UNIVERSITÀ DEGLI STUDI DI PADOVA

DEPARTMENT OF INFORMATION ENGINEERING

Master Course in COMPUTER ENGINEERING

Master Course Thesis

An Adaptive Cross-Site User Modelling Platform for Information Exchange Techniques

Graduate:

Stefano MARCHESIN

Supervisor:

prof. Maristella AGOSTI
UNIVERSITÀ DEGLI STUDI DI PADOVA

Co-supervisors:

prof. Vincent WADE

prof. Séamus LAWLESS
TRINITY COLLEGE DUBLIN

12 December 2016

Academic Year 2015-2016

Abstract

This thesis discusses an adaptive Cross Site User Modelling platform for information exchange techniques. The objective is to identify and evaluate different mechanisms of information exchange that can be subsequently used by websites to provide tailored personalisation to users that request for it. This is done by implementing a third party User Model Provider that, through the use of an API, interfaces with custom-built module extensions of websites based on the Web-based Content Management System (WCMS) Drupal. The approach is non-intrusive, not hindering the browsing experience of the user, and has a limited impact on the core aspects of the websites that implement it. This is achieved by implicitly tracking user activities on the websites and by allowing the user to decide when to trigger the information exchange mechanisms. The design of the API ensures user's privacy by not disclosing personal browsing information to non authenticated users. The user can enable/disable the Cross Site service at any time. The evaluation of the information exchange paradigms was conducted on a cultural heritage domain as if it were an open domain use-case. The thesis introduces the architecture and a prototype implementation of the service, providing some encouraging initial results.

Contents

1 Introduction and Motivation

1.1 Motivation

The exponential increase in Web usage[1] has allowed modern societies to access, create, manage and distribute massive amounts of information. This fact has led to an increase in the rate at which the information is created and consequently uploaded to the Web. The phenomenon is known as 'Information Explosion' and it results in an increased difficulty in organising digital information to meet users' information needs. Therefore, to tackle this impending issue systems have to implement novel methods and techniques.

A variety of systems have been created to try to address the problem by assisting users' information needs. These systems include, among others: keyword-based search engines, recommender systems, and Web Personalisation techniques that adapt different aspects of the web experience to the needs and preferences of the individual user. The latter have gained great popularity in recent years, in fact, techniques like

[1] According to `http://www.internetlivestats.com/internet-users/`.

personalised information retrieval, where the search result list is re-ranked based on the user's individual search history, or social graphs[2] have been widely adopted.

Within this context, however, most current approaches are unable to assist users in more complex conditions than just providing results for simple information needs, such as providing recommendations for a retail website. For example, user's information needs that span different subject domains from different independently hosted websites represent a challenge which these 'traditional' techniques cannot answer. In addition, the Cross Site browsing process carries with it two phenomena that are known as 'Lost in Hyperspace' and 'Information Overload', which can both negatively affect the fulfilment of user's information needs [ABH$^+$08], [Ber97], [EM00], [HT85], [SVV99]. Therefore, to support the user and prevent them being affected by these phenomena, a browsing experience is required that exploits all the content together of the different websites the user is browsing/has browsed to tailor the content to be provided. In this way, although the traditional way of browsing is left unchanged, that is when a user browses individual websites across the Web moving from one to another, the conceptual vision of the browsing experience can be redefined as a unified and seamless browsing experience, not simply within, but also across different websites. Even websites, and consequently web publishers, would benefit from this new way of seeing the browsing

[2] Social graphs socially connect users with content and products that 'peers' like. Social media applications have a major impact on how individual users find and access information with several applications using the social interactions of users to filter information [DDH$^+$00].

experience, getting more insight on each individual user landed on them and therefore being able to provide better content to users, thus prolonging their stay in the website.

Currently, most websites do not offer any Cross Site communication, leading to a 'silo-ed' browsing experience, from one 'silo-ed' website to another, that negatively impacts on what should be a unified and seamless experience. For example, a user who wants to gather information about a specific topic that spans different independently hosted websites will experience a repetition of their actions, such as keyword based search queries and navigational patterns throughout the different websites, due to the lack of communication between websites and their personalisation techniques.

Thus, it becomes clear that there is a need for a consistent Cross Site support mechanism that ensures effective assistance to users in the websites they browse across, by means of personalisation techniques such as link highlighting or content recommendation. A concrete representation of such a Cross Site support mechanism, which has been designed and implemented in this thesis work, is a third party Cross Site User Modelling service. According to Fischer and Gerhard [Fis01], a User Model is the collection and categorisation of personal data associated with a specific user, which sets the basis for any adaptive changes to the system's behaviour. Which data is included in the model depends on the purpose of the application; it can include personal information such as users' personal identity data, their interests, goals, plans, preferences and their interactions with the system. Therefore, a third party User Model Provider, held and maintained by a single website, is able to take specific aspects of each user together from different websites in order to provide Cross Site Personalisation to target websites that require for it. However, such a support mechanism has to ensure that the user is able to freely browse without any limitations or control, providing them with non-intrusive guidance which provides assistance, but does not limit the user's browsing experience. Concerning this matter, two interconnected web domains can be identified. The former, 'closed domain', is introduced as a browsing space that is limited to a specific number of websites on strict editorial guidelines. The latter, 'open domain', introduces a browsing space not limited to a specific number of websites or a specific user group. Moreover, the content is heterogeneous and mostly not limited to editorial guidelines or structures.

To offer a more seamless experience across different websites it is also necessary to investigate limited-impact techniques allowing Web Personalisation to be introduced to websites more flexibly and in later stages than design-phase.

Therefore, providing Web Personalisation techniques that assist users across the Web and not only on single websites appears to be a major objective. To address this problem, one of the key aspects on which to focus is the effective exchange of information between the third party User Model Provider and the current website the user is browsing. For the third party User Model Provider, the exchange should provide novel user's information to enhance the User Model. For the target website, the exchange should provide tailored Cross Site user information that can be used by personalisation techniques that assist users during their stay in the website. The service should be simple and cost effective to integrate into existing websites and should limit the flow of user's information from the third party User Model Provider to target websites, thus honouring users' privacy needs.

1.2 Research Question and Objectives

1.2.1 Research Question

How can an exchange of information be provided between a target website and the third party User Model Provider that is satisfying both for the target website and the third party service, which also limits the flow of user's information from the third party User Model Provider to the website thus honouring users' privacy needs?

By "satisfying" it is meant "to what extent the information exchange can provide to the target website relevant Cross Site user information that can be potentially used to address user Cross Site information needs and, on the other hand, enhance the User Model held by the service".

The primary objective of the thesis is to identify the most effective information exchange paradigms for a Cross Site User Modelling service.

1.2.2 Research Objective, Goals and Challenges

To address the above research question, the objective of this thesis is to identify and evaluate the most effective information exchange paradigms for a third party Cross Site User Modelling service, which provides information that supports websites in assisting users to address information needs that span independently hosted websites.

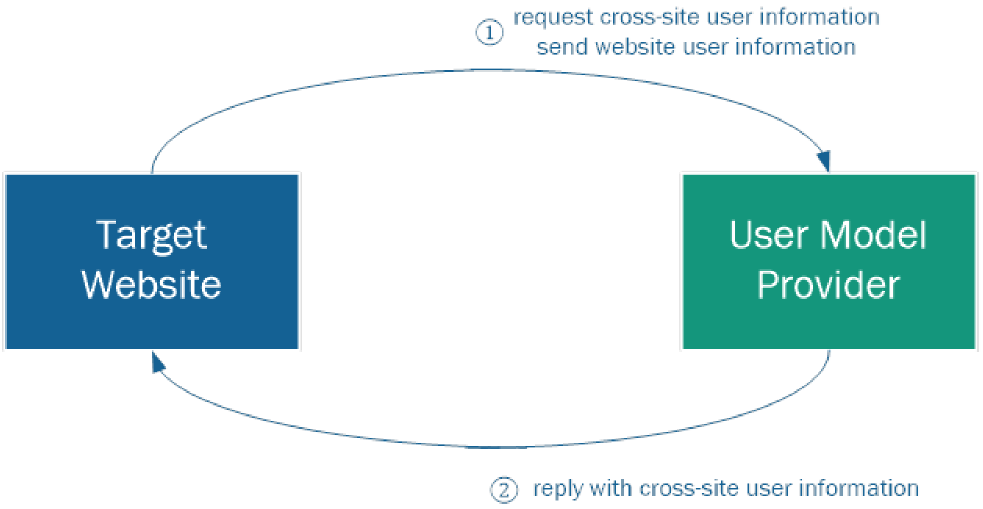

① request cross-site user information
send website user information
Target Website
User Model Provider
② reply with cross-site user information

Figure 1.1. *Information exchange prototype*

Therefore, three main goals can be identified:

1. To build an API that provides to the target website the ability to access and question the User Model offered by the User Model Provider;

2. To identify users' activities on websites and the information needs that drive those users to browse the website;

3. To decide which information about the user the third party User Modelling service should provide to the target website and whether to offer this information in its entirety or in smaller chunks, depending on the context of the application that utilises the Cross Site service. By context it is meant the privacy policy the organisation, which holds the website, has on the treatment of users' data. Depending to this, various information exchange methods can be deployed that provide different amounts of users' sensible information to target websites.

The approaches designed to address the research goals, which will be analysed in depth in section 3.3, can be grouped into two main classes: privacy-insensitive techniques and privacy-aware techniques. These classes, defined by different levels of user's privacy, can be summarised as:

- Privacy-Insensitive Techniques: the techniques belonging to this class are characterised by a massive flow of user's information out from the service, thus ignoring or not sufficiently considering the user's privacy concerns. Such techniques tend to reward web publishers[3] over web users, tipping the balance towards the organisation (or the person) that controls the website and allowing the website to gather massive amount of information about users.

- Privacy-Aware Techniques: the techniques belonging to this class are characterised by a greater focus on preserving user's privacy, thus avoiding the provision of huge amounts of information to websites or, in extreme cases, not providing it at all. Such techniques tend to reward web users over web publishers, tipping the balance towards users that hence experience website's personalisation with their privacy better preserved. Furthermore, privacy-aware techniques tend to be more advantageous even from the service perspective, which does not give away the only real value it holds – user data.

The resulting design and research challenges are:

- Interaction

[3] By web publisher it is meant a person or organisation that controls websites and therefore provides means to enable assistance.

  - How does a website request user information from the third party User Model Provider?
  - How does the Cross Site User Modelling service get value? What information does each website pass back to enhance the User Model?
  - How is privacy implemented? At which level?

- Representation
  - What vocabulary is used to describe the characteristics in the User Model?
  - What type of ontology can be used? Generic or Domain Specific[4]?
  - What is the nature of the user information?

Moreover, particular attention has to be paid to how the service affects both accuracy and relevance. Assessments of performance and efficiency should be focused either on number of interactions between the target website and the third party User Model Provider (qualitative measure) or on payload round-trip-time (quantitative measure).

1.3 Methodology

This research focuses on identifying and evaluating the most effective information exchange techniques between target websites and a third-party User Modelling service, which provides information that supports websites

[4] The type of ontology utilised may be dictated by the use-case defined.

in assisting users to address information needs that span independently hosted websites.

This is achieved through the design and implementation of an adaptive Cross Site User Modelling platform that allows to test and evaluate the aforementioned information exchange techniques. For confirmation, comparative methods from the Information Retrieval domain are used.

The conducted research can be categorised as applied research based on the investigation and development of both a Cross Site User Modelling platform and a set of information exchange techniques that are designed with the goal of providing relevant user information to target websites, while preserving user's privacy and enhancing the Cross Site User Model held by the third party User Model Provider.

Qualitative considerations and quantitative research techniques were applied to address the research question of the thesis. Qualitative considerations were applied to understand the underlying motivations that drive users to browse target websites (i.e. their information needs) and to detect, through an indicative evaluation based on simulative case studies, how good the information exchange techniques perform in relation to the topics of information overload and Web Personalisation. Quantitative research techniques were applied to study the effect the introduced information exchange techniques have on the satisfaction of both the website (in terms of effectiveness) and the overall system (in terms of efficiency). The underlying approach for these quantitative studies is experimental, by controlling the results through Information Retrieval formulas and payload round-trip-times in pre-set browsing environments based on real-world use-cases and real-world content.

1.4 Thesis Overview

The thesis starts with the State of the Art review (chapter 2) discussing different research and technological approaches related to key personalisation techniques, User Modelling approaches and Cross Site Personalisation. The State of the Art review is focused on identifying and discussing challenges in the area of Web Personalisation, trying also to define the gap between what has been done by systems so far and what the novel designed system proposes instead.

The State of the Art discussion is followed by the design and implementation of the system prototype and the information exchange techniques (chapter 3). First the requirements are set, then the use-case is defined and finally the design and the implementation of the key technical elements are discussed. Chapter 4 of the thesis evaluates the prototype in three user-focused case study evaluations. The final chapter (chapter 5) discusses conclusions and future work related to Cross Site Personalisation.

2 State of the Art - Related Work

2.1 Introduction

The chapter analyses the State of the Art in personalisation techniques and User Modelling approaches. Both subjects are then analysed in the context of Cross Site Personalisation, through the study of a Cross Site Personalisation system that has the objective of assisting users in addressing Cross Site information needs.

Web Personalisation can be defined as any action that tailors the web experience to a particular user, or set of users, by providing the information users want or need without expecting them to ask for it explicitly, as described in [Koi13], [KL05], [MCS00], [MAB00]. Therefore, it can be easily guessed that methods and techniques that provide Web Personalisation are important not only for research works, but also for commercial applications [Lie], [MCS00]. The reason why Web Personalisation has gained so much importance in recent years is related to the huge amount of information the user has to face every time they browse the Web (this event is often referred as 'information explosion [RR86], [Doy01]. Phenomena as 'Information

Overload' and 'Lost in Hyperspace', which have been already mentioned in section 1.1, need to be addressed in order to allow the user to fulfil their information needs. The former referring to the user being overwhelmed by the amount of information choices, while the latter referring to orientation difficulties of the user within and across different websites. Hence the need for Web Personalisation, which assists users in managing the access to vast amounts of rapidly growing information spaces [BC92], [NDRV09].

It can be stated that, in order to better assist users across the Web, a unified and seamless browsing experience, not simply within, but also across different websites is required. Hence, Web Personalisation techniques and methods are developed with the objective in mind to assist users in addressing information needs that span across independently hosted websites, thus trying to reproduce such a seamless browsing experience.

Therefore, it is clear that a key challenge lies in providing Web Personalisation techniques that assist user across the Web and not only on isolated websites. To address this problem, Web Personalisation methods have to balance both the needs of website users and website publishers. For website users, the introduced Web Personalisation technique should provide assistance across different websites, but in doing so honour the user's privacy needs and browsing freedom. For the website publisher, the introduced Web Personalisation technique should ensure simple and cost effective integration of Web Personalisation to existing websites and honour the website owner's control over the website. To make things clear on the website owner's control over the website, a web publisher might want to ensure specific content is promoted based on commercial goals, even if the content does not reflect the current interest of the website user.

Figure 2.1 summarises the concept of Cross Site Personalisation, representing the intersection space that Cross Site Personalisation techniques require to influence both the needs of website users and web publishers, trying to find a balance that allows for non-intrusive assistance for users (not limiting the users browsing paradigm) [BZ09], [Kay06] and limited-impact integration (ensuring simple and cost-effective integration) for websites [KCW09].

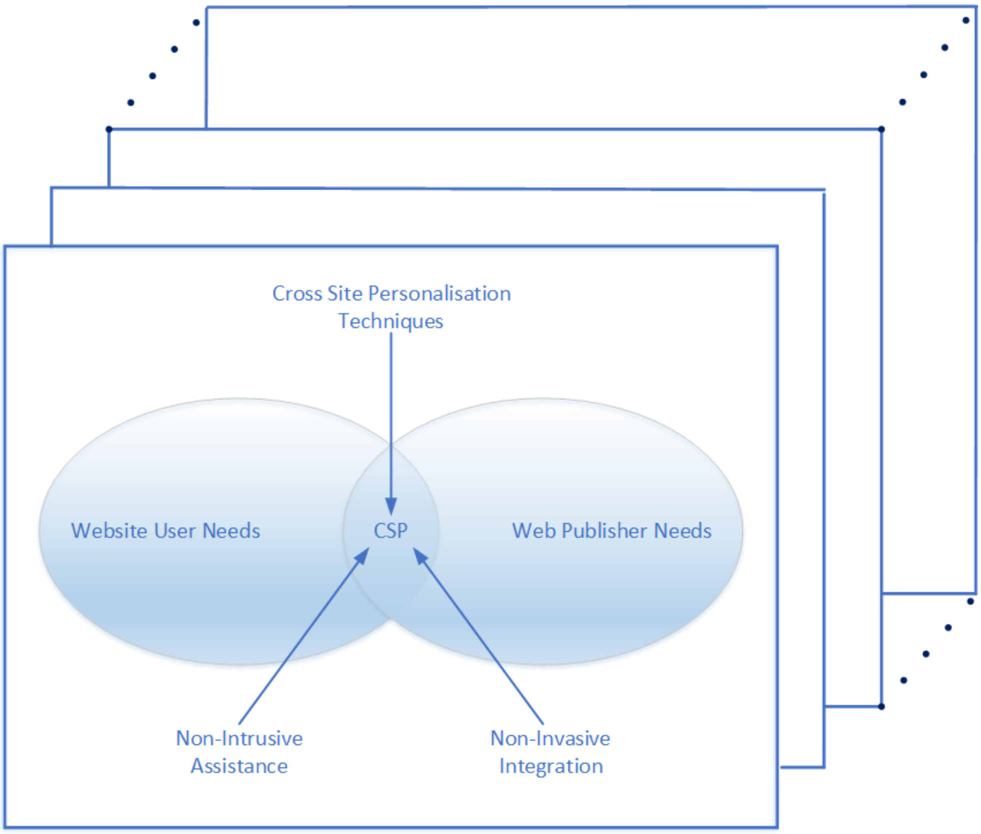


Figure 2.1. *Cross Site Personalisation concept*

Within this framework, it is important to consider both sides in order to obtain a service that is able to provide effective Web Personalisation to user, while resulting not invasive from a web publisher point of view. Hence, effective non-intrusive and limited-impact personalisation techniques need to be sought. Among the others, Information Retrieval [KL05] and Recommender Systems [LSY03] techniques stand out. Both the techniques have proven to be successful in personalisation environments, with limited impact on the users free browsing paradigm. However, the main shortcoming of non-intrusive approaches is the lack of focus on specific user needs and preferences. In fact, the majority of the methods apply a more generalised data driven approach resulting in a 'User that liked this item also liked these items' personalisation scenario. With regard to limited-impact integration of personalisation techniques in websites, two approaches can be identified:

1. Design-Time Integration: Limited-impact integration of Web Personalisation techniques at design-time.

2. Run-Time Integration: Limited-impact integration of Web Personalisation techniques at run-time.

1. Design-Time Integration: Limited-impact integration at design-time is related to the field of Web Engineering [GM01], which addresses the integration of techniques and methods to existing websites. The advantage of Web Engineering is in relying on strict design methodologies. For instance, OOHDM (Object Oriented Hypermedia Design Method) has the following phases: requirements gathering, conceptual design, navigational design, abstract interface design and implementation. The result is UML like diagrams and re-usable design patterns. Moreover,

most Web Engineering approaches include functionalities that assist the user in finding and exploring information. This includes techniques discussed above, such as Information Retrieval and Recommendation Systems [SMB07]. However, the main shortcoming of Web Engineering approaches, as stated in [Koi13], is the lack of flexibility in extending websites functionalities beyond what has been stated in the websites' setup design. Moreover, any extension towards Cross Site Personalisation would have to be considered in the early conceptual phase with limited flexibility at run-time.

Limited-impact aspects of Personalisation techniques are related to the needs and preferences of the web publishers (whereas non-intrusive are related to those of users). The integration should be introduced as minimal as possible, where minimal refers to the invasiveness of the integration. Moreover, limited-impact integrations do not have to aversively interfere with the look and feel of the websites and the control of the web publishers.

As a final note, it can be stated that the level of invasiveness to extend an already deployed website, based on Web Engineering principles, is high. To introduce any additional functionality, the underlying website needs to be re-designed and re-deployed. However, this limitation is balanced by the fact that a website, that has been designed and implemented based on Web Engineering principles, underlies strict design and documentation guidelines that can reduce costs and times.

Summarising, approaches that are deeply integrated into websites in the design-phase, lead to an increased engineering effort to apply Web Personalisation to existing websites.

2. Run-Time Integration: On the other hand, the integration of

personalisation methods and techniques at run-time has its main challenge in understanding the interdependencies within a website. An approach, that has recently gained popularity, to address this integration challenge is the use of website frameworks. Such frameworks are often referred to as Web Information System (WIS) and allow the implementation of an entire website as out-of-the-box deployment [AF03], [IBV98]. However, WIS are traditionally proprietary software products limiting the extensibility without the vendors help and/or approval. Therefore, a promising solution to such limitation is the usage of Web-based Content Management Systems[1] (WCMS), which allow a simple and more flexible implementation based on their Open-Source nature (WCMS allow the implementation of websites ranging from blogs to more complex e-commerce websites). The main feature of WCMS is the extendable framework, which allows external modules to influence different levels of the functional layer of the website without the need of re-design or re-deployment. Moreover, applying Cross Site Personalisation to WCMS provides the additional benefit of giving web publishers control in deciding what area(s) of the website should be affected by the user's guidance provided by the Personalisation service. The main shortcoming, instead, is the lack of strict and coherent coding guidelines and documentations (typical of integrations in the design-phase), which is balanced thanks to the active support of the developing community. As stated in [Koi13], the integration of Web Personalisation techniques into WCMS is limited-impact due to the extensible and pluggable design of

[1] Currently more than 45% of the top 10 million traffic sites on the Web are implemented as Web-based Content Management System according to `https://w3techs.com/technologies/overview/content_management/all` sourced on the 16/11/2016.

WCMS. This allows the web publisher to control the extension (if necessary disable) and also to decide how much the extension should influence the overall website look and feel. However, this only applies if the added Personalisation techniques are limited to certain areas (e.g. widgets).

Furthermore, it should be noted that it is possible to design and enable more invasive Personalisation techniques at run-time. An example can be deep link recommendations that change the presentation and delivery of content within a website. Such a technique, however, would have to be balanced with the needs and requirements of the web publisher – as always.

It can be argued that the overall motivation for investigating Cross Site Personalisation techniques is to assist the user across the Web. With this objective in mind, Adaptive Hypermedia (AH) systems can be noted as one of the main research fields addressing the challenge of assisting users in addressing information needs, especially domain specific needs. In fact, AH techniques and methods have been proven to be successful in educational hypermedia applications (e.g. e-learning platform [BHMW02], [Bru01], as well as on-line information and help systems). However, this domain focus has also led to an issue known as the open-corpus problem and described as: 'The problem to provide adaptation within a set of documents that is not known at design-time and, moreover, can constantly change and expand' [BH07].

The open-corpus problem has been further extended, introducing the open-web problem of personalisation. The open-web problem of personalisation extends the open-corpus problem of Adaptive Hypermedia systems by adding: '[. . .] and which may reside in individually hosted websites on the open web' [Koi13]. It is to address the open-web problem of personalisation that the concept of Cross Site Personalisation (CSP) has been introduced.

Cross Site Personalisation can be defined as a process in which a web user is individually assisted in addressing information needs that span independently hosted websites. In the context of this thesis, information needs that span independently hosted websites are referred to as Cross Site information needs and are described as: A perceived lack of information that the user requires to complete a task or intent to pursue an interest that requires browsing several independently hosted websites [Koi13].

To study and investigate the challenges and requirements of CSP, the State of the Art chapter is structured as follows: (2.1) The remainder of the Introduction provides a brief discussion about the limitations of the current search paradigms, with the objective in mind to investigate problems related to the open-web problem of personalisation. A discussion on the identified gap in the State of the Art is also presented. (2.2) Key personalisation techniques are discussed, along with (2.3) different types of User Modelling approaches. (2.4) A subsequent overview on the State of the Art of Cross Site Personalisation is hence provided. The outcome of the State of the Art discussion is the identification of the main challenges and requirements to be addressed by the design, implementation and evaluation of an adaptive Cross Site User Modelling platform for information exchange techniques.

2.1.1 Limitations of Current Search Paradigms

Due to the abundance of data provided by web technologies to users, that allows them to easily create, share and publish information, it has become increasingly difficult for information systems to effectively assist users in addressing information needs. The information explosion phenomenon resulting from this massive amount of data provided led to the ‘information overload’ effect, which can result in user’s disorientation and the risk of misinterpretation of information [CS98].

To address the consequences of such negative effect, information access technologies have been introduced, such as Information Retrieval [SM86] following the ‘needle in the haystack notion’, Text Mining [hT99] to extract meaningful information and behaviour patterns from massive amount of data, Recommendation Systems [AT05] using mostly user data correlation techniques to recommend relevant content, and finally Adaptive Hypermedia [Bru01], [KL05] providing adaptive assistance based on the user’s needs and preferences.

Among them, the most applied approach in addressing information overload is Information Retrieval. However, although the approach is highly effective in addressing user’s partial information needs, which are expressed in a sequence of keywords provided by the user, it has three significant limitations:

1. The user’s expression of need is based on the interaction, within the website, with the search engine. Hence, it is difficult for a search engine to assist a user outside the boundaries of its domain of applicability;

2. The search query of the user is only a partial representation of the user’s overall information needs;

3. Search engine technology can lead to a filtering effect that narrows the choice and the informational browsing freedom of the user [O'C11].

In addition to search related approaches, browsing related techniques have been introduced. The overall idea of browsing assistance mechanisms is to guide the user to the relevant content through manipulating the websites link structure [Bru01], [TAAK04]. Popular approaches, among the others, are link highlighting and content recommendation, which range from simple colour overlays to more intrusive techniques, such as reordering and hiding mechanisms [Bru08], [PB07]. However, being deeply integrated in the application, these techniques result difficult to be applied across different applications, thus providing isolated personalisation. Therefore, in a Cross Site context as the one investigated in this thesis work, the shortcoming becomes even more relevant since users need to address overarching information needs across several different applications and/or websites, leading to an increase of the user's disorientation and frustration [FAJ10].

It seems clear that the classic techniques borrowed by Information Retrieval and Recommender Systems are not sufficient to address the challenge of information overload in information needs that span independently hosted websites. Therefore, this gap in the State of the Art, which needs to be addressed in order to overcome the issues outlined in this introduction, has to be discussed before moving on to the next section.

2.1.2 Gap in the State of the Art

The main objective of Web Personalisation is to influence the interaction between the website and the user through means of information access and presentation, with the goal to assist the user in experiencing the most relevant content. Therefore, Web Personalisation has become popular in mass user web applications, such as social media [SDAN$^+$09], retail [SKR99] and search platforms [KL05]. Furthermore, challenges related to user identification [CC09], [PB97], understanding user behaviour [BS00], [RK11], user modelling [CCG11], [DN03], entity extraction [JG12], [MCS00], content modelling [OK01], web engineering challenges (related to the integration of Web Personalisation techniques [GM01], [KKZB08]) and privacy concerns [LC11] represent some of the main technical and non-technical issues that need to be addressed to tailor the user's web experience in Cross Site browsing contexts.

Currently, the majority of Web Personalisation techniques provide to users an isolated experience, only effective on the current website the user is browsing but non across different websites. Reasons can be found in the continuing isolation of web experience referred to as 'walled gardens' [BL10], [KCW09]. Such a silo-ed experience can potentially deteriorate the overall browsing experience of the user, especially when their information needs span independently hosted websites. As already pointed out in the introductory part of the section, this liability is highly related to the open-corpus problem identified by various research groups [BHMW02], [BH07], [BM02], [HN01] and, therefore, to the open-web problem of personalisation introduced by Koidl [Koi13].

Following the work of Koidl [Koi13], this thesis designs, implements and evaluates an adaptive Cross Site User Modelling platform for information exchange techniques. Such a Cross Site platform can be potentially used as a personalisation testing environment to implement and evaluate Cross Site Personalisation techniques (which use the information exchange methods here designed as underlying communication pattern between target websites and the third party User Model Provider) that assist users in addressing information needs across individually hosted websites. Therefore, this thesis work serves as a starting point for the overcoming of the gap in the State of the Art and it is developed with that final objective in mind.

In the following, three sections concerning the State of the Art techniques, methods and approaches for Web Personalisation are presented.

2.2 Key Personalisation Techniques

This section categorises the State of the Art techniques and methods for Web Personalisation. The discussion is divided in two separate subsections, each related to a different type of approach. First, Information Retrieval techniques (2.2.1) are presented. Then Recommender Systems approaches (2.2.2) are introduced.

Both the approaches have proven to be effective in Personalisation case studies, such as:

- Information Retrieval in personalised search [FFS11];
- Recommender Systems for content and item recommendations [LSY03], [LdGS11].

Despite the separate subsections to describe the two approaches, it is important to clarify that these techniques are not distinct, indeed Content-based Recommender Systems are usually based on Information Retrieval techniques [PB07]. However, since their introduction, these approaches have been extended towards Personalisation. It is therefore within this context that both the techniques are discussed in the following.

2.2.1 Information Retrieval

Traditionally, Information Retrieval (IR) has a strong focus on technological advancements, such as performance, scalability, precision and recall [SM86]. Within this domain, however, specific aspects related to the Cognitive Information Concept [IJ05] have also been studied, leading to focus on the user's needs beyond a stated query. Therefore, more user focused research areas have been introduced in recent years, such as Personalised Information Retrieval (PIR).

The overall idea of PIR techniques is to seek to learn from the user's search history, in order to build a user model. This model is then used to perform query adaptation and result adaptation. Query adaptation is mostly divided into query expansion and query relaxation [XC96]. In a short query, query expansion can be applied to add related terms to the query without the user's explicit knowledge or participation [MSB98]. In a long query the technique of query relaxation is used to reduce the amount of terms that may be synonymous or confusing [ERW11], [ZGBN07]. In a set of experiments ran by Kumaran and Allan [KA08] it has been seen that the use of query adaptation can even provide performance improvements of up to 40%. On the other hand, result adaptation relies on a user model to assist the re-ordering of the result list [SG05].

The main shortcoming of Information Retrieval is to be limited to the expressiveness of a query. Indeed, it can be argued that a query only provides a partial representation (i.e. snapshot) of the user's overall information needs [HPPL98]. Nevertheless, for information needs in which the user is seeking a specific fact or piece of information an Information Retrieval based search engine can provide fast and accurate results.

In relation to the intrusiveness of PIR, it can be argued, that the introduced techniques are non-intrusive. This argument is based on the user not noticing the extension or relaxation of the query and therefore is not interrupted. On a higher level, however, it can be argued that search within a browsing task is in itself intrusive due to it requiring an interruption of the browsing process. Furthermore, it can be argued that query expansion and relaxation mostly happens without the consent of the user. Even though the intentions of PIR are to assist the user it may result in the filtering of content that is relevant to the user.

2.2.2 Recommender Systems

Recommender Systems follow a similar 'finding a needle in the haystack' notion as Information Retrieval. The main difference, however, is that Recommender Systems do not require an explicit query stated by the user. To recommend an item of interest, Recommender Systems rely either on in-depth knowledge of the content or on explicit feedback of the user (such as rating). The literature discusses different types of Recommender Systems. The main approaches discussed are Collaborative [LSY03], [SFHS07] and Content-based Recommender Systems [PB07]. In addition, Hybrid Recommender Systems [Bur02] and recently Social Recommender Systems [Guy15] are discussed.

The main advantage of Collaborative Recommendation Systems is that they do not require any machine-readable description of the underlying content to ensure effective use. Collaborative Recommender Systems require explicit ratings data provided by the individuals interested in the item. This type of recommendation can be described as ′people-to-people correlation′ resulting in statements such as ′People that like this item also like the following items′ [SFHS07], [SKR01].

Content-based Recommender Systems, conversely, do not require user ratings. They are based on a query that is automatically created based on the user information stored in the system [PB07]. Similar to result adaptation in Personalised Information Retrieval, the resulting list is used to inform the Recommendation System about the relevance of items in relation to the user′s preferences and needs.

Recently variations of Recommender Systems have been introduced, such as Hybrid Recommender Systems using both rating and information retrieval data [Bur02], as well as Social Recommender Systems, which only use ratings from friends and peers within the social network of the user [Guy15].

Recommender Systems have several shortcomings. The main shortcomings are known as the New User Problem (also known as the Cold Start Problem), the New Content problem and the Portfolio Effect:

- The New User problem relates to missing information about a new user (e.g. no rating or/and no click data) resulting in a higher probability for misinformed recommendations [BFG11].

- The

  New Content problem relates more to Content-based Recommender Systems and can occur if new content is added, but not incorporated into the underlying model/system that calculates recommendations. This problem is apparent in large and dynamic content base, such as news websites, which underlie rapid content updates [HKTR04].

- Finally, the Portfolio Effect relates to Recommendation Systems recommending items or content already consumed/purchased and therefore that are not relevant anymore. This shortcoming has been addressed by allowing the user to actively participate in the creation and updating of the user's profile [SKR02]. Other solutions have been introduced for example by adding demographical data, such as age and gender [WPB01], specifically to overcome the New User problem. Some strategies relate to implicit techniques, such as Web Usage Mining or Conceptual User Tracking [OBHG03]. Both techniques rely on analysing log data for unique evidence such as IP address, device preferences and navigational behaviour [MCS00].

The level of intrusiveness introduced by Recommender Systems varies and depends on how much the user is required to participate. Therefore, it can be argued that Content-based Recommender Systems are less intrusive than Collaborative Recommender Systems, which mostly rely on explicit rating. Both approaches have their use especially when trying to find specific items/facts, however in information needs that are not specific and possibly span over different sources it is required to the user to cognitively connect the partial information. This can result in browsing patterns that consists of back tracks leaving the browsing experience shallow [HJ04].

A further concern in relation to the intrusiveness of Recommendation Systems and Information Retrieval techniques has been identified as 'over filtering' or 'over personalisation' and is also referred to as the filter bubble [O'C11], [Par11]. This issue can lead to the user receiving more of the same without any diversity or serendipity[2] in either the search results or the served recommendations.

2.3 User Modelling Approaches

The following section studies User Modelling approaches in Web Personalisation. Since User Modelling is a wide and complex research area, the discussion is specifically focused on identifying and addressing challenges - in the State of the Art - in assisting users across separate websites. Therefore, the topics discussed range from classification of User Models to model gaps and the addressing of separate needs, passing through implicitly modelling the user across separate websites, cold start problem, and modelling temporal aspects.

2.3.1 User Model Classification

User models are often implemented as an overlay model based on or laid over the concepts of the application [BSS05], [ACDNG10], [ACDN12]. The

[2] Serendipity is defined as finding something good or useful while not specifically searching for it.

main advantage of this implementation approach is an increase in accuracy and performance. On the other hand, the main shortcoming is a limitation of the models interoperability within other applications or subject domains [CCG11].

When user models allow a higher-level abstraction they are often referred to as user profiles. User profiles can be defined as a subclass of User Modelling, less sophisticated and more suited for applications that require a more general abstraction of information needs [KW01]. Furthermore, user profiles can be based on implicit or explicit data. Hence, due to the non-intrusiveness notion adopted for the Cross Site Personalisation techniques, the following discussion focuses on implicit user data.

The three main user profiles discussed are: keyword based profiles, semantic network profiles and concept profiles [GSCM07].

2.3.1.1 Keyword Based Profiles

Keyword based profiles are based on keywords extracted from web pages, such as tags, metadata and keywords explicitly provided by the user. This type of profile typically consists of keywords from websites associated with a form of weighting typically ranging from 0 to 1. Keyword based profiles are usually based on a single term vector and use a TF-IDF approach to extract terms from content [SM86]. Examples of approaches using TF-IDF based keyword extraction are Amalthaea [MM98] and Webmate [CS98].

The main drawback of keyword-based profiles is the lack of understanding in the meaning of the stored terms. This problem is known as the polysemy problem. For example, the term bank can relate to a riverbank or a financial institution. Approaches extracting keywords based on information retrieval mechanisms, such as TF-IDF, are strongly affected by this shortcoming [Mea92]. PSUN [SM95], as one example, addresses this shortcoming by using weighted n-grams[3] allowing the user to gain a better understanding of the terms meaning. Alipes [WYEN^{+}99], as a second example, assigns three keyword vectors to a higher-level interest. Finally, to extract the higher level interest implicitly PEA [MGH98] uses the user's bookmarks by indicating a bookmark as a higher level interest and associating a term vector to the bookmark.

2.3.1.2 Semantic Network Profiles

To allow the user to gain a deeper understanding in the meaning of the extracted terms and to overcome the problem of polysemy semantic term networks can be used. Semantic networks usually consist of a term structure, which entails an order or relationship. Based on the example above, the term Bank relates to the term pair (n-gram) Financial Institute. This allows a system to understand that the term Bank does not relate to a riverbank. A further advantage of Semantic Network Profiles over keyword based profiles is that the connection between the keywords and the higher-level concept can be maintained. This leads to a graph like structure in which the nodes represent keywords or concepts and the arcs connections. Following the example above the concept node Financial Institute may have several different keyword nodes associated with it, such as bank, coffer and

[3] By n-gram it is meant a contiguous sequence of n items from a given sequence of text or speech. The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles. [BGMZ97]

credit union. The main shortcoming of creating Semantic Network Profiles is that they can be time consuming to create and keep up-to-date.

2.3.1.3 Concept Profiles

Concept profiles are similar to term based profile with the difference that higher level topics are displayed, which allows the user to gain a better overview of the interest the system has inferred. For example, a concept profile would indicate to the user that an interest was identified related to higher-level n-gram Financial Institute, but not display the lower level term Bank. This higher-level topic extraction allows a profile to become more focused without showing the user all terms, such as in a purely term based profile. A different example of a concept-based profile is based on a thesaurus. For example, Wordnet[4] allows the extraction of term relationships and synonyms to facilitate the construction of a concept profile. Open Directory Project (ODP)[5] [WBC09] as a second example, allows the extraction of a yellow page based hierarchy of terms. The main shortcoming of concept profiles is that usually rely on taxonomy or ontology[6], which are time and cost consuming to create and update.

Further identified and discussed challenges are related to the modelling of the user's Cross Site information needs, the cold start problem, the modelling of temporal aspects and the model gaps.

[4] http://wordnet.princeton.edu/

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

[6] A taxonomy can be defined as a concept or term structure [MS01]. A more advanced concept to represent relationships of terms is an ontology, which, in addition to only relying on the structure i.e. connection between the concept and terms, adds relationships to allow logical reasoning.

2.3.2 Modelling the User Across Separate Websites

Within the CSP domain, modelling the user's needs and preferences across independent websites creates a complex challenge, in which it is necessary to assist the user not only in one website, but across several websites with different subject domains. To address this challenge, a possible approach is the introduction of a shared conceptualisation of the overall browsing space of the user. A shared conceptualisation could consist of an overarching vocabulary reflecting the user's browsing space. This would allow the user to receive a constant and unifying browsing experience towards addressing Cross Site information needs [Gru93].

2.3.3 Cold Start Problem

A further challenge in the context of modelling user needs is known as the cold start (or new user) problem. This problem relates to the necessity of bootstrapping the model of new users [Fis01]. Recently social media have been introduced as a solution to the cold start problem [AHHK10]. However, it can be argued that such an approach is not ideal for browsing that may include spontaneous and unpredictable actions. In fact, social media based solutions bear the challenge of figuring out what information provided in the social stream of the user is relevant.

An alternative approach to the cold start problem is the use of collaborative techniques known as item-to-item collaborative filtering. The main idea of the algorithm is to try to correlate the navigational behaviour of groups of users. This approach is very successful with retail websites where the application uses the purchase information of their users to correlate purchase tendencies. This allows recommendations, such as 'customers that bought x also bought y' [LSY03]. However, most approaches target a specific need that has been expressed through an explicit interaction, such as a purchase. If this interaction is missing, the cold-start problem may occur.

2.3.4 Modelling Temporal Aspects

The main challenge implicit user profile creation faces is to ensure the information of the user is described accurate and up-to-date [SK92]. A negative example of missing temporal aspects is a system using a user profile that represents a need that has passed, such as assisting a user in exploring holiday locations after the holiday is over.

User information needs can be split into two separate classes, which can be used to address temporal aspects:

1. Indefinite information needs: relates to needs that does not present clear starting and ending points;

2. Definite information needs: relates to interests with a definite start and end (e.g. exploring possible presents for an upcoming birthday).

The main difference between the two classes is that there is no factual evidence of when indefinite information needs starts or ends. The only thing that can be observed is that the intention in addressing a need will vary over time. Therefore, addressing this aspect may require the introduction of a deprecation factor, which can be used to either increase or decrease the impact of the current need within the user's profile. An example system using a deprecation factor is Letizia [Lie95]. Here the idle time is used to gauge an understanding if the user is not sure what to do next or if the user is simply not interested.

2.3.5 Model Gaps

It may occur that users, e.g. due to privacy concerns, decide not to use assistance on every website they browse. Therefore, the understanding of the user's information needs may lack in accuracy due to this perceived gap in the user's model. It appears clear that model gaps represent an additional challenge in relation to modelling user needs across websites. A possible solution, which is out of the scope of this research work, is related to the research field that discusses topics in user model scrutiny [Kay06]. Even though this research field is focused on offering scrutiny related to collected and modelled user information it opens the discussion to a wider user participation in the collection and usage of collected user data.

Many other challenges, related to User Modelling on the Web, can be identified; challenges that are, however, outside the scope of this thesis work. For example, the modelling of different types of information needs which relates to the shift of user needs within one browsing session. In this case, a Cross Site User Model requires a good understanding of the user's browsing behaviour in order to model sudden switches in the need of the user. Such an understanding can either result from explicit user control (e.g. informing the assistance mechanism about the change in information needs) or implicitly, that requires a deep understanding about the user's browsing behaviour and the user's browsing space (e.g. a sudden shift from online clothing stores to websites related to high-tech customer service can be interpreted by the assistance mechanism as a change in the information needs).

The following section concludes the chapter discussing the State of the Art techniques in Cross Site Personalisation.

2.4 State of the Art of Cross Site Personalisation

This section describes the State of the Art in Cross Site Personalisation. The approach presented has the objective to assist users in addressing information needs that span independently hosted websites. It addresses the existing gap in the State of the Art related to the lack of seamless assistance across independently hosted websites by introducing a third party service and Web-based Content Management System extensions, that allow users to receive consistent (based on the user's Cross Site information needs) assistance across independently hosted websites by means of navigational guidance. [KCW09] [KCWS11] [KCW13]

Major contributions of the approach are:

1. To ensure the Cross Site Personalisation applied is non-intrusive for the user;

2. To ensure the Cross Site Personalisation applied is limited-impact for the website;

3. To ensure the Cross Site Personalisation applied honours the privacy needs of the user by introducing the concept of informed decision.

4. To introduce interoperability in personalisation techniques within the Web Personalisation area.

In the following, the architecture of the Cross Site Personalisation approach is presented. The architecture is designed to assist users in complex online tasks which are short-term and Cross Site, such as in Online Customer Care (e.g. solving PC security issues), through a third-party Adaptive Feature Service (AFS). This is achieved by the usage of a unified term space based on the Cross Site browsing experience of the user. The main advantage for the user of choosing a service and not client-side approach is to receive personalised Cross Site recommendations both browser and device agnostic. The advantage for the content provider is a central interface to request personalised Cross Site recommendations. The overall architecture utilises a number of service-side repositories and services as well as an extendible interface layer and client side Web-based Content Management Systems modules.

2.4.1 Term Identification Service

The main purpose of the Term Identification Service is to identify terms related to the current webpage the user is viewing. Terms can either be retrieved directly from the interfacing website (e.g. through an existing taxonomy or folksonomy) or through external term extraction tools, such as Yahoo JQL table[7]. In a second stage, the extracted terms can be used to receive related terms through external knowledge services, such as WordNet[8]. Once an initial term-based taxonomy is created, the service can

[7] http://developer.yahoo.com/search/content/V1/termExtraction.html

[8] http://wordnet.princeton.edu/

train text analytics tools like Weka[9] to create related taxonomy terms for websites which either have no related terms or for which the terms are not in the scope of the previously collected terms.

2.4.2 User Model Repository

The User Model consists of terms related to the current task of the user and that were identified by the Term Identification Service. The User Model can be enriched with properties from external User Modelling Services such as preferred content type or language.

2.4.3 Strategy Repository

Depending on the task and User Model specific properties, the Strategy Model Repository can identify suitable strategies to create Cross Site personalised recommendations. An example would be the use of different strategies depending on the device the user is currently.

2.4.4 Scrutiny Interface

To ensure user trust in the recommendations provided, the user is allowed to access their user profile and to influence that profile. Furthermore, the user can receive information on where and how the user related data was collected and used. A basic implementation allows the visualisation of the terms that correspond to the webpages the user has accessed. In addition, size of the terms can indicate the relevancy of the terms based on the users' engagement with the different websites.

[9] http://www.cs.waikato.ac.nz/ml/weka/

2.4.5 RESTful Service Layer

The architecture implements a RESTful service layer for client communication.

2.4.6 WCMS Module Extensions

Based on the mostly flexible and simple extensibility mechanisms of Web-based Content Management Systems, different models are offered to be deployed at run-time to enable Cross Site personalised recommendations.

In the following chapter, the influences of the State of the Art are extended towards specific design requirements which are subsequently used throughout the design, implementation and evaluation of this thesis work. These requirements are introduced and discussed in section 3.1.

3
Design and Implementation

This chapter introduces the design and implementation of the Cross Site User Modelling platform, along with the Information exchange techniques that perform the exchange of user data between the service and the target websites. It is based on the outcome of the State of the Art review chapter and it is influenced by the PhD thesis work of Koidl [Koi13]. This research focuses on information exchange patterns, and explicitly highlights what has been done differently and what has been improved in terms of flexibility, performance and adaptability. The design and implementation are subsequently applied to three separate case-studies in the Evaluation chapter of this thesis. Each related to a different user behaviour.

The chapter is structured as follows: First, the introduction and the requirements are presented (section 3.1). Then the architectural design and the use-case are defined (section 3.2). After this, the design of key technical elements is proposed (section 3.3). Finally, a use-case implementation is provided (section 3.4).

3.1 Introduction and Requirements

The state of the Art review provided in chapter 2 of this thesis addressed various aspects related to Cross Site Personalisation, either connected to User Modelling approaches or to personalisation techniques. The main goal of this section is to provide an overview of the influences from the State of the Art review resulting in design requirements. The requirements, introduced and addressed below, are categorised into three 'Web Dimensions': The Web User Dimension, the Web Content Dimension and the Web Service Dimension.

After the analysis of the Web Dimensions, a brief summary concludes the discussion over the requirements.

3.1.1 Web User Dimension Requirements

With regard to the Web User Dimension, the main challenge identified is related to ensuring the information exchange mechanism maintains the right balance between providing relevant data and preserving user privacy[1]. Therefore, the following design requirement was identified:

- An information exchange technique should provide a trade-off between web publisher's needs and user's needs. It should therefore balance the amount of user data exchanged with target websites in a way that is satisfying for both entities. [WU1]

[1] Privacy concerns related to the user's trust are not addressed in this thesis work. However, studies based on this in relation to user model scrutiny are [Kay06] and [KB06].

Then, related to the modelling of user information needs, two issues identified are the need for a shared conceptualisation[2] of the user's browsing space and the challenges in modelling the users' information needs across independently hosted websites (e.g. cold-start users, modelling temporal aspects to identify the freshness of the need, shifts in users' interests etc.). Hence, the following design requirements were identified:

- A Cross Site information exchange service should have a unified understanding of the user's browsing space by creating a shared conceptualisation of this understanding. [WU2]
- A Cross Site information exchange service should apply a user modelling technique that does not interrupt the user's browsing experience. It should hence be implicit, by not requiring the user to explicitly participate in the identification, collection and management of information needs. [WU3]

3.1.2 Web Content Dimension Requirements

Within the Web Content Dimension two areas can be identified. The first related to web content that underlies a pre-set structure and the second related to web content that is unstructured. The following requirements were identified:

[2] The concept of 'shared conceptualisation' is used in ontology based research and semantic web. In the context of this thesis work, a conceptualisation is intended as an open and evolving representation of the user's information needs across the Web. Ontology (or semantic web) based approaches usually rely on a pre-set structure or standard, such as JSON or RDF, but can also be saved in RDBMS systems – as in this thesis work.

- A Cross Site information exchange service should be able to utilise existing content models to create a shared conceptualisation across different websites. [WC1]

- A Cross Site information exchange service should be able to use additional services and tools for term identification. [WC2]

- Depending on the subject domain, a Cross Site information exchange service should be able to use different ontologies to better identify representative terms of content extracted by the term identification component. [WC3]

- A Cross Site information exchange service should be able to identify representative terms of content at run-time, therefore also on contents recently added or updated within a target website. [WC4]

3.1.3 Web Service Dimension Requirements

The Web Service Dimension can be split into three areas: Non-intrusive application of information exchange techniques, limited-impact application of information exchange techniques and interoperability of information exchange techniques. The overall objective is to identify challenges and requirements of communication patterns related to the Cross Site Personalisation domain.

For what concerns non-intrusive integration of information exchange techniques, the following main requirement was identified:

- A Cross Site User Modelling platform should provide non-intrusive information exchange techniques to ensure the user is not hindered in freely browsing. [WS1]

In relation to limited-impact integration of information exchange techniques to existing websites, the focus is on understanding how much the communication patterns can be applied without affecting the look and feel of the website. Based on this, the following design requirements were identified:

- A Cross Site information exchange service should affect the website the minimum possible, that is the exchange service should ensure that the website's look and feel is not aversively affected by the integration with it. [WS2]
- A Cross Site information exchange service should not negatively influence the loading time of the website. [WS3]

Finally, related to interoperability, the following design requirements were identified:

- A Cross Site information exchange service should be based on a design that allows simple integration and interfacing with existing websites. [WS4]
- A Cross Site information exchange service should ensure that the communication between the Cross Site service and the target websites is flexible and does not depend on the websites technology stack. [WS5]
- A Cross Site information exchange service should ensure device and browser independence. [WS6]

3.1.4 Summary

This section introduced the design requirements necessary to set a Cross Site information exchange service into the Web Personalisation Domain. Three separate Web Dimensions were identified, each of these connected with different design requirements. The requirements have been considered and used for the design and implementation of a use-case investigating information exchange techniques in the Cross Site Web Personalisation Domain.

The following section defines the architectural design and the use-case, making clear what is, and is not, in scope for this thesis work.

3.2 Architectural Design

Based on the identification of the design requirements (section 3.1), this section introduces a high-level design of an adaptive Cross Site User Modelling platform where information exchange techniques, between target websites and a User Model Provider, can be tested and evaluated. This design sets the foundation for a use-case where the prototype implementation is defined. The use-case investigates the information exchange techniques in open domains[3].

[3] By open domain it is meant a web domain that introduces a browsing space not limited to a specific number of websites or a specific user group (e.g. Enterprise level website federations). The content is heterogeneous and mostly not limited to editorial guidelines or structures (e.g. blogs).

3.2.1 High-Level Design Definition

The proposed Cross Site information exchange service is based on the design requirements discussed in section 3.1 of this thesis. The overall approach relies on a third party User Modelling platform that interfaces with the various websites the user is browsing. This ensures that the independently hosted websites can obtain relevant users' data from the service without a dependency related to the websites' technology stack. From a user perspective, a third party service approach to Cross Site communication patterns provides a central access point, ensuring more control over user data (and therefore over user privacy) and resulting in a more website-agnostic approach. Furthermore, device and browser independence is ensured.

To clarify the notion of a third party User Model Provider for Cross Site information exchange techniques, the following Figure 3.1 is shown. The components presented represent the main elements of the Cross Site system: The Cross Site Browsing Space consisting of websites that interface with a central third-party User Model Provider; the RESTful API offered by the service that can interface with the websites facilitating information exchange techniques to provide relevant user data based on the user's Cross Site browsing within the Cross Site Browsing Space; website extensions that allow a simple integration of the service in the websites.

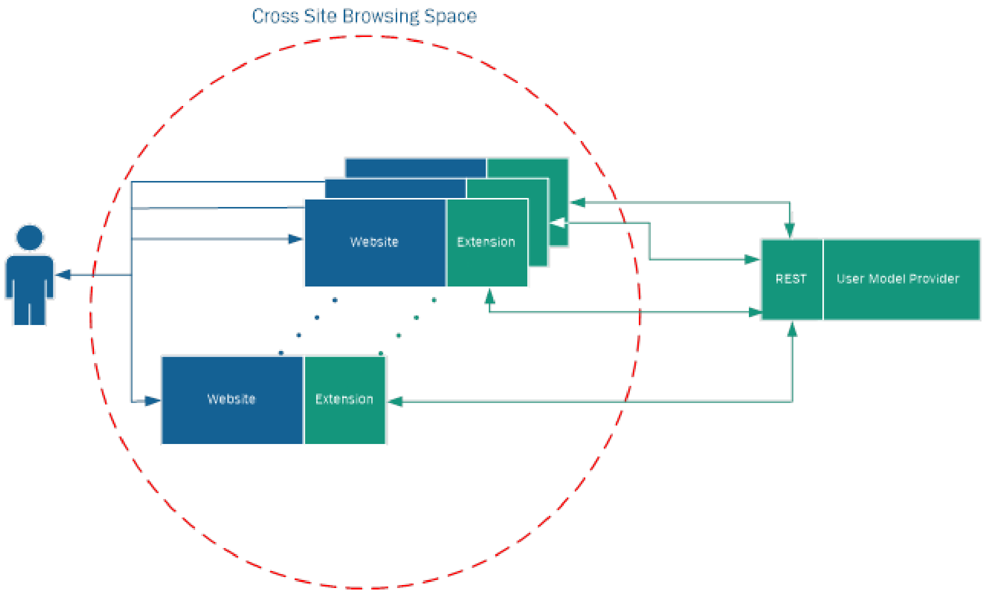


Figure 3.1. *High-Level Service Design*

To ensure a third party service approach is successful in providing relevant user information to target websites, it is important to address interface/communication challenges. The interfacing mechanism between the different websites and the User Model Provider should facilitate constant interaction. This interaction should be kept general to avoid limiting the approach to a specific website design, content subject or technology stack. Basically, the proposed Cross Site service consists of three high-level architectural components (Figure 3.1):

- A third party User Modelling platform to store user information from Cross Site browsing sessions and to facilitate accurate, user focused and privacy sensitive information exchange techniques.
- A service API independent from specific website designs, content subjects and technology stacks.

- Website module extensions to allow simple and cost effective integration of the information exchange techniques in a limited-impact manner.

The following subsection further analyses the high-level design, providing a conceptual architecture of the Cross Site information exchange service.

3.2.2 High-Level Architecture Overview

The considerations made in the previous subsection 3.2.1 about the high-level design of the service are here extended through the overview of a high-level conceptual architecture. The architecture reflects the requirements designed in section 3.1 and provides the conceptual foundation for the prototype implementation. As previously mentioned, the prototype implementation is defined within a use-case based on information exchange techniques in open domains. The goal of the introduced architectural overview is to illustrate the responsibilities and capabilities of the different service components proposed.

Therefore, according to the components depicted in Figure 3.2, the high-level service architecture is explored. Each component is introduced and described, considering the high-level design requirements defined in section 3.1. The proposed architecture is then used and adapted to the use-case set for the prototype implementation, which investigates the information exchange techniques in open domains.

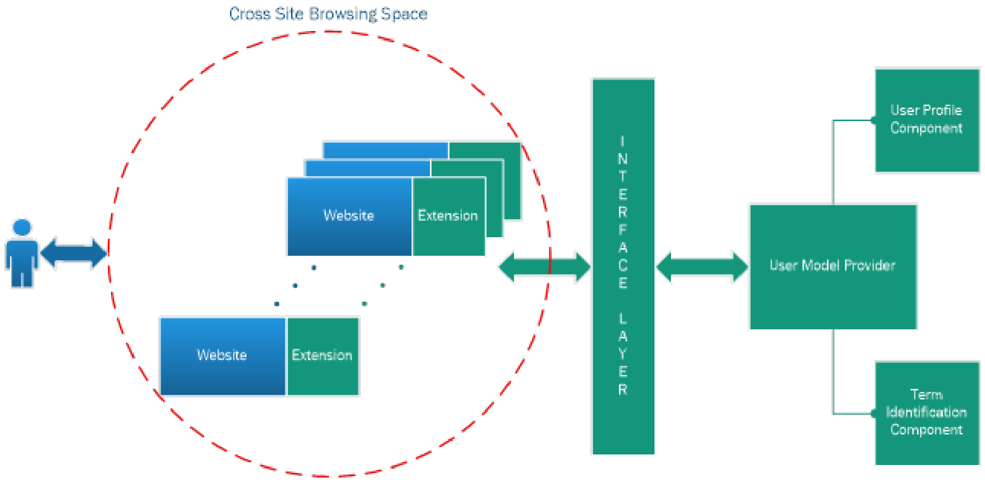


Figure 3.2. *High-Level Service Architecture*

3.2.2.1 Term Identification Component

The term identification component of the Cross Site User Modelling platform has the responsibility to facilitate the identification of text entities. The text entities indicate the meaning of the underlying content from websites within the Cross Site Browsing Space. This identification can be achieved by using additional services/tools to identify terms. Once the entities are identified they can be used by the user profile component to represent the user's information needs.

An additional responsibility of the term identification component is to ensure the creation of a shared conceptualisation of the user's Cross Site Browsing Space. This shared conceptualisation is represented as a text entity space and it is based on the contents the user is browsing within the websites of the Cross Site Browsing Space. Therefore, the term identification component needs access to the openly accessible contents on the different websites. Within the interface layer of this architecture, it is proposed that the website sends at run-time, to the term identification component, the contents the user is currently browsing. In this way, unlike the approach taken by Koidl in [Koi13], the only contents sent to the component are those browsed by the user. Hence, the bandwidth used by websites to send contents is reduced, thanks to not constantly sending new content to the term identification component, and the amount of data provided to the third party service is reduced too (limiting the leakage of information from the website).

3.2.2.2 User Profile Component

The user profile component is responsible for the correct storage and aggregation of text entities[4] related to the content the user is browsing. The text entities are provided by the term identification component and are stored together with the activity the user is conducting on the content. This activity can include mouse clicks, scrolls, cut & paste operations etc. Together the entities and the activity form the information block that is used by the information exchange techniques to provide relevant user data to target websites. Furthermore, the user profile component requires user's identification. This identification can be provided by a sign-in mechanism, such as OAuth[5]. With regard to privacy at the data level, there is no mechanism that provides protection to the API and therefore to the server

[4] A text entity in the context of this thesis can be considered any set of useful information connected to an extracted content related term. Examples of this information can be the topic (or context), the confidence level and the resource reference.

[5] `http://oauth.net/`

that stores user data. However, the issue is out of scope for this thesis work and won't be addressed.

3.2.2.3 Interface Layer

The interface layer provides an abstraction from the specific implementation of the different websites within the user's Cross Site Browsing Space. In order to do so, it implements a RESTful API (Representational State Transfer[6]) that facilitates the communication between the User Model Provider and the websites the user is browsing. A key responsibility of the API is to ensure that the interface with the User Model Provider is browser independent and does not depend on a specific technology stack. Furthermore, it should ensure fast and accurate interconnectivity between the interfacing websites and the Cross Site information exchange platform. Also, the API communicates with a website through provided website extensions. As a final note, this thesis work does not provide any management of the API, which is out of scope and will be considered as part of the future work.

The design requirements directly relating to the interface layer are simple interoperability and accurate and timely information exchanges.

3.2.2.4 Web-based Content Management System Module Extensions

Web-based Content Management Systems (WCMS) module extensions allow a simple and limited-impact integration of Cross Site information exchange

[6] `http://www.ics.uci.edu/~fielding/pubs/dissertation/rest_arch_style.htm`

techniques to existing website implementations. The responsibility of the WCMS module extensions is twofold: (1) To facilitate the communication between the website and the API of the Cross Site User Modelling platform and (2) to provide non-intrusive information exchange techniques to the user, within the website the user is currently browsing. The introduction of Cross Site information exchange techniques by ensuring non-intrusive and limited-impact integration in existing websites (at run-time) is a priority of this research. In fact, such integration can lead to providing non-intrusive assistance for the user and to ensuring limited-impact integration in existing websites of personalisation approaches. Therefore, it can be argued that WCMS module extensions are part of the high-level design of the use-case defined.

3.2.2.5 Cross Site Browsing Space

The application of information exchange techniques to independently hosted websites introduces a Cross Site Browsing Space. Within the browsing space target websites receive user data through information exchange techniques. The information exchange techniques are enabled through the aforementioned website extensions. Based on the back-end integration of the User Model Provider with the website, it is possible that the user decides when the information exchange service is enabled and when disabled. This allows the user to control the mechanism and ensures the user's privacy is honoured. The enabling/disabling of the service is performed through the user signing into the service.

The subsection introduced a high-level design of the Cross Site information exchange service, together with architecture components and responsibilities. The following subsection introduces a use-case, based on the design abstraction defined, which can be adopted for applying Cross Site information exchange techniques in open domains. The proposed use-case is then used in the implementation (section 3.4) and evaluation (section 4.2) of a service prototype.

3.2.3 Use-Case Definition

This subsection introduces considerations related to open domains. Compared to a closed domain, where the amount of websites within the Cross Site Browsing Space is limited and the content updates are usually controlled, the open domain can be considered as higher in complexity, due to the wide and open web space considered. Such complexity is based on a more flexible Cross Site Browsing Space in terms of websites members and content updates. Therefore, an open domain use-case can be categorised as a general-purpose focused use-case with the following characteristics:

1. An unclear problem space in which the user may gather information as part of a higher and not yet fully formed Cross Site information need.

2. User is not yet fully aware of what terms to use to identify the need.

3. Usually unknown set of websites to address the need.

4. Unpredictable content types.

5. Usually consists of many different interconnected Cross Site browsing sessions with different time frames.

6. Example real-world use-cases are among others: Research and planning, news, hobbies and activities etc.

Hence, in order to apply the high-level design defined in section 3.1 to the open domain use-case, it is necessary to introduce the following considerations on the high-level design itself:

a) To identify the meaning of website content by using an open and external API based service to allow the identification of content related text entities at run-time.

b) To ensure the identification of content meaning is applied to website content that is browsed by the user.

c) To use text entities extracted from the external service to populate the user profile, thus enhancing the user profile across different subject domains.

d) To store user activities (e.g. clicks, scrolls, cut & paste, etc.), along with content related text entities, in the user profile to enhance the level of awareness of user's information needs.

e) To use user profile data to perform an information exchange with a target website whenever a user triggers the exchange function within the website.

f) To ensure the communication between websites and User Model Provider is sufficient for real-time information exchange techniques.

g) To provide WCMS extension modules that enable non-intrusive and limited-impact integration to WCMS based websites.

The following section reflects the influences of the use-case just defined on the different components of the high-level architecture previously designed. Furthermore, it presents the design of a set of information exchange techniques that provide different levels of information enrichment for both websites and third party service.

3.3 Design of Key Technical Elements

This section describes the different components of the Cross Site information exchange service in relation to the open domain use-case. In addition, the section introduces the design of a set of information exchange techniques, whose most representative two are implemented and evaluated in sections 3.4 and 4.2 respectively.

3.3.1 Service Architecture Components and Capabilities in Open Domains

The subsection describes the different architectural components and capabilities in relation to the open domain use-case defined in section 3.2. Due to the open web nature of the use-case, the User Model Provider requires an open and flexible approach in the application of understanding content. The understanding of content should be simple, fast and performed in real-time for each web page the user is currently browsing.

3.3.1.1 Term Identification Component

The term identification component has responsibility for identifying the meaning of content extracted from websites through external services. Along with this, the component has also responsibility for assigning a context (or entity) to the terms extracted. In this way the service is able to store a representation of the web content that the user has browsed, with a contextualisation of the topics related to that content. In an open domain use-case, where the web space is wide and generic, an additional responsibility of the term identification component is to identify the meaning of content extracted from websites at run-time, that is when the user is browsing a web page. Therefore, differently from what has been done by Koidl [Koi13], there is no need to create an index of the publically available content of the websites within the Cross Site Browsing Space to generate a representation of the web content, the representation can be extended based on extracted text entities and which identify the meaning of the underlying content. The only concern is to provide meaningful text entities, extracted at run-time by the external term identification service, to the user profile component.

3.3.1.2 User Profile Component

The user profile component is responsible for storing text entities related to content within the Cross Site Browsing Space. In addition to the storage of text entities, user activities tracked on the web pages the user has browsed can be collected. The reason for this is to support information exchange techniques that rely on the user activities to provide more relevant and specific user data to target websites. Moreover, since user activities also record where the activity has originated, their storage allow an understanding of where the user information comes from.

3.3.1.3 Interface Layer

The interface layer is responsible to enable the communication between the User Model Provider and the different websites the user browses. The communication is based on distinct API endpoints, depending on the type of request performed, and relies on the REST architectural style to ensure a high-level of abstraction. REST provides a set of architectural constraints that, when applied as a whole, emphasises scalability of component interactions, generality of interfaces, independent deployment of components, and intermediary components to reduce interaction latency, enforce security, and encapsulate legacy systems. However, in order to avoid the RESTful API being limited to a specific subject or technology, it is important to ensure that the level of abstraction is maintained also in the open web, which by its nature requires integration with different websites technologies and contents.

3.3.1.4 Web-based Content Management System Module Extensions

WCMS extensions should ensure non-intrusive and limited-impact integration of information exchange techniques at run-time. Due to the open nature of the web space in the use-case applied, WCMS extensions need to operate at run-time, thus leading to considerations related to time response efficiency and scalability. It is also important to highlight that, in cold-start user[7] situations, the target website that performs the information exchange technique does not receive anything from the service which, on the other side, still obtains user activities and text entities from the website. Therefore, such a particular situation can be seen as unattractive from a website point of view. However, in the long run this initial lack of information will be rewarded and the website will benefit from this mutual exchange.

3.3.1.5 Open Domain Browsing Space

The open web browsing space is based on the notion that all webpages are not known at design-time and that the browsing space constantly changes based on the amount of websites adopting the Cross Site information exchange service. Furthermore, the content's topic may change rapidly based on the change of user information needs. Therefore, the Cross Site information exchange service has to ensure that the provided user data is up-to-date and compliant to the type of user profile implemented (i.e. long-term user profile or short-term user profile).

3.3.2 Design of Information Exchange Techniques

In this subsection the information exchange techniques between the target websites and the Cross Site User Modelling service are described. The techniques range from highly generic to highly specific, providing different levels of information enrichment for both the websites and the service.

[7] In this thesis work we consider as a cold-start user a brand new user of the Cross Site information exchange service. Hence, the user profile of this new user lacks information related to them.

The former tend to be more satisfactory from a web publisher perspective allowing a massive flow of user's information to leak from the service, thus ignoring or not sufficiently considering the user's privacy concerns. The latter, on the other hand, are more focused on preserving user's privacy, thus avoiding the provision of huge amounts of information to websites or, in extreme cases, not providing it at all, tend to be more satisfactory both for the user, who sees their privacy more respected, and the service itself, as it doesn't give away the only real value it holds – user data. This second category of techniques, however, while theoretically can be considered the most user-sensitive, when deployed in real-world applications have a tendency to be viewed with suspicion by web publishers[8]. The reason behind this is the fact that websites (i.e. web publishers) are reluctant to give their data away without anything satisfactory in return, whether it is money or information. Therefore, a trade-off between the two opposite trends described above must be sought.

The techniques described below are listed from the highly generic to the highly specific. For each technique, pros and cons related to privacy concerns are pointed out along with a technical review focusing on the

[8] The problem is due to a limited trust in the central service. So if a user browses a first website and read about something and then browses to a second website they should get content recommendations about the content read on the first website. This is valuable for the second website but does nothing for the first one, unless the user came in from somewhere else too. Thus, a solution has been proposed by two companies called Outbrain and Taboola. In this example, the second website places a content recommendation link on the first website. Should the user click on the link, they will be sent to the second website and the second website, that is the target website, will pay money to the first website for that click (it basically follows the same concept of advertisement).

effectiveness and efficiency of the exchange. The goal of this survey is to present a range of exchange paradigms rather than finding the best one, in fact all of them present both advantages and shortcomings that may be relevant depending on the context and the scope of the application that deploys them.

3.3.2.1 Privacy Insensitive Technique

The first technique, which is also the most generic, presents a really simple behaviour. When a user triggers the function that enables the exchange of information between the currently browsed website and the Cross Site service, an HTTP GET request is sent from the website to the service in order to retrieve all the user information stored in the service. The GET request is then processed by the service which returns everything it has on that user to the website, that is the user profile data (i.e. relevant terms and their associated entities) and the past user activities (e.g. browsed pages, scroll counts, click counts, etc.). After receiving the user data, the website sends an HTTP POST request back to the service containing all the user activities, along with the relevant text entities extracted by the term identification component, tracked since the last 'personalisation request'. With the last POST request the exchange is finished and the interaction between website and service stops until the user requests personalisation again by clicking on the 'personalise' button.

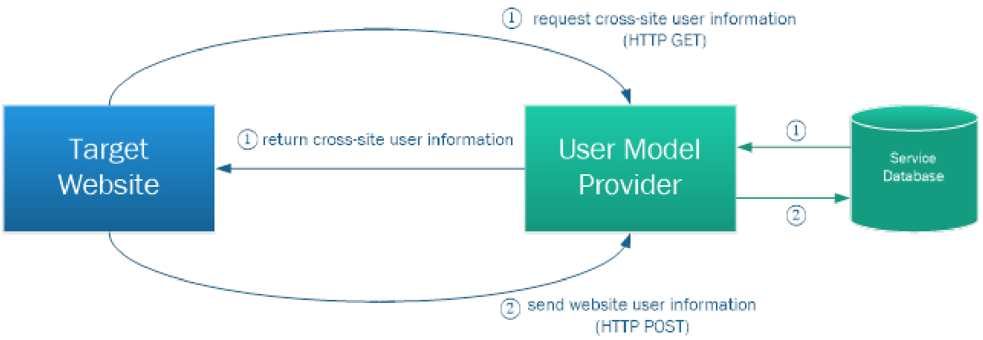

① request cross-site user information
(HTTP GET)
Target Website
① return cross-site user information
User Model Provider
①
②
Service Database
② send website user information
(HTTP POST)

Figure 3.3. *Privacy Insensitive Exchange Mechanism*

Effectiveness: due to the high amount of data transferred from the service to the target website, the effectiveness of the technique is considerably high, although it doesn't discern between relevant and non-relevant information.

Efficiency: the exchange consists of two interactions only: one HTTP GET request and one HTTP POST request. On service side, the database containing the user data is accessed twice, one to retrieve all the related user information and one to pass back new user data that the service integrates into its user model. Therefore, this technique presents the least number of interactions required to successfully perform the exchange proving to be one of the most efficient ones.

Privacy Concerns: from a web publisher perspective, this technique is the most attractive since it returns all the user data to the target website. On the other hand, although such a solution might be effective, it doesn't take into account the privacy implications related to users. In fact, such a massive flow of user data from the service leaves users undefended against the over-use of their personal information by websites, which collides with the objective of keeping user information as much as possible preserved while trying to provide effective ‘info-for-personalisation’. Moreover, even from the Cross Site service perspective, the unrestricted stream of user data produced by the method is seen as a serious issue. The reason is due to the fact that the privacy insensitive paradigm gives away completely the service only valuable resource – the user’s data, quickly making it an optional asset for target websites.

3.3.2.2 Threshold Technique

This technique, unlike the previous one, tries to temper the flow of sensitive information coming out of the service setting a threshold on the amount of returned data. In this way, although pretty simplistically, the service mitigates the over-use of personal data by websites and makes a first step towards the preservation of privacy. For this paradigm the exchange works similarly to the previous one, with a HTTP GET request to obtain the data stored in the service and a HTTP POST request to send new user related data (i.e. user activities and relevant entities connected to browsed pages) to the service. The main difference to the previous approach is the aforementioned introduction of a threshold which limits the number of items retrieved from the service. In practice, this reduces to set a limit to the number of rows retrieved from the service’s database, filtering out rows with a ‘timestamp’ inferior to the threshold one. Therefore, the service still returns every characteristic it knows about the current user, but it does that giving back only the ‘n’ most recent activities (i.e. the ‘n’ most recent searches, page clicks, page scrolls, etc.) together with the term/entity pairs related to those activities.

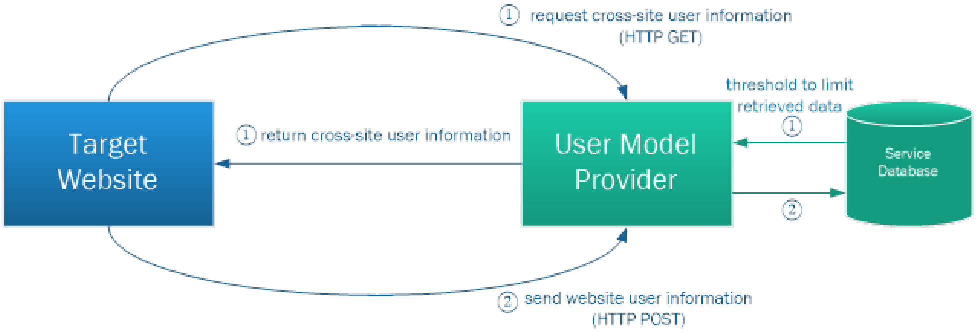

① request cross-site user information (HTTP GET)
threshold to limit retrieved data
Target Website
① return cross-site user information
User Model Provider
①
Service Database
②
② send website user information (HTTP POST)

Figure 3.4. *Threshold Exchange Mechanism*

Effectiveness: due to the simplistic approach of the technique, the threshold can heavily influence the effectiveness. In fact, whereas the previous approach provided all the user information regardless, this one needs to set the threshold in a way that it constrains the flow of personal information while returning enough information to be considered useful. In plain words, setting too high a threshold (i.e. reducing too much the number of items returned) might lead to a depleted set of useful information, which breaks the equilibria to the detriment of web publishers or, vice versa, setting too low a threshold (i.e. passing too many items to target websites) might make this technique too similar to the previous approach, to the detriment of user privacy. The key point here is to find a balance between returning information and preserving privacy.

Efficiency: since the only thing changed is the introduction of a retrieval threshold on service side, the exchange, like the previous paradigm, is performed with the least number of interactions required. Therefore, this technique also proves to be very efficient.

Privacy Concerns: as already mentioned in the 'Effectiveness' section, the privacy concerns are heavily influenced by the threshold. However, the method is still quite privacy insensitive since it only limits the number of items returned but it doesn't filter sensitive information that may not be interesting for the current context. Therefore, even if it is a step forward compared to the 'privacy insensitive' technique it still lacks filtering components that allow other techniques to provide only contextualized information, thus avoiding the sharing of non-relevant personal data with target websites.

3.3.2.3 Top-Feature Selection Technique

The technique presented here represents the first real detachment from the 'give-everything' approach. The idea is to classify the term/entity pairs in user profiles into three categories: highly relevant, relevant and poorly relevant. On service side this results in mapping the categories using the weights[9] associated with each term/entity pair for the specific user. Thus, thresholds weights are set for the highly and poorly relevant pairs and the service will retrieve items above or below the specified

[9] The weights are computed for each term/entity pair inserted into the service's database. A default value of 5 is assigned to the first occurrence of each pair, then the pair's weight is increased by 5 every time a new occurrence of the same pair is provided to the service (i.e. whenever we try to insert a duplicate pair, an update operation is performed instead). It is clear that the weights increase depending on the frequency with which the associated pairs are extracted. An alternative option, not investigated in this thesis work, is to adopt, in addition to an increase of recurrent pairs, a decrease of pairs that don't appear for a long time. Therefore, a different time threshold to evaluate the freshness of weighted pairs is set depending on the type of user profile adopted: long-term or short-term.

values according to what the website requires. Therefore, when the user triggers the information exchange function, the website sends an HTTP GET request specifying a term/entity attribute as 'highly relevant' in the query string. If the returned values are not enough (or not satisfying enough) then a second HTTP GET request is sent with the term/entity attribute set to 'relevant'. The process then is iterated a third time, with the 'poorly relevant' specification, in the eventuality that the values are still not satisfying enough. After the retrieval of user information from the service, the website sends the usual HTTP POST request containing the user's tracked activities, along with the relevant text entities extracted by the term identification component since the last 'personalisation request'.

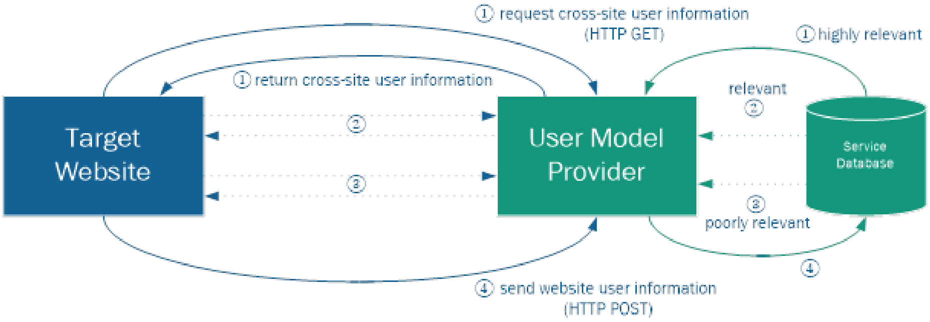


Figure 3.5. *Top-Feature Selection Exchange Mechanism*

Effectiveness: since this method returns the most relevant term/entity pairs and then, if necessary, pairs in decreasing relevance, the effectiveness, when compared to that of the preceding techniques, is improved. In fact, whereas the threshold technique returned a limited set of items based on a timestamp threshold, the top-feature selection technique returns a limited set based on relevancy. Moreover, this paradigm also has the advantage of providing items in a more controlled way, thus allowing the website to handle them more easily.

Efficiency: this paradigm has a major flaw in the number of interactions performed to complete the information exchange. In fact, in the worst case scenario (i.e. when neither the highly relevant nor the relevant term/entity pairs are enough) the method performs three HTTP GET requests, which resolves in three calls to the service's database. Compared to other approaches that only perform a single HTTP GET request, this one represents a decrease in performance due to the higher number of interactions required in all those cases where 'highly-relevant' pairs are not enough.

Privacy Concerns: compared to the previous one, this method is far more nuanced in terms of privacy control. Requesting user information by means of relevance levels, in fact, allows greater control over the flow of data out from the service than what a simple time threshold can allow. Whereas the threshold technique returns all the term/entity pairs, regardless of their relevance, above a certain time threshold, the top-feature selection technique gives back to websites more tailored information providing, in the best case, only 'highly-relevant' pairs. However, in the worst case scenario the method performs similarly if not worse (e.g. when the time threshold has been better optimised) than the threshold paradigm, returning user data ranging from 'highly-relevant' to 'poorly-relevant'.

3.3.2.4 Ranking Technique

This method, among the privacy insensitive paradigms, is the most refined since it combines the previous two approaches, limiting the number of term/entity pairs returned both by timestamp and by weight. Furthermore, the ranking technique is created in order to overcome the efficiency flaws related to the top-feature selection technique. Hence, instead of sending multiple HTTP GET requests (for less relevant term/entity pairs) when necessary, the method performs a single HTTP GET request that takes the form of an SQL statement of the type:

```
SELECT * FROM <table>

WHERE <timestamp> > threshold

  ORDER BY <weight> DESC;
```

In this way the term/entity pairs are still retrieved from the most relevant to the least relevant, but the number of database calls executed is reduced to one. Therefore, when the user triggers the function that performs the information exchange, the mechanism consists of one HTTP GET request to obtain user data from the service and one HTTP POST request to send new user data back to the service.

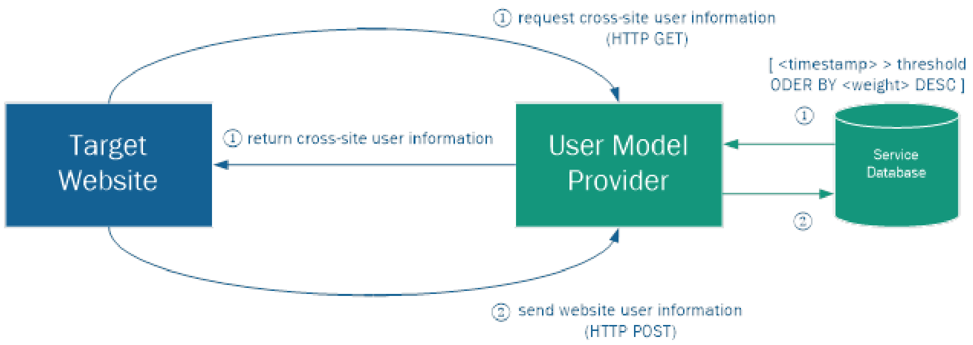

① request cross-site user information
(HTTP GET)
[<timestamp> > threshold
ODER BY <weight> DESC]
①
Target
Website
① return cross-site user information
User Model
Provider
Service
Database
②
② send website user information
(HTTP POST)

Figure 3.6. *Ranking Exchange Mechanism*

Effectiveness: the ranking technique further improves what done by the threshold and top-feature selection approaches. In fact, whereas the previous methods were only able to provide either time limitations or weight limitations, the ranking technique performs a combination of both which leads to websites being returned more focused user information. Therefore, it might be said that the ranking technique is the most effective among the privacy insensitive techniques.

Efficiency: by ranking the term/entity pairs by weight, the method is able to produce the least number of interactions required for the exchange to be performed. Therefore, since one of the criteria to assess the efficiency of the different exchange paradigms is based on the number of interactions needed to carry out the exchange, the ranking technique establishes itself as one of the most efficient, bringing back to one the number of HTTP GET requests required to perform the exchange.

Privacy Concerns: by combining the two previous approaches this technique is able to mitigate the problems, related to the uncontrolled flow of user data out from the service, that both of the methods suffer. Thus, even though the amount of term/entity pairs returned to websites is still considerably high, it is reasonable to identify the ranking technique as the most privacy sensitive of the privacy insensitive techniques.

3.3.2.5 Entity-Oriented Technique

The following technique departs from previous methods due to its more privacy-sensitive approach. Although this is still a first rough attempt to provide an effective exchange while preserving the privacy needs of users, it is nonetheless a step forward towards the ultimate objective of providing an exchange technique that is able to return relevant information to websites without compromising sensitive user data. In this approach, when a user triggers the exchange function and the HTTP GET request is sent to the service, a database call is made to retrieve the number of 'search' and 'browse' attribute occurrences. If the number of 'search' occurrences is greater than that of 'browse' occurrences, then the service will label the user as 'impatient' and will return, along with user data, a 'search-list personalisation' suggestion string which alerts the website about the user's leading behaviour. Vice versa, if the 'browse' occurrences are higher in number, the service will label the user as 'explorative' and will return a 'navigational-browse personalisation' suggestion string. Then, a second call to the database is performed, this time retrieving the entities connected to the specific user but not the terms (threshold operators can also be set in order to constrain the number of results). Finally, the results are sent back to the website and the HTTP POST request is performed as usual.

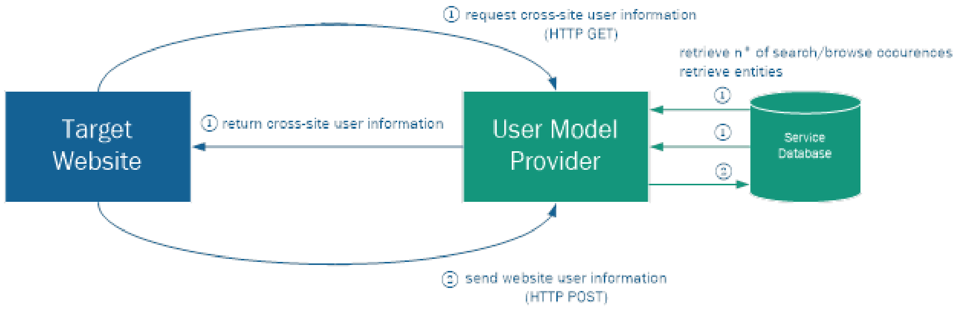

① request cross-site user information
(HTTP GET)
retrieve n° of search/browse occurences
retrieve entities
Target Website
① return cross-site user information
User Model Provider
①
①
②
Service Database
② send website user information
(HTTP POST)

Figure 3.7. *Entity-Oriented Exchange Mechanism*

Effectiveness: due to its attempt to preserve user privacy, the method does not return the same volume of data when compared to more privacy insensitive techniques. However, since it returns suggestions for personalisation (e.g. 'search-list personalisation' string), along with user topic interests (i.e. the entities related to the terms extracted from browsed pages), it can provide more tailored information.

Efficiency: the method suffers in efficiency due to the double database call. Even though there is only one HTTP GET request made, the service performs two database queries that inevitably affect the overall performance. The method is less efficient compared to all the previous techniques except top-feature selection, which on average performs similarly to this approach and in the worst case scenario even worse (three HTTP GET requests).

Privacy Concerns: this method is the first to actually perform a primitive embodiment of privacy-awareness. Returning entities implies that websites can only infer user topic interests, but not the actual contents related to those topics. Therefore, sensitive user information is better preserved and privacy is more respected at the cost, of course, of a deterioration in the volume of the returned data.

3.3.2.6 Activity-Oriented Technique

The activity-oriented technique extends the entity-oriented technique, introducing a rule-based method that attempts to find the most relevant term/entity pairs by evaluating the user activities. The result set is then ranked by weight and timestamp, placing the most up-to-date term/entity pairs at the top of the list. Therefore, when a user triggers the exchange function, a HTTP GET request is sent to the service which performs a database call to retrieve all the activities for that particular user up to a specific timestamp. The result set is then analysed by means of the aforementioned rule-base method:

$$activity\_weight = 5\times\langle clicks\rangle+5\times\langle ctrl{+}f\rangle+10\times\langle cut\_copy\_paste\rangle+1/10\times\langle scrolls\rangle \tag{3.1}$$

where:

- $\langle clicks\rangle$ is the number of clicks made by the user in the web page;
- $\langle ctrl + f\rangle$ is the number of browser searches performed by the user in the web page;
- $\langle cut\_copy\_paste\rangle$ is the number of cut/copy/paste actions performed by the user in the web page (cut_copy_paste is assigned with a higher value based on this activity indicating a high relevancy of the content for the user);

- $\langle scrolls \rangle$ is the number of scrolls done by the user in the web page (caveat: since the JavaScript function to track scrolls tends to be very sensitive, it is necessary to contain it in order not to bias the final value).

Hence, the top-n activities are kept and a second call is performed to retrieve, ranked in descending order from the most recent, the term/entity pairs related to these activities. Finally, the result is returned to the target website, together with the 'personalisation suggestion' string which is computed alongside the rule-base method, and the HTTP POST request is then sent.

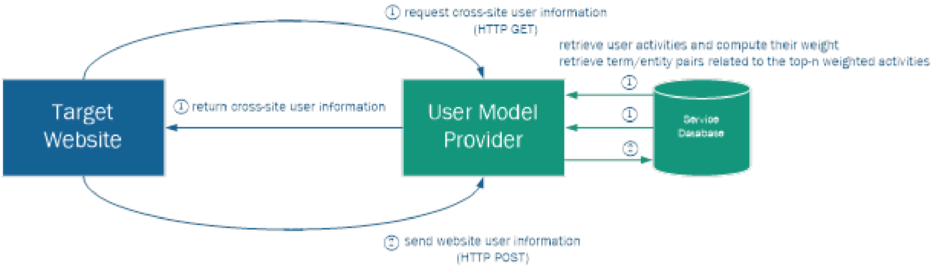

① request cross-site user information
(HTTP GET)
retrieve user activities and compute their weight
retrieve term/entity pairs related to the top-n weighted activities
Target Website
① return cross-site user information
User Model Provider
①
①
②
Service Database
② send website user information
(HTTP POST)

Figure 3.8. *Activity-Oriented Exchange Mechanism*

Effectiveness: the current method improves the effectiveness of the previous one by attempting to infer the relevancy of the term/entity pairs from the user activities, providing 'personalisation suggestion' strings as well. Compared to the previous paradigms, the activity-oriented technique is the most effective of those that do not give all user information back to requiring websites. To find more effective methods that are also mindful of privacy, query expansion techniques or learning algorithms are required. Otherwise, the level of effectiveness will remain similar to this approach.

Efficiency: the efficiency considerations are identical to those achieved by the entity-oriented technique.

Privacy Concerns: this technique tries to maximize the effectiveness while preserving user data. In fact, thanks to the additional emphasis placed in the selection of the items to be returned to websites, the method is able to diminish the returned quantity of sensitive information unrelated with the current context, even though the lack of semantic awareness precludes it from being highly effective in this sense.

3.3.2.7 Knapsack Technique

This technique is the first that implements a learning algorithm in order to improve the effectiveness and the privacy-awareness of the information exchange. When the user triggers the function that activates the information exchange, a HTTP GET request is sent to the service, which utilizes a feature weighting algorithm to assign weights to the different attributes that form the user profile. The service then returns a list of features, along with their newly calculated weights, that the website can request from. The website has a certain amount of 'points' (which needs to be defined a priori) that can be spent to request features and therefore the problem can be modelled as the 'Knapsack 0/1' problem where each feature has unit size. After that a solution algorithm for the 'Knapsack 0/1' problem could be run on the website (client-side) in order to find the optimal combination of features to request, and a HTTP GET request with the list of the chosen attributes then sent back to the service which hence returns the corresponding values. Finally, a HTTP POST request is performed to finish the information exchange and update the service with new user activities.

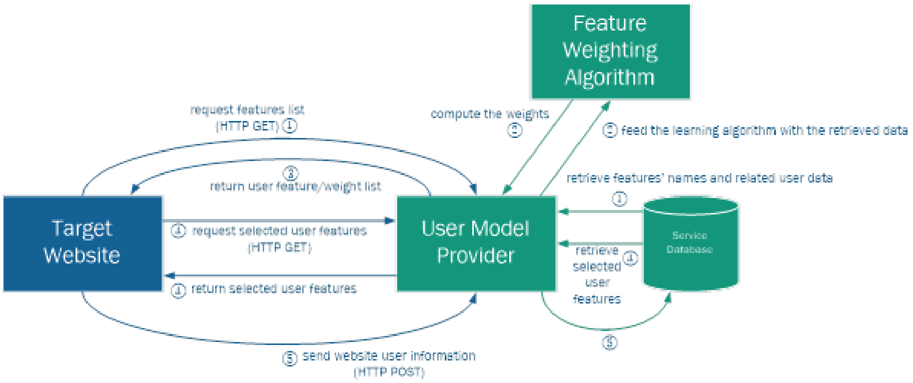

Feature Weighting Algorithm
request features list (HTTP GET) ①
compute the weights ②
② feed the learning algorithm with the retrieved data
③ return user feature/weight list
retrieve features' names and related user data ①
Target Website
④ request selected user features (HTTP GET)
User Model Provider
Service Database
retrieve selected user features ④
④ return selected user features
⑤
⑤ send website user information (HTTP POST)

Figure 3.9. *Knapsack Exchange Mechanism*

Effectiveness: thanks to the learning algorithm that is able to weight the features based on their relevance, the method can provide more tailored information. Furthermore, with the implementation, on the client side, of an optimization algorithm the method is able to request the optimal combination of features that allows for more effective personalisation.

Efficiency: with its computation overheads due to the feature weighting technique and the optimization algorithm, the method performs slightly worse than the other techniques that present two HTTP GET requests.

Privacy Concerns: the strength of the knapsack method is the fact that only a reduced number of features can be selected by target websites for the exchange and therefore, although not completely, the flow of sensitive data out from the service is limited.

3.3.2.8 Semantic Technique

The semantic technique involves the use of a semantic recommender. As will be explained in section 5.3, the semantic recommender can be seen as an additional component of the system that helps to enhance the personalisation effectiveness. However, in the context of the exchange techniques, when the user triggers the function that activates the information exchange and the HTTP GET request is sent to the service, the service calls the semantic recommender which provides back the most semantically relevant term/entity pairs. At this point, the service sends those pairs to the website which then performs the usual HTTP POST request with the new user activities.

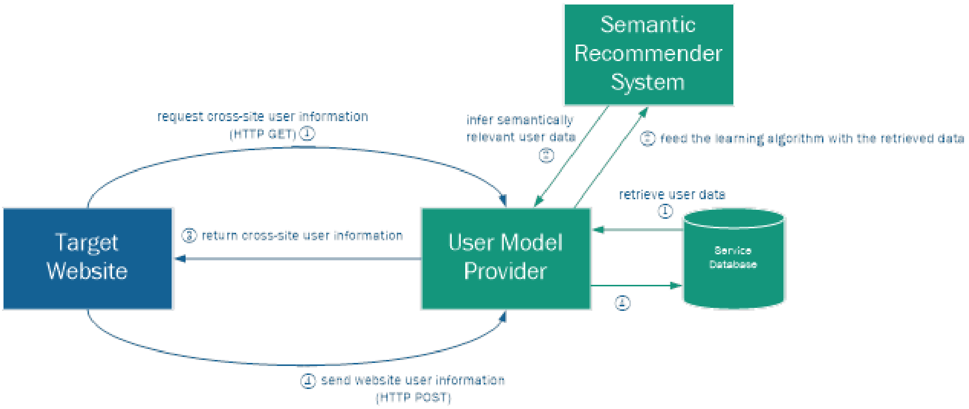

Semantic Recommender System
request cross-site user information (HTTP GET) ①
infer semantically relevant user data ②
② feed the learning algorithm with the retrieved data
retrieve user data ①
Target Website
③ return cross-site user information
User Model Provider
Service Database
④
④ send website user information (HTTP POST)

Figure 3.10. *Semantic Exchange Mechanism*

Effectiveness: this method, among the privacy-sensitive ones, is the most effective since it is able to infer the most relevant pairs based upon semantic reasoning. Therefore, the semantic technique succeeds where the others fail, in providing user data that is contextually relevant to the user's current interest.

Efficiency: this method performs one HTTP GET request and one HTTP POST request, therefore, even when the overhead of the recommendation algorithm is taken into consideration, it is one of the most efficient as well as the most effective.

Privacy Concerns: the considerations for the privacy are very similar to those achieved by the entity-oriented technique. However, the semantic technique, obtaining contextualised term/entity pairs from the semantic recommender, returns to the target website additional information than the only entities, thus resulting slightly less careful to privacy than the entity-oriented technique.

3.3.2.9 Suggestion-Oriented Technique

This technique, first presented by Koidl [Koi13], provides an informed decision to the website the user is currently browsing. An informed decision assists a website to identify relevant links based on the user's Cross Site browsing. The method implements a decision process that facilitates elements of the user's current user profile and relates them with the current website the user is browsing. The resulting informed decision allows a website to receive relevant information to provide adaptive guidance based on the user's overall information needs. Therefore, when the user triggers the function that activates the information exchange, a HTTP GET request is sent to the service which then calls the component responsible for the computation of the informed decision. To generate an informed decision, the component requests terms and activities from the user profile and sorts them based on the terms related to the most recent update first.

Then it computes the relevancy (boost) of each term, based on the same approximation described in the activity-oriented technique paragraph and used to weight user's activities:

- Multiply $\langle cut\_copy\_paste \rangle$ actions by 10;
- Multiply $\langle clicks \rangle$ actions by 5;
- Multiply $\langle scrolls \rangle$ actions by 1/10;
- Multiply $\langle ctrl + f \rangle$ actions by 5.

Which results in:

$$term\_boost = 5 \times \langle clicks \rangle + 5 \times \langle ctrl + f \rangle + 10 \times \langle cut\_copy\_paste \rangle + 1/10 \times \langle scrolls \rangle \tag{3.2}$$

Then it adds a deprecation factor to the calculated boost[10] and generates a query, using the terms and the related weights, on the index of the website that requested the informed decision. Depending on the amount 'n' of links requested the first 'n' results of the result list would be used. Results are ranked according to decreasing relevancy. Hence, the links and the calculated relevancies are added to a list, each link and its Cross Site relevancy representing an informed decision. Finally, the list is sent back to the website which then performs the usual HTTP POST request with the new user activities.

[10] The deprecation factor is calculated by dividing the calculated boost by the update position to the power of 2. Therefore, the boost of the terms related to the most recent activity has the smallest divider with a denominator of 2. The boost factor is divided by 4, the next by 16 etc. This ensures that the most recent activity has the highest impact.

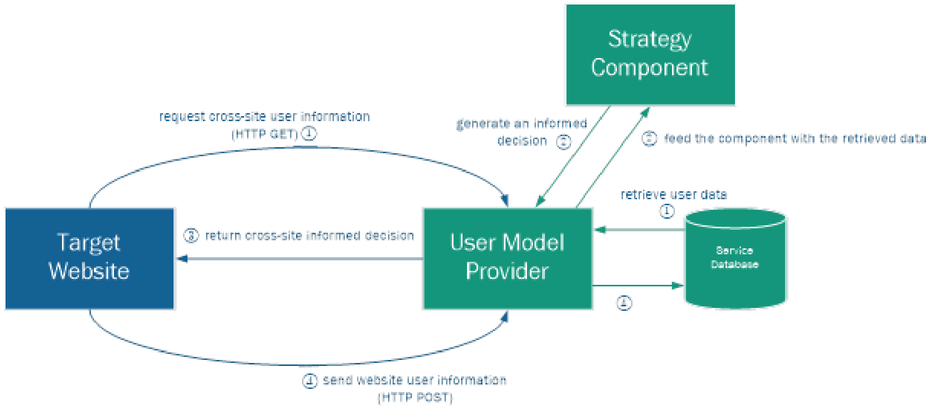

Strategy Component
request cross-site user information (HTTP GET) ①
generate an informed decision ②
② feed the component with the retrieved data
retrieve user data ①
Target Website
③ return cross-site informed decision
User Model Provider
Service Database
②
① send website user information (HTTP POST)

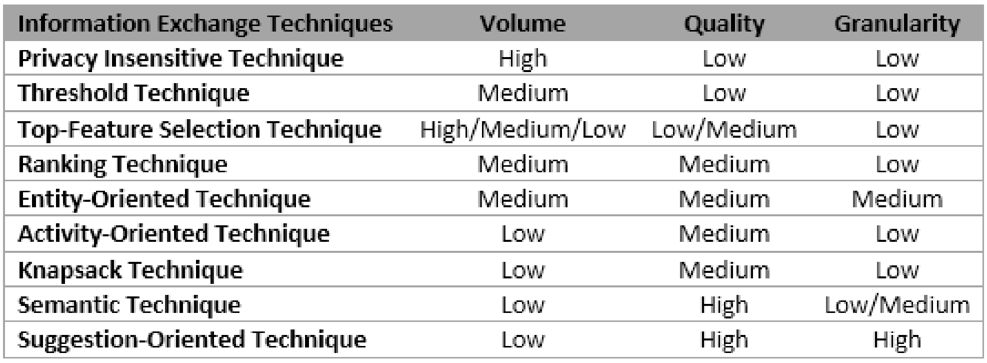
Figure 3.11. *Suggestion-Oriented Exchange Mechanism*

Effectiveness: this technique is one of the most effective, second only to the semantic paradigm (which performs semantic reasoning). Its ability to provide tailored suggestions, as informed decisions, to target websites ensures an enrichment of the user browsing experience without the need to give away any personal information about them.

Efficiency: the considerations for the efficiency are very similar to those achieved by the semantic technique.

Privacy Concerns: no terms are sent to the different websites which only receive a list of links with related relevancy indications. Therefore, the method represents the highest point of the privacy-aware techniques and the best embodiment of the concept of data privacy. However, it is important to stress that, even though the suggestion-oriented is 'theoretically' the best paradigm for providing personalisation suggestions without giving any actual information about the user to target websites, it lacks applicability in a real world context, where web publishers are not willing to share their data without a significant amount of relevant user information in return.

To conclude the survey on the various information exchange techniques, a table (3.1) is presented below that summarises the comparison of the exchange techniques in relation to their effectiveness. Hence, for each technique three aspects related to the effectiveness are considered: Volume, quality, granularity. Volume refers to the amount of data returned to the target website by each information exchange technique. Quality refers to the ability of each information exchange technique to provide tailored information to target websites, therefore exchange techniques that return huge amounts of information tend to have a lower quality in the information they provide. Granularity refers to the level of detail of the information returned to the target website, therefore information exchange techniques that return only entities or informed decisions tend to have a higher level of granularity. Each aspect can take on the values 'High', 'Medium' and 'Low', keeping in mind that a 'Low' granularity refers to a high level of detail.



Information Exchange Techniques	Volume	Quality	Granularity
Privacy Insensitive Technique	High	Low	Low
Threshold Technique	Medium	Low	Low
Top-Feature Selection Technique	High/Medium/Low	Low/Medium	Low
Ranking Technique	Medium	Medium	Low
Entity-Oriented Technique	Medium	Medium	Medium
Activity-Oriented Technique	Low	Medium	Low
Knapsack Technique	Low	Medium	Low
Semantic Technique	Low	High	Low/Medium
Suggestion-Oriented Technique	Low	High	High

Table 3.1. *Effectiveness Comparison*

3.3.3 Summary

The section adapted the key technical elements of the service to an open domain use-case and introduced the communication patterns to exchange information between target websites within the Cross Site Browsing Space and the third party User Model Provider. The characteristics of the open domain use-case introduced are: an unclear problem space in which the user may gather information as part of a higher and not yet fully formed Cross Site information need, a usually unknown set of websites to address the need, unpredictable content types and a multitude of distinct interconnected Cross Site browsing sessions performed in different time frames. As discussed, the term identification component relies on a third party term identification service. This enables a higher level of flexibility in terms of webpage content usage when compared, for example, to closed domain spaces where a pre-extracted representation of a website is required and needs to be re-assessed whenever an extension, update or modification to the content of the website within the Cross Site Browsing Space occurs. It should be also noted that this design can be equally applied within closed domains. However, since closed domains are often spaces for closed enterprise website federations, it might be necessary to use specific content description models (as the ones obtained by pre-extracted representations of websites within the Cross Site Browsing Space) in order to maintain data security by not exposing content to third party term identification services.

The second part of the section described a set of information exchange techniques between target websites and the Cross Site User Modelling platform. These techniques range from highly generic to highly specific, providing different levels of information enrichment for both the websites and the service. For each technique, pros and cons related to privacy concerns were pointed out along with a technical review focusing on the effectiveness and efficiency of the exchange. A range of exchange paradigms were presented, however, no claim was made regarding an ideal or optimal solution, rather it presented both advantages and shortcomings that might be relevant depending on the context and the scope of the application that deploys them.

The chapter concludes with the following section, which describes the use-case implementation taking into account all the considerations made for open domains.

3.4 Use-Case Implementation

Whereas the previous sections (section 3.2 and section 3.3) described the design of a Cross Site information exchange service that can provide relevant user information to target websites to potentially address user information needs that span independently hosted websites, this section introduces the implementation of a prototype service, set in the cultural heritage domain and approached as if it were in the open domain space.

3.4.1 Prototype Implementation for the Cross Site Information Exchange Service

The subsection illustrates the implementation of a Cross Site information exchange service prototype within an open domain. First, the characteristics of open domains are briefly recalled. Then, the use-case defined in 3.2

follows with the relative considerations. The use-case is consistent with the user behaviours depicted in 4.1. After this, content considerations are analysed along with the introduction of the technological architecture and its components.

3.4.1.1 Open Domain Characteristics

Based on the use-case discussion presented in section 3.2, the following main features of open domains are listed:

- Users may not know the website they are browsing and may not be visiting it on a regular basis.
- The information needs of the user are usually generic and broad, leading to explorative and information-gathering behaviours.
- The type of websites within open domains range from large-scale enterprise websites to small-scale websites, like forums and blog based websites.
- Changes to the websites structure or content are usually un-planned and frequent.

3.4.1.2 Use-Case Considerations

The specific domain chosen for the use-case presented in this thesis is cultural heritage[11]. Two instances of the CULTURA 1641 website were

[11] The content for the use-case belongs to the CULTURA project, in particular the '1641 Depositions' section. It is used on two naive instances of the CULTURA 1641 website, where only the browse and search tools were implemented.

implemented, that provide witness testimonies mainly by Protestants, but also by some Catholics, from all social backgrounds, concerning their experiences of the 1641 Irish rebellion. Furthermore, the two instances are independently hosted.

The use-case process is as follows:

1. The user browses to a website related to an information gathering task in the cultural heritage domain.

2. The user authenticates a first time with the third party service.

3. The website tracks all user activities in the webpages along with the relevant text entities identified by the term identification component.

4. The user triggers the information exchange function, in anticipation of a subsequent personalisation by the target website, which provides relevant user data (based on the selected communication pattern) to the website and newly tracked user information to the service.

5. The user browses to a second website and, depending on whether they are already authenticated or not, authenticates or directly triggers the information exchange function, which should provide more tailored information to the website.

6. The process is then re-iterated many other times, without a strict order of execution.

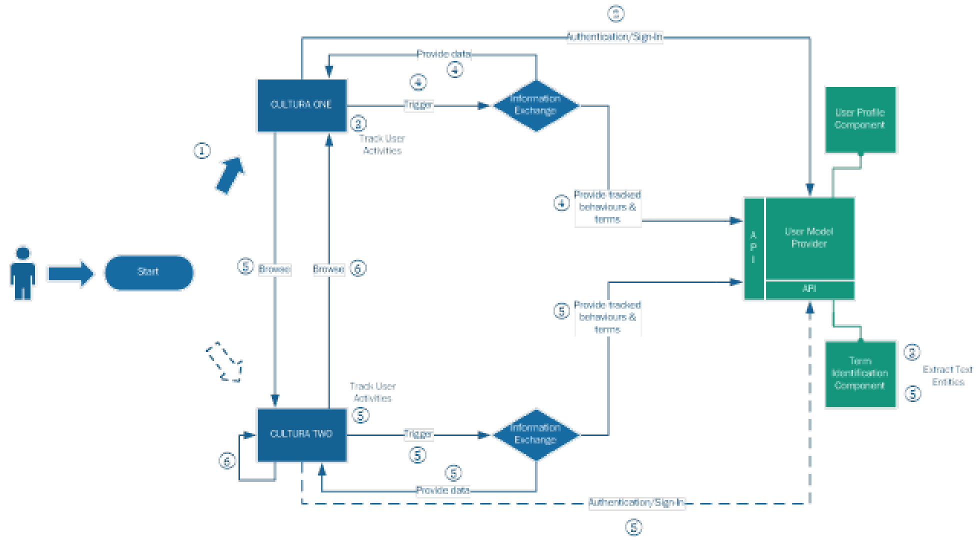


Figure 3.12. *High-Level Use-Case Process*

The use-case approach described above requests the following features in order to be applied:

1. Websites enable/allow third party sign-in/authentication by using OAuth[12].

2. The website needs to communicate with the User Model Provider. The communication includes:

   a. Sending of content browsed by users for term identification to a third party service.

   b. Enable/Allow browsing behaviour tracking.

   c. Sending of tracked user activities and extracted text entities to the User Model Provider.

[12] `http://tools.ietf.org/html/rfc5849`

3. The website has to use custom-built WCMS module extensions in order to access the Cross Site information exchange service RESTful API.

4. User authentication via website with the third party service is required in order to enable the information exchange mechanism and to protect users from unregulated treatment of their data.

The subsection which follows the prototype implementation of the Cross Site information exchange service is related to content considerations. After that, the technological architecture with its components is discussed.

3.4.2 Content Considerations

In an open domain use-case the content is usually not known at design-time. Therefore, since content can also change rapidly, it is necessary to implement a flexible approach to term identification. Such an approach has to consider content updates at run-time to ensure the text entities extracted by the term identification component are up-to-date and reflect the current state of the target websites. Then, to ensure accurate and just-in-time identification of representative terms of websites' content, a third party term identification service is triggered every time a user is browsing a web page within the Cross Site Browsing Space. For this, the WCMS extension module that tracks user behaviours within the currently browsed website, whenever a user lands on a new web page of a website belonging to the Cross Site Browsing Space, extracts the body of that web page and sends it to the third party term identification service, which analyses it and identifies the representative content-related text entities.

Moreover, it is also important to point out that the term identification service, like all the other tracking functions implemented by the WCMS extension module, gets activated only after the user has authenticated with the User Model Provider. The approach hence deploys a pluggable architecture that allows simple additions/replacements of external API based term identification services. This ensures higher flexibility, leading to a better quality of the content related text entities in the case of particular subject domains.

3.4.3 Service Architecture Implementation

The technological architecture for the Cross Site information exchange service is here presented. The implementation follows the guidelines set by the high-level architecture design in section 3.2 and it is depicted in Figure 3.13. For each architectural component, specific implementation details are discussed in the following. Furthermore, the implementation of two of the various information exchange techniques presented in section 3.3, that have been used for the evaluation experiment in section 4.2, is also discussed.

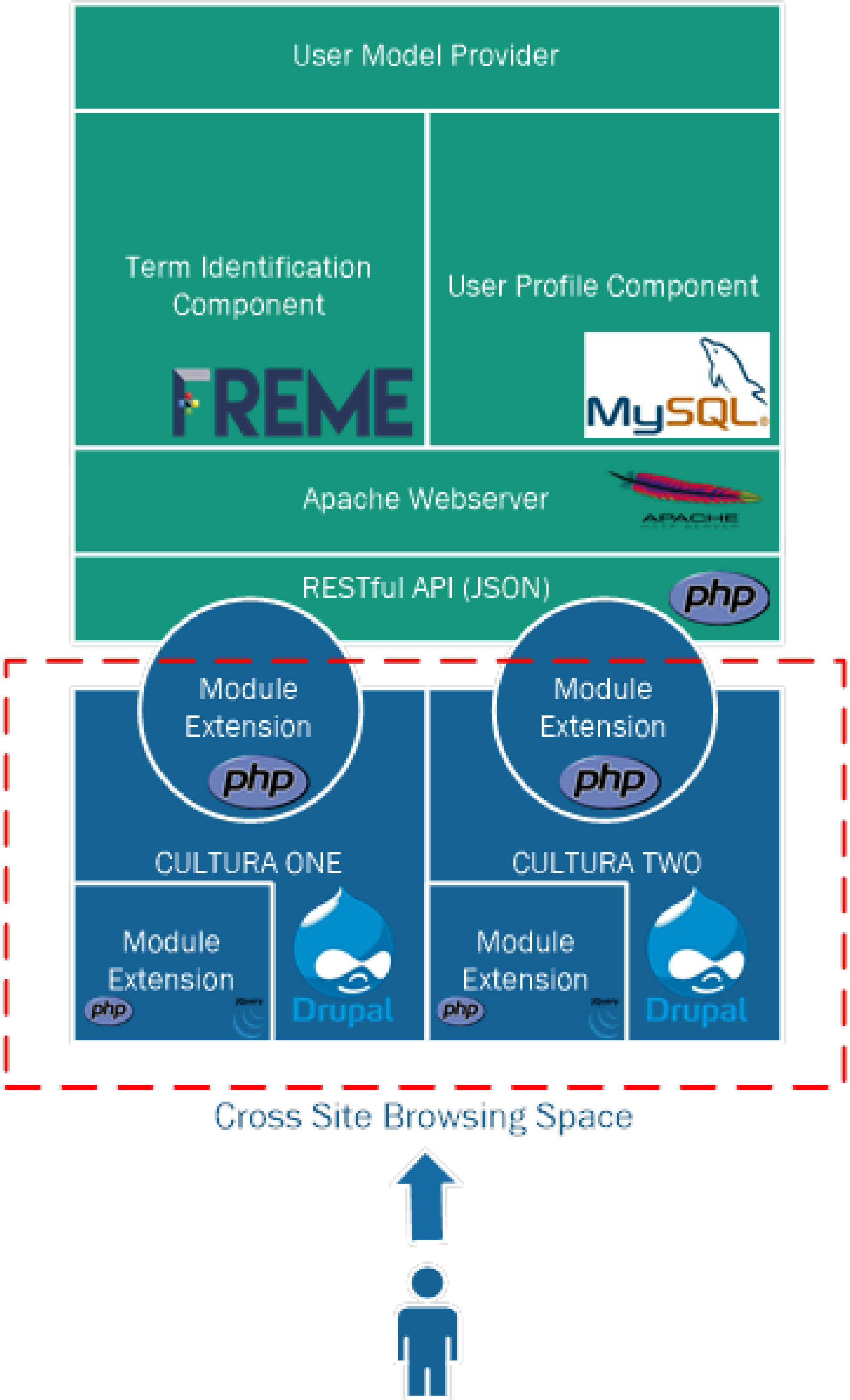


Figure 3.13. *Technical Service Architecture*

3.4.3.1 Term Identification Component

Since the WCMS module that tracks user activities and extracts text entities from related content needs to do so at run-time, the term identification component for the Cross Site information exchange technique has to enable the term identification as a fast and constant process. Furthermore, it should allow the identification of text entities across several subject domains due to the open domain nature of the use-case considered. As a future work notice, such a component should also allow the identification of text entities across different languages, enhancing the effectiveness of the system in multi-cultural domains. Based on this, a pluggable architecture was implemented. This allows the plugging of additional external term identification services. The current instance of the overall system used in the open domain use-case defined in section 3.2 interfaces with FREME[13] due to its open source nature. The FREME project aims at validating integration of multilingual and semantic technologies for digital content enrichment. These technologies are capable of processing (harvesting and analysing) content, capturing datasets, and adding value throughout content and data value chains across sectors, countries, and languages.

The result set of FREME includes term/entities pairs, indicating terms related to high-level entities such as Location, Person, Company etc. Associated with these term/entities pairs, a confidence parameter is also returned that should be interpreted as "the probability that a given term is of entity type XX", ranging from 0 to 1. Since the implementation uses two instances of the CULTURA 1641 website, the text entities extracted from

[13] http://www.freme-project.eu/

the content, which belongs to the cultural heritage domain, overlap. This overlap is important for the Cross Site interaction since ensures a shared conceptualisation and allows for a strengthening of the most extracted term/entity pairs.

The process of term identification is initialised as soon as a user authenticates with the third party User Model Provider. After the user authenticates, the WCMS module responsible to track user activities on the website starts to send the content body of the web page the user is currently browsing to the FREME API. The API then returns the term/entity pairs with their related confidence level in JSON format. The pairs are temporarily stored in the website as part of the user activity on the webpage. Whenever a user triggers the information exchange function that leads to the exchange of user data with the third party User Model Provider, the user activities stored in the website are sent to the third party service and the term/entity pairs are extracted and stored in a table representing all the text entities identified from the content the user has browsed since their last activation of the exchange technique, along with their confidence level and the timestamp of the user's landing in the webpage. In addition to these items, other ones, less relevant for the current scope of this thesis, are stored in the table too. Examples include resource references of the identified terms in the ontology used by the external term identification service[14] and a list of other term associated entities retrieved by FREME.

Other solutions can be implemented in order to extract text entities. One notable example, used by Koidl [Koi13] due to its wide usage [RT11],

[14] In this implementation use-case the Ontology used is DBpedia.

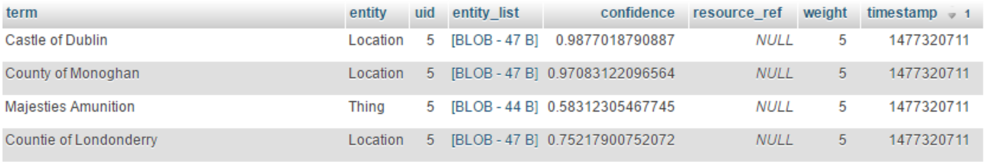
is OpenCalais[15]. In future work, OpenCalais could be implemented and compared with FREME in order to identify the best term identification component to use, depending on the subject domains addressed.

3.4.3.2 User Profile Component

The user profile is stored in MySQL[16] and stores the term/entity pairs related to the user's browsed pages (extracted using the term identification component), along with their weights based on the frequency with which the associated pairs are extracted, and the related browsing activities on the page (Figure 3.14). User activities are represented by a unique activity id, the IP address of the user, a unique user id, a unique website id, the content path within the website, the page scope (either visit or search), the referrer, the type of navigation to reach the page (either search or browse), the searched terms, the content title, a list of the inferred text entities related to the content browsed, click counter, scroll counter, cut and paste counter, a list of cut/copied/pasted sentences, key pressed counter, time spent on the page and timestamp. The timestamp is important to indicate when the last page access has taken place, allowing some information exchange techniques to ensure that more up-to-date activities are prioritised. Extracted text entities and user activities are sent by the website to the User Model Provider via the WCMS module extension endpoint of the API.

[15] http://www.opencalais.com/

[16] http://www.mysql.com/



term	entity	uid	entity_list	confidence	resource_ref	weight	timestamp ▾ 1
Castle of Dublin	Location	5	[BLOB - 47 B]	0.9877018790887	NULL	5	1477320711
County of Monoghan	Location	5	[BLOB - 47 B]	0.97083122096564	NULL	5	1477320711
Majesties Amunition	Thing	5	[BLOB - 44 B]	0.58312305467745	NULL	5	1477320711
Countie of Londonderry	Location	5	[BLOB - 47 B]	0.75217900752072	NULL	5	1477320711

Figure 3.14. *Snapshot of term-entity pairs from the User Profile*

As already mentioned, for user identification OAuth was implemented. The mechanism allows the website to remain open by allowing the user to sign-in to the Cross Site information exchange service via the website. Furthermore, the authentication mechanism warns the user (by means of a footnote in the authentication form) that by signing-in to the service they authorise the Cross Site information exchange service to track their behaviours, for the time they stay connected to it, and to store these tracked data permanently or until they decide to unsubscribe from the service.

3.4.3.3 Integration of Web-based Content Management System Module Extensions

In order to enable target websites to exchange user information with the User Model Provider, two interdependent custom-built Drupal 7[17] module extensions were deployed. The former was used to track all the user activities within web pages and to extract the content related text entities

[17] At the time of the implementation we started using Drupal 8. But, due to its short life (most module extensions were not yet compatible and many useful functions were still in development) we decided to move to Drupal 7 which is more grounded and stable thanks to its long life.

(PHP[18] and JavaScript[19] were used for the implementation). Hence, its priorities can be summarised in:

1. To enable term identification within target websites by sending openly accessible content to the term identification component of the User Model Provider once the user has signed in.
2. To enable user behaviour tracking once the user has signed in.
3. To delete all the stored user information in the website once the user logs out or their session ends.

The latter was used to activate the information exchange mechanism with the User Model Provider (PHP was used for the implementation). The role of the module is:

1. To provide a sign-in mechanism for users to enable the Cross Site service.
2. To request relevant user data from the third party service which can be subsequently used for personalisation by the website once the user is signed in and has authorised the treatment of their data.
3. To provide tracked user activities and extracted content related text entities to the User Model Provider once the user is signed in and has authorised the treatment of their data.
4. To delete all the tracked user information from the website's storage system after every activation of the information exchange mechanism.

[18] https://secure.php.net/

[19] https://github.com/tc39/ecma262

Drupal[20] was used for the system implementation based on its wide adoption. The motivation for using it and its module extensions is related to considerably low development times and flexibility in providing non-intrusive techniques that allow users to browse freely the websites, without being aversively guided.

The WCMS module extension that allows users to trigger the information exchange mechanism is focused on a block content element that provides a "personalisation" button which, when pressed by the user, triggers the information exchange mechanism (i.e. authentication process if the user is not signed in and communication pattern between the website and the User Model Provider). Therefore, prior to rendering the webpage after the user has clicked the button, the module requests/sends user information from/to the Cross Site information exchange service through one of the techniques presented in section 3.3. The returned user related term/entity pairs are rendered within a webpage of the website and are ready to be used for personalisation purposes.

3.4.3.4 Interface Layer Implementation

The interface layer is one of the central components of the Cross Site information exchange service. Its role is to provide a central access point for the different websites interacting with the User Model Provider. Therefore, for websites to communicate with the third party service and to ensure abstraction (i.e. not relying on specific website implementations or subject domains) the communication layer is implemented as a RESTful service

[20] `https://drupal.com/`

using PHP. The reason for this is that PHP is widely used in open domain development frameworks.

The interface between the third party User Model Provider and the different websites was facilitated through HTTP GET and HTTP POST requests. Websites (client-side) can use the HTTP GET method to request resources, that is the term/entity pairs and their related weights from previous browsing experiences, whereas the third party User Model Provider (service-side) can receive the current content browsed by the user, that is term/entity pairs extracted by the term identification component along with the related user activities, through a HTTP POST request instantiated by the website. Methods of the PHP PECL Oauth[21] package were used to access the third party User Model Provider API after the user authenticated in the service.

The term/entity pairs sent to the service are added to the user profile, and their weights are updated depending on the presence or not of the pair in the profile, via the user profile component. An example of the communication process can be observed in Figure 3.12 that presents the high-level use-case process.

3.4.3.5 Information Exchange Paradigms Implementation

The two information exchange techniques selected for integration are the ranking technique and the activity-oriented technique.

Client-side, the implementation is quite similar for both cases. Each method performs a single HTTP GET request to the service and the only real difference is the introduction, or not, of the personalisation suggestion (e.g.

[21] `https://pecl.php.net/package/oauth`

search-list personalisation). Service-side, however, the situation is different. Whereas the ranking technique performs a single database call, the activity-oriented technique needs to call the database twice: once for evaluating the activities (i.e. their weights) and once to retrieve the term/entity pairs related to the top-n activities.

It can be concluded that all the requirements related to the high-level design defined in section 3.1 have been addressed by the introduced use-case design and the subsequent prototype implementation.

In particular, regarding the Web User Dimension:

- An information exchange technique should provide a trade-off between web publisher's needs and user's needs. It should therefore balance the amount of user data exchanged with target websites in a way that is satisfying for both entities. [WU1]
  - Addressed by the design and implementation of the information exchange paradigms in sections 3.3 and 3.4.
- A Cross Site information exchange service should have a unified understanding of the user's browsing space by creating a shared conceptualisation of this understanding. [WU2]
  - Addressed by the term identification component and by the user profile component in sections 3.3 and 3.4.
- A Cross Site information exchange service should apply a user modelling technique that does not interrupt the user's browsing experience. It should hence be implicit, by not requiring the user to

explicitly participate in the identification, collection and management of information needs. [WU3]

  - Addressed by the WCMS module extension that tracks user behaviours and by the user profile component in sections 3.3 and 3.4.

Regarding the Web Content Dimension:

- A Cross Site information exchange service should be able to utilise existing content models to create a shared conceptualisation across different websites. [WC1]
  - Addressed by the term identification component and by the user profile component in sections 3.3 and 3.4.
- A Cross Site information exchange service should be able to use additional services and tools for term identification. [WC2]
  - Addressed by the term identification component in sections 3.3 and 3.4.
- Depending on the subject domain, a Cross Site information exchange service should be able to use different ontologies to better identify representative terms of content extracted by the term identification component. [WC3]
  - Addressed by the term identification component in section 3.4.
- A Cross Site information exchange service should be able to identify representative terms of content at run-time, therefore also on contents recently added or updated within a target website. [WC4]

  - Addressed by the WCMS extension module that tracks user behaviours in sections 3.3 and 3.4.

Regarding the Web Service Dimension:

- A Cross Site User Modelling platform should provide non-intrusive information exchange techniques to ensure the user is not hindered in freely browsing. [WS1]
  - Addressed by the WCMS module extension that triggers information exchange mechanisms in sections 3.3 and 3.4.
- A Cross Site information exchange service should affect the website the minimum possible, that is the exchange service should ensure that the website's look and feel is not aversively affected by the integration with it. [WS2]
  - Addressed by the WCMS module extension that triggers information exchange mechanisms in sections 3.3 and 3.4.
- A Cross Site information exchange service should not negatively influence the loading time of the website. [WS3]
  - Addressed by the interface layer in sections 3.3 and 3.4.
- A Cross Site information exchange service should be based on a design that allows simple integration and interfacing with existing websites. [WS4]
  - Addressed by the interface layer in sections 3.3 and 3.4.

- A Cross Site information exchange service should ensure that the communication between the Cross Site service and the target websites is flexible and does not depend on the websites technology stack. [WS5]
    - Addressed by the interface layer in sections 3.3 and 3.4.
- A Cross Site information exchange service should ensure device and browser independence. [WS6]
    - Addressed by the interface layer in sections 3.3 and 3.4.

3.4.4 Summary

This section discussed the implementation of a use-case based prototype for a Cross Site User Modelling platform for information exchange techniques. The use-case considerations were derived from the design requirements introduced in section 3.1. Furthermore, following the open domain approach defined in section 3.2 to design the service, the implementation relies on a content generic term identification method that utilises an external term identification service. In relation to WCMS module extension, two module extensions were introduced: one that tracks user activities on the webpages of websites within the Cross Site Browsing Space and one that triggers information exchange mechanisms whenever a user presses the 'personalisation' button in the website. Two instances of the CULTURA 1641 website were used for the implementation, based on Drupal 7. Finally, a sign-in mechanism based on OAuth was implemented to allow user identification within an open website.

In the following chapter the evaluation of the service prototype is performed and the derived results and findings are discussed.

4 Experimental Design and Evaluation

The objective of this thesis, derived from the overall research question introduced in chapter 1, is to identify effective information exchange paradigms for a Cross Site User Modelling platform that hosts user information gathered across different websites. Therefore, the following two goals were introduced: (1) identify the key requirements for the evaluation and (2) evaluate the appropriateness of the two most class-representative information exchange techniques designed in section 3.3 and implemented in section 3.4.

4.1 Introduction

To address the research objective, goals and challenges of this thesis a Cross Site User Modelling platform to test information exchange techniques was designed and then implemented, as described in chapter 3 (Design and Implementation). In this chapter, from the information exchange techniques (section 3.3), which represent the core aspect of the thesis itself, two techniques were selected for the implementation and testing. The two methods chosen are representative of the two classes defined in section 1.2: The Privacy-Insensitive class and the Privacy-Aware class. The evaluation of such techniques involves the following steps:

1. A preliminary identification of the criteria to be used to assess the designed techniques.

2. A preliminary definition of the case studies adopted for the evaluation phase. All the case studies were based upon browsing sessions in two instances of the CULTURA 1641 website.

3. The implementation of two selected information exchange techniques and their evaluation in terms of effectiveness and efficiency, using the criteria identified in subsection 4.2.1 and the formulas described in subsection 4.2.3. For this, three case studies were defined in subsection 4.2.2: the experienced user, the search oriented user and the browsing oriented user.

4. An analysis of the results and findings from these case studies using the implemented techniques.

The goal of this chapter is to provide detail on the evaluations conducted, as well as a discussion of the outcomes and findings. It is structured in the following manner: first is "Experimental Design" (section 4.2), which includes the evaluation criteria used, and the implementation of the selected information exchange techniques, then "Results and Findings" (section 4.3), which includes a discussion on the results and findings and a final comment.

4.2 Experimental Design

The following discussion is structured into three distinct subsections: (4.2.1) Assessment Criteria, (4.2.2) Definition of the case studies and (4.2.3) Experimental Framework.

4.2.1 Assessment Criteria

It is necessary to clearly define the assessment criteria which will be used to determine the effectiveness and efficiency of each information exchange technique.

4.2.1.1 Efficiency

The evaluation criterion is related to the number of HTTP requests, or to be more precise, the number of database calls from the target website to the third party service in those HTTP requests. In fact, the exchange mechanism performs a series of calls from/to the service database; what is important to remark is that, whereas the HTTP POST request always performs exactly one database call for all the information exchange methods, depending on the type of technique implemented, the number of HTTP GET requests required (or simply the number of database calls) may vary. Therefore, it is clear that the bottleneck of the mechanism lies in the number of requesting operations. The efficiency criterion can then be defined as "the number of HTTP GET requests performed from the target website or, more precisely, the number of database calls required to retrieve Cross Site user information from the third party service". In addition to this, the performance evaluation needs to take into account the possible overhead due to the learning techniques that some of the methods implement.

This criterion also allows for scalability considerations. Indeed, the techniques that perform a low number of requests are supposed to scale better compared to the others. The reason behind this must be sought in those situations where the number of concurrent requests from different users and/or target websites is considerably high (see Figure 4.1). A high number of requests for each single interaction might lead to performance decay due to the synchronization overhead with the service database. Therefore, it becomes imperative to find effective techniques that are able to keep the number of calls to the database to a minimum.

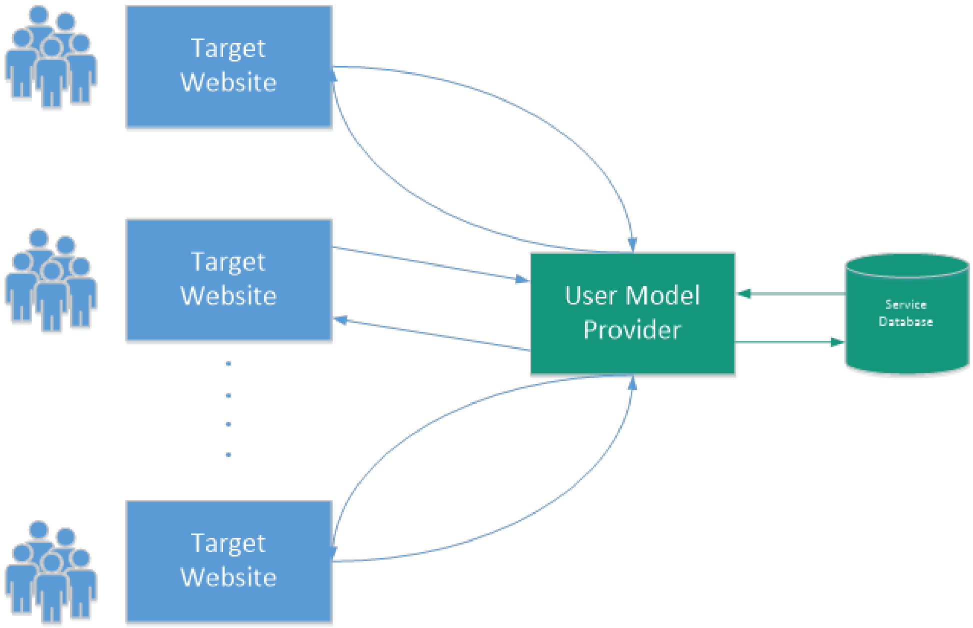


Figure 4.1. *Concurrent requests*

4.2.1.2 Effectiveness

There is no canonical formula that can be used to measure the effectiveness of the information exchange paradigms. Since the exchange provides Cross Site user information to target websites for personalisation, even assessing the effectiveness of website personalisation techniques wouldn't be appropriate. In fact, it couldn't be possible to separate the evaluation of the personalisation from the evaluation of the information exchange, leading to an assessment criterion that can be summed up as "Does the website have sufficient information to personalise?" and which simply requires a binary "yes/no" answer. However, the implementation of Web Personalisation is out of the scope of this thesis and won't be considered.

Obviously, the most effective exchange method would be to provide the website with all possible user data; however, this gives no thought to user privacy or, for that matters, to the business model of the third party service (the research work is not focused on real world deployment and hence it does not consider business aspects – if not slightly). Therefore, it is reasonable to consider providing all user data to the website as the gold standard for effectiveness, everything we try to do on top of that which honours user privacy will likely impact effectiveness to some extent, the key will be finding the balance. To summarize, the effectiveness criterion can be defined as: "Set the Privacy-Insensitive methods as the reference point, the effectiveness of the Privacy-Aware techniques is evaluated by comparing the quality of the information provided. Where quality means, the relevance of the user data provided".

4.2.2 Definition of Case Studies

As already mentioned, all the case studies base their browsing sessions on the two instances of the CULTURA 1641 website, which belongs to the cultural heritage domain. For the purposes of the evaluation, three case studies were defined: the experienced user (4.2.2.1), the search-oriented user (4.2.2.2) and the browsing-oriented user (4.2.2.3). Each is described in detail, below.

4.2.2.1 Experienced User

The user has browsed using the service for a long time, therefore the user profile has gathered a considerable amount of information related to that particular user. In this way the type of information exchange technique implemented heavily impacts on the user's privacy concerns, either the user can be considered search-oriented or browsing-oriented.

For experienced users, thanks to the great amount of data stored in the Cross Site service, evaluation results are expected to be significant both in the case the user is search-oriented and in the case the user is browsing-oriented.

4.2.2.2 Search-Oriented User

The user tends to prefer searching over browsing, therefore the amount of search operations is far greater than the amount of browsing operations. In this way the user is labelled as 'impatient' and, for those techniques that provide personalisation suggestions (see section 3.3: "Design of Information Exchange Techniques"), the exchange paradigm should perform a suggestion indicating to the website to personalise the search-results list for subsequent user interactions. Such suggestion will be returned as a 'search-list personalisation' string.

Search-oriented users tend to be more aware of their information needs, performing a series of focused search operations that narrow down the space of term/entity pairs retrieved by the information exchange techniques. Thus, evaluation results are expected to be satisfying, thanks to the aforementioned 'aware of the information needs' user behaviour, regardless of the volume of user data held by the service, which is not as significant when compared, for example, to that of experienced users.

4.2.2.3 Browsing-Oriented User

The user tends to prefer browsing over searching, therefore the amount of browsing operations is far greater than the amount of search operations. In this way the user is labelled as 'explorative' and, for those techniques that provide personalisation suggestions (see section 3.3: "Design of Information Exchange Techniques"), the exchange paradigm should perform a suggestion indicating to the website to personalise web pages (e.g. introducing personalised links, different type of contents, etc.) for subsequent user interactions. Such suggestion will be returned as a 'navigational-browse personalisation' string.

Browsing-oriented users tend to be less aware of their information needs, browsing the website in a more aimless way that leads to a plethora of unrelated term/entity pairs retrieved by the information exchange techniques. Thus, evaluation results are expected to be less satisfactory when compared to previous case studies, due to the generic and vague user behaviour and to the scarcity of Cross Site user information held by the service.

The information exchange techniques implemented will be evaluated and assessed according to each user-focused case study. The results are expected to be different according to the behaviour defined by the specific case study.

4.2.3 Experimental Framework

Following section 3.4, the two information exchange techniques selected to perform the evaluation are the ranking technique and the activity-oriented technique. The former can be considered the best method belonging to the class of the Privacy-Insensitive techniques while the latter can be considered the best method among the feasible ones of the Privacy-Aware class (the implementation of techniques that use learning algorithms is out of this thesis's scope). In this way, it is possible to compare the best approaches from both classes and make considerations on some of the issues posed in the research question, that is about the achievement of effective results while trying to preserve data privacy.

To compute the total execution time of the exchange, a timer, which starts right before the first HTTP GET request and finishes right after the HTTP POST request, has been introduced on the client-side. The calculation of the execution time is made in order to verify the efficiency criterion introduced in subsection 4.2.1, that is if the number of HTTP requests (or, better to say, database calls) performed against the service affects the overall performance of the information exchange mechanism. Theoretically, the total execution time of the exchange should be less in the case of the ranking algorithm, which performs a single database call against the two of the activity-oriented approach.

With regard to effectiveness, the activity-oriented technique has been compared with the ranking technique in order to evaluate the effectiveness based on the number of 'relevant' term/entity pairs retrieved. As previously mentioned, the ranking method, belonging to the class of privacy insensitive paradigms, can be considered as the technique that provides the most relevant term/entity pairs to target websites. Therefore, with the assumption that privacy insensitive techniques provide the most relevant user information (at the price of not preserving user data privacy) to websites, the evaluation is reduced to assess how close, in terms of effectiveness, the Privacy-Aware techniques perform to the aforementioned ones.

In the context of this experiment, the pairs to be evaluated have been defined as the top-n pairs (n has been set to 15[1]). Hence, in order to obtain consistent results, three information retrieval evaluation measures have been used: Average Precision [Zhu04], Precision and Recall [Rij79]. Moreover, the top-n pairs of the ranking technique have been set as the Recall Base (i.e. the total number of relevant term/entity pairs that can be retrieved); thus the top-n pairs of the activity-oriented technique are matched against the Recall Base to see how well the method performs compared to the ranking one. Since both the techniques return the results ranked in descending order from the most relevant, the computation of

[1] A value of 10 for n has been considered too strict (some high-weight terms were left out from the evaluation), whereas a value of 20 for n has been considered too loose. Hence, the value 15 has been selected as a trade-off.

Average Precision have been done using the term/entity pairs retrieved by the activity-oriented technique but positioning them as if they were retrieved by the ranking technique. In this way, we evaluate the pairs considering the actual position they should have in a perfect result list (i.e. the result list from the ranking method).

It is also important to note that, since the amount of pairs considered for the evaluation is 15, the number of retrieved pairs is equal to the recall base (except in extraordinary cases where the number of retrieved pairs is less than 15 due to a lack of content in the top-n activities[2]) and Precision and Recall achieve the same result, even though the meaning of that result has different implications. In fact, Precision is the probability that a retrieved pair is relevant, whereas Recall is the probability that a relevant pair is retrieved. Thus, the formulas are:

$$Prec = \frac{|P^{*} \cap P|}{|P|}, \tag{4.1}$$

$$Rec = \frac{|P^{*} \cap P|}{|P^{*}|}. \tag{4.2}$$

Where $|P|$ is the number of retrieved pairs, $|P^{*}|$ is the number of relevant pairs (RB) and $|P^{*} \cap P|$ is the number of retrieved relevant pairs.

Precision and Recall are single-value metrics based on the list of pairs returned by the service. Since the service returns a ranked sequence of pairs, it is desirable to also consider the order in which the returned pairs would

[2] However, these exceptions can be safely considered as outliers and hence excluded from the evaluation.

be in the ranking technique result list, so as to obtain more grounded results. This is where comes in Average Precision:

$$AvePrec = \frac{1}{RB} \sum_{i \in R} Prec(i) \tag{4.3}$$

Where RB is the Recall Base, R is the set of the rank of the relevant pairs and $Prec(i)$ is the Precision at level i[3].

Hence, for each of the three case studies defined in subsection 4.2.2, the two techniques have been tested, the term/entity pairs have been compared, by means of Precision, Recall and Average precision, and the execution times have been computed. It is important to stress that the execution times presented are equivalent to the arithmetic average of a hundred iterations.

The results are shown separately for each user-focused case study:

4.2.3.1 Experienced User

First, the execution times are compared.

[3] Precision at n: $Prec(n) = \frac{1}{n} \sum_{m=1}^{n} a_m$, where a_m is the relevance of the pair p_m (0/1 binary relevance judgement).

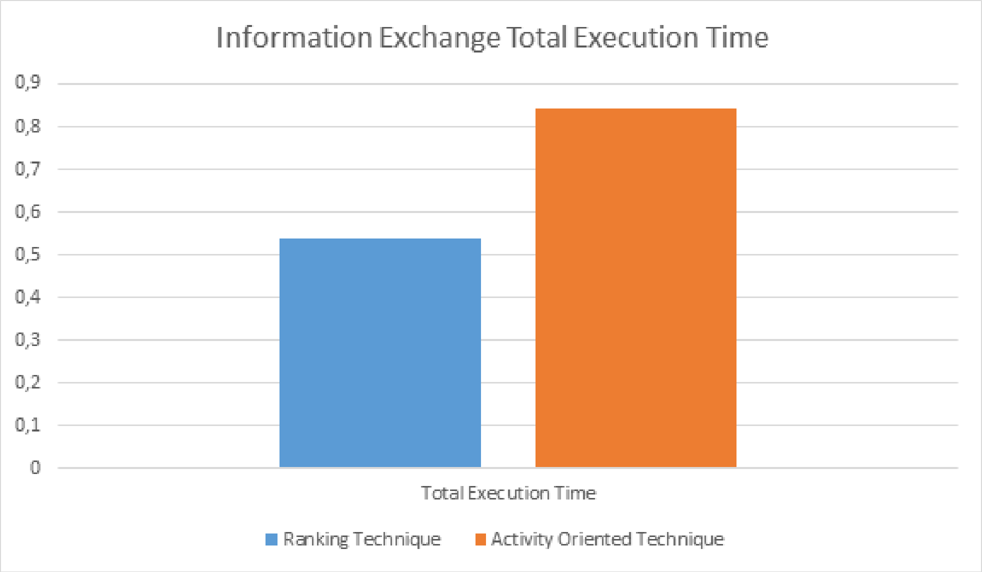
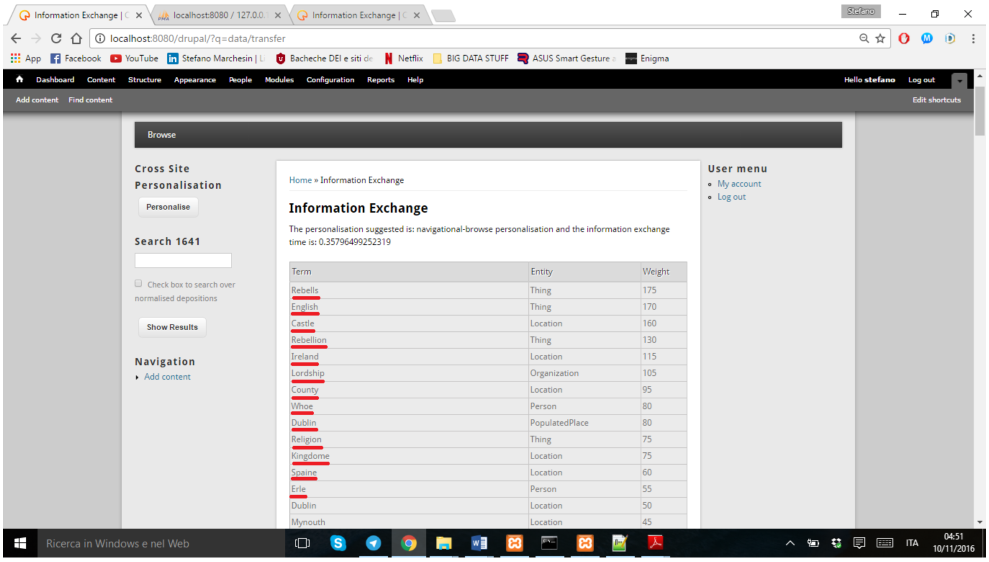


Figure 4.2. *Experienced User: Execution time comparison*

As expected, the activity-oriented technique performs worse due to the higher number of database calls operated (two against one). However, such a worsening is almost imperceptible from a user perspective, being approximately 300 milliseconds.

Then, the term/entity pairs are analysed.



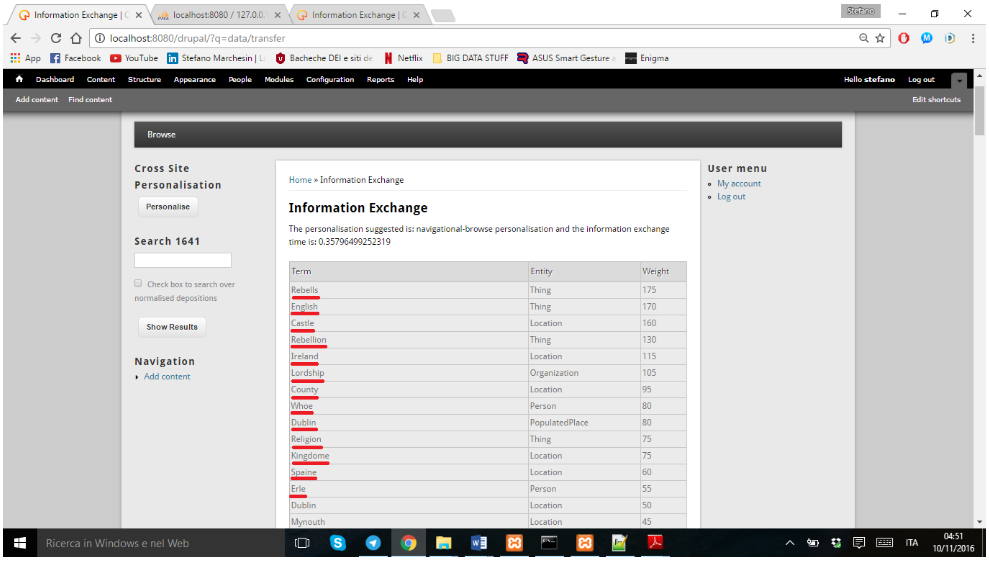
Home » Information Exchange

Information Exchange

The personalisation suggested is: navigational-browse personalisation and the information exchange time is: 0.35796499252319

Term	Entity	Weight
Rebells	Thing	175
English	Thing	170
Castle	Location	160
Rebellion	Thing	130
Ireland	Location	115
Lordship	Organization	105
County	Location	95
Whoe	Person	80
Dublin	PopulatedPlace	80
Religion	Thing	75
Kingdome	Location	75
Spaine	Location	60
Erle	Person	55
Dublin	Location	50
Mynouth	Location	45

Figure 4.3. *Experienced User: Activity-Oriented Technique*



Home » Information Exchange

Information Exchange

The personalisation suggested is: navigational-browse personalisation and the information exchange time is: 0.35796499252319

Term	Entity	Weight
Rebells	Thing	175
English	Thing	170
Castle	Location	160
Rebellion	Thing	130
Ireland	Location	115
Lordship	Organization	105
County	Location	95
Whoe	Person	80
Dublin	PopulatedPlace	80
Religion	Thing	75
Kingdome	Location	75
Spaine	Location	60
Erle	Person	55
Dublin	Location	50
Mynouth	Location	45

Figure 4.4. *Experienced User: Ranking Technique*

From the top-15 results returned by the exchange methods, it can be seen that the user is interested in the 1641's Irish rebellion, with terms like "Rebellion", "English", "Ireland" that appear in the result lists of both the techniques. In particular, as can be seen from the ranking technique, the focus of the user is directed to the counties of Dublin and Mayo. However, the information about the county of Mayo gets lost by the activity-oriented technique that retrieves a number of 'relevant' pairs equal to 13 out of 15, but it misses precisely the terms related to Mayo. Hence, the activity-oriented method is able to provide data to address the general information need of the user but it lacks some of the more specific information required to fully answer all the nuances related to this need.

From a technical perspective, Average Precision and Precision/Recall formulas have been computed to evaluate how good the activity-oriented paradigm performs compared to the ranking technique (which is assumed to produce the "perfect run"):

$AvePrec = \frac{1}{15} \times (1+1+1+0+\frac{4}{5}+\frac{5}{6}+\frac{6}{7}+\frac{7}{8}+\frac{8}{9}+\frac{9}{10}+\frac{10}{11}+\frac{11}{12}+\frac{12}{13}+0+\frac{13}{15}) = 0.7847$

$Prec/Rec = \frac{13}{15} = 0.8667$

All three measures return values above 0.7, implying that the activity-oriented technique provides relevant user information to target websites. Moreover, since the experienced user has been labelled as 'explorative' by the technique (the personalisation suggestion is "navigational-browse personalisation"), the results also confirm that, when the volume of Cross Site user information held by the service is considerable, the effectiveness of the exchange paradigm in relation to browsing-oriented users is high.

4.2.3.2 Search-Oriented User

First, the execution times are compared.

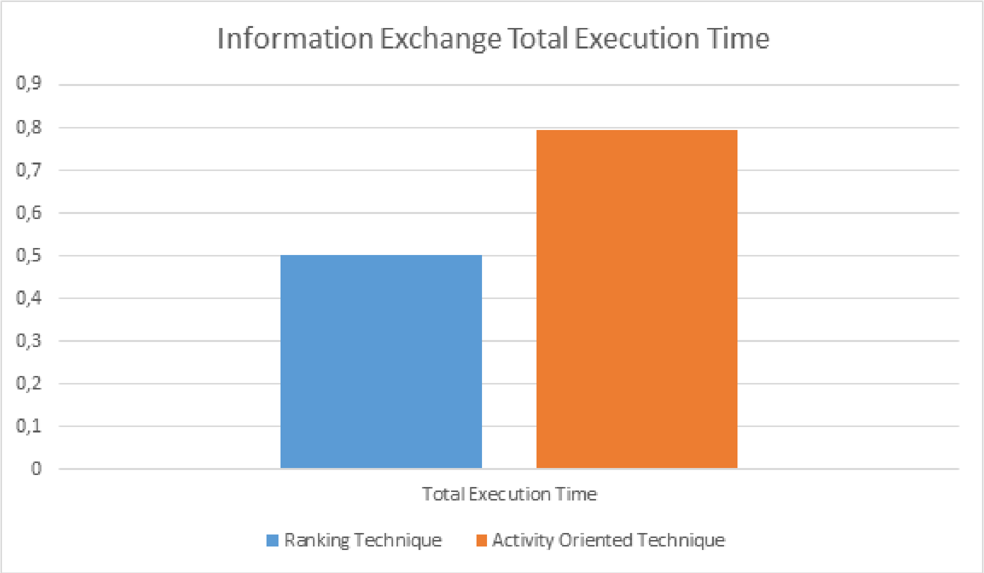
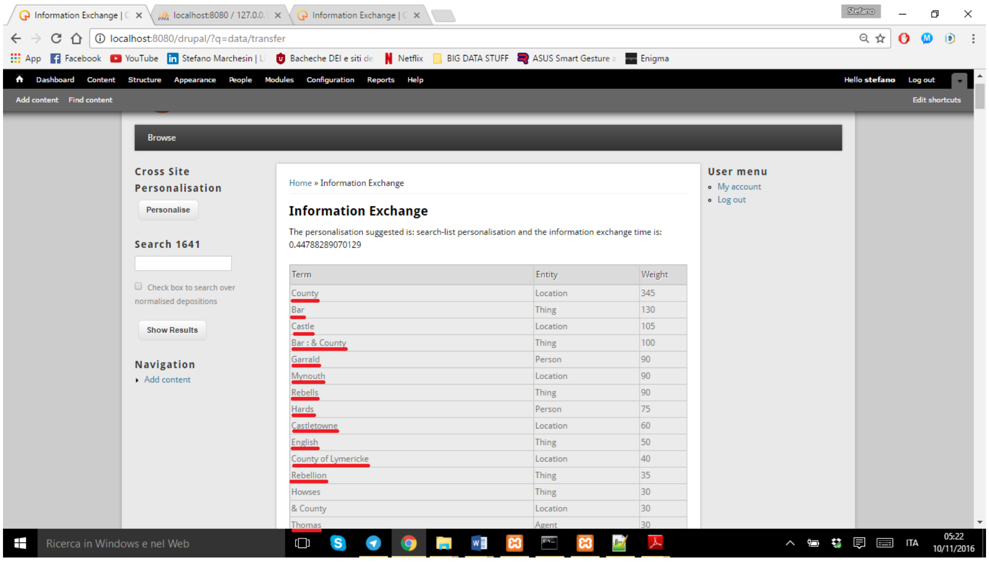


Figure 4.5. *Search-Oriented User: Execution time comparison*

As before the activity-oriented technique performs worse than the Ranking Technique, confirming once again that a higher number of calls to the service's database pauperises performances. Nevertheless, the worsening can still be considered imperceptible from the user perspective (nearly 300 milliseconds).

With regard to the term/entity pairs,



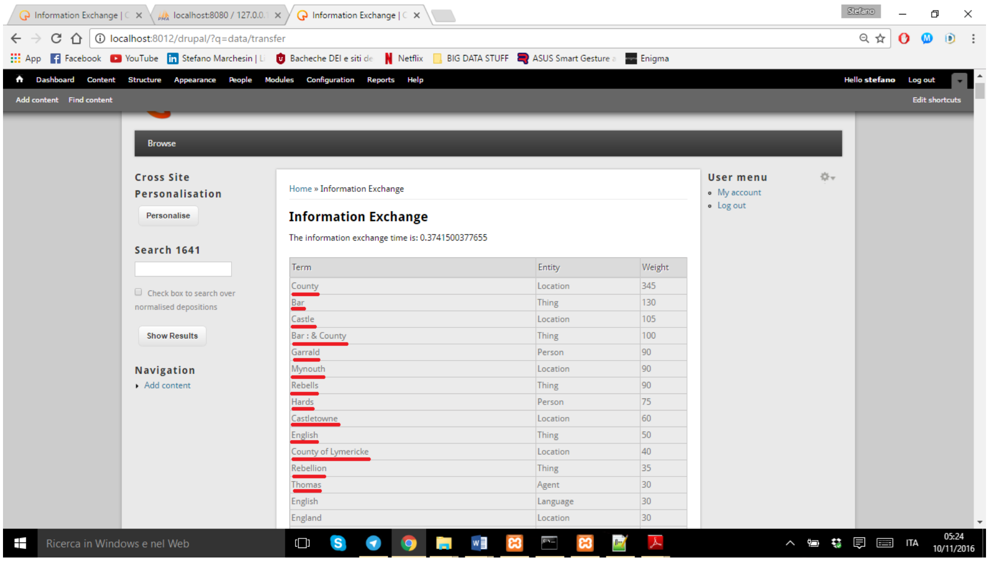
Home » Information Exchange

Information Exchange

The personalisation suggested is: search-list personalisation and the information exchange time is: 0.44788289070129

Term	Entity	Weight
County	Location	345
Bar	Thing	130
Castle	Location	105
Bar : & County	Thing	100
Garrald	Person	90
Mynouth	Location	90
Rebells	Thing	90
Hards	Person	75
Castletowne	Location	60
English	Thing	50
County of Lymericke	Location	40
Rebellion	Thing	35
Howses	Thing	30
& County	Location	30
Thomas	Agent	30

Figure 4.6. *Search-Oriented User: Activity-Oriented Technique*



Home » Information Exchange

Information Exchange

The information exchange time is: 0.3741500377655

Term	Entity	Weight
County	Location	345
Bar	Thing	130
Castle	Location	105
Bar : & County	Thing	100
Garrald	Person	90
Mynouth	Location	90
Rebells	Thing	90
Hards	Person	75
Castletowne	Location	60
English	Thing	50
County of Lymericke	Location	40
Rebellion	Thing	35
Thomas	Agent	30
English	Language	30
England	Location	30

Figure 4.7. *Search-Oriented User: Ranking Technique*

it can be observed that the top-15 term/entity pairs returned are mainly related to Irish counties and public places (e.g. "Bar" and "Castle"). Moreover, retrieved terms like "Mynouth" or "Castletowne" refer to specific locations in the county of Kildare. Therefore, it can be assumed that the user information need is related to the socio-cultural context of the Irish counties during the 1641's rebellion, hypothesis reinforced by retrieved terms such as "Rebellion", "Rebells" and "English" (which, however, are only boundary to what the real topic is). In this second case study, even though the total number of 'relevant' term/entity pairs returned is still 13 out of 15, all the relevant terms required to identify the main topic and its nuances are retrieved. In fact, the last two pairs of the ranking technique result list, which have not been retrieved, are not relevant for the topic outlined from all the other pairs. Hence, the activity-oriented method is able to address both the general and the specific information needs of the user, providing the same qualitative results of the ranking method while better preserving user privacy.

Moving to the technical side, Average Precision and Precision/Recall formulas have been computed:

$AvePrec = \frac{1}{15} \times (1+1+1+1+1+1+1+1+1+1+1+1+1+0+0) = 0.8667$

$Prec/Rec = \frac{13}{15} = 0.8667$

All three measures return values above 0.8, implying that the activity-oriented technique provides relevant user information to target websites. Moreover, since the user has been labelled as 'impatient' by the technique (the personalisation suggestion is "search-list personalisation"), the results also confirm that the user is aware of what is looking for, performing more tailored searches and therefore obtaining more focused term/entity pairs (the high value of the first position pair is an indicator for this) back from the service, even without a vast amount of information stored in it. The high quality of the results leads to state the soundness of the technique.

4.2.3.3 Browsing-Oriented User

Hence again, the execution times are compared.

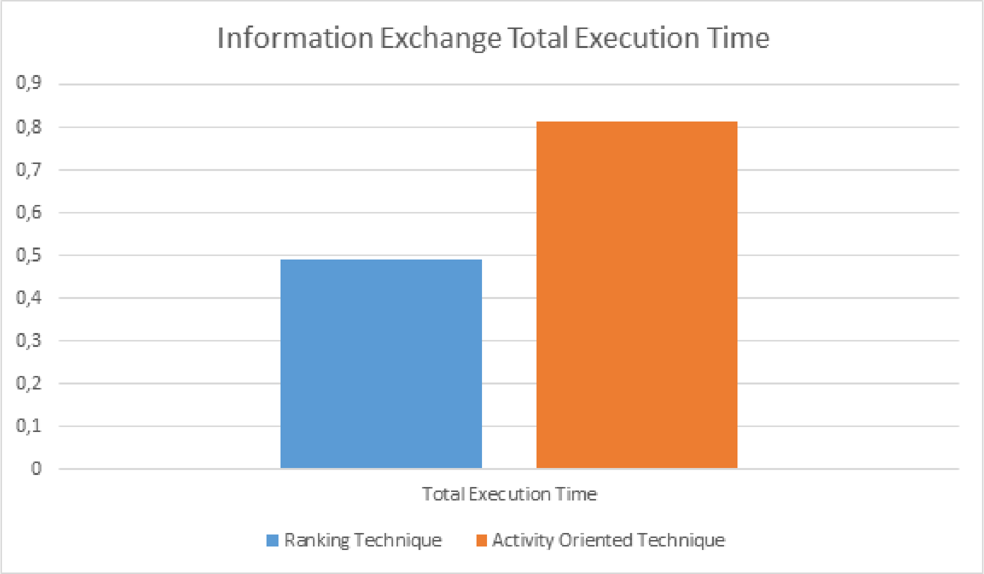
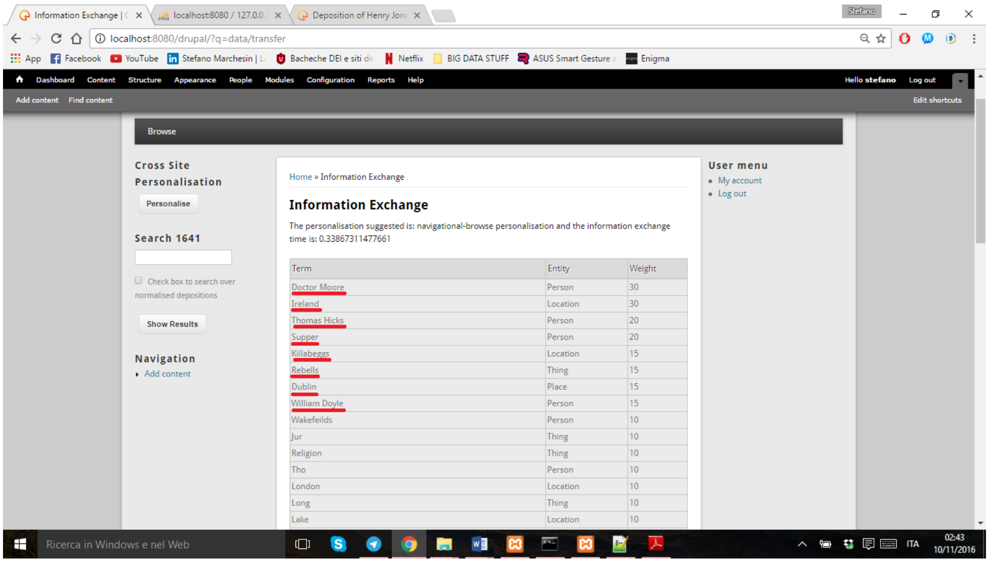


Figure 4.8. *Browsing-Oriented User: Execution time comparison*

Also in this third case, the same type of considerations that apply for the first two user-focused case studies can be done.

For what concerns the term/entity pairs,



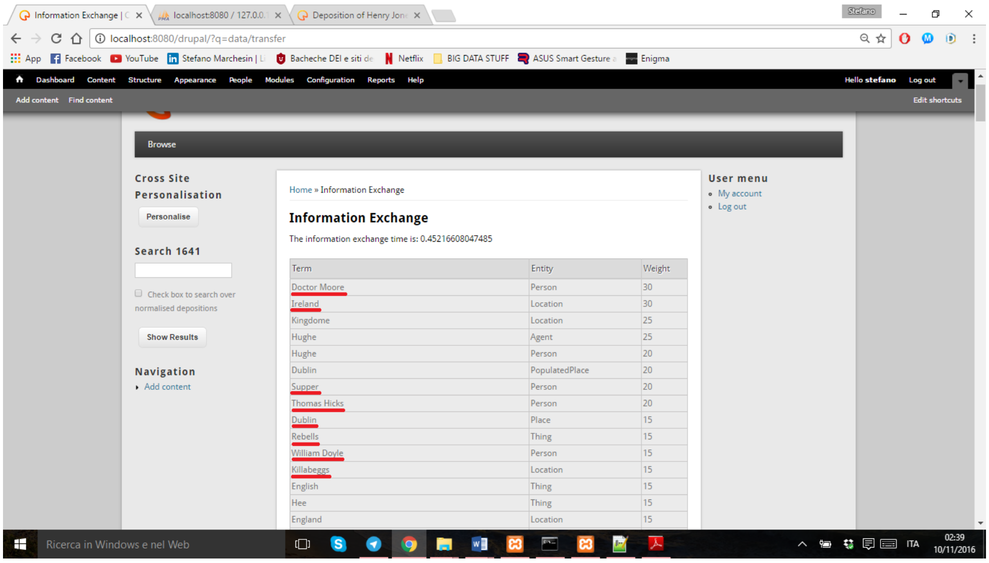
Home » Information Exchange

Information Exchange

The personalisation suggested is: navigational-browse personalisation and the information exchange time is: 0.33867311477661

Term	Entity	Weight
Doctor Moore	Person	30
Ireland	Location	30
Thomas Hicks	Person	20
Supper	Person	20
Killabeggs	Location	15
Rebells	Thing	15
Dublin	Place	15
William Doyle	Person	15
Wakefeilds	Person	10
Jur	Thing	10
Religion	Thing	10
Tho	Person	10
London	Location	10
Long	Thing	10
Lake	Location	10

Figure 4.9. *Browsing-Oriented User: Activity-Oriented Technique*



Home » Information Exchange

Information Exchange

The information exchange time is: 0.45216608047485

Term	Entity	Weight
Doctor Moore	Person	30
Ireland	Location	30
Kingdome	Location	25
Hughe	Agent	25
Hughe	Person	20
Dublin	PopulatedPlace	20
Supper	Person	20
Thomas Hicks	Person	20
Dublin	Place	15
Rebells	Thing	15
William Doyle	Person	15
Killabeggs	Location	15
English	Thing	15
Hee	Thing	15
England	Location	15

Figure 4.10. *Browsing-Oriented User: Ranking Technique*

it is possible to see that the total amount of retrieved relevant pairs, for the activity-oriented technique, is only 8 out of 15. The result is coherent with the user behaviour which, in this third case study, is labelled as 'explorative' (the personalisation suggestion is "navigational-browse personalisation") and therefore is less focused and aware than the 'impatient' behaviour, leading the service to store generic and vague user information. In fact, the pairs retrieved by the techniques have low weights and are uncorrelated, making hard to provide effective exchanges.

As further evidence of what has been said, Average Precision and Precision/Recall formulas have been computed:

$AvePrec = \frac{1}{15} \times (1+1+0+0+0+0+\frac{3}{7}+\frac{4}{8}+\frac{5}{9}+\frac{6}{10}+\frac{7}{11}+\frac{8}{12}+0+0+0) = 0.3591$

$Prec/Rec = \frac{8}{15} = 0.5333$

The values here demonstrate that in the case of a relatively new user who shows an explorative behaviour, the technique performs considerably worse than before. The reason is the low volume of user data held by the service. In fact, whereas an impatient user already knows what to search, an explorative one tends to browse in the website without having precisely in mind what to look for. Therefore, when the explorative user is not experienced, the quality of the few data stored in the service is heavily affected by their behaviour, leading to less effective exchanges.

As a last note, the reason why Average Precision performs worse than Precision is due to the fact that it also considers the order in which the pairs are returned from the service, penalising more result-lists that omit pairs in high positions.

4.3 Results and Findings

Overall the performance results of the experiment were encouraging. Related to the issues posed in the research question the following findings are relevant:

- The number of HTTP (GET) requests, or more precisely, the number of database calls impacts on the performances of the overall system. However, the impact is minimal and in the local context where the experiments have been tested does not pose a tangible problem to the efficiency of the system. Therefore, more high-scale experiments, that are out of scope for this thesis work, need to be executed in order to verify if the current imperceptible delay can turn into a major issue when multiple concurrent requests are sent to the service. Nevertheless, it is still reasonable to affirm that in order to avoid possible problems related to scalability it is necessary to implement techniques that operate a low number of service requests, thus decreasing the overhead due to concurrent operations on the system. So far, as already mentioned, only suppositions about the service behaviour on high scale can be made – actual implementations and experiments are left as future work. In conclusion, it can be said that on small-scale the system performs as expected even though such behaviour doesn't really affect the performances, being it sensed as imperceptible by the user and thus leading to a delay that is immaterial.

- The Privacy-Aware technique performs above 0.7 in the cases of experienced and search-oriented users, providing relevant user information to target websites, while trying to protect user privacy. However, as already stated, in the browsing-oriented case study, due to the low volume of data held by the service, the technique performs below 0.5 not being able to provide relevant information to target websites as effectively as in the other cases studies. These results confirm the different user behaviours and the impact that different volumes of data held by the service have on each case study. Moreover, formulas' values above 0.7 answer the issues of the research question related to the relevancy of the results returned, while honouring the user privacy mitigating the flow of personal information out from the service.

- The enhancement of the User Model held by the service is not covered by the evaluation since it has been answered by the implementation of the system itself. In fact, by providing to the service, each time that the user triggers the information exchange function, all the user activities tracked in the target website and the term/entity pairs extracted from the browsed pages, makes the flow website-service satisfying, thus avoiding the need for evaluation.

Hence, we can safely say that the evaluation answers the problems posed in the research question related to privacy concerns and effectiveness of the exchange, thanks to the use of IR-style techniques for the performance analysis, and also provides a starting point for future studies on the efficiency of the system when deployed on large scale. Moreover, it is capable of giving a first glimpse of the system's potential, especially of its central role in what can be a richer customisation environment that includes semantical-aware learning algorithms (service-side) and web personalisation techniques (client-side).

5 Conclusions

This thesis introduced an adaptive Cross Site User Modelling platform for information exchange techniques. The approach introduced served as a starting point to address an identified gap in the State of the Art relating to the research field of Web Personalisation. Precisely, the gap identified in section 2.1 relates to the fact that current approaches assist users only within single websites and not across independently hosted websites. Thus, the approach fits in the Cross Site Personalisation context, through the use of a third party User Model Provider and Web-based Content Management System (WCMS) extensions.

The chapter discusses the objectives and achievements of this thesis (sections 5.1 and 5.2). Then, a discussion on future work (section 5.3) and a brief summary (section 5.4) conclude the thesis.

5.1 Restatement of Research Question and Objectives

Introduced in chapter 1 (section 1.2), the research question that drove this research work was:

How can an exchange of information be provided between a target website and the third party User Model Provider that is satisfying both for the target website and the third party service, which also limits the flow of user's information from the third party User Model Provider to the website thus honouring users' privacy needs?
By "satisfying" it is meant "to what extent the information exchange can provide to the target website relevant Cross Site user information that can be potentially used to address user Cross Site information needs and, on the other hand, enhance the User Model held by the service".

The objective related to this research question was also defined in chapter 1 (section 1.2):
To address the above research question, the objective of this thesis is to identify and evaluate the most effective information exchange paradigms for a third party Cross Site User Modelling service, which provides information that supports websites in assisting users to address information needs that span independently hosted websites.

Therefore, three main goals were identified to address the research objective:

- To build an API that provides to the target website the ability to access and question the User Model offered by the User Model Provider; (G1)
- To identify users' activities on website and the information needs that

drive those users to browse the website; (G2)

- To decide which information about the user to provide to the target website and whether to offer this information in its entirety or in smaller chunks, depending on the context of the application that utilises the Cross Site service. (G3)

The resulting design and research challenges were:

- Interaction
  - How does a website request user information from the third party provider?
  - How does the third party User Modelling service get value? What information does each website pass back to enhance the User Model?
  - How is privacy implemented? At which level?
- Representation
  - What vocabulary is used to describe the characteristics in the User Model?
  - What type of ontology can be used? Generic or Domain Specific?
  - What is the nature of the user information?

In the next section, a discussion on each objective and how well it was achieved, based upon the State of the Art and the Evaluation chapters, is presented.

5.2 Achievements

5.2.1 Research Question

The research question that drove the entire research work was addressed by introducing the design, implementation and evaluation of an adaptive Cross Site User Modelling platform for information exchange techniques. The platform consists of a third party User Model Provider which interfaces with target websites within the Cross Site Browsing Space to exchange useful information both for the websites (Cross Site user information) and the service (tracked user activities and content related text entities), that can be potentially used to address user's Cross Site information needs. Module extensions for Web-based Content Systems were developed with the goal to enable non-intrusive information exchange techniques. The overall system (service and WCMS extension modules) was evaluated through the assessment of the objective posed in section 1.2, relying both on an indicative qualitative measure and on quantitative research techniques. The results obtained from the evaluation, accordingly with the case studies introduced in section 4.2, confirmed that the introduced communication paradigms (i.e. the information exchange techniques) can provide useful information to both sides.

5.2.2 Research Objective

The objective of this thesis was achieved through the evaluation of three user-focused case studies set in an open domain use-case. All the three case studies were based on real-world content and user's behaviours. The outcome of the case studies related to the experienced user and the search-oriented user proved that it is possible to provide useful information to target websites which can be used to assist users in addressing information needs that span across independently hosted websites. On the other hand, the outcome of the case study related to the browsing-oriented user proved that in order to actually provide useful information, it is necessary to either have users with a great amount of personal data stored in the third party User Model or users that are highly aware of their information needs – which is reasonable.

5.2.3 Research Goals

(G1): To build an API that provides to the target website the ability to access and question the User Model offered by the User Model Provider.

To address this first goal, the design requirements introduced in section 3.1 were considered. Based on these design requirements, a high-level design of a Cross Site User Modelling service for information exchange techniques was introduced. The high-level design was based on a centralised third party User Model Provider that interfaces with the backend of the different websites the user browses within the Cross Site Browsing Space. Within this framework, the interface layer, that is the service component responsible for the communication between the websites and the User Model Provider, was based on a RESTful API thus to ensure an abstraction from the websites specific technology stack [WS5] (3.1.3) and a preservation of core areas of the website, such as loading time [WS3] (3.1.3). In addition, a RESTful API approach ensured ease of interfacing with existing websites [WS4] (3.1.3), along with device and browser independence [WS6] (3.1.3). Strongly connected to the RESTful API are the custom built WCMS module extensions. These extensions were implemented to support non-intrusive [WS1] (3.1.3) and limited-impact [WS2] (3.1.3) integration of the information exchange techniques in the websites within the Cross Site Browsing Space.

The high-level design conceived was used in an open domain use-case. Within this use-case, three separate user-focused case studies were defined in order to evaluate the appropriateness of the information exchange techniques.

(G2): To identify users' activities on websites and the information needs that drive those users to browse the website.

To address this second goal, the design requirements introduced in section 3.1 were considered. Since the goal was to identify users' activities on websites and the users' information needs, two interconnected components were introduced. The first component, which is the term identification component, addressed the identification of the terms related to the users' information needs together with the relevancy of the terms. Within the open domain use-case considered, this component was designed as to use external term identification tools [WC2] (3.1.2). The selected one, FREME, allows to use different ontologies to better identify representative terms of content depending on the subject domain [WC3] (3.1.2). Moreover, the storage in the third party User Model of the text entities extracted by the term identification component allowed for a shared conceptualisation, represented as a text entity space and based on the contents the user browsed within the websites of the Cross Site Browsing Space [WC1] (3.1.2) [WU2] (3.1.1). The second component, which is the Web-based Content Management System extension module, tracked all the activities of the authenticated users in the website and had the responsibility to trigger the term identification component every time a user landed on a new content page (i.e. at run-time) [WC4] (3.1.2). Furthermore, the tracking of the users' activities along with the extraction, and subsequent storage in the User Model of the Cross Site User Modelling service, of the text entities was performed without hindering users in their browsing experiences. Therefore, the User Modelling technique applied was an implicit User Modelling technique [WU3] (3.1.1).

The importance of this goal was to identify the Cross Site information needs that drive users to browse independently hosted websites. The identification was possible through the use of users' activities and content related text entities, which allowed the creation of a shared conceptualisation useful to materialise users' interests and needs.

(G3): To decide which information about the user to provide to the target website and whether to offer this information in its entirety or in smaller chunks, depending on the context of the application that utilises the Cross Site service.

To address this third goal, the design requirements introduced in section 3.1 were considered. Based on these design requirements [WU1] (3.1.1), a set of information exchange techniques between the target websites and the Cross Site User Modelling service were designed. The techniques ranged from highly generic to highly specific, providing different levels of information enrichment for both the websites and the service. The former tend to be more satisfactory from a web publisher perspective allowing a massive flow of user's information to leak from the service, thus ignoring or not sufficiently considering the user's privacy concerns. The latter, on the other hand, being more focused on preserving user's privacy, thus avoiding to provide huge amount of information to websites or, in extreme cases, not providing it at all, tend to be more satisfactory both for the user, who sees their privacy more respected, and the service itself, as it doesn't give away the only real value it holds - user data.

Therefore, the survey presented one of the core aspects of this thesis work: the information exchange paradigms. The initial design of the communication patterns was followed by an evaluation experiment, based on three user-focused case studies in an open domain use-case. The exchange techniques implemented for the experiment represented the two classes of information exchange techniques, which are the privacy-insensitive class and the privacy-aware class. The experiment confirmed the hypothesis made on the effectiveness of privacy-aware techniques when compared to privacy-insensitive techniques in relation to the three case studies analysed.

5.2.4 Research Challenges

The research challenges were divided into two main areas: Interaction and representation. To address challenges related to interaction, a RESTful API that allows any type of website/application to interface with the Cross Site User Modelling service was implemented. Along with it, an authentication system built on top of the service was introduced, so to enhance privacy at user level. On the other hand, to address challenges related to representation, the design and implementation of a term identification component and of a WCMS extension module that tracked users' activities was required.

5.3 Future Work

Several potential areas were identified where the research work described in this thesis could be extended and advanced. The main areas are discussed below:

5.3.1 Distributed User Modelling

The Cross Site User Modelling service designed in this thesis presents a centralised third party User Model, which can be accessed by websites to request relevant user's information. Even though the WCMS extension modules provides non-intrusive information exchange techniques, the fact that the user necessitates to be authenticated in the service in order to trigger the exchange techniques may hinder the user's browsing experience. In fact, the action of signing into an external service to obtain personalisation (or, in our case, useful Cross Site user data to be used for personalisation by the website) may be perceived by the user as a nuisance and may lead the user not to use the Cross Site Personalisation provided, even if it goes against its interests. A possible solution to address this shortcoming is to shift from a centralised User Modelling to a distributed User Modelling. This can be achieved using Enigma[1] Enigma is a decentralized computation

[1] `http://www.enigma.co/`

platform with guaranteed privacy. Its main goal is to enable developers to build 'privacy by design', end-to-end decentralized applications, without a trusted third party. Enigma is private, using secure multi-party computation (sMPC or MPC), data queries are computed in a distributed way, without a trusted third party. Data is split between different nodes, and they compute functions together without leaking information to other nodes. Specifically, no single party ever has access to data in its entirety; instead, every party has a meaningless (i.e., seemingly random) piece of it [ZNP15]. In this way, the need for users to authenticate into a centralised third party service gets removed, allowing the Cross Site distributed User Modelling service to provide Cross Site Personalisation to target websites without hindering the user at all. Moreover, being distributed, the approach also reduces the eventuality of malevolent attacks that can expose user personal data to unauthorised applications.

5.3.2 User Profile Management and Scrutiny

The Cross Site User Modelling service presented in this thesis does not consider any form of user profile management or user scrutiny. In order to address this liability, which impacts on trust and control over the user needs, a way to allow users to actively engage with their user profile is required. Therefore, the following aspects can be considered for future work:

1. Allow the user to view terms relating to the user's Cross Site information needs (model scrutiny);

2. Enable users to add and delete terms within the user profile;

3. Provide insight on where the information was collected and used.

As an example, the CULTURA project provided some form of model scrutiny, allowing the user to visualise their interests as a word-cloud where the terms could be enlarged or diminished by the user depending on the relevancy that the user gave them [SAO+12], [SSA+13].

Regarding the user profile management and in particular the new user problem, user profile sharing could be introduced as a solution. Allowing the sharing of user profiles would enable users to share common information needs, such as exploring information about holidays, with friends and peers, thus fastening the filling process of the user profile. In addition to direct social sharing, public sharing could be introduced to allow users to make parts of their profiles public and available for others to use. However, this makes sense in mature user profiles that have been created over a longer time period and across several websites. Finally, social media applications can be used to address this shortcoming. In fact, by providing the social graph of the user the pre-population of the user's profile might be possible [AHHK10].

5.3.3 Semantic Personalisation Techniques

The Cross Site User Modelling service presented in this thesis performs information exchange techniques that provide to target websites relevant user information that can be potentially used to address Cross Site information needs, by means of personalisation techniques. Therefore, the natural direction this research work should take, in order to evolve, is the implementation of personalisation techniques. Due to the semantic nature of the information stored in the third party User Model, that is the text entities extracted by the term identification component 3.4.3.1, it is natural to prefer personalisation techniques that are able to understand and handle semantic data. Indeed, the use of semantic personalisation techniques might improve the quality of the personalisation compared to non-semantic techniques, thanks to a deeper understanding of the user's interests and needs due to the awareness of the meaning of stored terms. Hence, Semantic Recommender Systems can be used as personalisation approaches. Semantic Recommender Systems infer semantically relevant items by using semantic inter-relations defined between concepts of the ontology used. An example of Semantic Recommender by Fard et al. [FNS13] proposes to use a semantic similarity measure to find a set of k nearest neighbours to the target user, or target item. Three types of semantic similarity measures were introduced, which calculate the similarities between items serving as ontology-based metadata instances that are defined as three types of Taxonomy Similarity (TS), Attribute Similarity (AS) and Relation Similarity (RS). For each pair of items, the above semantic similarity measures were used by obtaining the weighted values of these measures. Hence, the semantic similarity was calculated as the weighted arithmetic mean of the three similarity measures defined above.

5.3.4 Long/Short Term User Profiling

Since the third party User Model of the Cross Site User Modelling service keeps stored all users' data since their first subscription to the service without ever deleting them, it is possible to discuss about dynamic adaptation strategies for short-term and long-term user profile for personalisation. As stated in [LYWK07], dynamic adaptation strategies are devised to capture the accumulation and degradation changes of user preferences, and adjust the content and the structure of the user profile to these changes. Therefore, it could be possible to adapt the personalisation for the user depending on these changes, thus tailoring the personalisation not only on the meaning of the content but also on its freshness and weight.

5.4 Conclusions

The scope of this thesis was to introduce and develop an adaptive Cross Site User Modelling platform for information exchange techniques. The reason derived from a perceived need for a consistent Cross Site support mechanism, of which the Cross Site User Modelling service is an instance, that ensures effective assistance to users in the websites they browse across, by means of personalisation techniques.

Within this framework, this research work served as a starting point in addressing those challenges related to the fulfilment of the users' Cross Site information needs. A careful review of the State of the Art ensured a deep understanding of the current approaches in the fields of personalisation techniques, User Modelling and Cross Site Personalisation. Such understanding was then used to design and implement all the components of the Cross Site service. The platform was then tested and evaluated, obtaining results consistent with the hypothesis and assumptions made. Finally, a brief discussion, which concluded the research work, on the future possible implications was provided.

Bibliography

[ABH+08] Jae-wook Ahn, Peter Brusilovsky, Daqing He, Jonathan Grady, and Qi Li. Personalized web exploration with task models. In *Proceedings of the 17th International Conference on World Wide Web*, WWW '08, pages 1–10, New York, NY, USA, 2008. ACM.

[ACDN12] Maristella Agosti, Franco Crivellari, and Giorgio Maria Di Nunzio. Web log analysis: A review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Min. Knowl. Discov.*, 24(3):663–696, May 2012.

[ACDNG10] Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio, and Silvia Gabrielli. Understanding user requirements and preferences for a digital library web portal. *Int. J. Digit. Libr.*, 11(4):225–238, December 2010.

[AF03] David Avison and Guy Fitzgerald. *Information systems development: methodologies, techniques and tools (3rd edition)*.

McGraw Hill, 2003.

[AHHK10] Fabian Abel, Nicola Henze, Eelco Herder, and Daniel Krause. Interweaving public user profiles on the web. In *Proceedings of the 18th International Conference on User Modeling, Adaptation, and Personalization*, UMAP'10, pages 16–27, Berlin, Heidelberg, 2010. Springer-Verlag.

[AT05] Gediminas Adomavicius and Alexander Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 17(6):734–749, 2005.

[BC92] Nicholas J. Belkin and W. Bruce Croft. Information filtering and information retrieval: Two sides of the same coin? *Commun. ACM*, 35(12):29–38, December 1992.

[Ber97] Hal Berghel. Cyberspace 2000: Dealing with information overload. *Commun. ACM*, 40(2):19–24, February 1997.

[BFG11] Robin Burke, Alexander Felfernig, and Mehmet H Göker. Recommender systems: An overview. *AI Magazine*, 32(3):13–18, 2011.

[BGMZ97] Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. Syntactic clustering of the web. In *Selected Papers from the Sixth International Conference on World Wide Web*, pages 1157–1166, Essex, UK, 1997. Elsevier Science Publishers Ltd.

[BH07] Peter Brusilovsky and Nicola Henze. *Open Corpus Adaptive Educational Hypermedia*, pages 671–696. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[BHMW02] Christopher Bailey, Wendy Hall, David E. Millard, and Mark J. Weal. *Towards Open Adaptive Hypermedia*, pages 36–46. Springer Berlin Heidelberg, Berlin, Heidelberg, 2002.

[BL10] Tim Berners-Lee. Long live the web. *Scientific American*, 303(6):80–85, 2010.

[BM02] Peter Brusilovsky and Mark T Maybury. From adaptive hypermedia to the adaptive web. *Communications of the ACM*, 45(5):30–33, 2002.

[Bru01] Peter Brusilovsky. Adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 11(1-2):87–110, March 2001.

[Bru08] Peter Brusilovsky. *Adaptive Navigation Support for Open Corpus Hypermedia Systems*, pages 6–8. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[BS00] Bettina Berendt and Myra Spiliopoulou. Analysis of navigation behaviour in web sites integrating multiple information systems. *The VLDB Journal*, 9(1):56–75, 2000.

[BSS05] Peter Brusilovsky, Sergey Sosnovsky, and Olena Shcherbinina. User modeling in a distributed e-learning architecture. In *Proceedings of the 10th International Conference on User Modeling*, UM'05, pages 387–391, Berlin, Heidelberg, 2005. Springer-Verlag.

[Bur02] Robin Burke. Hybrid recommender systems: Survey and experiments. *User Modeling and User-Adapted Interaction*, 12(4):331–370, November 2002.

[BZ09] Fabian Bohnert and Ingrid Zukerman. *Non-intrusive Personalisation of the Museum Experience*, pages 197–209. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[CC09] Francesca Carmagnola and Federica Cena. User identification for cross-system personalisation. *Inf. Sci.*, 179(1-2):16–32, January 2009.

[CCG11] Francesca Carmagnola, Federica Cena, and Cristina Gena. User model interoperability: a survey. *User Modeling and User-Adapted Interaction*, 21(3):285–331, 2011.

[CS98] Liren Chen and Katia Sycara. Webmate: A personal agent for browsing and searching. In *Proceedings of the Second International Conference on Autonomous Agents*, AGENTS '98, pages 132–139, New York, NY, USA, 1998. ACM.

[DDH$^+$00] A. Dieberger, P. Dourish, K. Höök, P. Resnick, and A. Wexelblat. Social navigation: Techniques for building more usable systems. *interactions*, 7(6):36–45, November 2000.

[DN03] Peter Dolog and Wolfgang Nejdl. Challenges and benefits of the semantic web for user modelling. In *Proc. of AH2003* *workshop at 12th World Wide Web Conference, Budapest, Hungary*, 2003.

[Doy01] P. Doyle. *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*. North Holland, 2001.

[EM00] Angela Edmunds and Anne Morris. The problem of information overload in business organisations: A review of the literature. *Int. J. Inf. Manag.*, 20(1):17–28, February 2000.

[ERW11] Shady Elbassuoni, Maya Ramanath, and Gerhard Weikum. *Query Relaxation for Entity-Relationship Search*, pages 62–76. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[FAJ10] Henry A. Feild, James Allan, and Rosie Jones. Predicting searcher frustration. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 34–41, New York, NY, USA, 2010. ACM.

[FFS11] Martin Feuz, Matthew Fuller, and Felix Stalder. Personal web searching in the age of semantic capitalism: Diagnosing the mechanisms of personalisation. *First Monday*, 16(2), 2011.

[Fis01] Gerhard Fischer. User modeling in human&ndash;computer interaction. *User Modeling and User-Adapted Interaction*, 11(1-2):65–86, March 2001.

[FNS13] Karamollah Bagheri Fard, Mehrbakhsh Nilashi, and Naomie Salim. Recommender system based on semantic similarity. *International Journal of Electrical and Computer Engineering*, 3(6):751, 2013.

[GM01] Athula Ginige and San Murugesan. Web engineering: An introduction. *IEEE multimedia*, 8(1):14–18, 2001.

[Gru93] Thomas R. Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, June 1993.

[GSCM07] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. The adaptive web. chapter User Profiles for Personalized Information Access, pages 54–89. Springer-Verlag, Berlin, Heidelberg, 2007.

[Guy15] Ido Guy. Social recommender systems. In *Recommender Systems Handbook*, pages 511–543. Springer, 2015.

[HJ04] Eelco Herder and Ion Juvina. Discovery of individual user navigation styles. 2004.

[HKTR04] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Trans. Inf. Syst.*, 22(1):5–53, January 2004.

[HN01] Nicola Henze and Wolfgang Nejdl. Adaptation in open corpus hypermedia. *International Journal of Artificial Intelligence in Education*, 12(4):325–350, 2001.

[HPPL98] Bernardo A Huberman, Peter LT Pirolli, James E Pitkow, and Rajan M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.

[HT85] Starr R. Hiltz and Murray Turoff. Structuring computer-mediated communication systems to avoid information overload. *Commun. ACM*, 28(7):680–689, July 1985.

[hT99] Ah hwee Tan. Text mining: The state of the art and the challenges. In *In Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases*, pages 65–70, 1999.

[IBV98] Tomás Isakowitz, Michael Bieber, and Fabio Vitali. Web information systems. *Commun. ACM*, 41(7):78–80, July 1998.

[IJ05] Peter Ingwersen and Kalervo Järvelin. *The Turn: Integration of Information Seeking and Retrieval in Context (The Information Retrieval Series)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.

[JG12] Faustina Johnson and Santosh Kumar Gupta. Web content mining techniques: a survey. *International Journal of Computer Applications*, 47(11), 2012.

[KA08] Giridhar Kumaran and James Allan. Adapting information retrieval systems to user queries. *Inf. Process. Manage.*, 44(6):1838–1862, November 2008.

[Kay06] Judy Kay. *Scrutable Adaptation: Because We Can and Must*, pages 11–19. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.

[KB06] Sherrie YX Komiak and Izak Benbasat. The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS quarterly*, pages 941–960, 2006.

[KCW09] Kevin Koidl, Owen Conlan, and Vincent Wade. Non-invasive adaptation service for web-based content management systems. 2009.

[KCW13] K. Koidl, O. Conlan, and V. Wade. Towards cross site personalisation. In *2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT)*, volume 1, pages 542–548, Nov 2013.

[KCWS11] Kevin Koidl, Owen Conlan, Lai Wei, and Ann Marie Saxton. Non-invasive browser based user modeling towards semantically enhanced personlization of the open web. In *Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference on*, pages 35–40. IEEE, 2011.

[KKZB08] Nora Koch, Alexander Knapp, Gefei Zhang, and Hubert Baumeister. Uml-based web engineering. In *Web Engineering:*

Modelling and Implementing Web Applications, pages 157–191. Springer, 2008.

[KL05] Kevin Keenoy and Mark Levene. *Personalisation of Web Search*, pages 201–228. Springer Berlin Heidelberg, Berlin, Heidelberg, 2005.

[Koi13] Kevin Koidl. Cross-site personalisation, 2013.

[KW01] Nora Koch and Martin Wirsing. Software engineering for adaptive hypermedia applications. In *8th International Conference on User Modeling, Sonthofen, Germany*. Citeseer, 2001.

[LC11] Chung Hun Lee and David A Cranage. Personalisation–privacy paradox: The effects of personalisation and privacy assurance on customer responses to travel web sites. *Tourism Management*, 32(5):987–994, 2011.

[LdGS11] Pasquale Lops, Marco de Gemmis, and Giovanni Semeraro. *Content-based Recommender Systems: State of the Art and Trends*, pages 73–105. Springer US, Boston, MA, 2011.

[Lie] Scott Liewehr. Understanding best practices for profiling, personalizing, and targeting next generation engagement.

[Lie95] Henry Lieberman. Letizia: An agent that assists web browsing. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'95, pages 924–929, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[LSY03] Greg Linden, Brent Smith, and Jeremy York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003.

[LYWK07] Lin Li, Zhenglu Yang, Botao Wang, and Masaru Kitsuregawa. Dynamic adaptation strategies for long-term and short-term user profile to personalize search. In *Advances in Data and Web Management*, pages 228–240. Springer, 2007.

[MAB00] Maurice D. Mulvenna, Sarabjot S. Anand, and Alex G. Büchner. Personalization on the net using web mining: Introduction. *Commun. ACM*, 43(8):122–125, August 2000.

[MCS00] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava. Automatic personalization based on web usage mining. *Commun. ACM*, 43(8):142–151, August 2000.

[Mea92] Charles T. Meadow. *Text Information Retrieval Systems*. Academic Press, Inc., Orlando, FL, USA, 1992.

[MGH98] Matthew Montebello, W. A. Gray, and Stephen Hurley. An evolvable personal advisor to optimize internet search technologies. In *Proceedings of the 9th International Conference on Database and Expert Systems Applications*, DEXA '98, pages 531–540, London, UK, UK, 1998. Springer-Verlag.

[MM98] Alexandros Moukas and Pattie Maes. Amalthaea: An evolving multi-agent information filtering and discovery system for the www. *Autonomous Agents and Multi-Agent Systems*, 1(1):59–88, 1998.

[MS01] Alexander Maedche and Steffen Staab. Ontology learning for the semantic web. *IEEE Intelligent Systems*, 16(2):72–79, March 2001.

[MSB98] Mandar Mitra, Amit Singhal, and Chris Buckley. Improving automatic query expansion. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '98, pages 206–214, New York, NY, USA, 1998. ACM.

[NDRV09] Nikolaos Nanas, Anne De Roeck, and Manolis Vavalis. *What Happened to Content-Based Information Filtering?*, pages 249–256. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

[not]

[OBHG03] Daniel Oberle, Bettina Berendt, Andreas Hotho, and Jorge Gonzalez. *Conceptual User Tracking*, pages 155–164. Springer Berlin Heidelberg, Berlin, Heidelberg, 2003.

[O'C11] Tiffany O'Callaghan. Eli pariser: The dark side of web personalisation. *New Scientist*, 211(2822):23 –, 2011.

[OK01] Douglas W Oard and Jinmook Kim. Modeling information content using observable behavior. 2001.

[Par11] Eli Pariser. *The filter bubble: What the Internet is hiding from you*. Penguin UK, 2011.

[PB97] Michael Pazzani and Daniel Billsus. Learning and revising user profiles: The identification of interesting web sites. *Machine Learning*, 27(3):313–331, 1997.

[PB07] Michael J. Pazzani and Daniel Billsus. *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

[Rij79] C. J. Van Rijsbergen. *Information Retrieval*. Butterworth-Heinemann, Newton, MA, USA, 2nd edition, 1979.

[RK11] Ian Ruthven and Diane Kelly. *Interactive information seeking, behaviour and retrieval*. Facet Publ., 2011.

[RR86] Mary J Rudd and Joel Rudd. The impact of the information explosion on library users: Overload or opportunity?. *Journal of Academic Librarianship*, 12(5):304–6, 1986.

[RT11] Giuseppe Rizzo and Raphaël Troncy. Nerd: evaluating named entity recognition tools in the web of data. 2011.

[SAO$^+$12] Mark S. Sweetnam, Maristella Agosti, Nicola Orio, Chiara Ponchia, Christina M. Steiner, Eva-Catherine Hillemann, Micheál Ó Siochrú, and Séamus Lawless. *User Needs for Enhanced Engagement with Cultural Heritage Collections*, pages 64–75. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[SDAN$^+$09] Avaré Stewart, Ernesto Diaz-Aviles, Wolfgang Nejdl, Leandro Balby Marinho, Alexandros Nanopoulos, and Lars Schmidt-Thieme. Cross-tagging for personalized open social networking. In *Proceedings of the 20th ACM Conference on Hypertext and Hypermedia*, HT '09, pages 271–278, New York, NY, USA, 2009. ACM.

[SFHS07] J Ben Schafer, Dan Frankowski, Jon Herlocker, and Shilad Sen. Collaborative filtering recommender systems. In *The adaptive web*, pages 291–324. Springer, 2007.

[SG05] Micro Speretta and Susan Gauch. Personalized search based on user search histories. In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '05, pages 622–628, Washington, DC, USA, 2005. IEEE Computer Society.

[SK92] Irene Stadnyk and Robert Kass. Modeling users' interests in information filters. *Commun. ACM*, 35(12):49–50, December 1992.

[SKR99] J. Ben Schafer, Joseph Konstan, and John Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, pages 158–166, New York, NY, USA, 1999. ACM.

[SKR01] J Ben Schafer, Joseph A Konstan, and John Riedl. E-commerce recommendation applications. In *Applications of Data Mining to Electronic Commerce*, pages 115–153. Springer, 2001.

[SKR02] J. Ben Schafer, Joseph A. Konstan, and John Riedl. Meta-recommendation systems: User-controlled integration of diverse recommendations. In *Proceedings of the*

Eleventh International Conference on Information and Knowledge Management, CIKM '02, pages 43–51, New York, NY, USA, 2002. ACM.

[SM86] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[SM95] Humphrey Sorensen and Michael McElligott. Psun: a profiling system for usenet news. In *Proceedings of CIKM*, volume 95, pages 1–2. Citeseer, 1995.

[SMB07] Ahu Sieg, Bamshad Mobasher, and Robin Burke. Web search personalization with ontological user profiles. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 525–534, New York, NY, USA, 2007. ACM.

[SSA+13] Mark Sweetnam, MO Siochru, Maristella Agosti, Marta Manfioletti, Nicola Orio, and Chiara Ponchia. Stereotype or spectrum: Designing for a user continuum. In *the Proceedings of the First Workshop on the Exploration, Navigation and Retrieval of Information in Cultural Heritage, ENRICH*, 2013.

[SVV99] Cheri Speier, Joseph S. Valacich, and Iris Vessey. The influence of task interruption on individual decision making: An information overload perspective. *Decision Sciences*, 30(2):337–360, 1999.

[TAAK04] Jaime Teevan, Christine Alvarado, Mark S. Ackerman, and David R. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '04, pages 415–422, New York, NY, USA, 2004. ACM.

[WBC09] Ryen W. White, Peter Bailey, and Liwei Chen. Predicting user interests from contextual information. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 363–370, New York, NY, USA, 2009. ACM.

[WPB01] Geoffrey I. Webb, Michael J. Pazzani, and Daniel Billsus. Machine learning for user modeling. *User Modeling and User-Adapted Interaction*, 11(1-2):19–29, March 2001.

[WYEN+99] Dwi H Widyantoro, Jianwen Yin, M El Nasr, Linyu Yang, Anna Zacchi, and John Yen. Alipes: A swift messenger in cyberspace. In *Proceedings of Spring Symposium Workshop on Intelligent Agents in Cyberspace*, pages 62–67, 1999.

[XC96] Jinxi Xu and W. Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pages 4–11, New York, NY, USA, 1996. ACM.

[ZGBN07] Xuan Zhou, Julien Gaugaz, Wolf-Tilo Balke, and Wolfgang Nejdl. Query relaxation using malleable schemas. In *Proceedings of the 2007 ACM SIGMOD International Conference on Management of Data*, SIGMOD '07, pages 545–556, New York, NY, USA, 2007. ACM.

[Zhu04] Mu Zhu. Recall, precision and average precision. *Department of Statistics and Actuarial Science, University of Waterloo, Waterloo*, 2, 2004.

[ZNP15] Guy Zyskind, Oz Nathan, and Alex Pentland. Enigma: Decentralized computation platform with guaranteed privacy. *arXiv preprint arXiv:1506.03471*, 2015.

Acknowledgements

It has been a long way up to here and without the help of the people around me it would not have been the same.

I would like to thank my brother Alessandro, without whom I would not be who I am today. You are my role model and the most important person in my life.

I would like to thank my parents for their love, encouragement and guidance throughout all these years. Without you none of what has happened in my life so far would have been possible, I could not have been luckier.

I would also like to thank my supervisor, Prof. Maristella Agosti, who was more than a supervisor to me and whose knowledge and encouragement have guided me through these past two years.

I also would like to thank Prof. Vincent Wade and Prof. Séamus Lawless, my co-supervisors, whose support and advice were essential to complete this dissertation. Your hospitality made me feel like home.

Finally, I would like to thank all my dearest friends for the support and the patience throughout these years.

I sincerely thank you all