is leader of WP5 on User personalisation services – log file analysis and use of annotations. In this work package, TELplus has aimed at improving the functionalities for user-interaction by emphasizing on the user requirement analysis and the design of innovative search functionalities. Among these innovative search functionalities, a great deal of attention has been paid to the personalization services needed to make search more effective for the end user than it is with the current version of the TEL portal. A key issue of personalization is context, its dimensions and factors.

At the 2003 National Science Foundation (NSF) Information and Data Management Principal Investigator workshop, a group of researchers met to discuss greater use of context for information access. The researchers involved in this workshop came from distinct research communities; in spite of this,

[3] http://ec.europa.eu/information_society/activities/econtentplus/projects/cult/telplus/index_en.htm

[4] http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus/

there is broad agreement that context is important and not sufficiently exploited by current information systems.

In order to improve search effectiveness, it is commonly accepted in the Information Retrieval research community that the observation of the actions of the end users is a necessary task, especially if personalization has to be designed and implemented; indeed, the information need of a user in a certain moment and place is different from the need of the same person in another situation or from the need of another user in similar situations. To study the users' action, a log file recording the actions performed by each single user is a useful, and sometimes necessary, source of information since it provides, if it is well designed, the indicators of what the end user did when interacting the system thus implicitly providing clues about the usefulness, satisfaction or relevance of the retrieved documents. At this aim, TEL offers the opportunity to register: the registration allows to the users to benefit of personalized services, and allows to the researchers to dispose of further information about the users. Currently, the number of registered users is relatively small if compared to the total size, and it has been noticed that the wide part of them has used TEL portal only for a day or just for a search session. However, that number is quite high in absolute terms, thus permitting the analysis of the data.

To improve search effectiveness, the information retrieval techniques based on the interaction between user and system can be exploited; these techniques are called Relevance Feedback. The idea is to take the documents retrieved in response to a certain query, using the information given by the user implicitly or explicitly about which documents are relevant and which are not, extract key words or other data from the relevant documents, modifying the query and executing the modified query. Thus, we can distinguish two different kinds of Relevance Feedback: a first one explicit and a second one implicit, depending on the action made by the user to express the documents' relevance.

When Explicit Feedback is provided, the users are required to express relevance assessments to indicate their interest towards the retrieved pages. In the study reported in this thesis, it was decided not to apply Explicit Feedback because, as reported in the relevant literature, the user is unwilling to mark the relevant documents due to the cognitive overload. In addition, having to stop to enter explicit indicators can alter normal patterns of browsing and reading.

Implicit Relevance Feedback, on the other hand, utilizes users' actions as implicit interest indicators; this gives some advantages both to the users and the researchers: the implicit indicators remove the cost of the user evaluating documents; potentially, every user interaction can contribute to an implicit indicator. Furthermore, they can be gathered "for free" and they can be combined with other implicit indicators for a more accurate rating, or with explicit indicators for an enhanced rating.

The most analyzed behavior-based indicators include display time, saving, printing, selecting and bookmarking. An example of dataset to which Implicit Relevance Feedback techniques may be applied is a log file, that is, a file in which an Information Retrieval application records the operations ordered by execution.

There are substantially two ways to gather user relevance assessments: (1) by the observation of the actions of a users' group invited to develop some actions fixed by a certain protocol; (2) utilizing log files of an computer system which allows the access to its own files from the users. In the first way, it would be difficult to gather data from a significant number of people and the obtained results would be not generalisable, that is, the obtained model could not good fit to the general population composed by all the system's users; in addition, these experiments are difficult to replicate because the replication implies the recruitment of subjects with the same characteristic.

In this thesis, we report the analysis of the TEL portal action log file that refers to the period between the 1st November 2007 and the 28th February 2008; it contains 498,292 rows regarding 63,528 different sessions. The log file contains information about the TEL portal's users interaction-behaviors. In particular, each action the user performs during his seeking activity is recorded in the log file, and corresponds to a row of the dataset. The research like that in this thesis is repeatable due to the availability of the TEL action log file; in order to replicate it, the recruitment of any subject with particular characteristic is unnecessary.

The firt objective of this thesis is to provide an explorative analysis of TEL log file to individuate the most appropriate Implicit Feedback technique.

In this log file, relevance assessments are not present. However, it is worth pointing out that our dataset contains some indicators which can be seen as clues of relevance. These indicators are defined as "proxy", where proxy refers to a statistic indicator that describes a certain phenomenon not directly observable or not objectively measurable. In particular, it is our opinion that the actions which provide information about the user interest or relevance, that is, the proxies, are all those referring to operations of retain, such as the action of printing, saving and sending by e-mail a result record.

On the basis of the characteristics of the dataset and of the users emerged from the analysis of the log file, we have chosen to investigate an implicit feedback technique whose main property is to take the context in which each search activity is performed into consideration.

In particular, the analysis of the log file raised that the TEL portal's users constitute an heterogeneous group of subjects, and that they perform a set of search activities, each different from the others, thus, it is important to consider a methodology which distinguishes every seeking activity from the others through the description of the context. Furthermore, it was found that many implicit interest indicators can be obtained from the dataset. The implicit interest indicators allow to delineate the context in which every user performs his actions of search, thus we consider the major number of indicator as possible. If we could capitalize on multiple aspects of user interaction, rather than exploit only one indicator, more evidence about preferences would become available, and more robust IRF algorithms could potentially be created.

Therefore, we made use of all the implicit indicators deducible from the log file to define the context in which each user performs his search activity.

Essentially, if we could gather and exploit the context in which the information need is reached, then it could be possible to retrieve all the documents which match the query in the same context of the query itself, and consequently improve the information retrieval systems' effectiveness.

Finally, we present a geometric framework based on vector spaces that utilizes multiple implicit interest indicators to develop enhanced implicit feedback models personalized for each user. The intuition underlying this methodology is that a vector is generated by a basis just as an informative object or an information need is generated in a context. Previous studies conducted in literature demonstrate the effectiveness of the above-mentioned technique.

The thesis is structured as follows.

The first chapter is devoted to Implicit Relevance Feedback: first, we introduce this concept, then we present a review of the literature about Implicit Relevance Feedback techniques. We mention some experiments developed to identify implicit interest indicators and we address the experimental studies on IRF algorithms. In the following, some considerations about the applicability of the techniques described to TEL case are provided. The differences between the dataset used in the studies described and TEL action log file are illustrated. Finally, the findings of the mentioned studies useful to suggest a methodology to apply to TEL case are mentioned.

In the second chapter we provide an explorative analysis of TEL action log file. First, we provide relevant characteristics of library automation systems and digital library systems that need to be taken into account in addressing the study of log data in digital libraries. After this, we show the informative content of TEL log file; in particular, we make understand what happens in the log file when a user accedes to The European Library portal.

In the third chapter we investigate the methodology that can be fitted to the TEL case. We provide an illustration of the methodology for navigation and search in context, and then we propose a way to fit the technique illustrated to TEL log file by providing an application utilizing data from the log file.

CHAPTER 1

IMPLICIT FEEDBACK

1.1 Introduction to implicit relevance feedback (IRF)

Relevance Feedback aims at improving the effectiveness of an Information Retrieval (IR) system by removing non-relevant documents and adding relevant documents using relevance or non-relevance assessments obtained from the user who is then not expected to directly construct new search strategies.

Efthimiadis provides a description of the typical automatic relevance feedback operations in [14]. According to that paper, Relevance Feedback requires the user expresses a query which is processed by the system for retrieving an initial set of documents. Then, the searcher chooses some relevant documents from the list of the retrieved records. Those documents are used for reweighting the existing query terms and/or by adding terms which appear as useful or deleting terms which do not. This process creates a new query which resembles the relevant documents more than the original query does.

Many experiments have demonstrated that Relevance Feedback allows to retrieve a larger number of relevant documents than that of the relevant documents retrieved in response to the initial query. Thus, this technique improves the system effectiveness, both in precision and recall.

Relevance Feedback, generally, can be implemented in various ways depending on the retrieval model used, such as the vector space or the probabilistic model, and also on the methods used to select the terms for the post-feedback query. Efthimiadis distinguishes four term selection methods for query reformulation and expansion.

1. The first relies entirely on the original query and uses only those terms in the new one [37, 38, 45].
2. The second method uses terms from the original query and adds terms from some other source [41, 42].
3. The third method is a mixed method because it combines the terms derived from the original query and those derived from the documents retrieved and judged relevant [39, 48, 50].
4. The fourth method abandons the terms from the original query and uses only the terms found in the retrieved set of documents [12, 13].

Query reformulation and expansion is entrusted entirely to the retrieval system. Query expansion can be performed with or without term reweighting—if without term reweighting, query expansion may involve the addition of terms from a knowledge structure, such as thesauri or term classifications. Most research on Relevance Feedback and query expansion has been done using both query expansion and term reweighting.

The relevance assessments can explicitly be gathered from the user who has submitted the query or using other methods which can be automatic. The latter case is called Implicit Relevance Feedback (IRF). Using IRF, the system observe user's behaviour and modifies the retrieved document set with the aim of offering a larger number of relevant documents.

Let us look at the advantages referring to the above mentioned techniques, by distinguishing between explicit and implicit feedback. Claypool et al. provide an objective review about benefits and handicaps in reference to the two types of feedback in [11]. In that work, they study the correlation between various implicit indicators and the explicit indicator of usefulness for a single web page. They used the explicit indicators exclusively to show the validity of the implicit indicators in gathering the user interest through the measurement of the correlation.

Through the explicit indicators, the users tell the system what they think about some object or piece of information. Explicit indicators are well understood, fairly precise, and are common in everyday life. However,

- having to stop to enter explicit indicators can alter normal patterns of browsing and reading;
- unless users perceive that there is a benefit from providing indicators, they may stop providing them. Hence, users may continue to display, thus resulting in system use, but no ratings at all;
- some research has found that when giving explicit indicators, users were displaying a lot more articles than they were rating;
- collaborative filtering requires many indicators to be entered for every item in the system in order to provide accurate predictions.

Hence explicit indicators, while common and trusted, may not be as reliable as is often presumed.

On the other hand, some obvious advantages of the implicit indicators are:

- they remove the cost of the user examining and rating documents;
- potentially, every user interaction can contribute to an implicit indicator.

Although each implicit indicator is likely to be less accurate than an explicit indicator, they:

- can be gathered "for free";
- can be combined with other implicit indicators for a more accurate rating;
- can be combined with explicit indicators for an enhanced rating.

This chapter contains a review of the literature about IRF techniques. First, some experiments developed to identify implicit interest indicators are described—these works will help us to identify the indicators attainable starting from TEL action log file.

In the following, the experimental studies on IRF algorithms are addressed—these studies are useful to suggest a methodology for the TEL case.

To conclude, some considerations about the applicability of the techniques described to TEL case are provided. The differences between the dataset used in the studies described and TEL action log file are illustrated. Finally, the findings of the mentioned studies useful to suggest a methodology to apply to TEL case are mentioned.

1.2 Previous studies

1.2.1 Implicit interest indicators

The research works relevant to the selection of the implicit interest indicators are presented first. A set of papers were selected with the aim of presenting an overview of the different approaches to IRF. We have chosen this specific set of research works because the most part of the indicators mentioned in these papers can be obtained from TEL action log file. In the following, the illustration of these research works was organized in a way that the emphasis was given on the methodological issues of the implicit interest indicators.

1.2.1.1 Classifications of the implicit indicators

First of all, it is worth presenting a general classification of the implicit indicators. To be precise, three different categories of indicators will be presented: the first two are one the refinement of the other, and they are more general than the third, in the sense that they can also be applied to non-textual documents, while the third is valid only for textual documents.

Kim et al. provide a framework in which the behaviors are categorized according to two axes [26]:

- Behaviour Category refers to the underlying purpose of the observed behaviour,

- Minimum Scope refers to the smallest possible scope of the item being acted upon.

This framework is reported in Table 1.1:

- the segment level includes operations whose natural scale is a portion of a document, for example, viewing a screen,
- the object level includes behaviors whose natural scale is an entire document, for instance purchase,
- the collection level includes behaviors whose natural scale includes more than one document (subscription).

By "natural scale" the authors mean the smallest unit normally associated with the behavior.

The choice of segment, object and collection as labels is intentionally inclusive, since the ideas captured in the table would apply equally well to non-text modalities such as video or music with only minor variations.

Interestingly, when viewed from this perspective, explicit feedback is merely one type of user behavior observed.

		Minimun Scope		
		Segment	**Object**	**Class**
Behavior Category	**Examine**	View	Select	
	Retain		Bookmark Save Purchase Print Delete	Subscribe
	Reference	Copy-and-Paste Quote	Forward Reply Link Cite	
	Annotate	Annotate	Rate Publish Organize	

Table 1. 1

Kelly and Teevan provide a refinement to the framework presented by Kim et al. [24]

They added a fifth behaviour category called "Create" to the original; this new category describes those actions the user engages in when creating original information. The researchers also added some additional commonly investigated observable behaviours. The classification scheme is displayed, with example behaviors, in Table 1.2.

Behavior Category	Minimun Scope		
	Segment	**Object**	**Class**
Examine	View Listen Scroll Find Query	Select	Broswe
Retain	Print	Bookmark Save Delete Purchase Email	Subscribe
Reference	Copy-and-Paste Quote	Forward Reply Link Cite	
Annotate	Mark up	Rate Publish	Organize
Create	Type Edit	Author	

Table 1. 2

Many of the papers of the literature reviewed in this thesis can be classified according to the reported tables.

As Kelly and Teevan suggested, a preponderance of the research works falls into the "Examine Object" category. This fact can be explained because many indicators included in "Examine Object", like document selection and viewing time, are relatively easy to obtain and are available for every object with which a user interacts.

Other categories contain little or no work, thus suggesting possible categories of observable behaviour to explore. One likely reason for the lack of literature across the Minimum Scope categories of “Segment” and “Class” is that the unit with which the user interacts is the “object” for many systems. An exception to this is that many annotation systems consider segments, thus suggesting the reason why much of the annotation literature falls into this category.

A further categorization has been done by Kelly and Teevan; they examined some papers that fell into the “Examine Object” category and classified them along two additional axes. One axis represents the standard software lifecycle based on the spiral model of software development, and its possible values are: design, implementation, evaluation. Of course, all three of these categories overlap, particularly because the work with implicit indicators is still in its infancy. The other axis focuses on whether the research deals with user preferences on an individual or group level.

A different categorization has been done by Claypool et al. [11]. They divide the interest indicators into the following categories.

- Explicit Interest Indicators: This category includes the selection by the user of a value from a scale.
- Marking Interest Indicators: These comprise bookmarking a Web page, deleting a bookmark, saving the page as a file, emailing the page, or printing it.
- Manipulation Interest Indicators: This group includes actions such as cutting and pasting, opening a new browser window searching in the page for text, or scrolling.
- Navigation Interest Indicators: The spending time with the page opening, following, or not following a link are considered forms of navigation interest indicators.
- External Interest Indicators: These concern with the user's physical responses to information, such as heart-rate, perspiration, temperature, emotions and eye movements.

- Repetition Interest Indicators: The authors hypothesize that doing more of something means more interest, so the spending more time on a page, doing lots of scrolling through a page, and repeatedly visiting to the same page are called repetition interest indicators.
- Negative Interest Indicators: Claypool at al. think that the absence of an indicator might be considered to be a negative indicator. They recognize that it is very difficult to distinguish between, for example, deliberately not visiting a page, and merely just not visiting it. However, they say that one could accumulate evidence in order to increase the reliability of the indicator.

The researchers pointed out that some indicators may be context sensitive, depending on the user's task or the category of the page. In addition, different combinations of indicators might mean different things. For example, if a user does not display a document for very long, but he does bookmark it, the short time might suggest that he does not like the page, while the bookmark might suggest that he does. In this case, he probably bookmarked it for later reading and we do not yet know if he likes it or not. This interpretation of the indicator combinations will be carried on in Chapter 3 devoted to IRF in the TEL case and specifically when a methodology based on the vector spaces, matrix decomposition and principal component analysis which aims at extracting these combinations will be illustrated. As illustrated in Chapter 2, in our dataset these situations, namely actions of retaining are often associated to lower display times, were observed, thus confirming the considerations by Claypool et al..

1.2.1.2 Selecting implicit interest indicators

The modalities with which the experiments aimed to select implicit interest indicators are illustrated.

- In general, these experiments are designed by recruiting about ten subjects and requiring them to read some documents for a certain time period.
- The subjects are also required to express their relevance assessments about the documents.
- During the test, the researchers measured some indicators suspected to be a sign of interest.
- After having gathered the users' behaviours and their relevance assessments, the researchers measured the correlation between the relevance assessments and the implicit indicators.

Some key papers which cover a range of procedures are presented in the following.

Morita and Shinoda [35] carried out an experiment in which eight users were for six weeks required to read all articles that were posted to the newsgroups of which they were members and to explicitly rate their interest in the articles.

The authors measured the display time, saving or follow-ups of a story; they further examined the relationship of three variables on displaying time: the length of the document, the readability of the document and the number of news items waiting to be read in the user's news queue.

Golovchinsky et al. [16] and Budzik and Hammond in [9] suggested that evidence of context can be found in numerous other applications with which the user interacts. Budzik and Hammond [9] proposed a system that automatically retrieved documents and recommended URLs to the user based on what the user was typing. The authors asked ten researchers to submit an electronic version of a paper that they wrote and then asked these users to evaluate the documents that their experimental system had retrieved based on these texts.

Claypool et al. [11] studied the correlation between various implicit indicators and the explicit indicator for a single web page. The methodology used is as follow:

- A browser called "The Curious Broswer" was implemented for gathering data on as many implicit interest indicators as possible;
- a user study was conducted with many participants browsing the web with this browser;
- the correlation between implicit interest indicators and explicit interest was analyzed.

The first time each web page was visited, the Curious Browser stored the user name, the URL, the time and date, the explicit indicator and all implicit interest indicators. Subsequent returns to the same page were not recorded.

The Curious Browser was available from March 20, 2000 to March 31, 2000. During this time, 75 students visited a total of 2,267 web pages. They were instructed to open up the Curious Browser and browse the web for 20-30 minutes, but were not told the purpose of the experiments.

The implicit interest indicators analyzed were:

- the time spent on a page;
- the time spent moving the mouse;
- the number of mouse clicks;
- the time spent scrolling.

Initially, Claypool at al. analyzed the mean of each implicit interest indicator versus the explicit indicator. However, the mean of any of the implicit indicator proved to be a poor indicator of explicit interest because of some extreme outliers. Thus, they focus on the median and distribution of each indicator using a Kruskal-Wallis test (based on 0.05 level of significant) to examine the degree of independence of the medians among each explicit rating groups for each implicit interest indicator.

A study quite different from those presented above was described by Rafter and Smith [36]. The goal of the study was to evaluate the validity of the assumption that accurate user profiles can be generated by analysing user behaviour in the CASPER system. CASPER system investigates personalisation technologies such as case-based reasoning and collaborative filtering to “JobFinder”, which is an online recruitment service.

This paper is cited and a larger description is devoted to it in this thesis because the experiment differently from the previously cited works is not based on the recruitment of a certain number of subjects, but it is based on the analysis of server logs, which is the same type of dataset as that at our disposal from The European Library portal.

“JobFinder” server logs recorded details of the user interactions within the website. Essentially, each line of the action log file recorded a single job access by a user, and encoded details like the time of access, the job and user identifiers. In addition to this, any action that the user performed with respect to that job is recorded. To obtain a more detailed profile representation of the user relevance information were also need to discriminate between those jobs that the user looks at or considers, and those that he is truly interested in.

Graded profiles supplemented the basic profile representation with relevancy indicators. These indicators are essentially the set of grades that measure the relevance of each item for that user. In CASPER, these grades correspond to three main types of information: the number of revisits made to a job description, the amount of time spent for displaying a job description, and whether the user applied for a job or mailed it back to himself. A more detailed description of the indicators observed by CASPER is provided in the following.

The number of times that a user clicks on a job is thought to be an indicator of his interest in that job, in the sense that the users will often return for a second or third display of an interesting job description while they are unlikely to revisit uninteresting jobs after an initial display. However, the number of times a user clicks on a job may not be correlate with the number of times that user revisits the job. Indeed, many of these clicks are so called ”irritation clicks” due to a frustrated user repeatedly clicking on a job in the event of, say, bandwidth problems while waiting for the description to download, and therefore these clicks do not constitute accurate revisit data. In order to deal with this misleading revisit data, CASPER employs a thresholding technique that counts repeated clicks on the same job as irritation clicks.

The time a user spends displaying a job description has been shown to correlate with that user's degree of interest. Again, a suitable thresholding technique is necessary to eliminate spurious display times due to a user logging off or leaving his terminal. In order to prevent spurious display times interfering with the identification of relevant jobs within a profile Rafter and Smith adopt a two-step process. This process is designed to identify some average value for the time it takes to display a job, and then replace any display times that deviate wildly from the average. The approach involved used the median of median display time values per individual job access for both users and jobs to calculate a normal display time for the system. The second step was to find any display times (per job access) within the profiles that had a display time greater or equal to twice the system median. This produced a set of adjusted display times where all the display time values are reasonable.

The final and perhaps most reliable interest indicators were JobFinder online application or email facility. A JobFinder user can either email a job description to himself, for later consideration, or apply for the job directly online. These actions indicate a more serious interest in a job than a simple displaying of the job description. However, users tend to do this infrequently, or not at all, resulting in insufficient data to exclusively base relevancy predictions on. As a result, the researchers preferred not to use the activity data for the profiling in that paper, and rather used it as a way of measuring the accuracy of the other indicators.

The experimental study was based on the user profiles generated from server logs between 2/6/98 and 22/9/98, which contained 233,011 job accesses by 5,132

different users. As the authors assume that the action of a user applying for a particular job online is a reliable indicator of his interest in that job, they evaluated the display time and revisit data based on how well it correlates with this information. They also tested whether the improved indicators of revisit and display time data improve prediction performance, or not. The experiments were therefore restricted to the set of those users who applied to at least one job. Furthermore, they only took users with a profile size (number of jobs in profile) of 15 or greater. These

users numbered 412 in total and were used as the profile base for the experiments. For each user in the profile, the researchers produced two sets of predictions for the jobs that the user applied for, based on the two kinds of revisit data. For each set of predictions then, they produced 5 lists of the top *k* predicted jobs, for *k* = {1, 2, 5, 10, 15} for each user. They then measured the precision and recall of each list. The display time prediction experiments proceeded in a similar way to those for the revisit data.

The finding with which almost all the studies agree is that the display time is positively correlated with the user interest. Regarding other implicit indicators, there are discordant results. Morita and Shinoda [35] found that the length and the readability of the article and the size of the user's news queue do not influence display time. Their analysis suggested that display time was correlated with user interest; saving, follow-up and copying of an article were not found to be related to interests. Furthermore, they examined several display time thresholds for identifying interesting documents and found that the most effective threshold was 20 seconds, resulting in 30% of interesting articles being identified at 70% precision.

Budzik and Hammond [9] showed that the recommendation of URLs to the users based on what they was typing yielded encouraging results, with at least eight out of the ten users indicating that at least one of the retrieved results would have been useful.

Claypool et al. [11] show that

- the total time spent on a Web page is a good indicator of interest;

- there is a positive relationship between the time spent moving the mouse and the explicit indicator, but mouse movements alone appear only useful for determining which pages receive have the least amount of interest but are not accurate for distinguishing amongst higher levels of interest;
- the number of mouse clicks is not a good indicator of interest;
- the total time spent scrolling by the mouse and the keyboard is a good indicator of interest.

Another finding of the study is that time and scrolling give an accuracy of 70% against the 80% of accuracy provided by the explicit indicators. The subjects involved in the experiment provided explicit indicators about 80% of the URLs only, the others were "no comment". This confirms the difficult for the users to offer explicit evaluations.

In an analysis on an online recruitment service, Rafter and Smith [35] showed that there is a clear correlation between revisit data, that is, the number of times that a user clicks on a job, and activity data when sending a job description by e-mail or apply for the job directly online. These results demonstrated that there was only a loose relationship between raw display time and activity data. This is because of the large amount of noise that this type of data is subject to, such as a user who logs out or leaves his terminal. However, a significant improvement in the correlation between display time and job application is gained by refining the data into graded display times that eliminate some of the erroneous information. The authors believe the revisit data perform better because they are less subject to noise than the display time data which show little correlation to the activity data in their most raw form.

To conclude this section, it is useful to pick Kelly and Teevan's comments out [24].

They underlined that inferring information from user's behaviours is not easy and that what can be observed does not necessarily reflect the user's underlying intention. For instance, the amount of time that an object is displayed does not necessarily correspond to the amount of time that the object is examined. Further, the amount of time an object is actively examined does not necessarily correspond to the user's interest in that object.

We agree with these objections, namely, a long lasting display time does not necessarily implicate user interest, but the works reported above demonstrate that there is a correlation between the time spent on a page and the user interest in their regards.

The author, in addition, suggested that IRF should be understood within the larger context of the user's goals and the system's functionalities.

We are of the same opinion of the authors, but unfortunately the information about the user's goals are often not available. We believe that search engines should favour the making explicit of users' task, for instance by providing an interface that requires to the user to choose among alternative tasks before to begin the search activity.

The authors then suggested that to allow for the effective use of implicit feedback, more research needs to be conducted on understanding what observable behaviours mean and how they change with respect to contextual factors. They also noticed that not all implicit indicators are equally useful and some may only be useful in combination with others. It is likely, also, that how implicit indicators are collected influence their effectiveness. Finally, the authors encouraged to develop implicit indicators systems; in fact, actually there is a lack of literature on developing test-beds and evaluation metrics for implicit indicators.

1.2.2 Experimental methodologies

The IRF techniques employed in literature are investigated in this section. Generally, the experiments are led by

- recruiting a certain number of people,

- observing their interaction behaviors when performing a query and looking at the retrieved documents,
- asking them to express their relevance judgments about the documents,
- expanding their queries on the basis of the algorithm which is under examination,
- choosing some evaluation measure to assess the performance of the experimented technique.

Different experiments are based on simulations. To help perform their study and to enhance repeatability, Kelly and White [25] developed an evaluation framework which can help us to understand how the studies are generally carried out; the evaluation framework's structure is divided into six key components:

1. the *interaction model* is a characterization of what is important about the user interaction data that serves as feedback to the IRF algorithms. This is often composed of logs gathered during a naturalistic study in which a certain number of subjects are recruited and their interaction behaviors are observed over a determined time period.
2. *IRF algorithms* take user interaction as input, and use the content of the documents conforming to the relevance criteria to generate a set of candidate query expansions terms to add to the initial query.
3. *Ground truth* information contains relevance assessments or to be precise the judgments on the usefulness of documents viewed during a search, generally in relation to a pre-determined search topic.
4. The *document collection* contains all the documents for which any interaction logs was logged.
5. The *Information Retrieval system* retrieves set of documents from the collection in response to the expanded queries generated by the algorithms.
6. *Evaluation measures* compute a score for each algorithm.

Let us now present two papers chosen because they were representative of the literature about IRF techniques. Although they are based on the same dataset, they showed two completely different approaches which may suggest a methodology to apply to the TEL case.

Kelly and White [25] explored how individual and task differences impact the effectiveness of the IRF algorithms. Their study used a 2 × 2 factorial design where the independent variables were task information and user information. Each factor had two levels, that is, present or absent. The dependent variable is precision, measured as the proportion of relevant documents in the top ten retrieved, and across

all the document retrieved. The authors developed one algorithm for all combinations

of the two factors, thus resulting in four algorithms in total. The study aimed to determine whether:

- IRF algorithms personalized to users can outperform IRF algorithms that ignore personalization,
- IRF algorithms developed using task information can outperform algorithms that ignore such information,
- IRF algorithms developed using a combination of personalization and task information can outperform algorithms using either source.

All algorithms were compared against a baseline algorithm with a single display time threshold across all subjects and all tasks. The study was based on the dataset created by Kelly in her PhD thesis [22]. In the following, its characteristics are described.

- Seven Rutgers University graduate students were recruited to participate in the study. They were told that the study was a longitudinal, naturalistic observation of their online information-seeking behaviors and that it would last for a university semester.
- This study lasted for fourteen weeks. The study started during the week of 27th January 2003 and ended during the week of 12th May 2003.

- As participants in the study, each subject received a new laptop computer and printer. Upon completion of the study, subjects were allowed to retain the laptop and printer as compensation for their participation. Subjects who were unable to complete the study were required to return the laptop and printer and issued $20.00 for each completed week of the study. All subjects completed the study.
- The laptops were equipped with the WinWhatWhere Investigator client-side logging software. Subjects' online activities were also directed through a proxy logger.

  WinWhatWhere Investigator is a commercially available software for monitoring users' behaviour. It was launched automatically each time the subject's laptop was started and executed in stealth mode while the laptop was in operation. The software did not interfere with any of the subject's natural behaviors; instead, the software unobtrusively monitored and recorded subjects' interactions with all applications including the operating system, web browsers and word processors.
- The Entry Questionnaire gathered background and demographic information from subjects and questioned subjects about their previous computer and searching experiences. The information obtained from the Entry Questionnaire was used to characterize the subjects, but not in subsequent data analysis.
- The Task and Topic Questionnaires elicited the tasks and topics that were of current interest, or were expected to be of interest, to the subject during the study. Subjects were asked to think of their online activities in terms of tasks and topics. Example tasks and topics were provided to subjects. In the following, what the subject had to indicate in this questionnaire is reported.
  - Task endurance was the length of time the subject expected to be working on the task, and it was measured on an eight-point scale, whose eight points demarcated specific lengths of time.

  - Frequency was how often the subject expected to conduct online information-seeking activities related to a task. As with task endurance, frequency was measured on a eight point scale, whose eight points demarcated specific amounts of time.
  - Stage was subjects' assessment of their progress in completing the task. It was measured on a seven-point scale.
  - Persistence was the length of time the subject expected to be interested in information about a topic. It was measured on an eight-point scale.
  - Familiarity was the subject's current state of knowledge about a topic. It was measured on a seven-point scale.
- At weekly intervals subjects were asked to update the lists by eliminating tasks with which they were no longer working and topics in which they were no longer interested, and re-characterizing all other tasks and topics according to the attributes.
- For each week of the study, subjects were presented with a selection of the documents that they had requested during the previous week and were asked to:
  1. classify each document according to their tasks and topics;
  2. indicate the usefulness of the document as it related to that task and topic;
  3. indicate their confidence in the usefulness indicator that they assigned to the document. If subjects could not remember a document, they were instructed not to evaluate the document.

Usefulness was measured on a seven-point scale where the scale anchors were "not useful" and "useful". Confidence was the extent to which the subject believed that the usefulness rating that they assigned to a document reflected their opinion of the document's usefulness. Confidence was measured on a seven-point scale, where the scale anchors were "low" and "high."

This action log files were used to create a document collection, the contents of which served as stimuli for the IRF algorithms studied. The authors used interaction with Web documents only, thus, the collection obtained is relatively homogeneous and would not be biased by different interaction behaviours for different document types. The document collection contains 2,741 web documents; 15% of them was used to derive the thresholds display times for the four algorithms. The remaining 2,329 documents were used to test the algorithms performance.

The experiment described in [25] was concentrated on one single indicator, that is, document display time; this choice aimed at reducing the noise caused by interaction between indicators. For each of the seven subjects, and for each of the nine task groups, an initial "title" query was created from the top three most frequent terms in the union of the non-stopword terms in the task labels generated by that subject.

Kelly and White conducted an analysis of the level of kurtosis based on those usefulness scores to determine how best to collapse the usefulness data from a seven-point scale to a scale of less granularity, and hence more consistency between subjects. The result was a three binary divisions (Table 1.3). Although the division does not result in an even distribution of relevant and non-relevant judgments for all subjects, this was the most consistent distribution that was obtainable from the data.

Subjects	User group	Rating	
		Non-relevant	Relevant
1, 3, 5, 7	1	1, 2, 3, 4, 5	6, 7
2, 4	2	1, 2, 3, 4, 5, 6	7
6	3	1, 2, 3, 4	5, 6, 7

Table 1. 3

Relevance was determined based on whether viewing time equalled or exceeded a temporal threshold. There are four thresholds:

1. TaskAndUser: separate threshold document display times for each subject-task pair.
2. TaskOnly: separate threshold document display times for each task, across all subjects.
3. UserOnly: separate threshold document display times for each subject, across all tasks.
4. All: a single threshold document display time across all subjects and all tasks. This refer to the baseline algorithm.

The researchers led some analysis that indicated that the median was the most consistent indicator of relevance, thus, the median document display time was used in all algorithms as a relevance threshold value; documents viewed for that time or above were assumed to be relevant.

The IRF algorithms selected query expansion terms from documents assumed relevant. All the algorithms used the "wpq" method to rank terms for query expansion—this method is based on the probabilistic distribution of terms in relevant and non-relevant documents. It was used to select the six expansion terms to be added to the original query.

The evaluation measures adopted in this study are mean average precision (MAP) and precision at the top-10 documents retrieved (P10). The MAP and P10 values for the four algorithms were computed across a series of feedback iterations, where an iteration was defined as a document that met the relevance criteria. The following methodology was applied during the study:

1. Create initial set of queries from task labels.
2. For each algorithm, loop through the document set for each task and subject:
    a. If document display time equals or exceeds the pre-determined threshold for that algorithm, for the current task and subject:

   i. Pass the document to the algorithm and use it, and any previous seen relevant documents, to expand initial query.
   ii. Use expanded query to retrieve new set of documents using a best match tf.idf weighting scheme.
   iii. Use ground truth information to evaluate the documents retrieved, and score the current IRF algorithm.
3. IRF algorithms are ranked based on MAP and P10 averaged across all search tasks, users, and tasks to determine algorithm performance.

In [33], Melucci and White present a geometric framework that utilizes multiple sources of interaction between users and search system to develop enhanced IRF models personalized for each user and tailored for each search task. The authors thought that a way to gather feedback from users at minimal cost to them in terms of time or cognitive resources is to use the contextual information generated during the interaction between the user and the system as implicit relevance feedback. Contextual features such as document display time, document retention, and document interaction can be mined and used as the basis for relevance criteria in IRF algorithms.

Usually, IRF algorithms use just one implicit feature as relevance criteria. As mentioned above, the two most common features used are document display time or document visitation; however, it was already pointed out that there is mixed opinion about whether these elements are accurate and imply relevance [1, 11, 35, 47]. In addition, it has been shown through user experimentation that a single feature can vary greatly between users and search tasks [25]. This means that implicit evidence can be unreliable as there are usually only a small number of relevant documents available for each user, each task, and each user/task pair.

The authors suggest to capitalize on multiple aspects of user interaction, so that substantially more evidence about preferences becomes obtainable, and more robust IRF algorithms could be created.

The methodology illustrated below can be attractive because every user is influenced by the context in his seeking activity and the more indicators are considered, the better this context can be represented. The authors essentially stated that information seeking activities are affected by the context which can be described by the features characterizing users, time, places, or anything emerging from user-system interaction.

Melucci and White exploited the properties of the theory of the Vector Spaces for modelling this context in a way that can be leveraged by information retrieval

systems. They provided some definitions for understanding the mathematical framework:

- Variable: refers to either an entity of the context, for example, user, task, topic, or document, or a relationship between entities, for example, relevance or aboutness.
- Dimension: refers to a property of an entity, for example, user behaviour, task difficulty, topic clarity, document genre, or relevance.
- Factor: refers to a value of a property, for example, browsing, complex search task, difficult topic, relevant, non-relevant, or mathematical document.

When some evidence is gathered from context, IRF can be performed for expanding queries, reordering retrieval results, or re-searching. Once some variables and dimensions of context are selected from the domain for which a context-aware information retrieval tool is designed, the methodology presented in this paper can be summarized as follows:

1. for each dimension of context a set of orthogonal vectors is defined—each orthogonal vector of such a set models one factor of the dimension of context;
2. a basis is built for representing a context by selecting one or more factors from each dimension—one factor refers to one dimension;

3. an informative object is matched against a context by computing a function of the distance between the vector and the subspace spanned by the basis—the closer the vector to the subspace, the more the object is "in the context".

In general, factors of distinct dimensions are mutually linearly independent; the vectors corresponding to a given dimension of context are mutually orthogonal for signifying that the values taken by the dimension are mutually exclusive. Many distinct dimensions can co-exist in the same space. These dimensions model a document or a query from different point of view and each perspective corresponds to a dimension of context. As there is a one-to-one correspondence between a subspace spanned by a set of vectors and its projector, a projector can be taken as the algebraic operator for a contextual factor and a linear combination of projectors is a mathematical operator which refers to a mixture of contextual factors.

Mathematically, the most natural combination which can represent a context is the linear combination. Thus, the operator adopted in this paper is a linear function of projectors by using a predefined set of coefficients which measure the weight of each dimension of context.

Let $L(\{b_i\})$ be the subspace of the vectors which are obtained by multiplying b_i by a scalar. Therefore, the operator is

$$C_B = w_1 B_1 + \ldots + w_k B_k$$

where the w_i's are non-negative coefficients such that $w_1 + \ldots + w_k = 1$ and the B_i's are the projectors onto the subspaces $L(\{b_i\})$'s. C_B is called context matrix or context operator.

In their experiments, the researchers have computed the vectors which represent the contextual factors by Singular Value Decomposition (SVD) of the correlation matrix between the features observed from a set of documents seen by the user during the course of his search. The values of the eigenvector are scalars between −1 and +1; the further a value is from 0 the more the feature to which the value corresponds is a significant descriptor of the contextual factor represented by the eigenvector. The sign can express the contrast between features and then the presence of subgroups of features in the same contextual factor.

Melucci and White used an interaction logs of real subjects to simulate a user who accesses a series of documents and performs some actions. In particular, the real subject were the seven people recruited in Kelly's PhD thesis [22]. The document features of the dataset used in the study were:

- the unique identifier of the subject who performed the access;
- the unique identifier of the attempted task, as identified by the subject;
- the display time;
- a binary variable indicating whether the subject has added a bookmark for the webpage;
- a binary variable indicating whether the subject has saved a local, complete copy of the webpage on disk;
- the frequency of access, namely, the number of times a subject expected to conduct on-line information-seeking activities related to the task;
- the number of keystrokes for scrolling a webpage;
- the depth of the webpage, that is, the number of slashes in the URL.
- In addition to these features, the authors make use of the relevance assessments assigned by participants in the study.

The IRF algorithms under investigation was assumed to be part of a system that monitors subject behaviour and uses these interaction data as a source of IRF to retrieve and order the unseen documents. When the task or the subject are known, the

system records the data by subject / task and then retrieves and ranks the unseen documents for the given subject / task. The details of the simulation are as follows:

1. The features of all the documents seen by the user when performing a task and searching for information relevant to a topic are observed. *n*

documents from these are used for computing a representation of context.

2. The observed features of the n documents are used for computing the contextual factors as follows:
   a. the feature correlation matrix is computed;
   b. the eigenvectors are extracted from the correlation matrix.
3. The whole document collection is ranked by the ranking function. In the experiments reported in this paper no mixture has been investigated and therefore $C_B = w_i B_i$ where w_i = 1 and w_j = 0 for any $i \neq j$. Then, for each projector:
   a. The ten most frequent keywords of the n top-ranked documents are used for expanding the textual description of the topic, which is then considered as a new, expanded query.
   b. The expanded query retrieves a list of documents.
   c. The usefulness scores assigned to the documents are used as ground truth information for evaluating this query expansion-based retrieval.

To evaluate retrieval effectiveness it has been used Normalized Discounted Cumulative Gain (NDCG).

The projector-based method (PRJ) is compared with two algorithms: the QRJ and CTR.

In QRY, the topic description was expanded using the query expansion capabilities of MySQL and in CTR the computation of the projectors is replaced with the computation of the unique centroid vector of the cluster of n vectors of the documents seen by the subject when performing an information-seeking activity. That centroid vector has then been used for selecting the feedback documents – the inner product between the centroid vector and the unseen document vectors is then computed for ranking the unseen documents.

QRY was chosen as the baseline since it is one of the most successful RF techniques, and IRF is a viable substitute for PRF in operational environments.

CTR allow to determine the value of utilizing multiple factors.

1.2.3 Experimental findings

Kelly and White presented in [25] the findings of their study for the 2,329 documents of the test collection, over 1, 2, 5, 10, 15, and 20 iterations. Indeed, as mentioned in section 1.2.2, the values of the evaluation measures (MAP and P10) for the four algorithms were computed across a series of feedback iterations, where an iteration was defined as a document that met the relevance criteria.

All the new queries generated were expansions of the original set of queries. Parametric statistical testing is used at a 0.05 level of significance where appropriate.

From the evaluations' measures obtained during the analysis it appears that UserOnly performs worse than any of the other algorithms, including the baseline algorithm, where task and user information are ignored. In contrast, using information about the

search task (in TaskOnly) appears to enhance retrieval performance, especially in later iterations. Further analysis suggested that there was a large variability in algorithm effectiveness between users and perhaps a fewer degree of variation between tasks groupings.

The more surprising finding of the study was that tailoring display time thresholds to the individual user appeared to worsen retrieval performance. White and Kelly hypothesized that this should have been due to the indicator selected, or the way in which they derived threshold display. Besides, there was a lot of variability between subjects, and consequently a lack of consistency in document display times between users; this may have been related to the small number of users involved in this study. Furthermore, there were large variations in how much evidence is available to tailor algorithms to individuals: some subjects viewed many pages, while others viewed only a few.

Additional data can be obtained employing a larger subject pool or using data gathered from other information sources as interaction logs. The first alternative is difficult and costly, whereas interaction logs are limited in that they provide only partial access to information about the task the users are attempting or about the users themselves.

An important finding of this work was that multiple subjects interacted more consistently within a single task group than one subject did within multiple tasks. This suggest that grouping users and developing algorithms based on groups rather than individual users may be one way to improve the consistency of IRF algorithm performance. That can be done given adequate interface support; the challenge is how to offer this support in a lightweight way that will be easy to use by the broader user population, which has become accustomed to minimal interaction with search systems. However, further research is needed for to automatically identifying tasks. A useful result of this study was the framework developed to automate aspects of the experimental process; it allows one to vary the IRF measures used, to determine the most effective measure for a given algorithm.

A limitation of this study was that only one IRF indicator was used to address the research questions.

Melucci and White compared the NDCG value of the projector-based method with the NDCG of the two others algorithms, calculated across all subjects and all tasks for variations in the number of visited documents and for variations in the number of ranked documents used for computing NDCG [33]. They noted that PRJ and CTR are on average comparable with each other and QRY is much less effective than PRJ and

CTR. However, CTR was based on the centroid of vectors which is actually an average vector and does not distinguish among the diverse factors by which context may impact on interaction and then on retrieval effectiveness. In order to establish the role played by the projectors, an analysis was conducted to compare the effectiveness of CTR with the effectiveness of PRJ by varying the projector. That is, one projector was fixed at a time and the documents were ranked using this projector.

The results show that each projector produces a different NDCG. The projector which achieved the highest average NDCG of PRJ was selected over all the projectors. Further analysis suggested that for each subject or task a projector which is more effective than the centroid exists thus indicating that PRJ has the potential of being more effective than a cluster-based method. When PRJ performed better than CTR, the highest weights of the best performing projector (BPP) correspond to the features of the documents which provided the most effective query expansion terms. This result suggests that a relationship between the BPP and query expansion terms exists.

However, the relationship between BPP and these terms requires further investigation because as it may by symptomatic of a complex interaction between PRJ and pseudo-relevance feedback. The BPP is no consistent across tasks.

The results obtained suggest that tailoring projectors to users and tasks leads to improved performance over algorithms that do not use such information. An important finding of this paper is the existence of an algebraic operator for each subject-task pair which can be used for tailoring document rankings to the user attempting the task.

1.3 Implicit Relevance Feedback and TEL

Let us make some considerations about the adaptability of the algorithms reported in [25, 33] and described above to The European Library portal. These two papers are considered because they illustrate IRF techniques differently from the works presented in the previous section.

In particular, we list the differences between the dataset used in both the papers and the action log file that have been analyzed in this thesis.

- The papers above refer to a naturalistic study in which seven subjects have been observed in their information-seeking activities.

  On the other hand, the action log file at our disposal contains 498,292 rows regarding 63,528 different sessions, thus, it regards a completely different size of data. Furthermore, in our action log file anomalous values occur quite frequently; this fact is certainly more frequent in dataset of big size, than in dataset obtained from studies as that of Kelly [22]. Moreover, the very large size of the action log file prevented us for correcting these anomalous values by hand, thus requiring semi-automatic yet prone-to-error procedures.
- The seven subjects of the studies reported in [25, 33] were Rutgers University graduate students; this means that they had a similar cultural background, they probably were all of the same age, it is likely that they were well-selected, in other words they constituted an homogeneous and well controlled group.

  On the contrary, TEL portal's users are certainly members of an heterogeneous group: the portal is potentially reachable any user world-wide, of any age, hardly controllable and with different cultural background. In that sense, TEL portal user population is much more similar to search engine user populations than the subjects recruited by Kelly. This suggests the idea that the results gained from the investigations conducted using the TEL action log file can be generalized to search engines, and vice-versa, the research on search engine user interactions can provide useful findings for improving the TEL portal.
- The seven subjects were each given a laptop computer equipped with a particular software that unobtrusively monitored and recorded users' interactions with all applications.

  In our work, the only information's source is done by the action log file; we have no information about the interaction of the users with other applications, nor could any "gadget" be given to the users.

- Upon completion of the study, the subjects of Kelly's study were allowed to retain the laptop and printer as compensation for their participation. Subjects who were unable to complete the study were required to return the laptop and printer and issued a money retribution for each completed week of the study.

  TEL portal's users do not receive any compensation for connecting to the portal and consequently to offer us their interaction behaviors.
- The subjects were asked to think about their online information-seeking activities in terms of tasks and topics, to create labels for each task and topic, and to classify the pages that they viewed according to these tasks and topics. Furthermore, for each task, subjects were asked to indicate the task endurance, frequency and stage; for each topic, subjects were asked to indicate the topic persistence and their familiarity with the topic.

  In the dataset at our disposal, no one of these information were recorded; when a user accedes to the portal he doesn't find no request of expressing his task. A different interface should allow to the subjects to choose between a set of tasks without interfering on their search activity. This additional information might be very useful to improve users' searches.
- The seven subjects were also asked to express their preferences for each document by how useful they believed the document to be in helping them to complete and/or understand the particular task and topic in which they classified the document. In addition to usefulness, subjects were asked to indicate the navigational usefulness of the documents that they viewed. Subjects' were finally asked to indicate their confidence with respect to the usefulness indicators that they assigned to the documents.

  In our action log file relevance assessments are not present. This is quite a limitation since the explicit relevance assessments are a precious source of evidence when IRF are to be designed. However, the absence of explicit interest indicators, as that occurred in the TEL action log file, is by far more realistic than the log files built during naturalistic, controlled user studies.

- Essentially, the dataset created in Kelly's PhD comes from a controlled experiment, with recruited people eager to participate to the study, also if it requires them to employ time and pledge.

  As a consequence, the experiments conducted with that dataset are not repeatable, because, in order to replicate them, other subjects with the same characteristics of the seven Rutgers University graduate students should be found and subjected to the same proceeding we have described above.

  The research works like that in this thesis instead is repeatable due to the availability of the TEL action log file. In order to replicate it, the recruitment of any subject with particular characteristic is unnecessary.

Notwithstanding the differences between the experimental settings, the findings of the two papers summarized can be useful to suggest an effective methodology to apply at TEL project.

In particular, Kelly and White [25] provided an evaluation framework which can partially help us to investigate a methodology to apply at TEL case. Furthermore, they suggested that grouping users and developing algorithms based on groups rather

than individual users may be one way to improve the consistency of IRF algorithm performance. Considering the TEL case, the users can be grouped by some criteria, for example a criterion can be the choice of a determined collection of documents, another principle can be the performing of a certain type of search.

The results reported in the related literature about the difference of the display times threshold to discriminate between relevant and non relevant documents in basis on the user and the task suggest that we cannot rely only in this indicator. Kelly and White further admitted that a limitation of this study was only one IRF indicator was used to address the research questions. As it has been stated by Melucci and White [33], utilizing a single indicator can be unreliable, indeed, there is mixed opinion about whether the display time is accurate and imply relevance.

Melucci and White started from this consideration and presented a methodology that take into consideration more than one implicit indicator. The results obtained in their experiment suggest that tailoring projectors to users and tasks leads to improved performance over algorithms that do not use such information. In addition, they have demonstrated the existence of an algebraic operator for each subject-task pair which can be used for tailoring document rankings to the user attempting the task.

Starting from TEL action log file we can extract various indicators that give us information about the context in which the users are, thus we think that the methodology described in this paper can be fitted to our case.

The absence of explicit relevance indicators in our dataset can be viewed as an handicap, but we have to consider that it is definitely more difficult to achieve data including explicit relevance indicators because people are reluctant to offer their assessments. Furthermore, in our action log file we can find more than one indicator which we can view as proxy of the relevance[5].

[5] The concept of ”proxy” is better explained in section 2.8

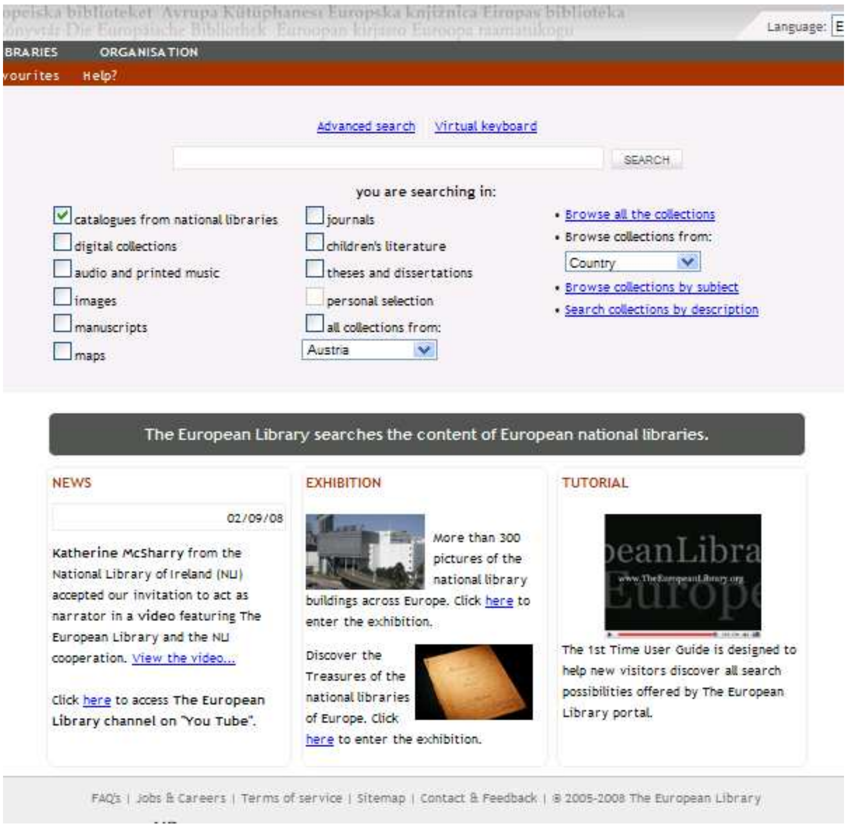
CHAPTER 2

A DESCRIPTIVE ANALYSIS AND METHODOLOGICAL DATA PREPARATION OF THE TEL PORTAL ACTION LOG FILE

2.1 Introduction

In this chapter, we provide an explorative analysis of the dataset at our disposal, that is, the TEL portal action file. Agosti provides the relevant characteristics of library automation systems and Digital Library (DL) systems that need to be taken into account in addressing the study of log data in digital libraries [3]. The author underlined that log data constitute a relevant aspect in the evaluation process of the quality of a DL system and of the quality of interoperability of DL services. Finally, she introduced a general approach for the analysis of log data generated in the use of services of DL systems.

The European Library portal (Figure 2.1) action log contains 498,292 rows (records) and 13 columns (attributes): there is a row in the log file for each user action and each column represents an information about the user or about his seeking. In the following, a detailed description for each of the variables is provided.


opeiska biblioteket Avrupa Kütüphanesi Europska knjižnica Eiropas bibliotēka
önyvtár Die Europäische Bibliothek Euroopan kirjasto Euroopa raamatukogu
Language: E
BRARIES
ORGANISATION
vourites
Help?
Advanced search
Virtual keyboard
SEARCH
Country
Austria

you are searching in:

☑ catalogues from national libraries
☐ digital collections
☐ audio and printed music
☐ images
☐ manuscripts
☐ maps

☐ journals
☐ children's literature
☐ theses and dissertations
☐ personal selection
☐ all collections from: Austria

- Browse all the collections
- Browse collections from: Country
- Browse collections by subject
- Search collections by description

The European Library searches the content of European national libraries.

NEWS

02/09/08

Katherine McSharry from the National Library of Ireland (NLI) accepted our invitation to act as narrator in a video featuring The European Library and the NLI cooperation. View the video...

Click here to access The European Library channel on "You Tube".

EXHIBITION

More than 300 pictures of the national library buildings across Europe. Click here to enter the exhibition.

Discover the Treasures of the national libraries of Europe. Click here to enter the exhibition.

TUTORIAL

The 1st Time User Guide is designed to help new visitors discover all search possibilities offered by The European Library portal.

FAQ's | Jobs & Careers | Terms of service | Sitemap | Contact & Feedback | © 2005-2008 The European Library

Figure 2. 1

2.2 The Log File

Log data are collected by a computer system to make a permanent record of events during the usage of the system. This is done to better support its operations, and in the case of operating systems or database management systems, its recovery procedures. Initially, these data were mainly used to manage recovery procedures of the system, but over time it became apparent that they could also be used to study the usage of the application by its users, and to better adapt the system its objectives. In the 1970s, the library automation systems were among the first system able to manage the permanent data of interest to libraries. These systems were only able to manage the catalogue data representing physical library objects that were held in a real and physical library. Thus, objects held in archives and museums were not represented at that time in those application systems.

In the 1980s, the first log data appeared; they were collected to manage the system itself, and especially to monitor the usage of system search facilities by users. The mentioned "search facilities" was designed for user search, and the access to catalogue data was called Online Public Access Catalogue (OPAC). An OPAC is a software application designed to allow final users to directly access to the catalogue, without the intervention of a professional subject, and to make available to all of them the data stored in the catalogue database managed by the software system. The catalogue database is constructed by professional librarians who use authority control rules in describing author, place names and other relevant catalogue data [19]. Over time, the librarians usually construct many authority files where the software application stores all lists of preferred or accepted forms of names and other relevant headings [7]. The log file of an OPAC system stores information on the specific queries which have been made by final users referring to the specific authority files from which the data were extracted. Therefore the analysis of the OPAC queries can be used to better understand the effective use the final user makes of the data stored by the library automation system. In traditional OPAC systems it was possible to trace each user-system interaction and each user session was identifiable. In [34], Mitev et al. provide a detailed record of the features which can be evaluated in an OPAC directly accessible by its final user. They are: technical performance, information retrieval performance, and user behavior, which includes studies of users and of use, user profiles, user search patterns, and user interaction success.

In late 1980s, it became evident that a library automation system could not only manage catalogue data or metadata describing physical objects, but also digital representations of some types of physical objects. Furthermore, some objects started to appear in a digital form, thus the collections were becoming increasingly diversified and complex. Previous library automation systems appeared to be limited in managing data related to such a diversified situation, therefore, a new type of systems was designed and named "Digital Library (DL) systems".

One significant aspect that still relates DL systems to OPACs is that the representation of the content of the digital objects that constitute the collection of interest is still done by professionals. The management of metadata can still be based on the use of authority control rules in describing author, place names and other relevant catalogue data. A DL system can exploit authority data that keep lists of preferred or accepted forms of names and all other relevant headings. This is a significant difference between DL systems and search engines, and it is usually overcome with the analysis of log data. Indeed, a search engine often becomes a specific component of a DL system, when the DL system faces the management and search of digital objects by content in the same manner as information retrieval systems and search engines [2]. In all other types of searches, the DL system makes use of authority data to respond to final users in a more consistent and coherent way through a search system that is a sort of a new generation OPAC system, or the system supports the full content search with a service that gives the final users the facilities of a search engine. Finally, it is worth pointing out that the access to each service a DL system provides is usually supplied through a Web browser, and not through a specifically designed interface. This means that the analysis of user interaction with systems that have a Web-based interface requires ways that support the reconstruction of sessions in a setting, like the Web, where sessions are not naturally identified.

2.3 The European Library Portal

The European Library (TEL) is a non-profit organization which provides the services of a physical library and offers search facilities for the digital or bibliographical resources of many of the European national libraries. TEL initiative aims at providing a "low barrier of entry" for the national libraries that should then be able to join the federation with only minimal changes to their systems [46]. This means that TEL exists to open up the universe of knowledge, information and culture of all European national libraries, where a national library is the library specifically established by a country to store its information database. Currently TEL gives access to 150 million entries across Europe, but the amount of referenced digital collections is constantly increasing. TEL Portal is constituted by the three components:

1. a Web server which provides access to the services to the users;
2. a central index which harvests catalogue records from national libraries, supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[6], and provides integrated access to them via Search/Retrieve via URL (SRU);
3. a gateway between SRU and Z39.50[7] which makes accessible through SRU also national libraries which would otherwise be accessible only through Z39.50[8].

In addition, the interaction between the portal, the federated libraries, and the user mainly happens on the client side by means of an extensive use of Javascript and Asynchronous JavaScript Technology and XML (AJAX)[9]. Once the client, which is a standard Web browser, accesses the service and downloads all the necessary information from the Web server, all the subsequent requests are managed locally by the client. The client interacts directly with each federated library and the central index, according to the SRU protocol, makes separate AJAX calls towards each federated library or the central index, and manages the responses to such calls in order to present the results to the user and to organize user interaction.

[6] http://www.openarchives.org/OAI/openarchivesprotocol.html

[7] http://www.loc.gov/z3950/agency

[8] "Z39.50" refers to the International Standard, ISO 23950: "Information Retrieval (Z39.50): Application Service Definition and Protocol Specification", and to ANSI/NISO Z39.50

[9] http://www.w3.org/TR/XMLHttpRequest

For all the different categories of users of a DL system, the quality of the services and documents the DL supplies are very important [5]. Log data constitute a relevant aspect in the evaluation process of the quality of a DL system and of the quality of interoperability of DL services [18]. In these evaluation processes the final user needs to be considered the guide of the system designers, prompting them to conceive and invent solutions of real use for the user himself.

TEL is one of the most relevant effective DL initiatives that can be studied and that constitutes a significant building block towards the common European Digital Library that the European Commission is promoting. In particular, the Europeana thematic network[10] is a project launched in July 2007 with the aim of addressing the interoperability issues among European museums, archives, audio-visual archives and libraries towards the creation of the "European Digital Library".

The framework that has to be designed and put in place is going to be a coherent infrastructure for the collection, storage, curation and management of relevant data which are derived from sources of different nature; among those sources two are most relevant:

1. the data collected through log systems, and
2. the data which are generated and collected through user studies.

The logging requirements of a relevant DL initiative suggest logging data throughout the whole portal, which means collecting data for the user navigation on both static and dynamic Web pages. Among those log data there are HyperText Transfer Protocol (HTTP) or Web logs, action logs, and static content logs. In particular, the structure of HTTP logs often conforms to the W3C Extended Log File Format [20]. This kind of log contains, among other things, the following useful information:

- the Internet Protocol (IP) address and the user-agent which allow the identification of single users [4]; and

[10] http://www.europeana.eu

- the referrer field, a Uniform Resource Locator (URL) address which communicates the last page viewed by the user, and this can be used to know how visitors get to TEL service.

Together with log data analysis, it is envisaged the necessity of collecting data generated by controlled studies which have to be performed on groups of users that freely crawl and navigate TEL portal and then fill in specifically designed questionnaires to report and describe their impressions. The goal of the controlled studies is to combine the data of the sessions of the people who have compiled the questionnaires, data which are present in the log data, with those that have been reported in the questionnaires. The final aim is to gain insights from data on user sessions and judgments in the questionnaires to generalize the results obtained. The insights gained by analyzing log data together with data from controlled studies are more informative than the results that can be derived by separately analyzing the groups of data. In this thesis, the TEL portal action log file will be investigated.

2.4 The TEL Portal Action Log File

Before analyzing each single column of the dataset in detail, the TEL action log file is described in the following. Tables 2.1 and 2.2 show an excerpt of the log file. Let us look at the first row of each table and show the information deducible from the dataset:

- the "id" is a progressive number useful to identify each row,
- the user whose the row refers to is not registered to the portal,
- its IP address is 147.91.249.1,
- the code associated to his session is 3n09267661nl5f26mekqaaq3f0,
- the user maintains the default language for the portal interface,
- he performs an action of displaying of a result record list referred to the query "platonov",

- the column “colid” indicates that the collection upon which the action is performed is identified by the code “a0001”,
- the result list contains 60 records,
- the column “recordPosition” indicates that the user is looking at the records from the 21th to the 40th,
- the variable “sboxid” has only missing values[11],
- the column “objurl” does not include any string because it regards only the objects reached through the actions “available_at” or “see_online”[12],
- finally, the user performs this action the 1st of November 2007, at 10:01:44.

The following row regards the same user, and it tells us that he looks at a single result record. The third row is referred to another user—both the session code and the IP address are changed. This user performs a simple search two seconds after that the first user has viewed the single record.

[11] See Section 2.5
[12] See Section 2.5

id	userid	userip	sesid	lang	query	action
1008583	guest	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief
1008584	guest	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_full
1008585	guest	81.159.36.9	o4526tgmj321rr2bptoj9lp820	en	("woodhead")	search_sim
1008586	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("pubilc law")	view_full
1008587	guest	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief
1008588	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	search_res
1008589	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_full
1008590	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_full
1008591	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	search_res
1008592	guest	81.159.36.9	o4526tgmj321rr2bptoj9lp820	en	("woodhead")	view_full
1008593	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_full
1008594	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_brief
1008595	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_full
1008596	guest	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief
1008597	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_brief
1008598	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_brief
1008599	guest	193.166.120.39	19shslpbm60vncnfhnou7383e6	fi	("legal aid")	view_full

Table 2. 1

colid	nrRecords	recordPosition	sboxid	objurl	date
a0001	60	21-40			2007-11-01 10:01:44
a0001	60			http://www.theeurop...	2007-11-01 10:01:51
	0	-			2007-11-01 10:01:53
	0				2007-11-01 10:01:56
a0001	60	41-60			2007-11-01 10:02:11
	0	-			2007-11-01 10:02:17
	107				2007-11-01 10:02:34
a0038	107	18		http://www.theeurop…	2007-11-01 10:02:36
	0	-			2007-11-01 10:02:44
	1000				2007-11-01 10:02:46
	107				2007-11-01 10:03:00
a0038	107	21-40			2007-11-01 10:03:02
a0038	107			http://www.theeurop…	2007-11-01 10:03:05
a0037	342	21-40			2007-11-01 10:03:08
a0038	107	41-60			2007-11-01 10:03:09
a0038	107	61-80			2007-11-01 10:03:20
a0038	107	65		http://www.theeurop…	2007-11-01 10:03:41

Table 2. 2

2.5 Analysis of the TEL Portal Action Log File

Before proceeding with the analysis, some considerations about the operations performed on the variables present in the original log file are made. This analysis aims to show the informative content of the dataset at our disposal, and also to prepare it so that an Implicit Relevance Feedback (IRF) technique can be designed. In particular, a methodology for which it is necessary that the variables are quantitative continuous or in alternative qualitative dichotomous [15] will be considered. A description of the variables follows.

Identifier ("id") is the identifier of each action, that is a progressive number associated to each row. This is the "key" of the dataset, that is a number that identify univocally each row.

User identifier ("userid") identifies each user: the users registered to the portal are assigned a number, and the not registered users are shown as "guest", which is the default value for this field. If we group the actions by session we can find the proportion of users registered respect to the users guest: there are 230 sessions referring to 158 registered users and 63,298 sessions referring to users guest, thus, 99.64% of the sessions refer to non-registered users (Figure 2.2). It is interesting to note that each of the 158 registered users accesses to TEL portal a few times only:

- 129 of them made use of the portal only once,
- 19 registered users accessed twice,
- six used the portal three times,
- one user accessed four times,
- only one user accessed 10 times,
- one user accessed 12 times and, finally,
- one user accessed 19 times.

This distribution can be interpreted as a signal of the dissatisfaction of these users; we believe that if a subject is were satisfied of a service, he would'll make use of it again. However, this is only an hypothesis; another alternative is that the users have found what they were looking for, so they did not need searching again. Furthermore, may be the registered users may access to the portal as guest, too; perhaps, they make the log-in only when they wanted to take advantage of some particular benefit as, for instance, saving in reference session's favourites.

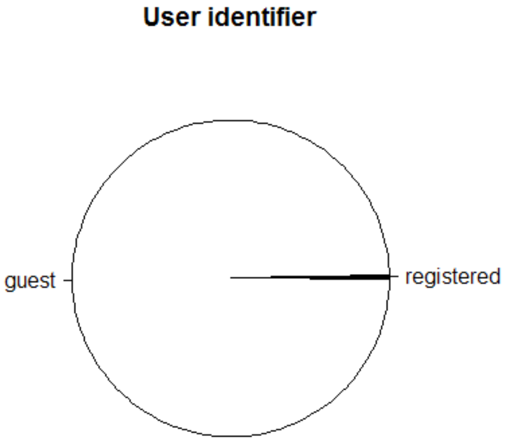


Figure 2. 2

User IP ("userip") is the IP address assigned to the computer or device of the user that performs every action. An Internet Protocol (IP) address is a numerical identifier that is assigned to computers or devices participating in a computer network utilizing the Internet Protocol for communication between its nodes. In the TEL portal action log file, there are 46,203 distinct IP address across 63,528 different sessions: 89.29% of the IP addresses appear a single time, 9.46% of them appear from two to six times, and only 1.25% of the IP addresses appear more than six times.

It is worth noting that these percentages do not necessarily indicate that the 89.29% of the users of TEL portal accessed only a single time in the period from 1st November 2007 until 29th February 2008, because a group of people might share a common IP address, and a subject can access to the Internet through different computers having distinct IP addresses.

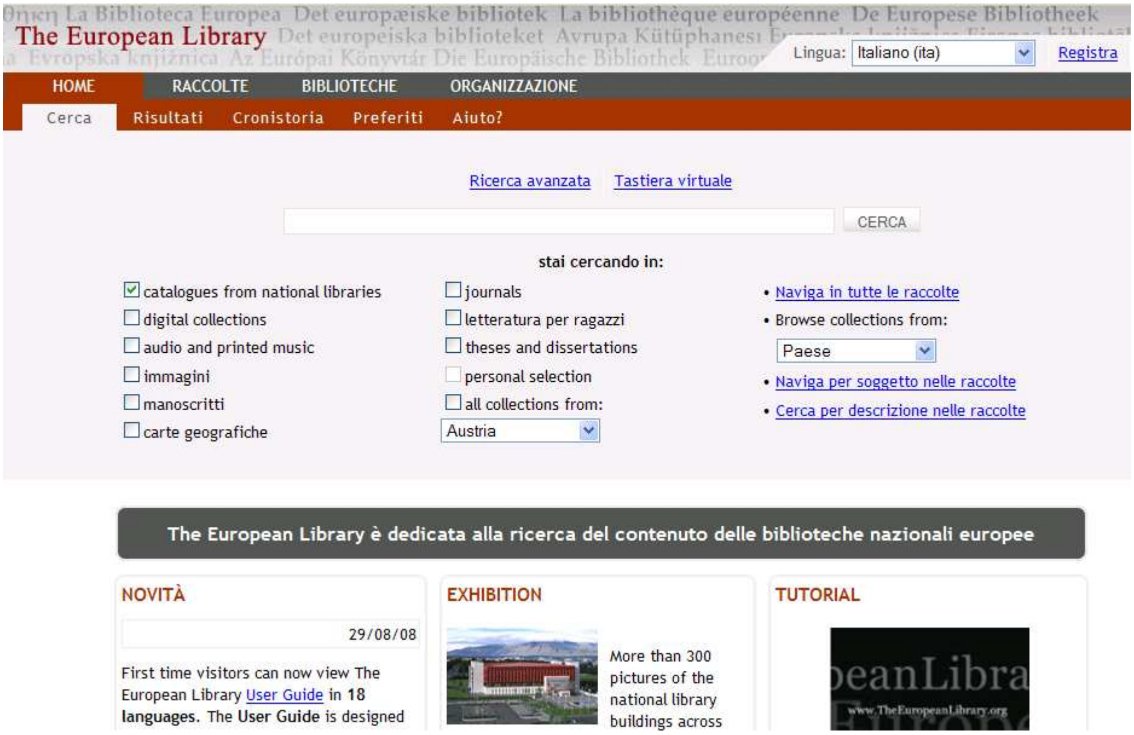
Session identifier ("sesid") is a code that identifies every session. A session is a time-contiguous sequence of actions performed by the same user. In our log file, how explained by Luxenburger et al. in [28], session boundaries have been found by relying on the PHP session ID and the additional requirement of no more than 5 minutes of inactivity between subsequent actions within the same session. In our log file there are 63,528 different sessions. 8,573 rows of the dataset store a string whose value is "null" in correspondence of the column "sesid".

Language ("lang") refers to the language of the portal interface, the default value is "English". As shown in the histogram, most of the users do not change the default interface language. In particular, 425,935 of the actions are performed in the default language, namely the 85.48% of the total actions. Table 2.3 reports the language and the respective number of users that put that language in the portal interface.

Country	Number	%
England	425935	85,48
France	13555	2,72
Italy	8682	1,74
Poland	7811	1,57
Spain	7781	1,56
Germany	6499	1,30
El	6367	1,28
Sierra Leone	4879	0,98
Hungary	3266	0,66
Slovakia	2378	0,48
Portugal	2369	0,48
Croatia	1459	0,29
Netherlands	1202	0,24
Suriname	1195	0,24
Cs	1123	0,23
Russia	738	0,15
Lithuania	475	0,10
Gabon	468	0,09
Latvia	403	0,08
Iceland	356	0,07
Finland	313	0,06
El Salvador	276	0,06
Da	243	0,05
Ethiopia	181	0,04
Norway	153	0,03
---	118	0,02
Malta	32	0,01
Tag	29	0,01
	5	0,001
Und	1	0,0002

Table 2. 3

Perhaps most of the users preserve the default language because, also if they should change it, a large part of the writing in the portal interface remain in English. For example, if "Italiano (ita)" is selected in the language menu, the interface reported in Figure 2.3. is returned to the end user.



The European Library

Lingua: Italiano (ita) Registra

stai cercando in:

- ☑ catalogues from national libraries
- □ digital collections
- □ audio and printed music
- □ immagini
- □ manoscritti
- □ carte geografiche
- □ journals
- □ letteratura per ragazzi
- □ theses and dissertations
- □ personal selection
- □ all collections from: Austria

- Naviga in tutte le raccolte
- Browse collections from: Paese
- Naviga per soggetto nelle raccolte
- Cerca per descrizione nelle raccolte

The European Library è dedicata alla ricerca del contenuto delle biblioteche nazionali europee

NOVITÀ

29/08/08

First time visitors can now view The European Library User Guide in 18 **languages.** The User Guide is designed

EXHIBITION

More than 300 pictures of the national library buildings across

TUTORIAL

Figure 2. 3

Query shows the text of the query performed by the user. In Table 2.4 the twenty most frequent words present in the queries with their respective frequencies are reported.

Frequency	Word	Frequency	Word
4065	mozart	454	floyd
995	gogh	452	international
789	meisje	448	pink
789	parel	439	music
776	harry	434	erasmus
771	potter	416	rembrandt
546	journal	410	nuremberg
522	european	401	world
484	europe	379	maps
473	history	342	library

Table 2. 4

Besides Table 2.4, Figures 2.5 and 2.6 show how the portal interface appears when a user performs an advanced search.

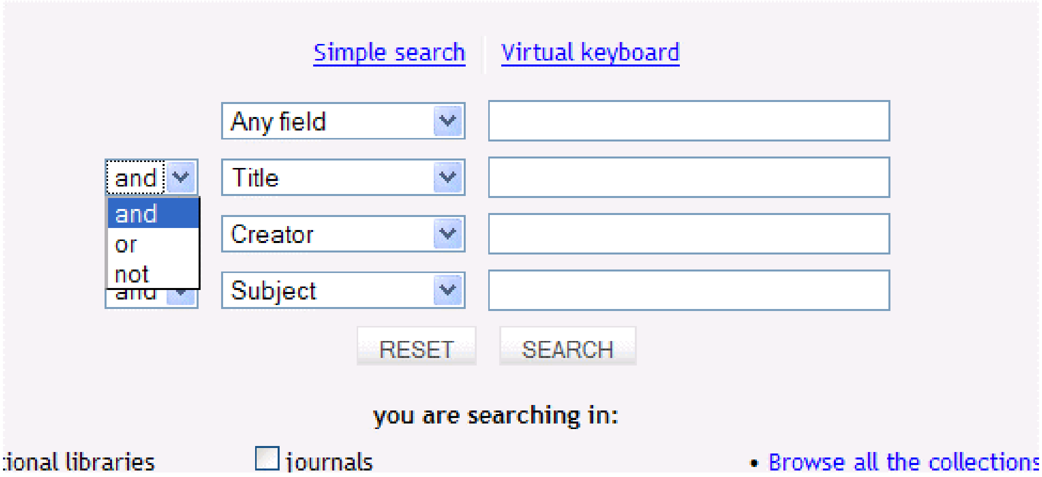


Figure 2. 4

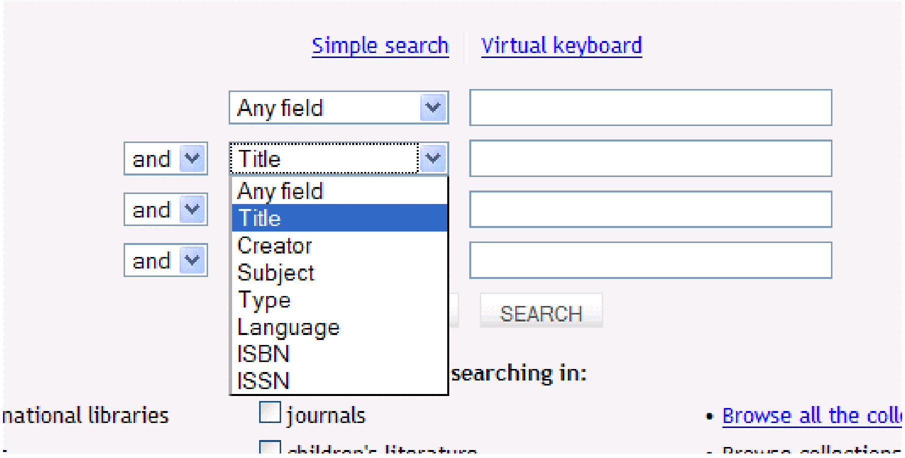


Figure 2. 5

When a user wants to refine his search, he can select one of the operator "and", "or", "not" from the menu in the left, and/or one of the items "title", "creator", "subject", "type", "language", "ISBN" from the menu beside. Let us show what the log file records when a user performs an advanced search; looking at Table 2.5, the column "query" contains also the operator and / or the items selected by the user from the menu. When counting the most frequent word, only the words really typed by the users was considered in the table.

sesid	query	action
00opv207t8bh0cgfdokp7ccsc2	(title all "vegetarian movement") and (creator all "twigg")	search_adv
00opv207t8bh0cgfdokp7ccsc2	(title all "vegetarian movement") and (creator all "twigg")	view_full
00opv207t8bh0cgfdokp7ccsc2	(title all "vegetarian movement") and (creator all "twigg")	view_full

Table 2. 5

In the log file each single query performed by each user is reported in much than one row. For example, in the portion of log file in Table 3.3, the query "(title all "vegetarian movement") and (creator all "twigg")" is executed once, through an advanced search, but the query is also reported in the two following rows. Therefore, to obtain only the real words frequencies we have selected the rows of the log file related to action of search.

It is interesting to observe the distribution of the frequencies of the words in regard to their rank. In Figure 2.6 graphical representation of the frequencies in function of their rank is depicted.

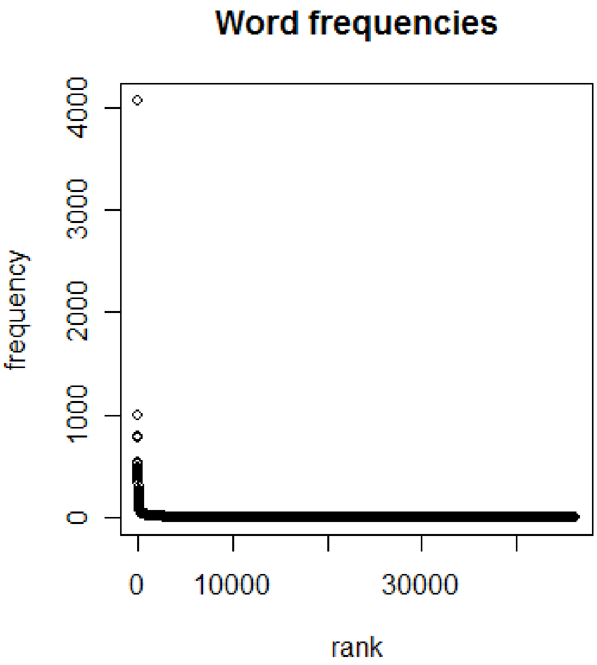


Figure 2. 6

The frequencies of each term are inversely related to their correspondent ranks. This fact remembers a manifestation of the Zipf's law; it states that, given a corpus of natural language utterances, the frequency of any word is inversely proportional to its rank in the frequency table.

Number of words per query ("word"). A useful information obtainable starting from this field is the number of words for each query, thus we create a new variable, called "word", that contains the number of words for each query. This variable has been computed only for the queries referred to not advanced searches; this is because, when a user performs an advanced search he can type some terms in a field, some other terms in another field (Figures 2.4 and 2.5), etc.. In other words, the query results refer to different queries, thus no one-to-one correspondence between the query and its length can be established. However, the advanced searches present in the dataset are 10.7% of the total number of search actions and then the contribution of those query lengths to the average query length would be negligible.

By "words" it is meant all the terms typed by the user in the query, articles and prepositions included; furthermore, each term linked to another by an hyphen is counted as two terms. Consider Table 2.6, as an example,: the query contains three terms separated by two hyphens, thus the variable "word" assumes a value equal to three.

query	word
("Reichs-Marine-Amt")	3

Table 2. 6

In Table 2.7 we report a summary of this new variable.

Min.	1st qu.	Median	Mean	3rd qu.	Max.
0	1	2	2.23	3	52

Table 2. 7

The minimum value is equal to zero—the queries it refers to are mistakes of the users; for instance, one can erroneously clicking on the search button before having typed any word.

In Figure 2.7 the distribution of the number of words per query is reported: the *x*-axis refers to the number of words and the *y*-axis refers to the number of queries which contain the correspondent number of words. As expected after observing the Table 2.7, the distribution is strictly asymmetric: most of the queries include less than three words.

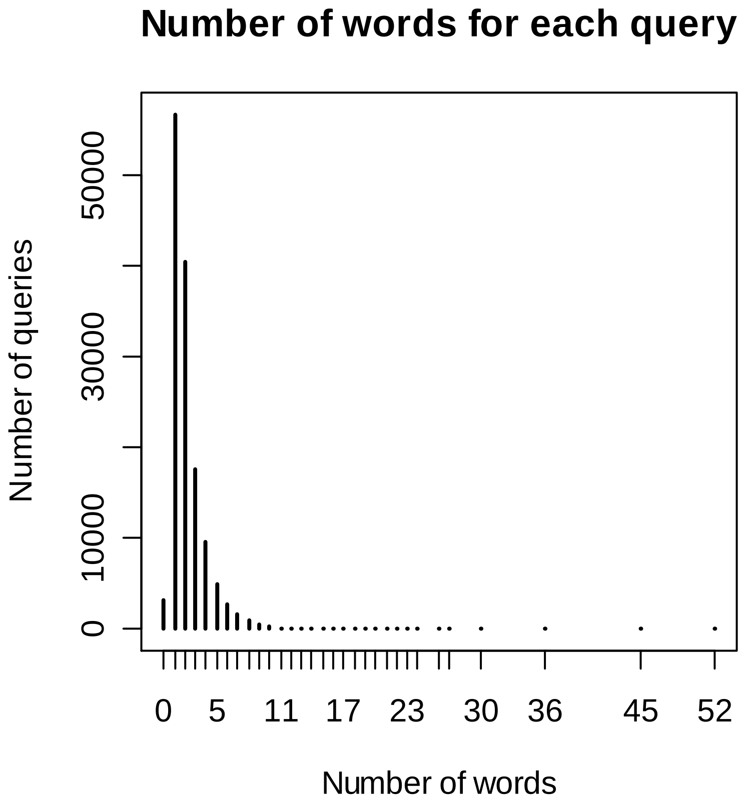


Figure 2. 7

To obtain the graph in Figure 2.7 all the queries related to actions of search (except the advanced searches) were used—as said above, the text of each query is recorded also for the action following that of search (Table 2.5).

The mean number of words per query found in this log file is close to other values reported in literature. In [43] Spink et al. find that the mean number of terms in unique queries was 2.4 as for 1999. They also noted that the average number of words per query increased over time; furthermore, those authors found that English language queries increased in length more quickly than European language queries. In [40] Silverstein et al. pointed out "the one contrast that was noted early in the history of web search is that searches on the web tend to have many fewer search terms than searches in more traditional information retrieval contexts [Jansen et al., 1998]". In their experiment, the authors found 2.35 words per query. Ussery stated that the average Google query now consists of 4 words, while before the average number of words per query was 3 [44].

Action ("action"). The dataset contains an entry for each single user action. The value that this variable can assume are:

- *search_sim* the search started with a simple search,
- *search_adv* the search started from an advanced search form,
- *search_res* the search started from a result record page,
- *search_res_rec_any / all* the search started from within a full record view by clicking on search (magnifying glass) icon in the record's available fields,
- *search_url* the search started from an URL[13] query string. This string may also have a domain name attached to it (search_url_www.domain.org) if it is coming from a remote TEL search-minitel (a marketing tool),
- *view_brief* the display of a short title list,
- *view_full* the display of an individual result record. It is activated when a user clicks on a title link in the list of brief records displayed (20 per page), or when a user clicks on the previous or next link when already viewing a full record individual result record,
- *jump_to_pag* the user entered a numerical value for skipping several pages of records from the brief title display,

[13] URL means Uniform Resource Locator and it's a compact string of characters used to represent a resource available on the Internet.

- *available_at* the user clicked the link "Available at Library" on the result page to view associated record in native interface,
- *see_online* the user clicked the link "See online" on the result page to view associated object in native interface,
- *col_set_X* the collection was chosen by method X where X can be:
    - *col_set_theme*: from theme list,
    - *col_set_theme_country*: from country list on homepage or results-page,
    - *col_set_country*: from all collections tab (collections listed by country),
    - *col_set_subj*: from subject list,
    - *col_set_desc*: by searching by description,
    - *col_set_default*: collection default list reinstantiated,
- *option_print* the result record is printed,
- *option_save_session_favorite* the result record is saved in reference session's favourites,
- *option_send_email* the record has been sent by e-mail,
- *option_save_reference* the record has been saved for reference manager use
- *service_<country>* the user utilized the full record service link to <country> for the currently viewed result record,
- *service_all* the user utilized the full record service link to other web services such as Google, Amazon, etc.,
- *show_help_<helpfilename>* the user clicked the "help" link.

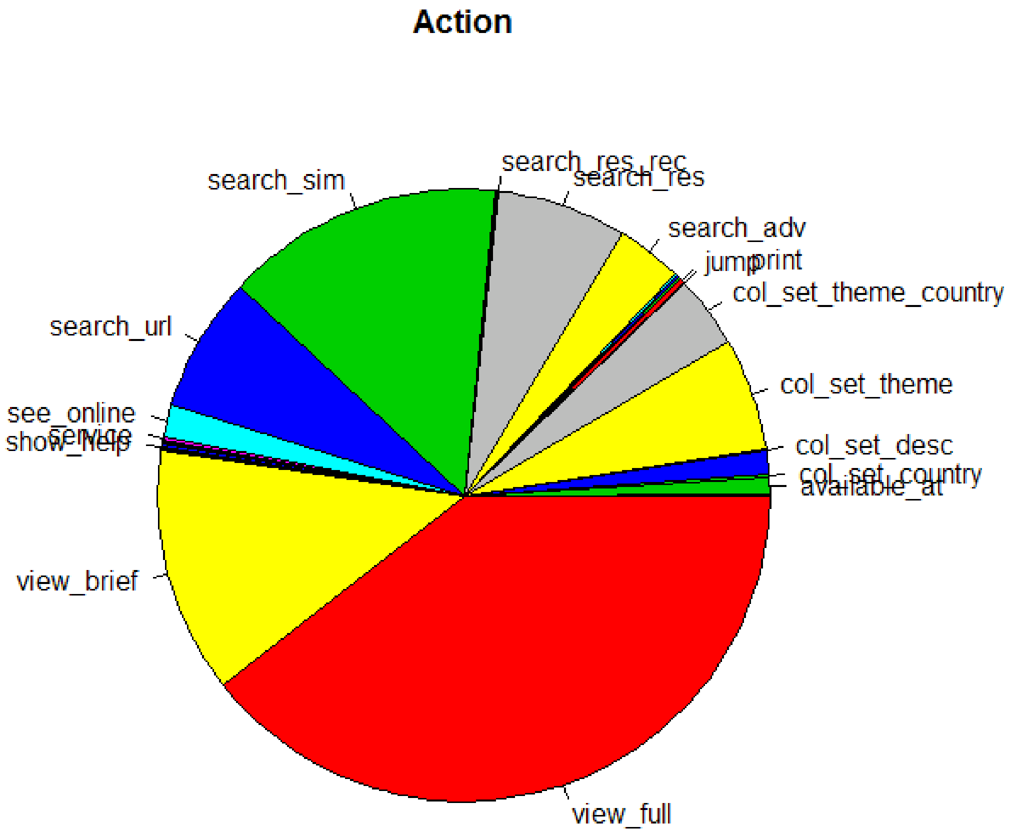


Figure 2. 8

In Table 2.8 the frequencies of each action are reported.

action	count
view_full	196428
search_sim	73100
view_brief	64429
search_url	35789
search_res	33771
col_set_theme	29855
col_set_theme_country	18894
search_adv	17247
see_online	8456
col_set_default	6144
available_at	4417
option_print	1384
col_set_country	1170
option_save_session_favorite	896
service_all	837
service_undefined	836
search_res_rec_any	597
search_res_rec_all	556
option_save_reference	537
option_send_email	490
service_uk_t	356
show_help_help/english/search_	343
service_denmark_t	321
jump_to_page	241

action	count
service_hungary_t	195
service_netherlands_t	192
service_portugal_t	131
	88
col_set_subj	84
show_help_help/english/collect	68
Search	62
col_set_desc	30
show_help_help/german/search_s	26
show_help_help/estonian/search	19
show_help_help/slovenian/searc	15
show_help_help/croatian/search	15
show_help_help/maltese/search_	15
show_help_help/hungarian/searc	15
show_help_help/polish/search_s	14
show_help_help/finnish/search_	13
show_help_help/german/collecti	13
show_help_help/french/search_s	11
show_help_help/latvian/search_	11
show_help_help/german/search_r	11
show_help_help/serbian/search_	11
show_help_help/danish/search_s	10
show_help_help/greek/search_si	9
show_help_help/portuguese/sear	9
show_help_help/danish/collecti	8
show_help_help/croatian/collec	7
show_help_help/polish/collecti	6
show_help_help/hungarian/colle	6
show_help_help/estonian/collec	6

action	count
show_help_help/german/search_h	6
show_help_help/czech/search_si	6
show_help_help/portuguese/coll	6
show_help_help/polish/search_r	6
show_help_help/czech/collectio	5
show_help_help/slovenian/colle	5
show_help_help/french/search_r	5
show_help_help/maltese/collect	5
show_help_help/greek/search_re	5
show_help_help/serbian/collect	4
show_help_help/czech/search_fa	4
show_help_help/finnish/collect	4
show_help_help/polish/search_h	4
show_help_help/french/collecti	4
show_help_help/greek/search_fa	4
show_help_help/greek/collectio	4
show_help_help/german/search_f	3
show_help_help/danish/search_f	3
show_help_help/latvian/collect	3
show_help_help/danish/search_r	2
show_help_help/czech/search_re	2
show_help_help/french/search_f	2
show_help_help/danish/search_h	2
show_help_help/greek/search_hi	1
show_help_help/french/search_h	1
show_help_help/czech/search_hi	1
show_help_help/polish/search_f	1
Total	498292

Table 2. 8

From Table 2.8 some considerations can be done. There is an empty field, meaning that there are 88 actions not classified. It is worth notice that this field is not constituted by missing values. Another field is that called "search": we do not know the exact meaning of the actions recorded in this field, but they are expected to refer to search actions. Moreover, we can see that the most frequent action regards the display of an individual result record. There are a lot of actions that involve a very little part of the users, these action regarding the clicking on the "help" link. Furthermore, the rows of Table are 80, signifying that there are 80 different types of action.

The sum of all the frequencies of the actions regarding the clicking on the "help" link is only the 0.15% of rows of the log file, thus it is reasonable to group all these actions in the variable "show_help_<helpfilename>". Then, we decide to merge in the variable "service" both the actions referring to the use of the full record service link to a certain Country and the actions indicating the use of the full record service link to other web services in the same group because they provide similar information. After the groupings illustrated above, we obtain the Table 2.9.

action	**count**	%	**action**	**count**	%
view_full	196428	39,42	option_save_session_favorite	896	0,18
search_sim	73100	14,67	search_res_rec_any	597	0,12
view_brief	64429	12,93	search_res_rec_all	556	0,11
search_url	35789	7,18	option_save_reference	537	0,11
search_res	33771	6,78	option_send_email	490	0,10
col_set_theme	29855	5,99		88	0,02
col_set_theme_country	18894	3,79	jump_to_page	241	0,05
search_adv	17247	3,46	col_set_subj	84	0,02
see_online	8456	1,70	search	62	0,01
col_set_default	6144	1,23	col_set_desc	30	0,01
available_at	4417	0,89	show_help_<helpfilename>	759	0,15
option_print	1384	0,28	service	2868	0,58
col_set_country	1170	0,23	**total**	498292	100

Table 2. 9

Display time ("display") is the time spent displaying a page. This can be viewed as an implicit measure of users interest. In the previous work reported in the literature, display time was examined for a variety of user tasks, such as news reading, web browsing, reading journal articles and web searching. Some researches showed that display time can be an effective measure of user interest when users are engaged in a news reading task or web browsing. Morita and Shinoda [35] conducted an experiment where users explicitly rated news articles and the time spent displaying was recorded. They found a strong tendency to spend a long time displaying articles they rated interesting as compared to articles they rated not interesting. A later experiment conducted by Konstan et al. [27] required users to explicitly rate UseNet articles while the time users spent on a page was recorded. The results showed a relationship between time spent displaying and explicit ratings. Claypool et al. [11] examined the correlation between time spent displaying and user interest for user directed web browsing. Users explicitly rated web pages and the time spent displaying was recorded. The time spent displaying was found to be a good indicator of interest. Kim et al. [26] evaluated time spent displaying for users reading academic journal articles. Users explicitly rated the articles and the time spent displaying was recorded. They also found that time spent displaying could be used to predict interest. Users tended to spend longer amounts of time displaying relevant articles than non relevant articles.

The effectiveness of display time for Information Retrieval tasks was examined, too, in the past, but the findings varied across the literature works. White et al. [49] recorded the time users spent displaying while users judged the relevance of a query to a document summary. They reported that the difference between time spent displaying relevant documents and non relevant documents was statistically significant. However, when Kelly and Belkin [23] attempted to replicate the results of Morita and Shinoda for web search tasks, they reported that they found no significant difference in the time spent displaying relevant and non relevant documents. These findings suggest that although time spent displaying may be good indicator of interest for some tasks, such as news reading and web browsing, it may not be a good indicator for all tasks.

It is our opinion that display time provide useful information about the users' search activity performed to access the TEL portal because the task performed by those users is similar to news reading and web browsing due to the link-based nature of the data presented by the interface and the short average size of the documents displayed. As a consequence, the time users spent displaying the individual result records was computed from the TEL portal action log file. Display time was computed as the difference between the timestamp referred to the action "view_full" and the timestamp related to the successive action, in the same session. We cannot compute the display time correspondent to all the displays of the individual records because the log file does not record the action for logging off the portal. To sum up, the display time only if the user performs another action inside the same session, after having seen a single result record. If the "view_full" is the last action the user executes inside a session, we cannot compute the time the user has spent looking at that page because the successive action that we find in the dataset is related to another user.

For the reasons explained, the 79.82% of the total displaying times could be computed. To obtain this percentage the number of not null elements in the column "display" was counted and divided by the total number of actions "view_full". To have an idea of the display time distribution a summary of this variable in Table 2.10 is provided.

Min.	1st qu.	Median	Mean	3rd qu.	Max.	Null
0	7	20	217.4	48	1805000	341495

Table 2. 10

The 341,495 null values that appear in Table 2.10 match to the "view_full" actions for which the display time could not be computed. To sum up, a missing value was found in the column "display" of the 301,864 rows referring to an action different from "view_full".

The minimum value this variable can assume is 0 seconds, the median is 20 seconds, while the mean is much higher: its value is 217.4 seconds; it is evident that this value is influenced by some outliers, that is, some very high values, indeed, the 3rd quartile is equal to 48 seconds. Further, the maximum display time is 1805000 seconds, that corresponds to almost 21 days! This value requires to be better investigated. After looking at Table 3.10 it can be noted that the third row match to the same session code of the second row, but the IP address is different, thus we can state that the action "view_full" of the second row does not belong to the same session of the action "search_res" in the third row. Thus we cannot compute that display time its value is replaced with a missing value.

userip	**sesid**	**action**	**date**	**display**
85.140.17.200	sjbe25llkbic7ks9klbjfaunn5	search_sim	2008-01-05 04:01:07	
85.140.17.200	sjbe25llkbic7ks9klbjfaunn5	view_full	2008-01-05 04:02:31	1804541
85.140.17.133	sjbe25llkbic7ks9klbjfaunn5	search_res	2008-01-26 01:18:12	
85.140.17.133	sjbe25llkbic7ks9klbjfaunn5	view_full	2008-01-26 01:19:23	

Table 2. 11

We inspected all the display time values greater than 86,400 seconds, i.e., 24 hours, and 31 cases were found:

- 4 cases were similar to that of the Table 2.11: the actions successive to the "view_full" were referred to different IP address, so we have replaced their value with a missing value,
- the remaining 27 cases seem correct, namely the following action was related to the same session and the same IP address. We also notice that
    - 40.74% (11/27) of these anomalous display time values match to the IP address "193.10.249.131" and
    - 14.81% (4/27) match to the IP address "194.171.184.19".

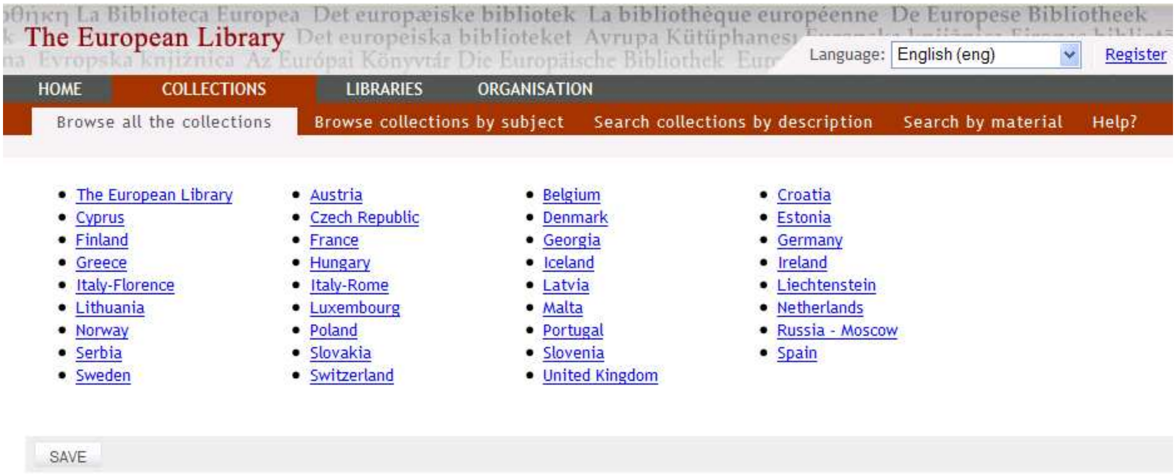
We report in Table 2.12 the summary of the variable after the replacement of the five anomalous values mentioned above with missing values.

Min.	**1st qu.**	**Median**	**Mean**	**3rd qu.**	**Max.**	**Null**
1	7	20	201.6	48	682200	341500

Table 2. 12

Although the maximum value is decreased, the variable distribution remains almost the same described above.

Collection identifier ("colid") is a code that identify the collection upon which the action is performed. As one can see in Figure 2.9, the portal offer the possibility to choose the collection.



Language: English (eng)

Figure 2. 9

330,420 rows of this field, that is 66.31% on the total, are empty.

Number of records ("nrRecords") is the total number of retrieved records from each collection.This field can be different from zero only when it refers to the action indicating:

- the display of the results of the search ("view_brief", "view_full"),
- the skipping of some page of records ("jump_to_page"),
- the clicking on the link "Available at Library" on the result page ("available_at"),
- the clicking on the link "See online" on the result page ("see_online"),
- the printing of the result record ("option_print"),
- the saving of the result record ("option_save"),
- the sending by e-mail of the record ("option_send_email"),
- the utilize of the full record service link ("service").

To have an idea of the distribution of this variable, the numbers of record referred to the action "view_brief" and "view_full" were selected and grouped, thus in Figure 2.10 it is represented the counting of rows which present an equal number of record, in function of the number of records. In this graph we restricted the range of the x-axis at 1000 records to better understand the decreasing trend.

The highest point placed on the left of the graph tells that there are about 64,000 search result display actions corresponding to zero retrieved records. This is probably due to some mistake in the encoding of the log file; an hypothesis can be that those value represent missing value, rather than zero.

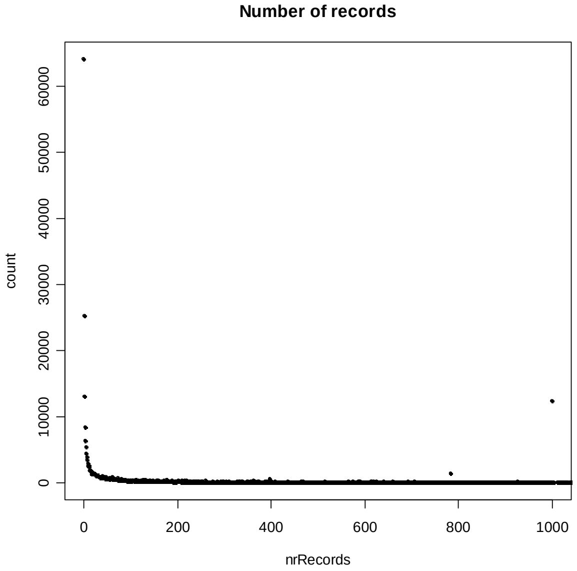


Figure 2. 10

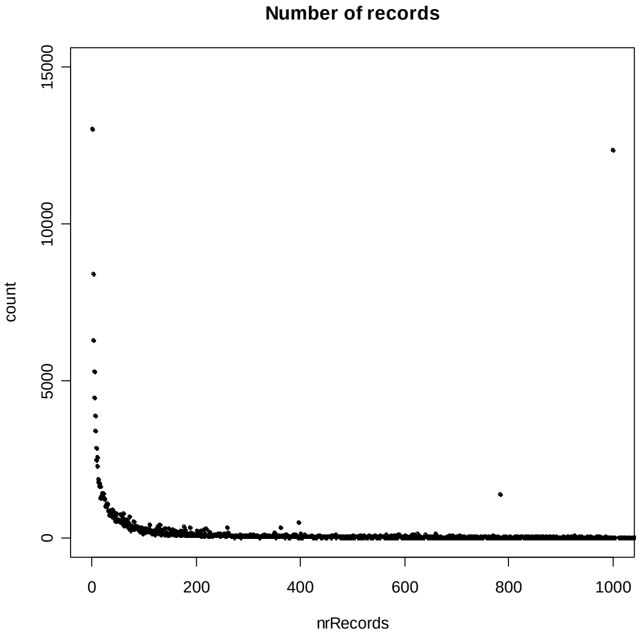


Figure 2. 11

In Figure 2.11 we restrict the range of the y-axis, too, and the decreasing trend is evident. Thus we can say that the most part of the searches gives in response a very small number of records; as the number of records increases, the counting of rows decreases. The highest points on the right of the Figures 2.10 and 2.11 correspond to the coordinates (783, 1676) and (1000, 13140); this means that there are 1,676 rows in the log file match 783 records in the result list, and 13,140 rows in the log file match 1000 records in the result list. It is likely that 1000 is a standard number of retrieved records.

Record position ("recordPosition") is a code that indicates the position of the viewed item in the total record list. The value of this variable correspond to an empty string or to an hyphen in 385,497 rows, that is the 77.36% on the total. This field contains a code only when it refers to the action indicating:

- the searching, only when the search's activity is initiated from an URL query string ("search_url"),
- the display of the results of the search ("view_brief", "view_full"),
- the skipping of some page of records ("jump_to_page"),
- the printing of the result record ("option_print"),

- the saving of the result record ("option_save"),
- the sending by e-mail of the record ("option_send_email"),
- the utilize of the full record service link ("service").

Search boxes identifier ("sboxid") is the identifier for remote search boxes which query the portal via URL. This field contains a missing value in each row, thus it has been taken off the dataset.

URL ("objurl") is the URL of the objects reached through the actions "available_at" or "see_online".

Date ("date") is the timestamp, in the format yy-mm-dd hh:mm:ss. The log file at our disposal contains information on users' accesses referring to the period from 1st November 2007 until 29th February 2008.

2.6 Variable Dichotomization

We decided to utilize dichotomous variables for different reasons, depending on each single case:

- some variables naturally adapt to the dichotomization because they essentially indicate whether a certain phenomenon occurs or not; these variables are "userid_dic", "lang_dic", "search_dic", "print_dic", "save_dic", "service_dic";
- "display_dic" is a quantitative continuous variable; we have looked for a reasonable threshold to discriminate between longer display times, sign of interest for the user, and shorter display times, sign of disinterest for the user;
- "word_dic" is a quantitative continuous variable; we have looked for a reasonable threshold to discriminate between longer and shorter queries.

User identifier dichotomized ("userid_dic"): Since the main information that gives this variable is about whether a user is registered an whether he is guest, it was dichotomized, namely, a new variable, called "userid_dic" was defined and its possible values are:

- 0 if the user did not perform the log in to the portal,
- 1 if the user accessed to the portal after the log in.

Language dichotomized ("lang_dic"). It is useful to distinguish between the users which maintain the default language and users that change it, therefore a new variable, called "lang_dic", whose value are:

- 0 whether the user maintain English language,
- 1 whether the user change the default language

was created.

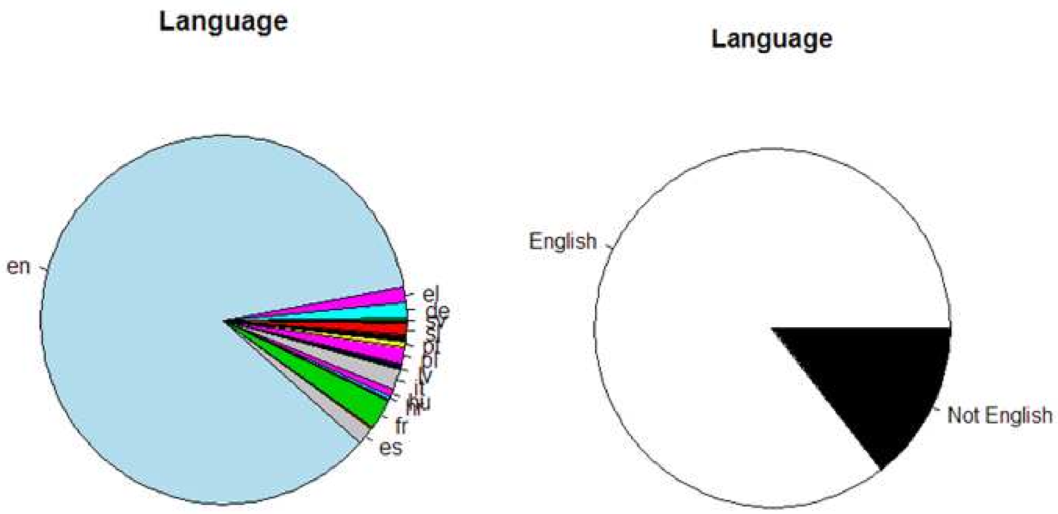


Figure 2. 12

Number of words per query dichotomized ("Word_dic"). A threshold which can discriminate between "long queries" and "short queries" is the median number of word per query. If the observations of a variable are ordered by value, the median value corresponds to the middle observation in that ordered list. The median value corresponds to a cumulative percentage of 50%, namely, half of the values are below the median and half are above it. The median has the property of being less sensitive to extreme values than the mean; this makes the median a better measure than the mean for highly skewed distributions. The distribution of the words in the queries of our log file is asymmetric, and we need a threshold that take into consideration this, so we choose a threshold equal to the median. A new variable, called "word_dic" whose values are:

- 0 whether the query contains one or two words,
- 1 whether the query contains more than two words

was defined in our analysis.

Display time dichotomized ("display_dic"). Morita and Shinoda [35] found that the most effective display time threshold to express user's interest was 20 seconds. Kelly and White [25] showed that the median display time was the most consistent indicator of relevance. We can summarize the reasons that led us to choose the median as threshold to distinguish between display time values indicating user interest and display time values not indicating it:

- the median is a statistic indicator with the property of robustness regard to the outliers,
- the median display time has been indicated as the most consistent indicator of relevance in other experiments,
- the threshold of 20 seconds has been found as the most effective threshold in another experiment.

The variable "display" was dichotomized and a new variable was defined the values of the latter being:

- 0 whether the display time is less than or equal to 20 seconds,
- 1 whether the display time is greater than 20 seconds.

Type of search dichotomized ("search_dic"). An interest information obtainable starting from the variable "action" is about which type of search the user performs. As said above, there are five types of search:

1. simple search,
2. advanced search,
3. search initiated from a result record page,
4. search initiated from within a full record view,
5. search initiated from an URL query string.

The greatest difference between these types of search is between the advanced search and all the others; in other words, the last three kinds of search are assimilable to a simple search. The advanced search is different from all the others because it is the unique which allows to refine the search; no one of the others did permit it.

Thus we create the variable "search_dic" whose value are:

- 1 whether it is referred to an advanced search,
- 0 whether it is related to another type of search.

This new variable is computed for each dataset row: for the rows related to actions of search, it reports the type of search, and, for the following rows, it describes to which type of search they refer to. In Table 2.13 a little part of the log file is reported with the addition of the column "search_dic": the first row refers to a simple search action, thus the new variable takes the value of zero, the second row is related to a display result action; since these results are obtained from the simple search, the "search_dic" variable takes the value of zero, too.

sesid	**action**	**search_dic**
006u22hebus1fu99gjuc8nbp76	search_sim	0
006u22hebus1fu99gjuc8nbp76	view_full	0
006u22hebus1fu99gjuc8nbp76	view_brief	0
006u22hebus1fu99gjuc8nbp76	search_adv	1
006u22hebus1fu99gjuc8nbp76	view_full	1

Table 2. 13

Print dichotomized "print_dic". Below the presence in this log file of actions, called "proxies" which can be considered as implicit indicators of relevance[14] will be discussed; one of these is that indicating the printing of a result record. As a consequence, a variable whose value are:

[14] See section 2.8

- 1 whether the user prints a result record,
- 0 whether the user does not print the result record

was defined and computed only for the rows referring to "view_full" actions, all the rows related to others actions present a missing values in the column "print_dic".

Save dichotomized "save_dic". Others actions signifying user interest are those indicating the saving of the result record; they are:

1. the saving in reference session's favourites,
2. the sending by e-mail,
3. the saving for reference manager use.

We have included in this group also the sending of the result record by e-mail because it is a form of saving.

A variable whose values are:

- 1 whether the user saves a result record,
- 0 whether the user does not save the result record

was then defined yet computed only for the rows referring to "view_full" actions, since all the rows related to others actions present a missing values in the column "save_dic".

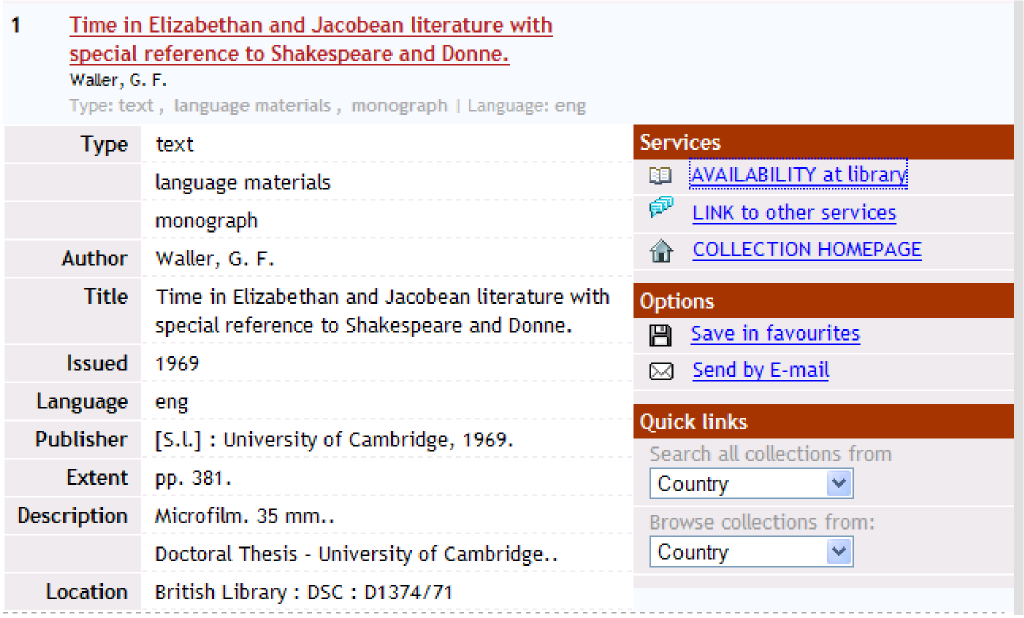
Service dichotomized "service_dic". When a user view a result record, he can utilize a set of services; in particular, he will be

- clicking the link "Available at Library" on the result page to view associated record in native interface,
- clicking the link "See online" on the result page to view associated object in native interface,
- utilizing the full record service link to a country for the currently viewed result record,
- utilizing the full record service link to other web services such as Google, Amazon, etc..

To better understand what these link are referred to, some examples are provided. In Figure 2.13 the portal interface when a user displays a result record is reported with the links "AVAILABILITY at library", "LINK to other services" on the left. Figures 2.14 and 2.15 show respectively what appears when the user click the first or the second link. When a user clicks on these links, he implicitly expresses an interest toward to the record viewed. Then, a variable whose values are:

- 1 whether the user clicks on at least one of these links,
- 0 whether the user does not click on any of these links

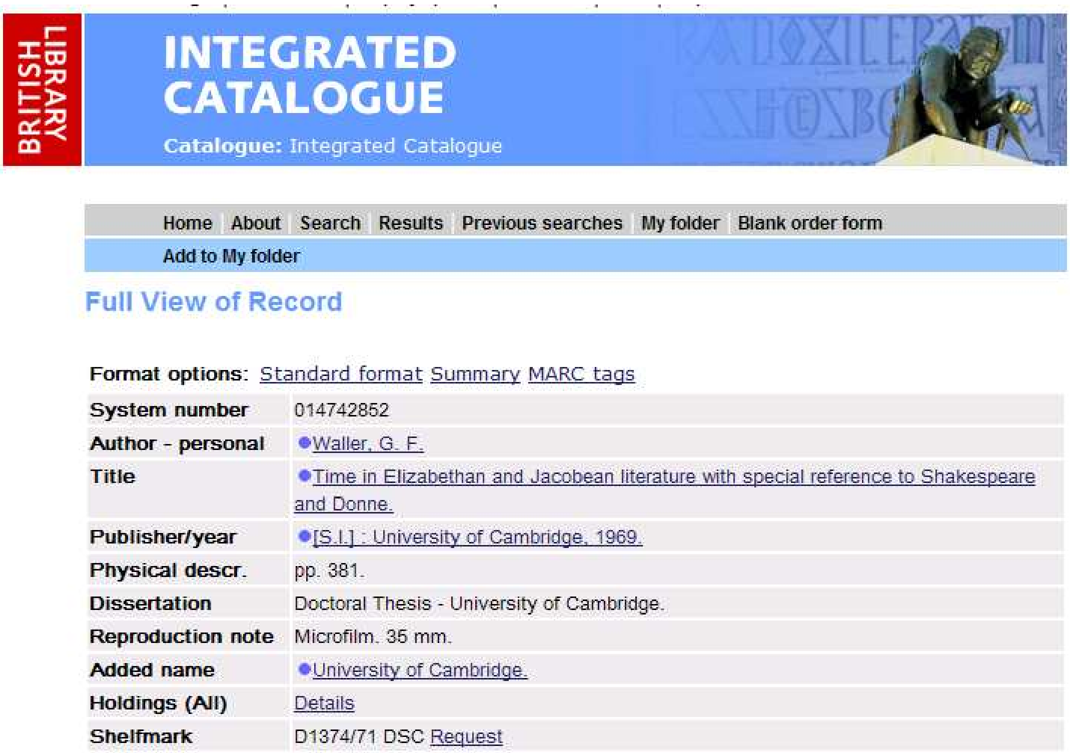
was defined yet computed only for the rows referring to "view_full" actions, all the rows related to others actions present a missing values in the column "service_dic".



1 Time in Elizabethan and Jacobean literature with special reference to Shakespeare and Donne.
Waller, G. F.
Type: text , language materials , monograph | Language: eng

Type	text
	language materials
	monograph
Author	Waller, G. F.
Title	Time in Elizabethan and Jacobean literature with special reference to Shakespeare and Donne.
Issued	1969
Language	eng
Publisher	[S.l.] : University of Cambridge, 1969.
Extent	pp. 381.
Description	Microfilm. 35 mm..
	Doctoral Thesis - University of Cambridge..
Location	British Library : DSC : D1374/71

Figure 2. 13

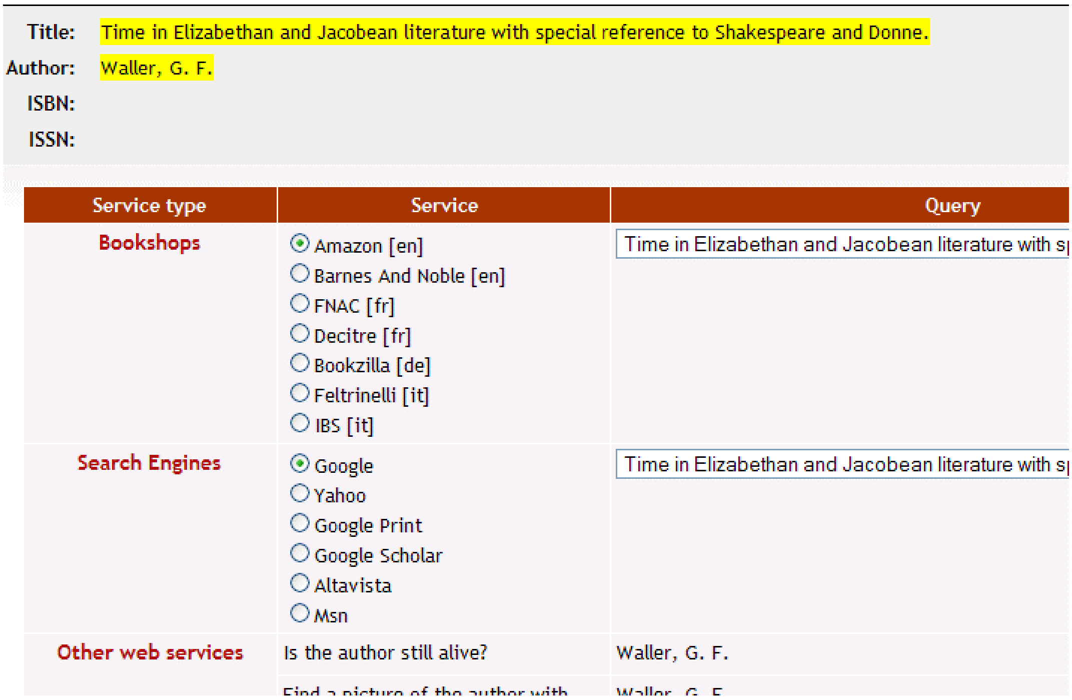


Full View of Record

Format options: Standard format Summary MARC tags

System number	014742852
Author - personal	●Waller, G. F.
Title	●Time in Elizabethan and Jacobean literature with special reference to Shakespeare and Donne.
Publisher/year	●[S.l.] : University of Cambridge, 1969.
Physical descr.	pp. 381.
Dissertation	Doctoral Thesis - University of Cambridge.
Reproduction note	Microfilm. 35 mm.
Added name	●University of Cambridge.
Holdings (All)	Details
Shelfmark	D1374/71 DSC Request

Figure 2. 14



Title:	Time in Elizabethan and Jacobean literature with special reference to Shakespeare and Donne.
Author:	Waller, G. F.
ISBN:	
ISSN:	

Service type	Service	Query
Bookshops	◉ Amazon [en] ○ Barnes And Noble [en] ○ FNAC [fr] ○ Decitre [fr] ○ Bookzilla [de] ○ Feltrinelli [it] ○ IBS [it]	Time in Elizabethan and Jacobean literature with s
Search Engines	◉ Google ○ Yahoo ○ Google Print ○ Google Scholar ○ Altavista ○ Msn	Time in Elizabethan and Jacobean literature with s
Other web services	Is the author still alive?	Waller, G. F.
	Find a picture of the author with	Waller, G. F.

Figure 2. 15

2.7 On the Impact of the Display Time

In Figure 2.16 there is the boxplot of the variable "search_dic" conditioned to the display time: display time variability associated to the group of advanced searches is larger than that associated to the group of the other types of searches. The "t test" and the "F test" confirmed that both the mean values and the variances are *significantly different.*

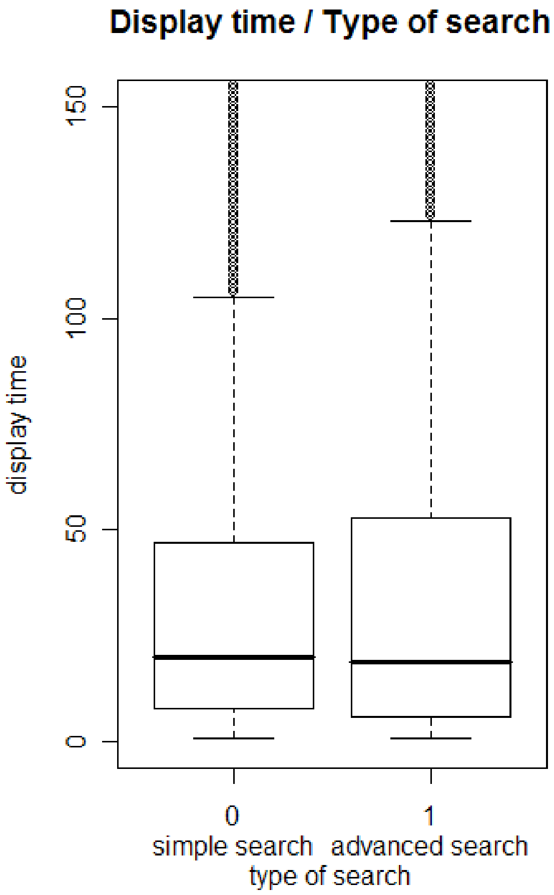


Figure 2. 16

In observing the boxplots in Figure 2.17 the actions of retaining are associated to lower display time values; furthermore, the boxes referred to printing, saving and use of services present lower display time variability than the other boxes. The tests performed with a 0.05 level of significance confirm that both the differences of the mean values and of the variances are *significant*.

This means that, in general, when a user retains a result record by saving or printing it, or when he looks for further information about the record, he spends less time looking at the page. This fact is reasonable because whether the user retains the result record he can look at it in another moment, and whether he looks for further information, probably he spends more time in the pages reached by the services links. To sum up, we do not know the real display time because the user will look at the result "off-line".

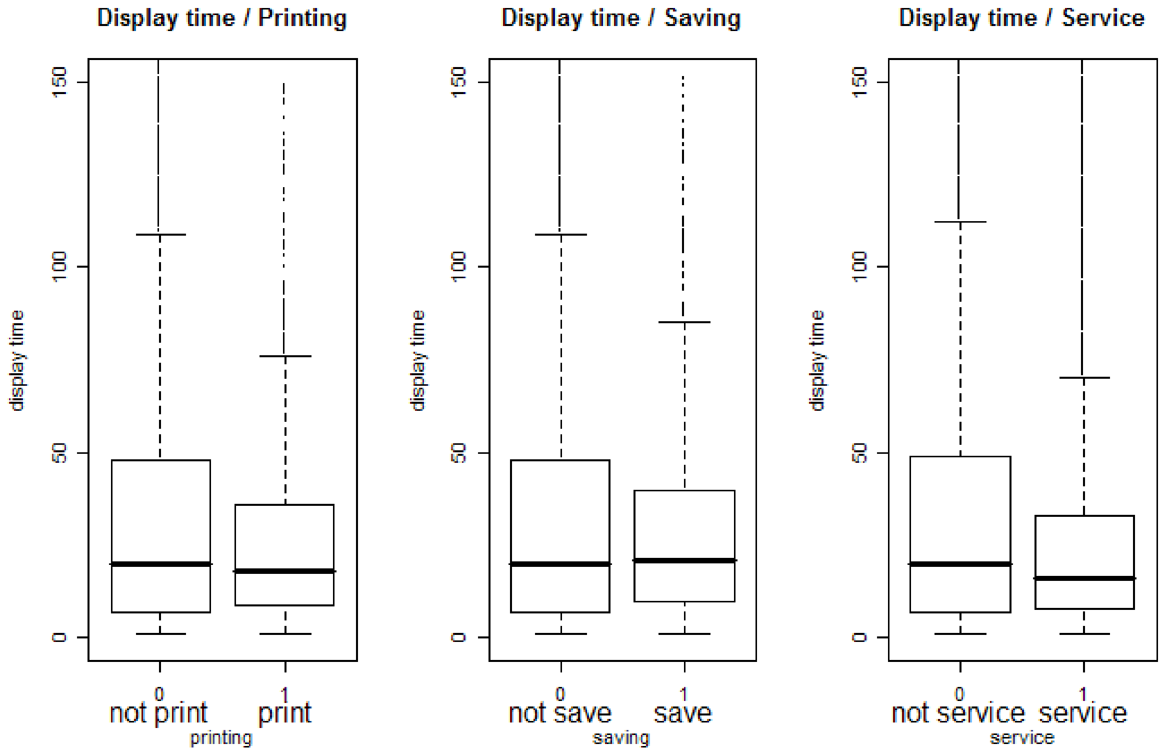


Figure 2. 17

In Figure 2.18 three boxplots are reported; they show the possible relation between the variable "display" and each of the dichotomized variables referred respectively to the user identifier, the language and the number of words per query. The variables "id", "userip" and "sesid" were not considered in this analysis because they are not about the users' behaviors.

Let us consider the first boxplot; apparently, there are not significant differences between the display time values of the registered users and those of the guest users: both user groups present the same median display time, represented by the horizontal line inside each box, and the variance of the display time inside each of the two users' group, represented by the length of the boxes, is almost equal. However, a "t test" was performed to verify the equality of the two mean values and a "F test" to verify the equality of the two variances, both at the 0.05 level of significance: both the means and the variances result to be *significantly different*.

Analogous considerations can be done regard to the second boxplot: the two groups seem to be equal, but *the tests refuse the null hypothesis of equality*.

Looking at the graph referred to the relation between the display time and the number of words, a slight difference between the two boxes occurs: to the short queries is associated a less variable displaying time; the median values inside the two groups of queries are equal.

However, the test performed show one more time that the differences between the display time values observed for different query lengths are statistically significant.

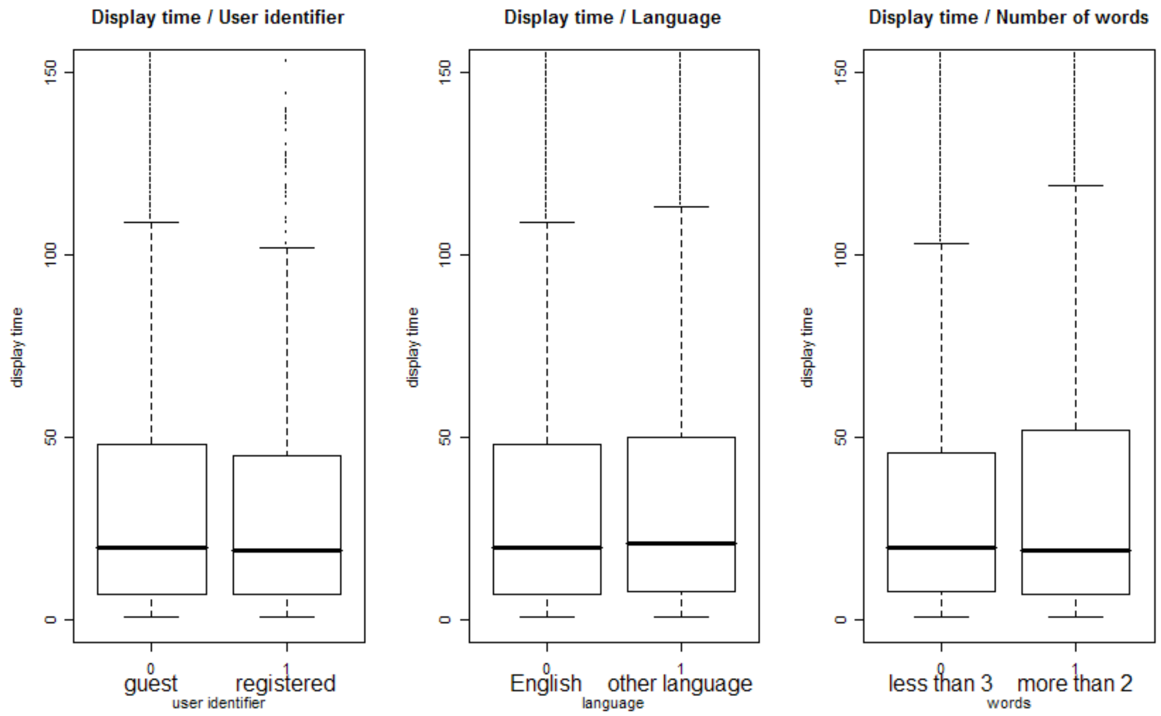


Figure 2. 18

2.8 The Proxy Variables

In our dataset there is not a variable explicitly indicating relevance. However, there are some actions that can be seen as clues of relevance. These actions are defined as "proxy", where proxy refer to a statistic indicator that describes a certain phenomenon not directly observable or not objectively measurable.

In literature, the proxy variables are used in the field of the services of public utility to the people; these type of services include all the performances of social interest in which the main result identifies with the effects that the service itself produces on the user. The mentioned result is named "outcome"; its definition is still object of debate. Gori and Vittadini [17] define it as the result, often in the long period, generated by the delivery of a benefit or by the distribution of a service, on a condition, state or behaviour of the user. It is evident how it is difficult to measure an outcome, thus we can recur to the use of proxy which are indicators that aim to describe the above mentioned results.

Relevance can be considered as a sort of outcome, because it is about the satisfaction of a need of a user which is not objectively measurable. Indeed, the relevance assessments can be subjective: for instance, if a rating in reference to our interest towards a certain document should be provided, an higher score than another person for which the document has been more interesting than for us could be used.

It is our opinion that the actions that provide information about the user interest or relevance, that is, the proxy, are all those referring to operations of retain:

- option_print,
- option_save_session_favorite,
- option_save_reference,
- option_send_email.

These actions have been recognized as implicit interest indicators in different experiments ([11, 21, 24]). Some researchers believe that the display of a result record is clue of relevance too; we have decide not to consider these action as proxy of relevance because we believe that these are much weak indicators; furthermore, they appear in the most part of the sessions in analysis, thus we probably would have evaluate as relevant documents not really relevant.

The actions listed above are not all equally significant in indicating relevance; in particular, printing is a stronger interest indicator than saving because entails physical paper consumption while saving may lead to forwarding the record to a colleague or friend who might be interested in the document, or indicate the intention of looking at it later, but not necessarily because it is relevant. However, these are still hypotheses which need to be confirmed after a further investigation.

2.9 The Dataset after the Analysis

In this section, a description of the dataset after the definition of the dichotomous variables is reported. To allow the observation of the consecutive actions inside a session, we have sorted all the log file by the column "session", while the original log file was ordered by the column "date". Tables 2.14, 2.15 and 2.16 report a small part of the dataset obtained after the definition of the variables described in this chapter. In particular, they show the information referred to the session "3n09267661nl5f26mekqaaq3f0". One can see that

- this session begins the 1st of November 2007, at 09:53:29, and it ends on the same day, at 12:48:00;
- the session begins with a simple search by typing the number 8684111079,
- then the user views a result record and remains on that page for 80 seconds; this time is larger than the median display time, thus the cell referred to the column "display_dic" contains "1".
- The variable "userid_dic" takes the value of zero, indicating that the user has accessed to the portal without logging in,

- the column “word” is full of “1”, signifying that all the queries typed by this user include a single term,
- the column ”print_dic”, “save_dic” and “service_dic” contain only zero or missing value, meaning that the user does not perform any retaining action, neither does he utilize links to services.

id	userip	sesid	lang	query	action	colid
1008541	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("8684111079")	search_sim	
1008542	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("8684111079")	view_full	
1008543	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("5722104027")	search_res	
1008546	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("5722104027")	view_full	
1008575	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	search_sim	
1008577	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_full	
1008583	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0001
1008584	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_full	a0001
1008587	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0001
1008596	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008602	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008612	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008615	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008616	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008617	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008618	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008620	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008621	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("platonov")	view_brief	a0037
1008659	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("582430582x")	search_res	
1008663	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("582430582x")	view_full	
1008665	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("582430582x")	view_full	a0037
1008975	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("strigin")	search_res	
1008984	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("strigin")	view_full	
1008985	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("strigin")	view_full	a0037
1009071	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("5885241228")	search_res	
1009074	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("5885241228")	view_full	
1009075	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("5885241228")	view_full	a0037
1009115	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("pivovavor")	search_res	
1009117	147.91.249.1	3n09267661nl5f26mekqaaq3f0	en	("pivovavor")	view_full	

Table 2. 14

nrRecords	recordPosition	objurl	date	userid_dic
0	-		2007-11-01 09:53:29	0
0			2007-11-01 09:53:39	0
0	-		2007-11-01 09:54:59	0
0			2007-11-01 09:55:23	0
0	-		2007-11-01 10:00:19	0
12			2007-11-01 10:00:53	0
60	21-40		2007-11-01 10:01:44	0
60		http://www.theeuropeanlibrar…	2007-11-01 10:01:51	0
60	41-60		2007-11-01 10:02:11	0
342	21-40		2007-11-01 10:03:08	0
342	41-60		2007-11-01 10:04:04	0
342	61-80		2007-11-01 10:05:01	0
342	81-100		2007-11-01 10:06:13	0
342	101-120		2007-11-01 10:07:55	0
342	121-140		2007-11-01 10:08:41	0
342	141-160		2007-11-01 10:09:24	0
342	161-180		2007-11-01 10:10:32	0
342	181-200		2007-11-01 10:11:04	0
0	-		2007-11-01 10:26:52	0
0			2007-11-01 10:27:15	0
1	1	http://www.theeuropeanlibrar…	2007-11-01 10:27:18	0
0	-		2007-11-01 12:16:09	0
0			2007-11-01 12:16:42	0
12	9	http://www.theeuropeanlibrar…	2007-11-01 12:16:42	0
0	-		2007-11-01 12:34:45	0
0			2007-11-01 12:35:10	0
1	1	http://www.theeuropeanlibrar…	2007-11-01 12:35:10	0
0	-		2007-11-01 12:47:39	0
0			2007-11-01 12:48:00	0

Table 2. 15

lang_dic	display	word	word_dic	display_dic	search_dic	print_dic	save_dic	service_dic
0		1	0		0			
0	80	1	0	1	0	0	0	0
0		1	0		0			
0	296	1	0	1	0	0	0	0
0		1	0		0			
0	51	1	0	1	0	0	0	0
0		1	0		0			
0	20	1	0	0	0	0	0	0
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0		1	0		0			
0	3	1	0	0	0	0	0	0
0	6531	1	0	1	0	0	0	0
0		1	0		0			
0		1	0		0	0	0	0
0	1083	1	0	1	0	0	0	0
0		1	0		0			
0		1	0		0	0	0	0
0	749	1	0	1	0	0	0	0
0		1	0		0			
0		1	0		0	0	0	0

Table 2. 16

CHAPTER 3

IMPLICIT FEEDBACK FOR TEL

3.1 Introduction

After having presented an overview of the implicit feedback techniques used in the literature of Information Retrieval (IR), and after having provided a description of TEL log file, a methodology fitted to TEL case is presented in this chapter. In particular, a technique which take the context in which the subjects are into consideration is illustrated. Since IR aim is to retrieve all and only the documents relevant to a specific user in a given moment and place, it is intrinsically linked to context. Indeed, what is relevant to a subject in a specific situation might no longer be relevant to another user or even to the same subject in another condition, that is users' searches are influenced by context.

As seen in the previous chapter, TEL users constitute a heterogeneous group of subjects who perform a set of search activities, each different from the others. Therefore, a methodology which takes the context in which each user is into consideration and distinguishes every seeking activity from the others is crucial. This methodology is investigated in this thesis by considering the greater number of indicators as possible, because the indicators would allow to delineate the context in which every user performs his search actions.

In this chapter, the methodology proposed by Melucci and White in [33] is described. In this paper, the authors suggest that an efficient IR system should be context-aware; then, they illustrate a model for navigation and search in context, that is, the navigation and search which adapts the retrieved results according to what the user does during his interaction with the system.

The chapter is structured as follow: it begins with a review on the concept of context, then, the Vector Space Model (VSM) is introduced since it has a strong relationship with the methodology we are going to investigate. Thereafter, the methodology for navigation and search in context is presented. Finally, we focus on our case study, propose a way to fit the illustrated methodology to TEL log file, present the records' indicators and provide an application using data from the log file.

3.2 Context

IR systems are designed to retrieve all and only the documents relevant to every specific information need, related to every user at every place and at every moment. By "relevant documents", we mean documents that contain information important, useful or necessary to satisfy the informative need of a user. Therefore, it is indisputable that relevance depends from the context: what is relevant to a user at a certain place and moment might no longer be relevant at another place or moment, to another subject or even to the same user. Furthermore, every query is ambiguous because it is difficult for a user to express the information need. The ambiguity mainly regards synonymity and polysemy:

- synonymity indicates the condition of substitutability of a linguistic element with an other in the context and in the situation given, without a consequent change in meaning;
- polysemy is the coexistence of different meanings in a word.

However, it is important to specify that the ambiguity also includes other problems, in addition to synonymity and polysemy. Consequently, the errors that an IR system commits when retrieving non-relevant documents or not retrieving relevant documents, is due to the ambiguity of the natural language used by the authors of the documents and by the user who express the query. If we could gather and exploit the context in which the information need is reached, then it could be possible to retrieve all the documents that contain the words of the query, in the same context of the query, and consequently improve the IR system effectiveness.

In this thesis, we refer to context as the whole of the features characterizing users, time, place, and everything emerging from the interaction between user and system.

Classical IR systems are context-unaware, since the most common models lack a formal representation of context. In the past, IR systems have been defined by assuming that there is one user, one information need for each query, one location, one time, one history and one profile, thus contextual features are not captured at indexing time, neither they are exploited at retrieval time.

However, the probabilistic model might offer the constructs for modelling context, as the probability of relevance can be updated by Bayes' theorem when context features are modelled as random events. This mechanism of probability revision is at the basis of relevance feedback.

The VSM provides the constructs for implementing relevance feedback too, but it remains only an application.

An approach to taking the context into consideration is to introduce space, histories, profiles, sensors data, clocks and calendars into indexing or retrieval algorithms.

The context can be acquired by observing the user's behaviors through his interaction with the system. Therefore, what the user does during the navigation and the search activities has to be monitored.

3.3 The Vector Space Model

This model makes a strong reference to linear algebra. The document and the query are imagined as points in a space, the space dimensions correspond to the descriptors and the space has at least one dimension because at least a descriptor exists.

At the beginning of the process of choice of the descriptors, the point which represents the document or the query corresponds to the origin of the space. Every time the author or the user chooses a descriptor, the point moves itself along the axis matching the descriptor; the length of the move is given by a coefficient. As the user chooses his query descriptors, the vector approaches and tend to overlap some of the document vectors and goes away from others. The coefficient c_i weighs the measure and the sense (positive or negative) of the descriptor importance in describing the informative contents of the object.

According to the VSM, the descriptors are vectors of a linear space of finite dimension and their linear combination are documents, queries or any other object which contains information. Thus, all the objects are represented by vectors.

Let's formalize these concepts:

$T = (t_1,...,t_k)$ is a set of k values in the space $\Re^n$.

T generates the vector $\vec{x}$ when

$$\vec{x} = \sum_{i=1}^{k} c_i t_i$$

which can be written as

$$\vec{x} = T \cdot c$$

where T is the matrix $n \times k$ with vectors column $(\vec{t}_1,...,\vec{t}_k)$.

T is independent if and only if

$$\vec{x} = \sum_{i=1}^{k} c_i t_i = \sum_{i=1}^{k} b_i t_i \Rightarrow c_i = b_i,\ i = 1,\ \ldots,\ k$$

T is a basis for $\Re^n$ when it is independent and it generates every vector in $\Re^n$, thus $k = n$. The set T is often the set of the versors $\vec{e}_1,...,\vec{e}_n$. In this case T is a orthonormal basis because the vectors are mutually orthogonal; if T is orthogonal, it is also independent, but not vice versa.

According to the VSM, given a basis *T*, every vector $\vec{t} \in T$ represents a descriptor; in the textual case, a descriptor is a term, but in general *t* can be utilized for every medium. Indeed, a basis vector is a set of numbers that have no reference to the medium utilized to build the documents or the queries.

As above mentioned, every document is expressed as a vector $\vec{d}$:

$$\vec{d} = \sum_{i=1}^{n} c_i t_i \ .$$

Analogously, a query is a linear combination of the vectors in *T*. The subspace generated by *T*, or rather the set of the linear combinations of the descriptors, represents a collection of documents, an interrogations' set, or any type of informative objects. The value of the coefficient c_i represents the weight of t_i in describing the document. The coefficients are often calculated by the TFIDF scheme.

It shall be noticed that the set *T* coincides conceptually with the index, that is, *T* includes one and only one vector $\vec{t_i}$ for every descriptor t_i of the index.

The idea underlying a system based on the VSM is that a document is more relevant to the information need expressed by a query, as the vector is closer to the vector of the query in the space.

Formally, $\vec{d}$ is the vector referring to the document *d* and $\vec{q}$ is the vector referring to the query *q*. *T* is a basis, thus

$$\exists \ c_1,\ldots,c_n : \ \vec{d} = \sum_{i=1}^{n} c_i t_i \ \text{ and } \exists \ b_1,\ldots,b_n : \ \vec{q} = \sum_{i=1}^{n} b_i t_i$$

A measure that gives an idea of the nearly of the document to the query is their inner product:

$$d^T \cdot q = c^T \cdot T^T \cdot T \cdot b = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i b_j t_i \cdot t_j$$

The higher is the inner product, the higher is the degree of relevance.

It is worth noting that the model requires the independence of *T* and not necessarily the orthogonality, neither the orthonormality. The assumption of orthogonality is necessary to limit the amount of computational resources.

Furthermore, there is no significant scientific evidence about the superiority of the retrieval algorithm which incorporates a non-diagonal matrix $R = T^T \cdot T$, rather than an algorithm in which *R* is diagonal.

Since the inner product is the length of the projection of the vector which represents the document on the vector which represents the query, the vector of a long document will probably have larger inner products than that of a short document. To allow that also the short documents, if relevant, are presented to the user, the measure of relevance should independent of the length of the documents. So, if we estimate the document length with the norm of the referring vector, the measure can be substituted by the cosine of the angle θ between the two vectors:

$$\cos\theta = \frac{\vec{d}^T \cdot \vec{q}}{\|\vec{d}\|\|\vec{q}\|},$$

where $\|\vec{d}\|^2 = c^T \cdot c$ and $\|\vec{q}\|^2 = b^T \cdot b$

3.4 Modeling context

3.4.1 Methodology

The intuition underlying the methodology we are going to present is that an object vector is generated by a basis vector as an informative object is affected by contextual factors. Therefore, a basis generates a vector subspace; this subspace includes all the vectors generated by the basis and it can be considered as the representation of a context.

We can visually think of a context as a plane in a three-dimensional space and all the vectors lying in the plane represents objects placed in the same context. As one vector spans a ray passing through the origin, the ray is an equivalent representation of the object; as infinite planes include a ray, the fact that one object belongs to many contexts can simultaneously be represented.

The indicators which characterize users, time, places and anything else emerge from the interaction between user and system form the notion of context. In [31], Melucci provides some useful definitions:

- Object: refers to either an entity of the context, for example, user, task, topic, or document, or a relationship between entities, for example, relevance or aboutness.
- Dimension: refers to a property of an entity, for example, user behaviour, task difficulty, topic clarity, document genre, or relevance.
- Factor: refers to a value of a property, for example, browsing, complex search task, difficult topic, relevant, non–relevant, or mathematical document.
- Feature: refers to the variables observed for implementing vectors of which it is an element.

By their interaction behaviors, users can describe the contexts in which they are. Once some objects and dimensions of context are selected from the domain for which a context-aware IR system is designed, the methodology presented can be summarized as follows:

1. for each dimension of context a set of orthogonal vectors is defined; each orthogonal vector of such a set models one factor of the dimension of context;
2. a basis is built for representing a context by selecting one or more factors from each dimension – one factor refers to one dimension;
3. an informative object is matched against a context by computing a function of the distance between the vector and the subspace spanned by the basis; the closer the vector is to the subspace, the more the object is "in the context".

To represent the properties of contextual factors and dimensions, the properties of Linear Algebra can be exploited. In particular, it is often assumed that the vectors corresponding to a given dimension of context are mutually orthogonal for signifying that the values taken by the dimensions are mutually exclusive.

Many dimensions can generate an object, for instance a query or a document can be represented by the infinite sets of coordinates which can be defined in the vector space. To formalize, the following vector space properties can be used:

- a vector $\vec{x}$ is generated by the contextual factors $\{u_1, u_2\}$ as

  $\vec{x} = p_1^2 u_1 + p_2^2 u_2$

  where $u_1 \perp u_2, p_1^2 + p_2^2 = 1$ and $p_i^2 \geq 0$.
- At the same time,

  $\vec{x} = q_1^2 e_1 + q_2^2 e_2$

  where $e_1 \perp e_2, q_1^2 + q_2^2 = 1$ and $q_i^2 \geq 0$.

Let us now looking for an algebraic operator for a contextual factor. We consider a set of vector $\vec{B} = \{b_1,...,b_k\}$ where b_i represents a contextual factor or a dimension of context. A projector is an operator that maps a vector to another vector which belongs to a given subspace; one projector can be computed from each vector. We remember below the main properties of this operator:

- a projector is symmetric: $B_i^T = B_i$
- a projector is idempotent: $B_i^2 = B_i$

Let $L(\{b_i\})$ be the subspace of the vectors which are obtained by multiplying b_i by a scalar. The projectors onto the subspace $L(\{b_i\})$'s are defined as $b_i \cdot b_i^T$; if $L(\{b_i\})$ is the ray containing b_i, then the projection of $\vec{y}$ onto $L(\{b_i\})$ is $B_i \cdot y$.

If b_i, b_j refer to the same dimension, $B_i \cdot B_j = 0$ when $i \neq j$, this is the definition of projector orthogonality.

Generally, two projectors B$_i$ and B_j are

- oblique: $B_i \cdot B_j \neq 0$
- non commutative: $B_i \cdot B_j \neq B_j \cdot B_i$

There is a one-to-one correspondence between a subspace and its projector, so a projector can be taken as the algebraic operator for a contextual factor, and a linear combination of projectors refers to a mixture of contextual factors.

The operator used here is a linear function of projectors formulated by using a predefined set of coefficients which measure the weight of each dimension of context:

$$C_B = w_1 B_1 + \ldots + w_k B_k \qquad (4.1)$$

where the w_i 's are non negative coefficients such that $w_1 + \ldots + w_k = 1$ and the B_i's are the projectors onto the subspaces $L(\{b_i\})$'s.

Since C_B depicts the context described by B, it is called context matrix or context operator.

3.4.2 Ranking function

If the objects are described by the $\vec{x}$'s, and C_B is the context operator, the ranking function is $x^T C_B x$ (4.2)

The 4.2 represents the averaged distance between the vectors and the contextual factors.

Taking into consideration the equation 4.1, the function becomes

$$x^T C_B x = w_1 x^T \cdot B_1 \cdot x + \ldots + w_k x^T \cdot B_k \cdot x \qquad (4.3)$$

As $B_i = b_i \cdot b_i^T$,

$$x^T B_i x = x^T \cdot b_i \cdot b_i^T \cdot x = \left(b_i^T \cdot y\right)^T \cdot b_i^T \cdot y = \left(b_i^T \cdot y\right)^2 \qquad (4.4)$$

and therefore

$$x^T C_B x = w_1 \left(x^T \cdot b_1\right)^2 + \ldots + w_k \left(x^T \cdot b_k\right)^2 \qquad (4.5)$$

The last equation illustrates the degree to which the object represented by $\vec{x}$ is close to the contextual factors of B; this degree is a weighted average of the size of the projections of x to the $L(\{b_i\})$'s.

3.4.3 The Choice of the Basis Vectors

The basis that generates an informative object is generally unknown. Therefore, it would be useful to develop a theory which maps a description of a context to a basis vector. When a user types a query, he chooses the descriptors depending on his capabilities in expressing the content of the documents searched; furthermore, the user selects a descriptor on the basis of the relationship with the other descriptors of the query, and on his Anomalous State of Knowledge (ASK).

Essentially, the use of a descriptor depends on context: when using a descriptor, a user is giving it a meaning that is different from the meaning given by an other subject to the same descriptor, or by the same user in other place and moment. So, context influences the selection of the descriptors, their semantics and inter-relationships.

The descriptors are represented by basis vectors which can change as context does; thus every descriptors has as many vector representations as there are contexts.

If we consider the keywords, we'll obtain a vector basis for each of them. If we consider the documents, the automatic approach to build a basis vector could be similar to the methodology adopted by the VSM: the *i*-th element of a document vector is the weight of the index term *i* in the document.

Whereas the manual definition of vector basis for textual document appears quite simple, it is much less simple referring to non textual document, or to other context dimensions, such as time or space.

How could work an automatic theory for building basis vector from context is not completely clear and it is matter of future research.

However, such theory could leverage matrix manipulation algorithms which extract sets of orthogonal vectors from a set of non – orthogonal vectors.

In [30], Melucci suggests that, to reach an automatic approach, the idea is to

1. collect some vectors which describe objects about a dimension of context,
2. manage the collected vectors for compiling a symmetric matrix, and

3. compute orthogonal matrices whose columns can be used as the vectors which correspond to the potential value of the dimensions of context.

To extract the basis vectors from the symmetric matrices we can exploit Singular Value Decomposition (SVD).

For example, we suppose to have at our disposal a set of documents, everyone of which is described by a vector, that represent a feature. We process this vectors for assembling a symmetric matrix *S* and then we decompose it into its eigenvectors through SVD.

In the following, we provide a brief description of the Singular Value Decomposition.

In linear algebra, the SVD is an important factorization of a rectangular real or complex matrix.

Let us suppose *S* is a symmetric $n \times n$ matrix defined over the real field; an example of such a matrix is a matrix whose elements are term – term correlations from a document set.

Let *S* be an Hermitian matrix ($S^T = S$), then it exist a factorization of the form

$$S = V \cdot \Lambda^2 \cdot V^T$$

where Γ^2 is a diagonal matrix and *V* an orthonormal matrix of which some columns can thus be used as a basis for the context of the document set.

Furthermore, using the Spectral Decomposition theorem,

$$S = \lambda_k^2 v_1 \cdot v_1^T + \ldots + \lambda_k^2 v_k \cdot v_k^T$$

where the λ_i^2's are the eigenvalues of *S* as well as the elements of Λ^2, the v_i's are the column vectors of *V*, and the $v_i \cdot v_i^T$'s are the projectors to the subspace spanned by v_i's.

This expression means that the relationships between the indicators are function of the contextual factors thus revealing that the IRF algorithm can discover more information than encapsulated by an average feature vector. The correlation matrices are usually small because the behavioral indicator do not need to be numerous. Therefore, the computational cost of SVD is quite limited.

Some algorithm that follow the lines described above have been implemented in various experiments, illustrated in [29, 32, 33]

3.4.4 Interpretation

Once we have extract the eigenvectors, we have to interpret their meaning.
In [33] the symmetric matrix is the correlation matrix between the indicator observed from a set of documents seen by the user during his search. Then, by SVD, the authors compute the eigenvectors; their values are scalars between –1 and +1, and the interpretation given to them is that the further a value is from 0 the more important it is. Whit "important" the authors means that the feature to which the value corresponds is a significant descriptor of the contextual factor represented by the eigenvector.
The authors provide then a further interpretation to the eigenvectors: they state that the first eigenvector extracted through SVD explains the largest fraction of the variance of the points around their mean vector. It is therefore an average vector interpolating a set of the points corresponding to the seen documents. The fraction of variance explained by the first eigenvector is the ratio between the first eigenvalue and the sum of all the eigenvalues.

3.5 Modeling TEL Context

3.5.1 Methodology

The methodology we have described above can be fitted to the TEL case using the procedure illustrated below.

1. The indicator of the last five records seen by the user when performing a search are observed. We believe that the last five records are sufficient

for infer the user behavior. These records are used for computing a representation of context.

2. The observed indicator of the records are used for computing the contextual factors as follows:
    a. the co-occurrence matrix is computed;
    b. the eigenvectors are extracted from the co-occurrence matrix.
3. The whole set of the records is ranked by the ranking function. Then, for each projector:
    a. The ten most frequent keywords of the five top-ranked records are used for expanding the original query.
    b. The expanded query retrieves a list of records.

3.5.2 Indicators

In Chapter 2 we have derived from the log file a set of indicators that can describe the context in which the users are.

To apply the methodology illustrated above, it is necessary that the indicator to insert in the context matrix are variables quantitative continuous or in alternative qualitative dichotomous [15].

Therefore, the indicator candidate to describe the context are:

- user identifier dichotomized,
- language dichotomized,
- word dichotomized,
- search dichotomized,
- display time dichotomized,
- print dichotomized,
- save dichotomized,
- service dichotomized.

To choose the indicator to insert in the context matrix we have to take in account that they refer to the single records, thus they have to distinguish each record to the other. Therefore there are some indicators from the list above that we can't utilize. These are:

- user identifier dichotomized,
- language dichotomized,
- word dichotomized,
- search dichotomized.

Indeed, when we consider a single search, they take the same value in relation to each of the record.

For instance, let us consider the search activity illustrated in Tables 3.1, 3.2 and 3.3: note that the columns "userid_dic", "lang_dic" and "search_dic" contain only zero, while "word_dic" include only "1". This is obvious if we remember the meaning of these indicators:

- usually the user doesn't change his state from "guest" to "registered" or vice versa, during a single search activity,
- the same discourse is valid for the language of the portal interface,
- the number of word of the query is necessarily the same for all the displaying of the result records,
- at the same way, the type of search is certainly the same for all the displaying of the result records.

In the follow we list the indicators we use.

1. Display time dichotomized ("display_dic") tells the length of the display time related to each single record. Its value are:
   - 0 whether the displaying time is not greater than 20 seconds,
   - 1 whether the displaying time is greater than 20 seconds.
2. Print dichotomized "print_dic" indicates if the result record has been printed. The values this indicator can assume are:
   - 1 whether the user prints a result record,
   - 0 whether the user doesn't print the result record.
3. Save dichotomized "save_dic" indicates whether the record has been saved; in particular a user can save a record by three different ways:

1. saving in reference session's favourites,
2. sending by e – mail,
3. saving for reference manager use.

The values of this indicator are:

- 1 whether the user saves a result record,
- 0 whether the user doesn't save the result record.

4. Service dichotomized "service_dic":

this indicator refer to the possibility for the user, when he looks at a result record, of utilizing a set of services; in particular, he can

- clicking the link "Available at Library" on the result page to view associated record in native interface,
- clicking the link "See online" on the result page to view associated object in native interface,
- utilizing the full record service link to a country for the currently viewed result record,
- utilizing the full record service link to other web services such as Google, Amazon, etc..

The values this indicator can assume are:

- 1 whether the user clicks on at least one of these links,
- 0 whether the user doesn't click on any of these links.

3.5.3 Example

In Tables 3.1, 3.2 and 3.3 we report an excerpt of the log file which includes the actions performed by a user during a search activity; in particular, we consider all the actions executed starting from a search began from a result record page, to the following search, that is another search began from a result record page.

id	userip	sesid	lang	query
1349049	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349051	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349052	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349053	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349054	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349055	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349056	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349058	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349059	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349060	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349062	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349063	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349065	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349067	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349073	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349075	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349084	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349087	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349089	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349094	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349106	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349107	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349112	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349114	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")
1349123	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	publisher all "ricordi, tito"
1349124	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	publisher all "ricordi, tito"
1349126	217.187.253.63	0bp72u8e6n7sv5e3urdai0uvq1	en	("rossini eduardo")

Table 3. 1

action	colid	nrRecords	recordPosition	objurl	date
search_res		0	-		2008-01-23 22:17:52
view_full		3			2008-01-23 22:18:21
view_brief	a0010	0	-		2008-01-23 22:18:21
view_brief	a0200	0	-		2008-01-23 22:18:21
view_brief	a0010	0	-		2008-01-23 22:18:22
view_brief	a0132	0	-		2008-01-23 22:18:28
view_brief	a0132	0	-		2008-01-23 22:18:28
view_brief	a0035	0	-		2008-01-23 22:18:31
view_brief	a0067	0	-		2008-01-23 22:18:35
view_brief	a0200	0	-		2008-01-23 22:18:36
view_brief	a0067	0	-		2008-01-23 22:18:42
view_brief	a0001	1	-		2008-01-23 22:18:46
view_brief	a0086	87	-		2008-01-23 22:18:58
view_full	a0086	87	1	http://www.theeuropean…	2008-01-23 22:19:17
view_brief	a0086	87	-		2008-01-23 22:19:58
view_full	a0086	87	6	http://www.theeuropean…	2008-01-23 22:20:34
option_send_email	a0086	87	6	http://www.theeuropean…	2008-01-23 22:22:26
option_print	a0086	87	6	http://www.theeuropean…	2008-01-23 22:23:01
view_brief	a0086	87	-		2008-01-23 22:23:08
view_full	a0086	87	20	http://www.theeuropean…	2008-01-23 22:24:02
option_print	a0086	87	20	http://www.theeuropean…	2008-01-23 22:24:43
view_brief	a0086	87	-		2008-01-23 22:24:46
view_brief	a0086	87	21-40		2008-01-23 22:24:55
view_full	a0086	87	21	http://www.theeuropean…	2008-01-23 22:25:09
view_brief	a0086	0	-		2008-01-23 22:26:06
view_brief	a0086	0	-		2008-01-23 22:26:15
search_res		0	-		2008-01-23 22:26:25

Table 3. 2

userid_dic	lang_dic	display	word	word_dic	display_dic	search_dic	print_dic	save_dic	service_dic
0	0		2	0		0			
0	0	0	2	0	0	0	0	0	0
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0	41	2	0	1	0	0	0	0
0	0		2	0		0			
0	0	112	2	0	1	0	1	1	0
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0	41	2	0	1	0	1	0	0
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			
0	0	57	2	0	1	0	0	0	0
0	0		2	0		0			
0	0		2	0		0			
0	0		2	0		0			

Table 3. 3

We provide now an example of application of the methodology illustrated in Section 3.5.1. It is based on the search activity described by Tables 3.1, 3.2 and 3.3.

1. The indicator of the last five records seen by the user can be represented by this matrix:

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

2. The following steps are performed:
   a. We compute the co – occurrence matrix

$$S = A^T \cdot A = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 4 & 2 & 1 & 0 \\ 2 & 2 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

   The interpretation we can give to this co – occurrence matrix is that, referring to this specific user,

   - the longest display time is performed four times (element [1,1]),
   - the longest display time together with the printing is performed twice (element [1,2]),
   - the longer displaying time together with the saving is performed once (element [1,3]),
   - the printing is executed twice (element [2,2]),
   - the printing together with the saving is executed once (element [2,3]),
   - the saving is executed once (element [3,3]),

- the making use of services is never performed (element [4,4]), for this reason all the element of the fourth row and of the fourth column are zero.

The matrix is symmetric thus the same interpretation can be given to the other elements.

b. We extract the eigenvectors by SVD

The eigenvalues are:

$\lambda_1 = 5.65, \lambda_2 = 1, \lambda_3 = 0.35, \lambda_4 = 0$

The eigenvectors are:

$$v_1 = \begin{pmatrix} -0.81 \\ -0.52 \\ -0.29 \\ 0 \end{pmatrix}, v_2 = \begin{pmatrix} 0.58 \\ -0.58 \\ -0.58 \\ 0 \end{pmatrix}, v_3 = \begin{pmatrix} 0.14 \\ -0.63 \\ 0.77 \\ 0 \end{pmatrix}, v_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

The first eigenvector explains the 80.71% of the total variance; it indicates that the most important feature is the longer display times, followed by the printing and the saving. The second eigenvector explains the 14.29% of the total variance, and it tells that the longer display time tends to be not performed when the user prints and saves. The third eigenvector describes the 5% of the total variance, and it tells that the longer display time and the saving tend not to be performed with the printing.

The fourth element of the eigenvectors is always equal to zero, meaning that the making use of services is not a significant descriptor. This is coherent with the co – occurrence matrix, in which we have seen that the making use of services is never executed. Really, in the fourth eigenvector, the fourth element is equal to "1", but the variance explained by this eigenvector is null, so we have not to take it into consideration.

CONCLUSIONS

In the following, some final considerations about the analysis conducted in this thesis are reported.

The methodology proposed in this thesis for the TEL portal action log file envisages to observe the indicators of the last five records seen by each user after having performed a search action. This implies that each user performs at least five actions “view_full” per every action of search. There are 161,122 rows of the dataset related to search action and the total of the actions “view_full” executed in the dataset is equal to 196,428. This means that there are, on average, 1.22 actions “view_full” for each search. Obviously, this does not mean that after every search action there is only one result record; in the dataset there are different situations: sometimes there are some ten of “view_full” in correspondence of an action of search, but often the user displays less than 5 single result records per search. The presentation of additional records can be favoured by an interface that, for instance, attracts the user to see the following single record.

In our review of the literature about IRF techniques, different works whose findings demonstrate the importance of knowing the task of each user were presented. In particular, Kelly and White found that using information about the search task appears to enhance retrieval performance [25]. Furthermore, Kelly found that there are different display time thresholds depending on the task [22]. We believe that search engines should make users’ task explicit, for instance by providing an interface that requires to the user to choose among alternative tasks before to begin the search activity.

Another consideration is that the implicit indicators are not all equally significant in indicating relevance. In our dataset, the following implicit interest indicators, all in relation to a single result record were used: printing, saving in a list of favourites, sending an e-mail, saving for reference manager use, and display time. We think, as an example, that printing is a stronger interest indicator than saving: in fact, the print implicates an expense, even though small, by the user; in addition, a printed page can be taken and observed everywhere, whereas a record saved can be utilized only with the support of a computer. Hence, we believe that one that print a document is certainly interested to it, while saving a document can indicate the intention of looking at it later, but not necessarily because it is relevant. Thus we suggest, as an issue reserved for the future works, that an interesting extension to the analysis reported in this thesis is an algorithm that provides a weighting of the indicators: this algorithm should assign an higher coefficient to the most significant indicators, and a lower coefficient to the less significant indicators.

REFERENCES

[1] E. Agichtein, E. Brill and S. Dumais. "Improving web search ranking by incorporating user behavior information". *Proceedings of SIGIR*, pp. 19 – 26, New York, NY, USA, 2006. ACM Press.

[2] M. Agosti Ed. *Information Access through Search Engines and Digital Libraries*. Berlin, Germany, 2008. Springer – Verlag.

[3] M. Agosti. "Log data in digital libraries." In print in *Post – Proceedings of the Fourth Italian Research Conference on Digital Library Systems (IRCDL 2008),* 2008.

[4] M. Agosti and G. M. Di Nunzio. "Web log mining: A study of user sessions," *Proceedings of the 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL 2007)*, pp. 70 – 74, Corfu, Greece, 2007.

[5] M. Agosti, N. Ferro, E. A. Fox and M. A. Gonçalves. "Modelling DL quality – a comparison between approaches: The DELOS reference model and the 5S model," *2nd DELOS Conference on Digital Libraries Working Notes*, C. Thanos and F. Borri, Eds. ISTI – CNR, Gruppo ALI, Pisa, Italy, 2007.

[6] M. Agosti and M. Melucci. *Reperimento dell'Informazione,* Libreria Progetto, 2008.

[7] B. Baldacci and R. Sprugnoli. *Informatica e Biblioteche: Automazione dei Sistemi Informativi Bibliotecari.* Roma, NIS, 1983.

[8] N. J. Belkin and J. Callan. "Context-based information access". *Report of the Discussion Group on Context-Based Information Access of the Workshop on "Information Retrieval and Database: Synergies and Syntheses"*, Washington, Columbia, USA, 2003. National Science Foundation. Available at http://www2.cs.washington.edu/nsf2003/discussionGroups.html

[9] J. Budzik, and K. Hammond. "Watson: Anticipating and contextualizing information needs". *Proceedings of the 62nd Meeting of the American Society for Information Science*, pp. 727 – 740, Washington, Columbia, USA, 1999.

[10] A. Camussi, F. Moller, E. Ottaviano and M. Sari Gorla. *Metodi Statistici per la Sperimentazione Biologica* Ed. Zanichelli, 1995.

[11] M. Claypool, P. Le, M. Waseda and D. Brown. "Implicit interest indicators". *Proceedings of the 6th International Conference on Intelligent User Interfaces (IUI '01)*, pp. 33 – 40, Santa Fe, New Mexico, USA, 2001.

[12] M. Dillon and J. Desper. "The use of automatic relevance feedback in boolean retrieval systems". *Journal of Documentation*, Vol. 36, No. 3, pp. 197 – 208, 1980.

[13] M. Dillon, J. Ulmschneider and J. Desper. "A prevalence formula for automatic relevance feedback in boolean systems". Information Processing & Management, Vol. 19, No. 1, pp. 27 – 36, 1983.

[14] E. N. Efthimiadis. "Query expansion". M. E. Williams (Ed.), *Annual Review of Information Science and Technology*, Vol. 31, pp. 121 – 187, 1996. Available at http://faculty.washington.edu/efthimis/pubs/Pubs/qe-arist/QE-arist.html visited the 24th July 2008

[15] L. Fabbris. "Lucidi per il corso di Statistica Sociale. Anno accademico 2007-2008"

[16] G. Golovchinsky, M. N. Price and B. N. Schilit. "From reading to retrieval: Freeform ink annotations as queries". *Proceedings of the 24th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '99)*, pp. 19 – 25, Berkeley, California, USA, 1999.

[17] E. Gori and G. Vittadini. "La valutazione dell'efficienza ed efficacia dei servizi alla persona. Impostazione e metodi". *Qualità e Valutazione nei Servizi di Pubblica Utilità,* a cura di E. Gori e G. Vittadini, pp. 121 – 241. ed. ETAS Libri, 1999.

[18] S. Gradmann. "Interoperability of digital libraries: Report on the work of the EC working group on DL interoperability," *Seminar on Disclosure and Preservation: Fostering European Culture in The Digital Landscape.* National Library of Portugal, Directorate-General of the Portuguese Archives, Lisbon, Portugal, 2007. Available at http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf visited the 15th September 2008

[19] M. Guerrini and L. Sardo. *Authority Control.* Roma, Associazione italiana biblioteche, 2003.

[20] P. M. Hallam-Baker and B. Behlendorf. "Extended log file format – W3C working Draft WD-logfile-960323." March 1996. Available at http://www.w3.org/TR/WD-logfile.html visited the 15th September 2008

[21] M. Kellar, C. Watters, J. Duffy and M. Shepherd. "Effect of task on time spent reading as an implicit measure of interest". *Proceedings of American Society for Information Science and Technology (ASIS&T),* pp. 168 – 175, Providence, Rhode Island, 2004.

[22] D. Kelly. *Understanding Implicit Feedback and Document Preference: A Naturalistic User Study*. PhD thesis, Rutgers University, The State of New Jersey, 2004.

[23] D. Kelly and N. J. Belkin. "Reading time, scrolling and interaction: Exploring implicit sources of user preferences for relevance feedback". *Proceedings of ACM Conference on Research and Development in Information Retreval (SIGIR '01)*, pp. 408 – 409, New Orleans, Louisiana, USA, 2001.

[24] D. Kelly and J. Teevan. "Implicit feedback for inferring user preference. A bibliography". *Proceedings of ACM Conference on Research and Development in Information Retreval (SIGIR '03)*, Vol. 37, No. 2, pp. 18 – 28, New York, NY, USA, 2003.

[25] D. Kelly and R. W. White. "A study in the effects on personalization and task information on implicit feedback performance". *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pp. 297 – 306, New York, NY, USA, 2006. ACM Press.

[26] J. Kim, D. W. Oard and K. Romanik. "User modeling for information access based on implicit feedback". *Proceedings of International Society for Knowledge Organization (ISKO)*, Nanterre, France, 2001.

[27] J. Konstan, B. Miller, M. Maltz, J. Herlocker, L. Gordon and J. Riedl. "GroupLens: Applying collaborative filtering to usenet news". *Communications of the ACM*, Vol. 40, No. 3, pp. 77 – 87, 1997.

[28] J. Luxemburger, E. van der Meulen and G. Weikum. "A user-interaction model for the european library portal." *10th DELOS Thematic Workshop on "Personalized Access, Profile Management, and Context Awareness in Digital Libraries"*, Corfu, Greece, 2007.

[29] M. Melucci. "A basis for information retrieval in context". *ACM Transactions on Information Systems (TOIS)*, Vol. 26, Issue n. 3, Article No. 14 New York, NY, USA, 2008.

[30] M. Melucci. "Modeling retrieval and navigation in context". *Information Access through Search Engines and Digital Libraries*, M. Agosti Ed., pp. 43 – 57, Berlin, Germany, 2008. Springer – Verlag.

[31] M. Melucci. "Vectors, planes and context" 10th DELOS Thematic Workshop *on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL 2007)*, Corfu, Greece, 2007. Available at www.dblab.ece.ntua.gr/persdl2007/papers/51.pdf visited the 15th September 2008.

[32] M. Melucci and R. W. White. "Discovering hidden contextual factors for implicit feedback". *Proceedings of the Second Workshop on Context-based Information Retrieval*, CEUR, Vol. 326, No. 114, pp. 69 – 80, Roskilde, Denmark, 2007.

[33] M. Melucci and R. W. White. "Utilizing a geometry of context for enhanced implicit feedback". *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*, pp. 273 – 282, Lisbon, Portugal, 2007. ACM Press.

[34] N. N. Mitev, G. Venner and S. Walker. "Designing an online public access catalogue: Okapi, a catalogue on a local area network," British Library, *Library and Information Research Report 39*, 1985.

[35] M. Morita and Y. Shinoda. "Information filtering based on user behavior analysis and best match text retrieval". *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '94)*, pp. 272 – 281, Dublin, Ireland, 1994.

[36] R. Rafter and B. Smyth. "Passive profiling from server logs in an online recruitment environment". *Proceedings of the IJCAI Workshop on Intelligent Techniques for Web Personalization (ITWP 2001)*, pp. 35 – 41, Seattle, Washington, USA, 2001.

[37] S. E. Robertson, J. D. Bovey, C. L. Thompson and M. J. Macaskill. "Weighting, ranking and relevance feedback in a front – end system". *Journal of Information Science*, 12, pp. 71 – 75, 1986.

[38] S. E. Robertson and K. Sparck-Jones. "Relevance weighting of search terms". *Journal of the American Society for Information Science*, Vol. 27, No. 3; pp. 129 – 146, 1976.

[39] G. Salton, E. A. Fox and E. Voorhees. "Advanced feedback methods in information retrieval". *Journal of the American Society for Information Science*, Vol. 36, No. 3, pp. 200 – 210, 1985.

[40] C. Silverstein, H. Marais, M. Henziger and M. Moricz. "Analysis of a very large web search engine query log". *Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '99)*, Vol. 33, No.1, pp. 6 – 12, Berkeley, California, USA, 1999.

[41] A. F. Smeaton and C. J. Van Rijsbergen. "The nearest neighbour problem in information retrieval. An algorithm using upper bounds". *Proceedings of the 4th International Conference on Information Storage and Retrieval: theoretical issues in information retrieval*, pp. 83 – 87, Oakland, California, 1981.

[42] A. F. Smeaton and C. J. Van Rijsbergen. "The retrieval effects of query expansion on a feedback document retrieval system". *Computer Journal*, Vol. 26, No. 3, pp. 239 – 246, 1983.

[43] A. Spink and J. L. Xu. "Selected results from a large study of web searching: The Excite study". Information Research, Vol. 6, No. 1, 2000. Available at http://InformationR.net/ir/6-1/paper90.html visited the 15th September 2008.

[44] B. Ussery. "Google – average number of words per query have increased!". Available at http://www.beussery.com/blog/index.php/2008/02/google-average-number-of-words-per-query-have-increased visited the 15th September 2008.

[45] C. J. Van Rijsbergen, D. J. Harper and M. F. Porter. "The selection of good search items". *Information Processing and Management*, UK, Vol. 17, No. 2, pp. 77 – 91, 1981.

[46] T. van Veen and B. Oldroyd. "Search and retrieval in the european library. A new approach." *D – Lib Magazine*, Vol. 10, No. 2, 2004.

[47] C. Vogt. "Passive feedback collection – an attempt to debunk the myth of clickthroughs". *Proceedings of TREC*, pp. 141 – 150, Gaithersburg, Maryland, 2000.

[48] S. Walker and R. De Vere. *Improving Subject Retrieval in Online Catalogues: 2. Relevance Feedback and Query Expansion*. British Library, *Research Paper 72*, London, UK, 1990.

[49] R. W. White, I. Ruthven and J. M. Jose. "The use of implicit evidence for relevance feedback in web retrieval". *Proceedings of 24th BCS – IRSG European Colloquium on IR Research,* Lecture notes in Computer Science 2291, pp. 93–109, Glasgow, UK, 2002.

[50] H. Wu and G. Salton. "The estimation of term relevance weights using relevance feedback". *Journal of Documentation*; Vol. 37, No. 4, pp. 194 – 214, 1981.