Exploiting available domain knowledge to improve the retrieval and recommendation of Digital Cultural Heritage materials

Mahmud Ahmed Usman

School of Science, Engineering, and Environment
University of Salford, Manchester, United Kingdom

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

August 2021

ABSTRACT

Cultural Heritage (CH) institutions, such as museums, have recently embraced computing techniques to digitise CH materials (artefacts, paintings, books etc) and to make accessible those digital representations through their online portals to millions of museum visitors (onsite and remote). This mass availability of digitised materials, however, can lead to information overload. Therefore, ordinary CH online users can find it challenging to access these materials, because they usually have no domain knowledge and also lack the experience of which precise keyword terms to use to search and discover new information.

As an attempt to mitigate the issues explained above, recommender systems and visual search interfaces have been used by millions of users to discover new and relevant to the users' interests CH materials. A CH recommender system is a system that uses knowledge — content and social — representations assembled from various domain knowledge sources, to generate personalised recommendations of CH materials. Social knowledge representations provide better recommendation quality than content knowledge representations when they have substantial social knowledge such as user-interactions and social tagging in the representation, but they suffer when available information is insufficient (cold start problem and sparsity of social knowledge).

Different approaches have been deployed to address these challenges, for example a hybrid approach that incorporated content directly into a social knowledge representation to provide a recommendation. But this hybrid approach only works well on domains for which specific content knowledge exists which can directly describe an item and is always available and meaningful. The CH domain does not have such rich specific knowledge that can directly describe the content of CH materials, thus limiting the ability to incorporate content directly into the social knowledge representation for CH recommendations.

To address these challenges, this Thesis starts with examining the strengths and weaknesses of content and social knowledge representations in the context of CH recommendations and how these knowledge representations can complement each other to improve the recommendations of CH materials. The identified knowledge gap is bridged through a new hybrid representation approach by integrating the content and social knowledge representations. The effect of knowledge integration is to increase the instances of quality recommendations and improved discovery, and to provide opportunities to users to discover unexpected and liked recommendations of CH materials.

The new integrated and social knowledge representations are used to further develop a dynamic hybrid CH recommender system. The dynamic hybrid system combines the learned integrated knowledge for each CH object with CH object's social knowledge, and assigns the weights to both integrated and social knowledge representations to control the contributions of each knowledge so that each representation could contribute based on the current user and search status.

A new visual search interface is also described in this thesis, developed as a part of the research work. The search interface enables visual search and exploration across large CH collections by providing an interactive visual summary of the recommended CH items, addressing the challenge of the lack of domain knowledge by online users. User satisfaction evaluation was conducted to measure the user satisfaction level for using a visual search interface for search and exploration of information from the vast collection when compared to the non-visual search interface. The evaluation showed that a user with no domain knowledge prefers using a visual search interface than one with no visual summary presentation, but the result also shows that there is no significant difference for users that have domain knowledge.

The challenges of evaluating CH recommendations are also addressed in the Thesis. The feedback provided by the users, both implicit and explicit, is exploited to measure and reflect the performance of the cultural heritage recommendation methods used. A user study to evaluate both the ground truth measures and integrated knowledge representations is conducted. Throughout the user study, the results obtained show that the hybrid representation produced a better quality recommendation of CH materials when compared with content and social knowledge representations. The social representation does not provide a better high-level recommendation quality compared to the hybrid representation, but it does outperform the hybrid representation in recommending novel CH materials.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Professor Apostolos Antonacopoulos, for his support and encouragement throughout this PhD journey. Apostolos has given free-hand on my research while keeping me on track by always reviewing my work – thank you. My appreciation also goes to the Pattern Recognition and Image Analysis (PRImA) research team for their support and sharing of ideas.

I also want to thank and acknowledge my parents (Alhaji Ahmadu, Hajiya Zainab, and Hajiya Fatuma), my wife (Safiya Abdulahamid) – I love you, my kids (Ahmad, Abdulrahman, Abdulhamid, and Khadija), and my brothers and sisters for their love, support, prayers, and encouragement. Other important persons that play a vital role in this journey are Senator Muhammad Danjuma Goje and Ahmed Isa – thank you for the support and encouragement.

During this PhD journey, I have come across a significant number of people, especially Dr. Sadiq Sani, Hamid A. Adamu (Komsiri), Murtala Muhammed (Dr Jazaks), Abubakar Halilu Gidado (Baban Murad), Abubakar Abbas (Abu Abbas), Dang Bwebu, Ali Dulla, and Faruq Sulaiman, who have contributed in one way or the other to this research.

LIST OF FIGURES

LIST OF TABLES

Chapter One: Introduction

This chapter discusses the motivation and research problem, research questions and objectives, research contributions, publications arising from work, and the thesis outline.

The chapter starts by highlighting the significance of Cultural Heritage (CH) to people's cultural and social values, then discusses the challenges that online CH users face as a result of information overload when exploring large collections of digital CH materials for new discoveries. The challenges of information overload differ depending on the knowledge representation and the user search interface presented. This chapter discusses background information on content and social knowledge representations assembled from different available domain knowledge sources. It also discusses the Visual Search Interfaces (VSIs) that enable the exploration of CH materials from large collections to address the challenges of CH recommendations, such as cold-start problems. Finally, the chapter presents the publications arising from the research and discusses the research contributions.

1.1 Motivation and research problem

The concept of CH provides people with a certain level of connection to their social values, beliefs, customs, and religions. It also provides a sense of belonging and unity within the community by providing a better understanding of their ancestors and history. In the past two decades, the mode in which CH information is accessed has experienced a tremendous change, from the era of physical materials to the current era of online access. Many factors have contributed to this remarkable achievement, including robust internet technologies and the digitisation of CH materials (Petras et al., 2017). Thus, the exploration of the digital CH materials contained in CH collections presents significant challenges to online users; information about hundreds of millions of objects from CH collections is available online, leading to **information overload**.

The number of online CH users continues to grow, especially non-professionals, such as tourists (Walsh et al., 2018). For instance, in the last quarter of 2017, over 85 million users left a review on social media of Europeana,[1] a popular online platform for researching digital CH material (Petras et al., 2017). However, despite this rapid growth, online users, especially those with no domain knowledge, find it difficult to explore digital CH collections and find CH objects of interest. Locating these materials and the information about them typically involves keyword searches; users enter search terms, and the results are presented via ranked lists of relevant information. Keyword searches are efficient if users have reasonable domain knowledge (Clough et al., 2017), but many users today generally access the internet, either for research or leisure, without specific domain knowledge (Walsh et al., 2018). To address these challenges, recommender systems (RSs) and VSIs have continued to become very popular (Aggarwal, 2016; Holzer et al., 2018).

RSs are artificially intelligent systems that provide alternative ways of discovering new information from a vast amount of data by providing personalised recommendations to the target users (Aggarwal, 2016). A VSI is an interface for visually searching and exploring information for new knowledge discoveries (Holzer et al., 2018). Despite the tremendous success recorded by RSs in various domains, e.g. e-commerce and movies, new challenges, such as **the cold-start problem, out of context recommendations, and most similar objects but bad recommendations,** continue to evolve in the CH domain because of its diversity (Amato et al., 2018). These challenges exist as a result of insufficient **knowledge representations** assembled from the available domain knowledge sources.

Knowledge representation describes the information from the real world represented for a computer to comprehend and then utilise to solve complex problems, such as personalised

[1] https://www.europeana.eu/portal/en

recommendations. CH recommendations rely on exploiting available knowledge that describes each CH object. A strong knowledge representation may define the relationships between CH objects and, thus, produce quality CH recommendations. Two core knowledge representations used for CH recommendations are **content knowledge representation**, which describes the content features of CH objects, and **social knowledge representation**, which describes CH users' interactions with CH objects.

Different recommendation techniques have been applied to knowledge representations to produce CH recommendations. The techniques include content-based, collaborative filtering, context-aware, knowledge-based, and hybrid approaches. Content-based techniques match users' content attributes with the objects' content attributes to recommend items of interest to the target users. Collaborative filtering uses social knowledge, such as users' reviews and ratings, without exogenous knowledge to recommend objects of interest to target users. Context-aware and knowledge-based techniques make recommendations based on context information and specific domain knowledge, respectively, while the hybrid approach combines two or more recommendation techniques. These techniques have strengths and weaknesses, depending on the source of knowledge available. For example, content-based techniques suffer from the over-specialisation or serendipity problem, and collaborative filtering techniques suffer from the cold-start problem. In most situations, a hybrid technique is used to address these challenges.

Therefore, this thesis focuses on **hybrid approaches** that combine knowledge representations assembled from available domain knowledge sources to address the challenges of the cold-start problem, out of context recommendations, and most similar objects but bad recommendations. The study's key challenge is developing mechanisms to dynamically control the contribution of each knowledge representation to reflect the anticipated CH recommendation performance.

1.2 Research questions and objectives

1.2.1 Questions

The motivation and research problem for this work has highlighted the significance of knowledge representations and VSIs for producing quality CH recommendations and enabling CH users with no domain knowledge background to explore a large CH objects collection for new information discovery. Therefore, the research questions are declared as follows:

I. How can a hybrid approach integrating content knowledge representation and social knowledge representation address the challenges of CH recommendations, such as the cold-start problem, out of context recommendations, and similar objects but bad recommendations?

II. How can a VSI help CH users with no domain knowledge explore a large collection of CH objects for new information discovery?

1.2.2 Objectives

To address the research questions, we define the research objectives as follows:

Objective I: Bridge the knowledge gap in the knowledge representations.

Bridging the knowledge gap requires integrating the content and social knowledge representations assembled from the available domain knowledge sources. The integrated knowledge representations should address the issues of the cold-start problem, out of context recommendations, and bad recommendations.

Objective II: Develop a dynamic hybrid cultural heritage recommendation and visual search interface.

The proposed approach should address the issues ascertained through literature review: the cold-start problem, out of context recommendations, and bad recommendations. The VSI should enable users with no domain knowledge to explore extensive CH collections for new information discovery of CH materials.

Objective III: Build a custom dataset for cultural heritage recommendations.

This requires exploiting any available domain knowledge sources, for example, online museums and social networks, to build a custom dataset that can be used to assemble content and social knowledge representations for CH recommendations.

1.3 Research contributions

In this section, the contributions made as a result of this study are presented as follows:

I. The major contribution derived from this work is the integration of content and social knowledge representations to bridge the knowledge gap and provide quality CH recommendations. This contribution has been Accepted for publication in the *IEEE Access Journal*.

II. A dynamic hybrid approach, a further combination of integrated and social knowledge representations, is another contribution derived from this work. This approach addresses the issues of the cold-start problem, out of context recommendations, and similar objects but bad recommendations.

III. Another contribution derived from this work is the design and development of an interface for the VIsual Search and Exploration (VISE) of CH collections to enable users with little or no domain knowledge to quickly discover new CH materials from extensive CH collections. This contribution has been published in the Association for Computing Machinery's (ACM) *Journal Of Computing and Cultural Heritage* (Usman & Antonacopoulos, 2019).

IV. The harvest of a custom CH dataset from the available domain knowledge sources is another contribution derived from this work. The current available open datasets of CH materials, for example, Europeana datasets, do not have all of the data required for recommendation techniques to make CH recommendations.

1.4 Publications arising from the work

The publications that have been derived from this work include the following:

I. Usman, M.A., & Antonacopoulos, A. (2019). VISE: An interface for visual search and exploration of museum collections. *Journal on Computing and Cultural Heritage (JOCCH)*, *12*(4), 25.

II. Usman, M.A., & Antonacopoulos, A. (2021). Integrating content and social knowledge representations to improve recommendations of cultural heritage materials [Submitted for publication]. *IEEE Access*.

1.5 Thesis outline

In this chapter, the research background and motivation have been discussed, focusing on online users' challenges with searching and exploring large collections of CH materials for new information discoveries. The research questions and objectives and the research contributions were also discussed.

Chapter Two discusses the recent work on RSs, from general recommendations to CH recommendations. Recent research on establishing knowledge representations through knowledge discovery for recommendations is presented, and the research gap from this work is identified. The literature on VSIs that enable search and exploration of CH materials and the evaluation of RSs are also described and examined.

Chapter Three presents the proposed methodology and experimental set-up. This includes a research process model, system design and development, and the evaluation of CH recommendations.

Chapter Four describes the datasets used to demonstrate the study, both an existing public dataset and a new, custom harvested dataset. The issues surrounding the domain knowledge sources used to extract knowledge for harvesting the custom dataset used in this study are discussed. Finally, the public dataset and the new harvested dataset are compared in terms of their social knowledge richness.

Chapter Five covers the essence of integrating content and social knowledge representation, assembled from the available domain sources, to bridge the knowledge gap within the collection for CH recommendations.

Chapter Six discusses the hybrid approach to producing CH recommendations, combining integrated and social knowledge representations. Finally, the chapter discusses the evaluation of a hybrid approach.

Chapter Seven presents the design and development of VISE for CH material collection and discusses its user satisfaction evaluation.

Chapter Eight discusses the user evaluation of the proposed system, and Chapter Nine presents the thesis conclusions and future work.

Chapter Two: Literature Review

2.1 Introduction

This chapter presents an overview and critical review of the ideas, techniques, and contributions provided by researchers in the literature relevant to the study. The beginning of the chapter discusses RSs in general and then focuses on the techniques involved, the knowledge discovery process for knowledge representations, and which recommendation techniques to exploit to produce recommendations.

The chapter later discusses CH recommendations from the context of metadata enrichment and level of personalisation, covering three major CH information retrieval (IR) systems. The importance of a VSI for online users to aid the discovery of new CH materials within large information collections is also discussed. Finally, the chapter concludes by discussing the key findings and way forward from the relevant reviewed literature.

2.2 Recommender systems

RSs use software tools and techniques to recommend items that are of interest and useful to a target user (Aggarwal, 2016; Bobadilla et al., 2013; Ricci et al., 2011). The recommendations relate to different decision-making procedures, such as what paintings to buy, what online books to read, or which museums to visit in the summer. Many notable online platforms have been using RSs to assist their users in selecting items from the vast amount (sometimes millions or billions) of options. For example, Netflix, an online movie streaming platform, uses an RS to suggest movies of interest to their users (Gomez-Uribe & Hunt, 2016). In recent years, RSs have become very popular with service providers because they increase sales, user satisfaction, and user fidelity, and they provide a better understanding of what the users want (Ricci et al., 2011). The next section discusses the techniques used for providing recommendations.

2.3 Recommendation techniques

Generally, there are six different approaches for recommendations (Ricci et al., 2011):

1. Content-based (Lops et al., 2011),
2. Collaborative filtering (Schafer et al., 2007),
3. Constraint-based (Felfernig et al., 2011),
4. Demographic (Al-Shamri, 2016),
5. Context-aware (Adomavicius & Tuzhilin, 2011), and
6. Hybrid (Burke, 2007b).

The limitations of each recommendation technique and the advantages of the hybrid approach over other approaches are further discussed in this section.

2.3.1 Content-based

The content-based approach matches the attributes of object contents or item descriptions with the user profile attributes to produce personalised recommendations (Pazzani & Billsus, 2007). It recommends items similar to those a target user previously liked, based on an item's content. For example, ACR News filters news that is similar to the target user's preferences based on content attributes (Mobasher et al., 2000). Figure 2.1 shows the architecture of a content-based RS.

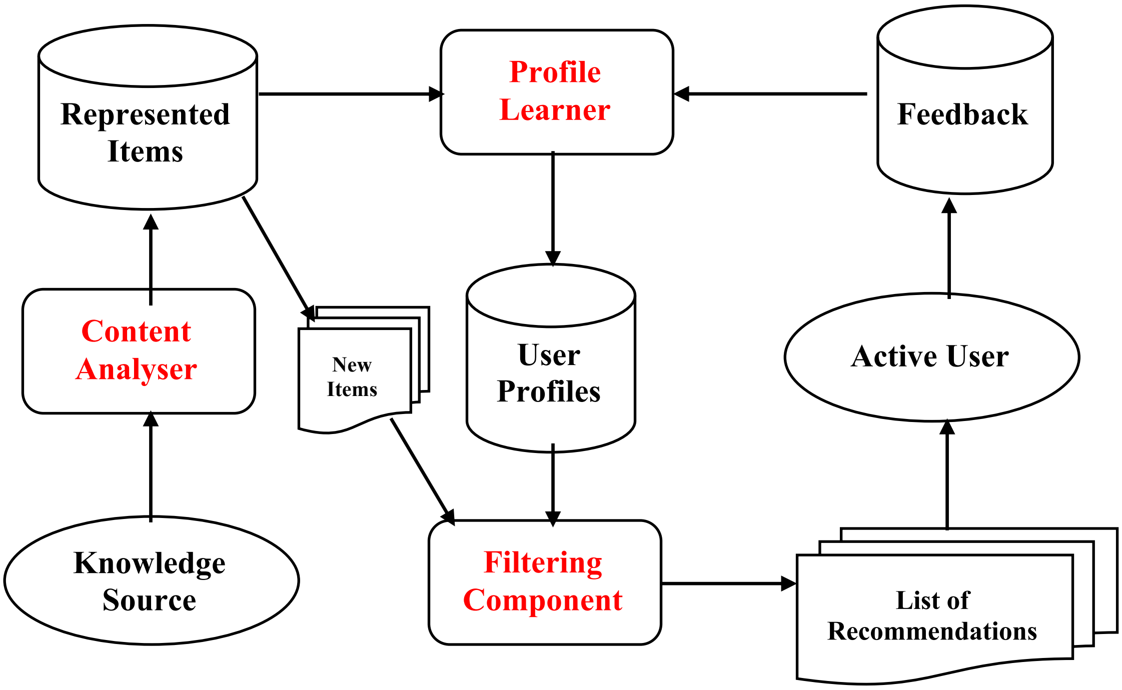


Figure 2.1: High-level architecture of a content-based recommender (Lops et al., 2011)

It can be noticed in Figure 2.1 that there are three important components (marked in red): a content analyser, profile learner, and filtering component. The first component, the **content analyser**, generates the important features extracted from various domain knowledge sources to describe each item. The **profile learner** constructs a user profile from the assembled user preferences data and data about items the user previously liked or disliked. The **filtering component** generates a list of recommendations using the input provided by the profile learner. The content-based approach is not user-independent when compared to collaborative filtering (see Section 2.3.2). One advantage of content-based over collaborative filtering is the capability of recommending new items without user interactions (Thorat et al., 2015). However, the content-based approach suffers from over-specialisation since recommendations are generated from the content of items previously liked by users (Pazzani & Billsus, 2007; Thorat et al., 2015).

Many works have been completed to address the over-specialisation limitations of content-based recommendations (Pereira & Varma, 2016), for example, using a genetic algorithm to filter the context of information (Jain et al., 2015), removing items that are too similar to those the user has seen before (Billsus & Pazzani, 2000), and measuring redundancy to identify whether an item deemed relevant contains vital information (Saat et al., 2018).

2.3.2 Collaborative filtering

As a result of the emergence of social media and the Netflix prize competition (Bennett & Lanning, 2007), which provided millions with access to social knowledge, this approach has recently enjoyed interest and progress from researchers. Unlike the content-based approach, collaborative filtering relies on other users with tastes similar to a target user to provide recommendations. Collaborative filtering measures users' taste similarities based on the similarity measures of previous ratings, which is why it is sometimes referred to as the 'people-to-people correlation' (Schafer et al., 2001). Collaborative filtering approaches can be grouped into neighbourhood- and model-based approaches.

Neighbourhood collaborative filtering approaches, whether heuristic-based (Adomavicius & Tuzhilin, 2005) or memory-based (Breese et al., 1998), directly predict ratings for new items from the user–item ratings available in the system. This can be carried out as a user-based recommendation or an item-based recommendation. User-based recommendations, for example, GroupLens (Resnick et al., 1994), Ringo (Shardanand & Maes, 1995), and Amazon (Zhao & Shang, 2010), exploit other users with rating patterns similar to the target user to predict new items. Meanwhile, item-based recommendations (Barkan & Koenigstein, 2016; Li et al., 2016; Sarwar et al., 2001), make predictions based on the target user's item ratings.

Model-based collaborative filtering uses ratings provided by users to learn a predictive model. The matrix factorisation model is a popular model-based approach for implementing collaborative filtering because of its accuracy and scalability (Hegde & Shetty, 2015). Other model-based approaches include neural networks (Salakhutdinov et al., 2007), Probabilistics

Latent Semantic Analysis (Hofmann, 2004), latent Dirichlet allocation (Blei et al., 2003), and the single value decomposition (SVD)-based model (Brand, 2003).

The collaborative filtering approach addresses the challenge of content-based over-specialisation. It also addresses the issue of making recommendations without content knowledge of an item. For example, items like images, music, and movies that have challenges with respect to content knowledge extraction for content-based recommendation can be used to make predictions with the collaborative filtering method of using user feedback (Hegde & Shetty, 2015). Despite having an edge over content-based filtering, collaborative filtering suffers from the sparsity problem and the cold-start problem (Cacheda et al., 2011).

The sparsity problem is a situation in which a dataset collection lacks the information required to produce recommendations. Much has been done to address the sparsity problem in various domains, but the problem continues to evolve in the CH domain (Pavlidis, 2018). Another challenge for collaborative filtering is cold-start (Lika et al., 2014a; Schein et al., 2002; Zhang et al., 2014), in which new users or items do not have a rating or opinion history. It is challenging for collaborative filtering to make CH recommendations without social knowledge from users.

2.3.3 Knowledge-based

Items like books, news, and movies are well-suited for knowledge exploitation via content-based and collaborative filtering for a recommendation because simple characteristics, such as age or time of production, may not determine the recommendation outcome (Bobadilla et al., 2013). However, products such as apartments and cars are not frequently bought, making it impractical to collect users' opinions. Thus, these products are not suitable for content-based and collaborative filtering approaches.

Knowledge-based techniques tackle these challenges by exploiting user requirements and using deep knowledge of the product domain before building recommendations. This approach relies on explicit knowledge provided by users as cases (Mirzadeh et al., 2005). Domain knowledge and user requirements are two aspects that distinguish the knowledge-based method from the content-based and collaborative filtering methods of exploiting knowledge for recommendations (Trewin, 2000). Case-based and constraints-based are two popular approaches for knowledge-based recommendations.

Case-based recommendations (Smyth, 2007) are created using case-based reasoning (CBR) techniques. CBR is the process of using similar past problems to solve a new problem. Fuzzy reasoning (Wu et al., 2008), for example, is one of the CBR techniques that generate new product or item ideas for enhancing the recommendation of new items. The case-based recommendation addresses the collaborative filtering challenges of dealing with objects' similarities.

Unlike case-based recommendations, constraint-based recommendations explicitly account for defined constraints. This approach establishes successful IR or item recommendations in domains where interactions between users and items do not frequently occur. Examples of a constraints-base recommender are VITA (Felfernig et al., 2007), a sales support environment for financial services located in Hungary, and the personalised conversational recommendation (Thompson et al., 2004).

One crucial advantage of knowledge-based recommendations over content-based and collaborative filtering recommendations is that it can recommend items that are not available or do not have a preference history.

2.3.4 Demographic

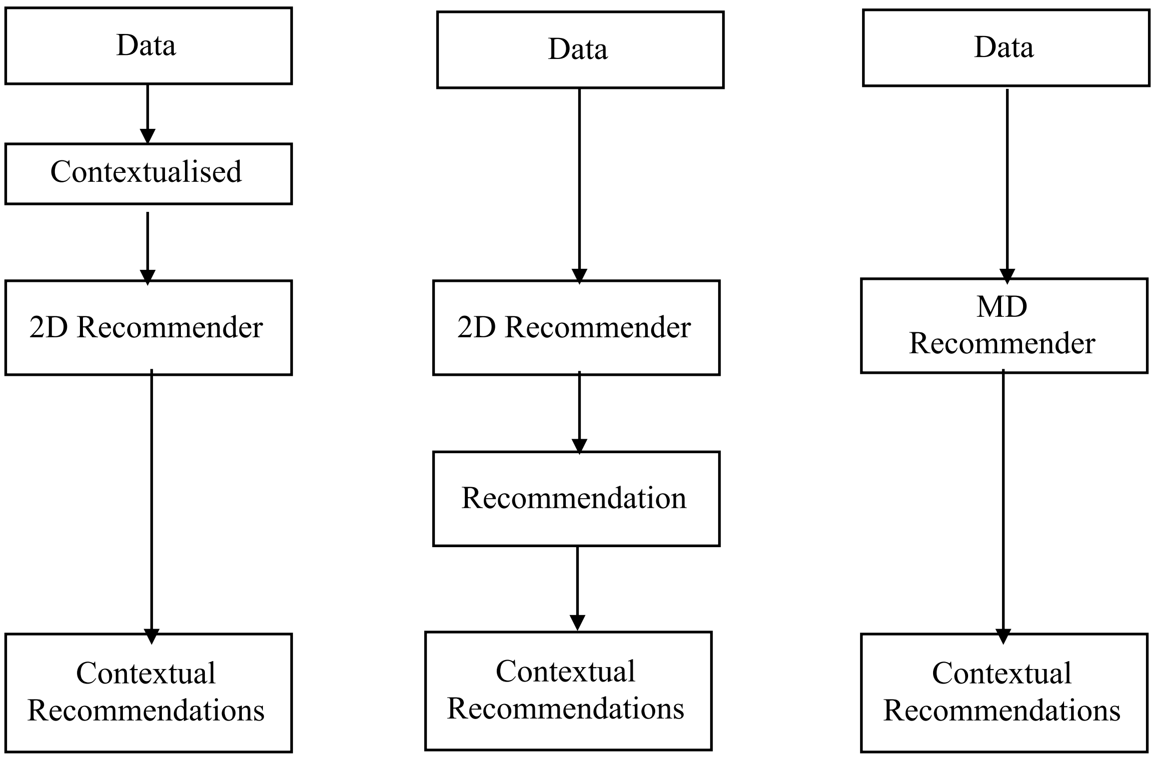
The demographic approach uses demographic information to make recommendations. The idea is that for every demographic niche, a different recommendation is expected to be generated.

Many e-commerce systems, for example, Amazon, sometimes adopt personalised solutions based on the demographic information available within the system.

The demographic exploitation of knowledge for recommendations is prevalent in the marketing literature. Despite its popularity, researchers have given little attention to this approach (Mahmood & Ricci, 2007).

2.3.5 Context-aware

Unlike the approaches discussed above, the context-aware approach is a comparatively underexplored research area because it is relatively new. The context-aware approach makes recommendations after considering contextual information. Time, place, geographical location, and similar users (users with similar taste) are types of context information that can be considered while making a context-aware recommendation. One well-known example of a context-aware RS is TripAdvisor (Adomavicius & Tuzhilin, 2011).



(a) Pre-Filtering (b) Post-Filtering (c) Contectual Modelling

Figure 2.2: Methods of incorporating context into recommender systems (Adomavicius & Tuzhilin, 2011)

There are three methods of incorporating context into RSs: pre-filtering, post-filtering, and contextual modelling (Panniello et al., 2009), as illustrated in Figure 2.2. Pre-filtering generates recommendations from contextual information selected or constructed from the most relevant two-dimensional (2D) data (users and items), as shown in Figure 2.2 (a). Post-filtering adjusts the recommendation result list generated from pre-filtering for each user, as shown in Figure 2.2 (b). The contextual model uses a multi-dimensional data (MD) RS, as shown in Figure 2.2 (c).

Panniello et al. (2009) presented a comparative study between contextual pre-filtering and contextual post-filtering in a context-aware RS. They found that good post-filtering is better than pre-filtering, but post-filtering is more expensive.

Online analytical processing (OLAP) is one of the MD models most widely used for context-aware recommendations (Adomavicius & Tuzhilin, 2011). OLAP analyses and manipulates MD for decision support, as shown in Figure 2.3. For example, the rating function *R (101,7,1) = 6* indicates that for Item ID 7 and User ID 101 (John), rating 6 was specified during the weekend. In this case, the term 'Weekend' is contextual information, time, added to the user and item for recommendations.

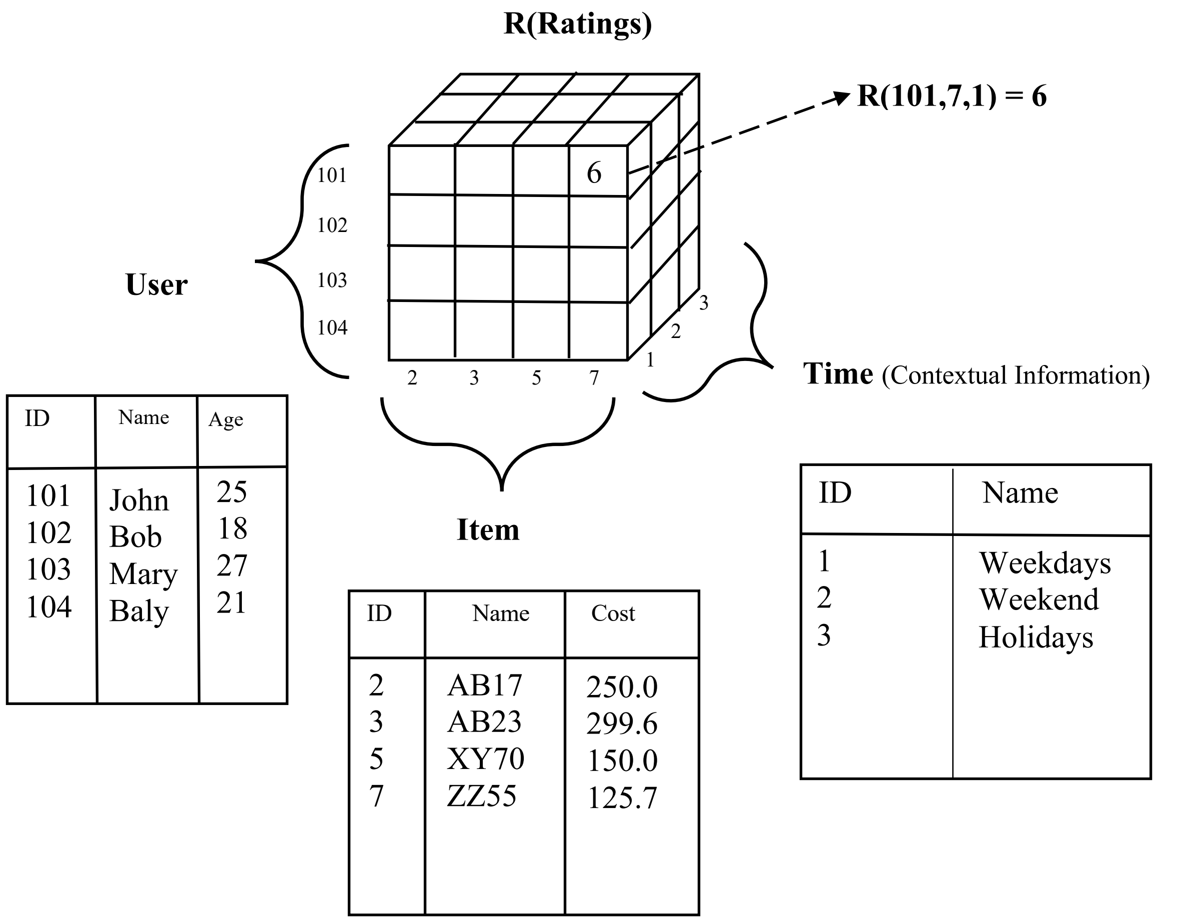


Figure 2.3: OLAP multidimensional data model for the User X Item X Time recommendation

2.3.6 Hybrid approach

The hybrid approach is a combination of two or more approaches. For example, a hybrid system combining content-based and collaborative filtering complements the limitations of both approaches. The three primary ways of creating hybrid RSs found in the literature are as follows:

- Ensemble design
- Monolithic design
- Mixed systems.

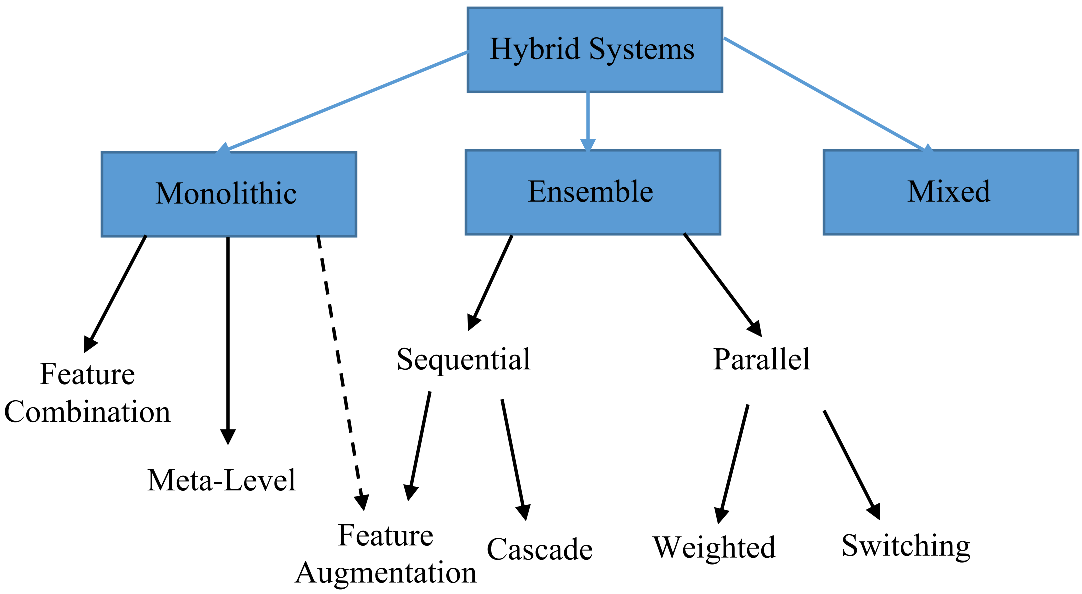


Figure 2.4: The taxonomy of the hybrid recommendation approach

Ensemble design: In this type of hybrid approach, off-the-shelf algorithm results are consolidated into a single and more robust result. For instance, one may join the rating outputs from a content-based approach and a collaborative approach into a single output. A critical variety exists as far as the specific approaches used for the combination procedure. The essential standard at work is not altogether different from the outline of ensemble techniques in numerous data mining (DM) applications, for example, classification, clustering, and exception investigation. One work that embraces this design is that of Bar et al. (2012), who introduced efficient methodologies for creating an ensemble of collaborative filtering models in light of a single collaborative filtering algorithm.

Monolithic design: In this situation, a coordinated recommendation algorithm is constructed by utilising different knowledge sources. Consequently, this approach tends to coordinate knowledge sources together more firmly and, thus, can recommend unexpected items or objects without having similar users' information. Examples of works that implement this design include feature combination (Zanker & Jessenitschnig, 2009), meta-level (Schafer et al., 2002), and feature augmentation (Burke, 2007a).

Mixed system: Like the ensemble approach to creating a hybrid RS, these frameworks utilise different recommendation algorithms as a black box of information. However, the items recommended by the different RS frameworks are exhibited side-by-side.

Burke (2007) reported on a set of experiments on 53 hybrid RSs to measure their performances. From his findings, monolithic hybrids were weaker and ineffective for sparse datasets. EntreeC (Burke, 2002) is an example of a monolithic hybrid system that does not work well with sparse datasets. It combines knowledge-based and collaborative filtering methods to recommend restaurants to improve the performance of collaborative filtering by extracting semantic ratings from the knowledge base.

Sparsity, scalability, and cold-start are issues that affect the recommendation quality (Ricci et al., 2011). A hybrid approach that combines content and metadata can reduce the problem of cold-start and produce a quality recommendation, but it cannot improve the individual knowledge exploitation approach (Zhang et al., 2010). Horsburgh et al.(2015) presented a hybrid representation of pseudo-tags learned from music content and tags from music listeners that addressed both the sparsity and cold-start problems in music RSs. The next section discusses more on CH recommendations.

2.4 Cultural heritage recommender systems

CH RSs have become a popular approach for users to explore large CH collections and make new discoveries. Online systems, such as Europeana,[2] the Google Art Project,[3] and the system of the Rijksmuseum,[4] are examples of the CH IR and RSs that provide millions of online users

[2] https://www.europeana.eu/portal/en
[3] https://artsandculture.google.com/
[4] https://www.rijksmuseum.nl/en

with access to over 500 million artefacts, books, films, and music from museums, galleries, libraries, and archives every day. These systems offer a gateway to steering large-scale multimedia collections of CH objects and materials. The growth of CH RSs is recent, as CH users, especially tourists and non-professionals, continue to engage and discover new CH materials online (Petras et al., 2017). Figure 2.5 shows this growth over the past 11 years using the results of academic papers. The blue bar in the figure indicates the number of articles that have either 'cultural heritage recommender system' or 'cultural heritage recommendation' in their title, and the brown bar presents the published articles from the annual ACM Recommender Systems (RecSys) conference that contain the term 'cultural heritage' in their title. These numbers were acquired through Google Scholar.[5]

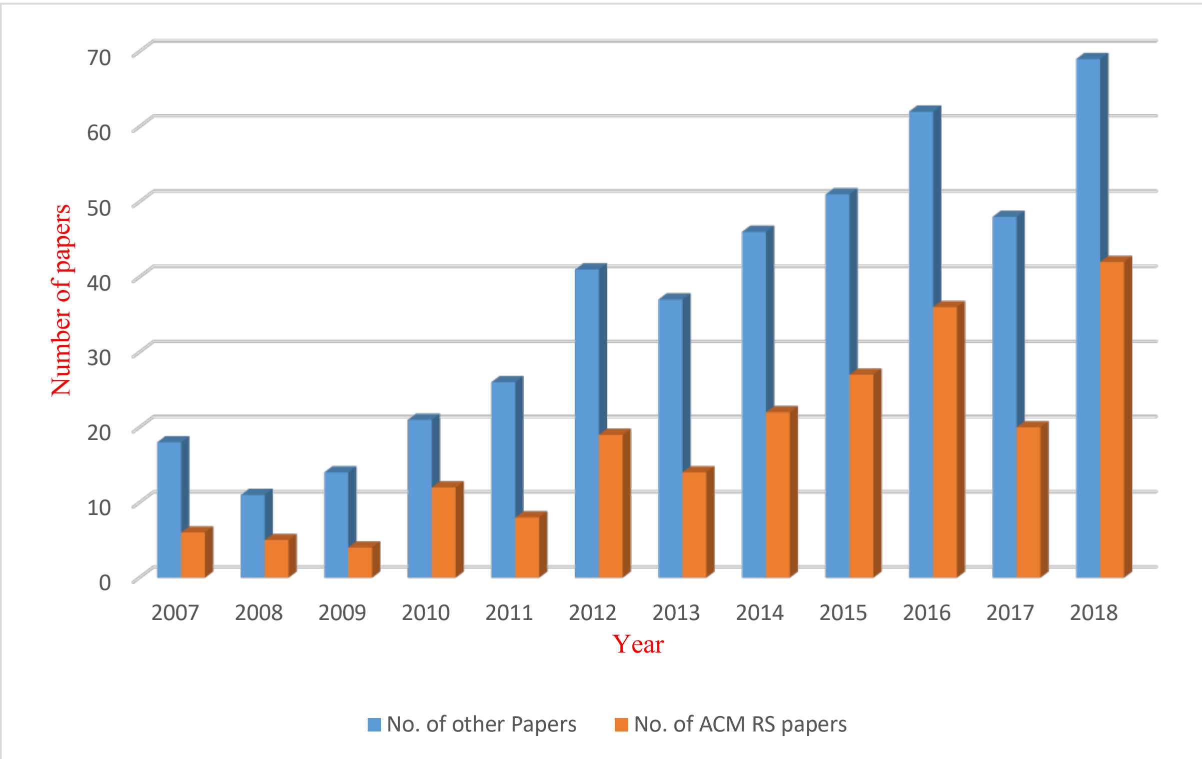


Figure 2.5: Articles related to cultural heritage recommender systems

[5] https://scholar.google.co.uk/

Early work on CH recommendations focused on semantic relations (Wang et al., 2008) and ontologies (Hyvönen, 2009) for quality recommendations. By the end of 2010, the direction had entirely shifted to exploiting social networks to bridge the semantic gap and evaluating users' study experiments and experiences with digital CH collections to provide personalised recommendations (Hampson et al., 2012; Knijnenburg & Willemsen, 2015; Vosinakis & Tsakonas, 2016). Despite this growth, CH recommendations continue to face challenges, such as the **cold-start problem** (Hong et al., 2017), **out of context recommendations** (Smirnov et al., 2017), and **most similar objects but bad recommendations** (Petras et al., 2017), which provide room for improvement in content and social knowledge representations. Evidence for this is summarised in findings from the literature directed at the three most popular CH RSs, in terms of knowledge representation, assembled from available domain knowledge sources (see Table 2.1).

Table 2.1: Summarised findings from the literature concerning the three most popular cultural heritage recommender systems

RS / **KR**	**Enrichment of Metadata**		**Personalisation**		
	Social	***Content***	***Individual***	***Social/Group***	***Content***
Europeana	☆	★	★	☹	☆
Rijksmuseum	☹	★	★	☹	☆
GAP	☆	★	☆	☹	☆

★ Provided by that facility
☆ Partially provided by that facility
☹ Not

The literature findings show that all three systems have a weak social knowledge representation. Thus, this study focuses on building a stronger social knowledge representation and further integrating the content and social knowledge representation to bridge the knowledge gap to address the CH recommendation challenges, such as the cold-start problem, out of context recommendations, and similar objects but bad recommendations. The next section discusses the challenges of CH recommendations identified in the literature.

2.4.1 Cold-start problem

Cold-start is a common problem with respect to CH recommendations (Sansonetti et al., 2019). It is a condition in which the system cannot produce quality recommendations without initial information from the user and/or CH objects. According to work presented by Lika et al. (2014b), the cold-start problem can be categorised into three areas: (a) new users, (b) new objects, and (c) both new objects and new users. All of these problems result from information sparsity for both new objects and users.

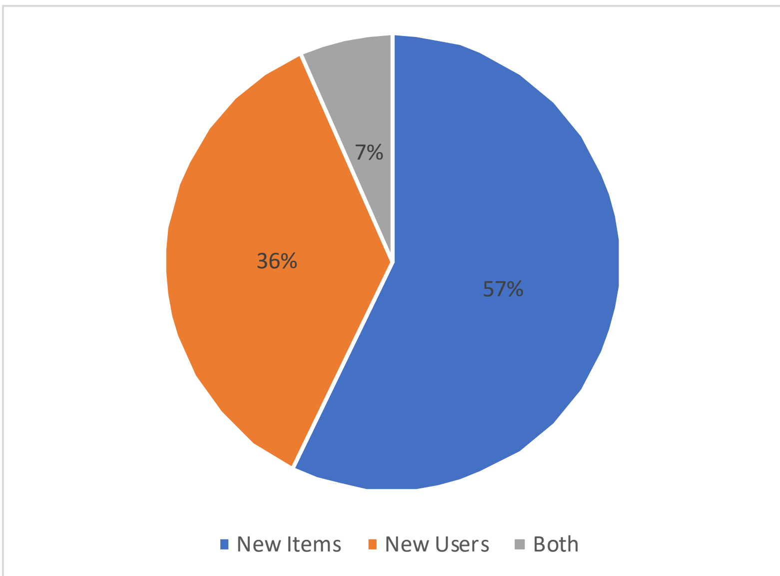


Figure 2.6: Reviewed articles related to the cold-start problem

This study reviewed approximately 180 research articles related to cold-start problems. It was discovered that the majority of the research carried out to address this problem mainly focused on the items or objects rather than the users, as indicated in Figure 2.6; 57% of the articles attempted to augment the sparsity of information on items or objects to produce quality recommendations, while 36% focused on integrating users' social knowledge to make recommendations. Only 7% concentrated on both items and users. This study focuses on addressing the cold-start problems from both angles: CH objects and users.

In the case of the CH domain, user interactions with CH objects are the variables used to assess the similarity of user profiles for the prediction of unseen CH materials. Therefore, when there are new CH objects and users that have no record of interactions with the objects, the system will suffer a cold-start problem. This can lead to poor quality recommendations because the system assumes that active users will largely react to popular CH objects that were rated highly by similar users with similar opinions (Wei et al., 2017). In most situations, users are not interested in reviewing or rating items previously seen, which results in providing low data quality required to produce good CH recommendations and, thus, the cold-start.

The literature reports various attempts to address the cold-start challenge in other domains. For example, in the music domain, pseudo-tags learned from the track's content have been introduced to augment social tags (Zheng et al., 2018) and the neighbourhood-based attribution technique (Lika et al., 2014b). However, these approaches do not address the problem completely, especially in the CH domain, due to poor knowledge representation. In most cases, the combination of two or more knowledge representations, for example, content and metadata or social data, are used to address this problem. This can be accomplished using a simple hybrid approach to content and social knowledge representations, which reduces the effect of cold-start. The content representation is considered when social knowledge is not available for requested queries and then switches to social representation as the metadata and social knowledge increase, as indicated in Figure 2.7.

The green dashed line in Figure 2.7 represents the quality of a content-based RS, which does not rely on metadata and social knowledge. For a system with low metadata and social knowledge, the content-based approach produces better recommendations, but as metadata and social knowledge grow, the system improves.

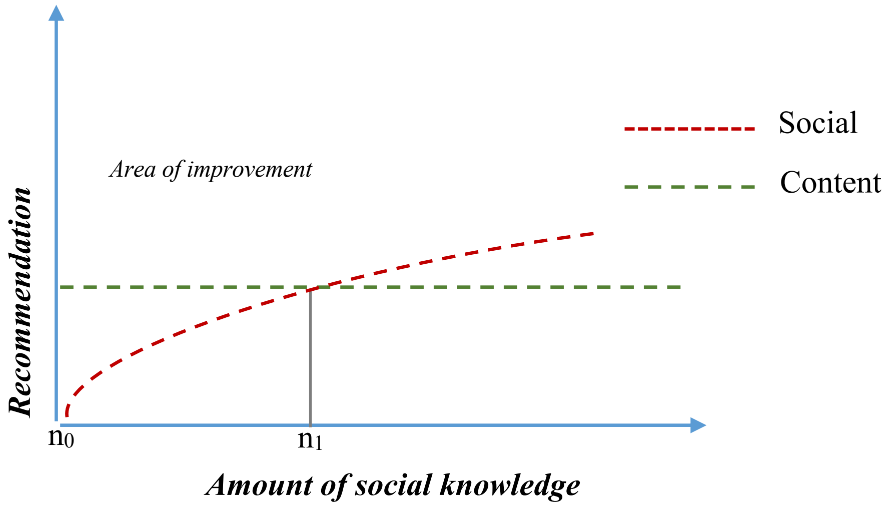


Figure 2.7: The cold-start problem

The challenge is determining the point at which the system should switch from content representation to social representation. A hybrid approach that assembles knowledge representations from a static combination of knowledge sources has been developed, but this study focuses on a hybrid approach that assembles knowledge representations from ***dynamic*** combinations of sources so that each knowledge representation can contribute to the current search status, which requires active maintenance of knowledge representations from available domain sources of knowledge. Chapter five and six discuss the integration of multiple knowledge representations from various knowledge sources.

2.4.2 Out of context recommendations

Out of context recommendations are those that users may not like because of the current situation (e.g. time or location); for example, users are not likely to want a jacket in the summer. The challenge of out of context recommendations results from low contextual information within the dataset collection. Contextual information, such as time, place, and geographical location, requires exploiting various available domain knowledge sources, which the current RSs lack. Under normal circumstances, recommendation algorithms identify target users' similar resources to provide personalised recommendations. However, in most cases, they fail to consider the user's current situation. As a result, the suggested resources may not meet the user's interest despite being similar. This problem often happens in tourism-related RSs (Borràs et al., 2014), which are within the CH domain.

2.4.3 Most similar objects but bad recommendations

Most similar objects but bad recommendations is another challenge of CH recommendations. Most research has concentrated on enhancing the accuracy of the RSs probability predictions (Ricci et al., 2011) but not recommending every similar object likely to be of interest to the users. For example, not all of the artwork featuring 'people' liked by similar users could be a good recommendation for the target user. Some research has highlighted this problem. For example, Thompson et al. (2014) discovered that most of the recommendations produced have similar tastes as the target user but are bad recommendations that do not satisfy the users.

Another scenario to describe this problem is a situation when a system recommends an object that is already familiar to the target user, demonstrating that even the most precise recommendations, as indicated by the standard measurements, are not good recommendations. For example, a user is looking to discover a new painting, but the system provides the user with paintings they have already viewed, which, unfortunately, is possible with the current RS, as highlighted by Sansonetti et al. (2019) in their work to enhance CH recommendations.

Most of the reported research attempts to address this problem by proposing to go beyond the traditional accuracy measurements and related evaluation methodologies by incorporating social knowledge that provides contextual information and user-driven directions for personalised recommendations, as indicated in Amato et al. (2018). However, this approach also has limitations.

The literature further reveals that these challenges lie within: (i) the domain knowledge sources available for knowledge discovery (Ristoski & Paulheim, 2016), (ii) choice of appropriate recommendation techniques (Aggarwal, 2016), and (iii) the user search interface (Dumas et al., 2014). The previous work on recommendation techniques was discussed in Section 2.2. Therefore, the knowledge discovery and user search interface for CH recommendations are discussed in the next two sections.

2.5 Knowledge discovery for cultural heritage recommendations

As discovered in the literature, the challenges of CH recommendations depend on the available domain sources in which the knowledge for CH recommendations is discovered (Ristoski & Paulheim, 2016). Therefore, this section discusses knowledge discovery in the context of CH recommendations, which includes knowledge extraction and knowledge representation.

In the context of CH recommendations, knowledge discovery is the process of discovering essential features that describe CH materials through knowledge extraction and preparation and knowledge representation, as shown in Figure 2.9. Knowledge extraction is the core component of knowledge discovery; knowledge extraction tools and algorithms are used to extract interesting and useful knowledge patterns from available domain knowledge sources for CH recommendations. Figure 2.9 presents the knowledge discovery process, which consists of two phases: knowledge extraction and knowledge representation. The arrows show the data flow. Knowledge extraction and knowledge representation of knowledge discovery are further discussed in Sections 2.5.1 and 2.5.2, respectively.

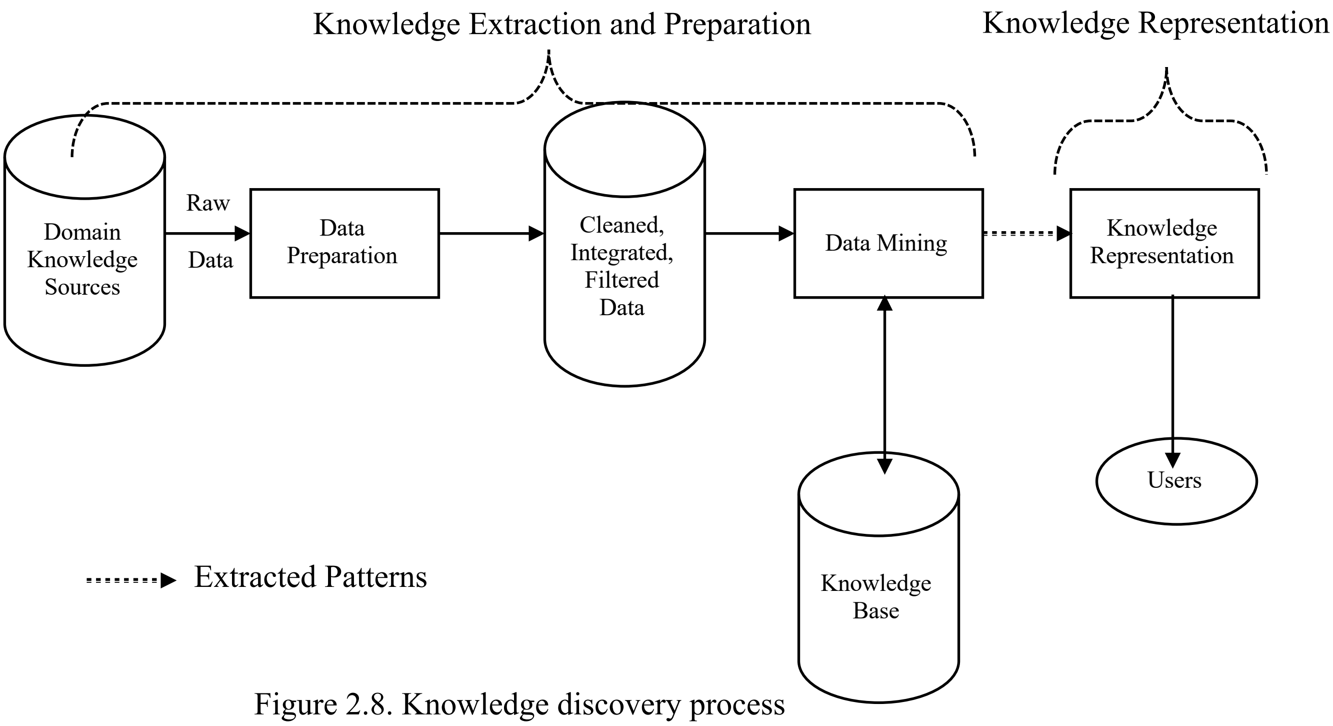


Figure 2.8. Knowledge discovery process

2.5.1 Knowledge extraction and preparation

As shown in Figure 2.8, the first phase of knowledge discovery, knowledge extraction, has four components: domain knowledge sources, data preparation, filtered data, and DM. Knowledge extraction is the process of building knowledge for CH recommendations from the available domain knowledge sources (Bandyopadhyay & Maulik, 2005). The knowledge sources can be either structured (extensible markup language, relational database) or unstructured (images, documents, text). The knowledge outcomes from the domain knowledge sources are prepared to be machine-readable and interpretable to facilitate inferences for knowledge representation, which is the second phase of knowledge discovery (see Section 2.5.2).

There are different knowledge extraction approaches, depending on the knowledge source available and the demand. Structured sources are relatively unchallenging for extracting knowledge because relevant terms are already labelled (Unbehauen et al., 2012). However, in this study, the domain knowledge sources are unstructured and heterogeneous. Digital CH materials come in different formats or types, which, by default, poses more challenges when compared with structured sources. The data sources for this study come from various CH institutions' webpages, for example, galleries, museums, libraries, and archives, which are unstructured sources. Therefore, important knowledge is located within documents that are images or text and in different languages. The issues and future direction of knowledge extraction from text and image sources are further discussed in detail.

2.5.1.1 Knowledge extraction from text sources

Knowledge extraction from unstructured text sources involves sentence segmentation, tokenisation, morphological and lexical analysis, syntactic analysis, and domain analysis, as shown in Figure 2.9.

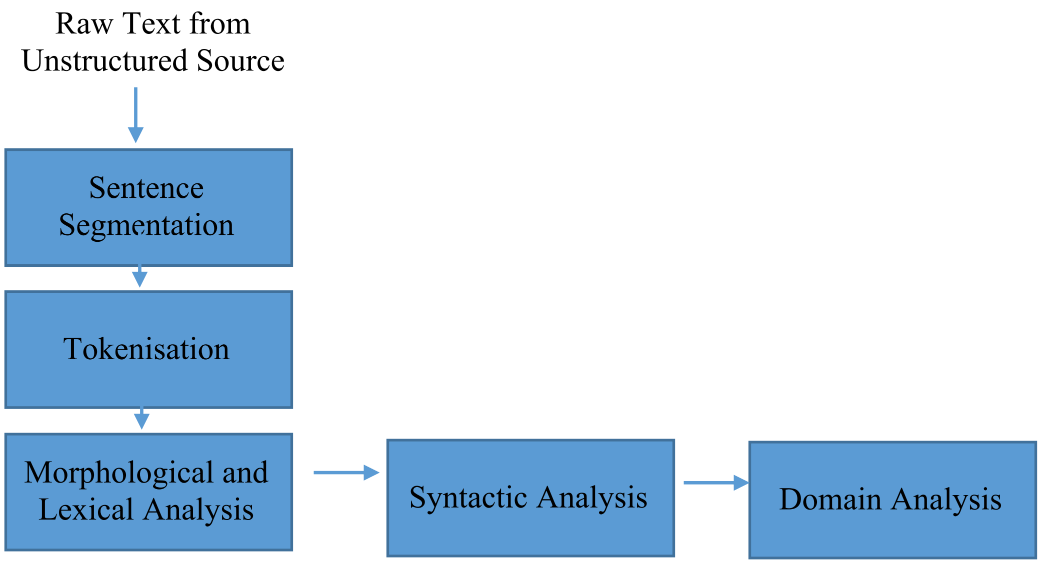


Figure 2.9: Knowledge extraction (text) architecture

The first component, sentence segmentation, divides the raw text into meaningful units. These units are further split into tokens (words). The morphological and lexical analysis component deals with sense disambiguation and parts of speech, tagging the words generated from the tokenisation component. Syntactic and domain analysis are critical components of knowledge extraction as they establish a relationship among objects within the collection. In some domains, extracting information that can form a relationship between objects can be difficult. For example, in CH collections that contain library documents, archives, and paintings, extracting relational information between objects is difficult because the objects are of different types and formats than text (Bandyopadhyay & Maulik, 2005).

A variety of tools to extract information from text sources have been proposed, for example, the general architecture for text engineering used for text extraction and analysis (Cunningham et al., 2011). Other tools include the Natural Language Toolkit for natural language processing (Bird, 2006), optical character recognition (Vesanto et al., 2017) for converting scanned documents into machine-encoded text, and SocialBus[6] for processing social network messages. Although tools exist to extract knowledge from unstructured sources, several obstacles and cost factors remain a challenge.

However, some tools are suitable for particular tasks. For example, the medical document from the seventeenth century in Figure 2.10 presents text describing a charms mixture for healing. Optical character recognition tools are applied to extract knowledge and recognise the patterns of the text from the scanned historical document (Laramée, 2019).

---

[6] http://reaction.fe.up.pt/socialbus/

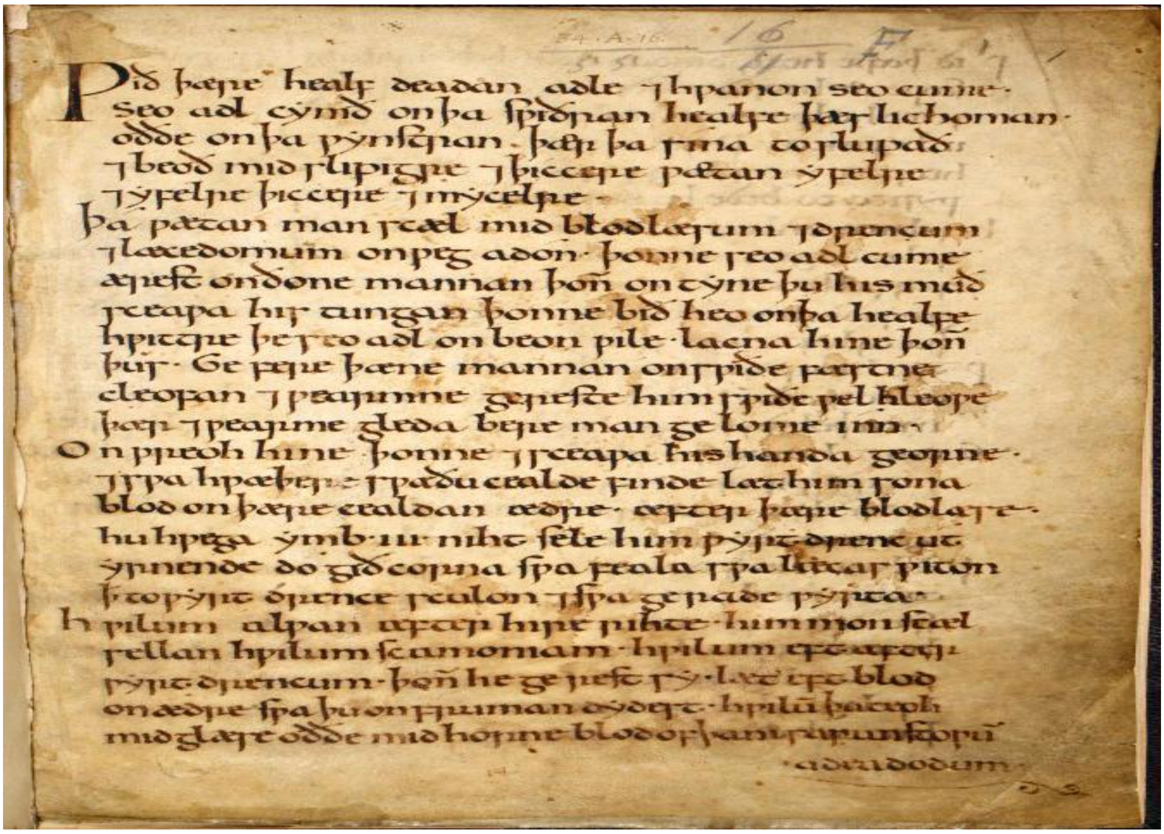


Figure 2.10: Anglo-Saxon medical recipes corresponding to book 2, chapter 59 of *Bald's Leechbook of 17th Century*

Gangemi (2013) presented a scenery analysis of knowledge extraction tools when applied to digital CH. Gangemi considered 14 different tools and investigated the feasibility of comparing tools when used for text document tasks but found that formal correspondences must be created between the text population and the natural language processing primary functions. Natural language processing enables the derivation of human or natural input meaning.

2.5.1.2 Knowledge extraction from image sources

Images as data differ from text in terms of their nature. Consider the digital painting in Figure 2.11; there is little or no text that describes the painting. Extracting the semantic contents from the image is challenging (Bandyopadhyay & Maulik, 2005). Recently, deep learning techniques, such as convolutional neural networks, have been used to extract knowledge from within images (e.g. paintings) (Bar et al., 2015). Bar et al. (2105) looked at the perceptiveness of the CH object's important features in recognising artistic styles in paintings, which are then used to augment the sparsity problem of the content knowledge representation and provide a personalised recommendation. This study applied a similar approach to image-type sources.

Figure 2.11: Digital painting from Rijksmuseum in the Netherlands

Typically, RSs model users and items by discovering the concealed-relation measurement that reveals the users' preferences toward items or objects. Critically, such measures were found based on user interactions (such as browsing history, reviews, and tags) and other information, such as object attributes and context-aware knowledge. However, in the case of digital paintings (a knowledge source for this study), *visual appearance* is one of the essential features that is ignored (Bar et al., 2015). Thus, in this study, visual appearance is considered by extracting it from the available domain sources and incorporate it into the new, harvested CH dataset (see Chapter Three).

2.5.2 Knowledge representation

After essential knowledge is extracted from various domain knowledge sources and knowledge outcomes are prepared, the next important task is presenting the knowledge, which is the second phase of knowledge discovery. The knowledge representation should support the assertion of new knowledge and inferencing. Generally, there are three models for representing knowledge:

- History-based (Palopoli et al., 2013),
- Classifier-based (Bobadilla et al., 2013), and
- Matrix-based (Hegde & Shetty, 2015)

The *history-based model* presents a list of objects with their respective user reviews, web browsing histories, and e-mail content boxes, in some cases, as a user profile. This model is commonly used in e-commerce. Amazon[7] and eBay[8] are two famous examples of e-commerce RSs that use this approach (Nguyen et al., 2014). WebSell, an e-commerce website, uses a similar method by providing two lists of purchased products, *uninterested* and *interested*, as a user profile (Cunningham et al., 2001). David et al. (2013) also represented knowledge using the history-based model in their e-mail filtering system that keeps track of e-mails and comments made by users.

The *classifier-based model* learns from knowledge extracted to represent the structure of the classifier as a user profile. Examples of this model include the Bayesian network (Jiang et al., 2018), decision tree (Yang et al., 2018), neural network (Covington et al., 2016), and inducted rules.

---

[7] https://www.amazon.co.uk/
[8] http://www.ebay.co.uk/

Unlike the history- and classifier-based models, *matrix-based approaches* represent knowledge as a matrix. This is popularly used in collaborative filtering RSs (Montaner et al., 2003). For example, GroupLens (Resnick et al., 1994) represents user ratings and movies as a matrix to predict scores based on heuristics for movie recommendations. Another example is Netflix, which utilises a user–movie rating matrix: each cell in the matrix represents the user's rating of a particular movie for movie recommendations (Takács et al., 2008). The location–activity matrix with GPS history data was represented in Zheng et al. (2010) as a user profile to recommend places based on previous locations visited by the user. In the **vector space model**, another form of matrix-based representation, items or documents and users are represented as a vector of features with associated values; these features are usually concepts or words. The association value can either be the relevance, frequency, or probability of the features. For example, each user can be represented as a vector such that users with similar vector contents or tastes have identical vectors. Burke (2002) used a similar approach by utilising multiple features of vector representations to make recommendations.

All of these knowledge representation models have strengths and weaknesses. Choosing which model to deploy depends on three aspects: *the problem to be addressed, the knowledge to be fulfilled, and how the knowledge solves the problem* (Bench-Capon, 2014). The issues and future directions of knowledge representations in the CH domain are further discussed in Section 2.6.

2.6 Knowledge representations in the cultural heritage domain: Issues and future directions

Representing CH content knowledge remains a major challenge because CH materials have different traditional ways of describing their content, making the structure format of CH content difficult to interchange and transform. For example, a CH object called a 'figure' might have different interpretations from different traditional backgrounds; some may consider it a ‘statue’, while others may consider it a ‘chieftain’. Therefore, CH knowledge representation requires metadata aggregation and continuous content refinement (Petras et al., 2017).

According to Alspaugh and Lin (2016), 70% of the search requests in CH domains are for named entities, for example, geographical, author, and object names, either in the language in which they were distributed or the word the user is most comfortable with. Most users prefer their native language for searching for and finding information on CH objects (Gäde, 2014). It is important to consider these attributes while representing CH knowledge for retrieval and personalised recommendations.

In the CH domain, RSs face challenges on the personalised retrieval of relevant information due to a lack of rich semantic and content knowledge in the collections (Wang et al., 2007). Some research has attempted to address these challenges; for example, Petras et al. (2017) and Mensink and Van Gemert (2014) aggregated metadata from various sources to enrich their databases using different vocabularies, as shown in Table 2.2. Vocabulary is an arrangement of familiar words inside a person’s dialect. A vocabulary, typically created with age, fills in as a valuable and crucial tool for correspondence and gaining information.

Table 2.3: Entities, target vocabulary, and number of enrichments in Europeana (Petras et al., 2017)

Entity	Vocabulary	Number of Enrichments
Place	Geonames	19,269,339
Concept	GEMET	14,633,522
Concept	DBPedia	6,022,071
Agent	DBPedia	889,152
Time Period	Semium Time	21,925,367

The issues highlighted in the literature review with regards to knowledge representations in the CH domain emphasises that a good knowledge representation requires active maintenance. This thesis explores rich semantic and social knowledge representations. The study exploits various domain knowledge sources to assemble content and social knowledge representations for CH recommendations. In the next section, the methods and DM used to assemble knowledge representations are further discussed.

2.7 Data mining methods for recommender systems

Typically, RSs apply methods and techniques adopted from related areas, such as IR and human–computer interaction. However, such related areas' core techniques are considered part of DM, which is the process of extracting meaningful information from large, pre-existing data collections. DM consists of three processing steps: data pre-processing, data analysis, and result interpretation (Witten et al., 2016). This section discusses different approaches to data pre-processing, data analysis, and interpretation, as shown in Figure 2.12.

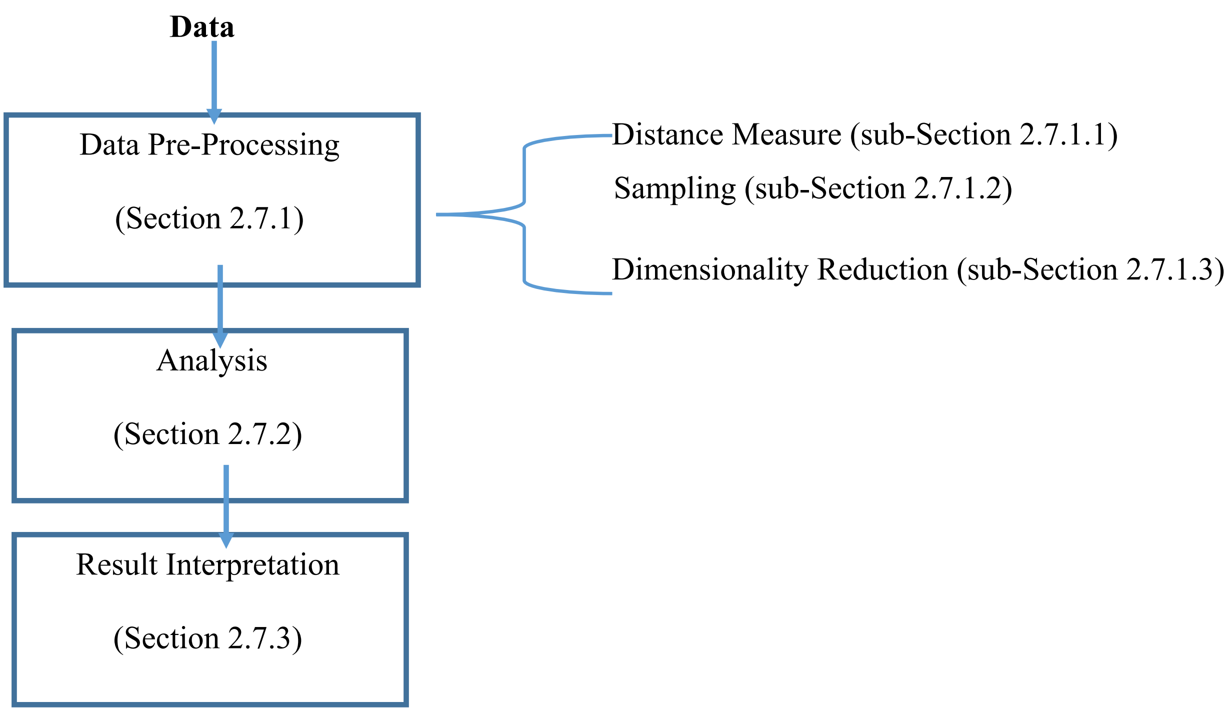


Figure 2.12: Steps and methods in data mining

It is important to note that this section does not have a thorough review of DM methods but, rather, highlights the role they play in the production of quality recommendations and discusses the successful DM techniques applied in the RS field.

2.7.1 Data pre-processing

In the context of this study, data refers to a collection of CH objects (e.g. digital paintings, archives, and historical medical documents) and their features. These features describe the characteristics or properties of a CH object collection. A primary choice for an artificial intelligence system implementer is the type and source of data that the system will employ. In the case of RSs, especially in the CH domain, users and CH objects are the two entities about which the system should have knowledge. The choice of the CH domain for this study restricts the kind of knowledge sources that an RS may exploit (see Section 2.2). Content and social knowledge are the two categories of knowledge sources considered for this study. Data extracted from such sources are required to be pre-processed to be used by analysing techniques in the analysis step.

In this section, three important issues in the data pre-processing stage of building CH RSs are discussed. For the first issue, *similarity measures*, this study focuses on reviewing the literature on different approaches to express similarity distance. The next issues discussed in this section are *sampling* and, finally, the different methods of reducing dimensionality.

2.7.1.1 Similarity measures (Distance measures)

This is the common method of measuring similarity distance between objects, users, or users and objects. One of the recommendation techniques, collaborative filtering (discussed in Section 2.4.2), uses the k-nearest neighbour (kNN) classifier for analysis to make CH recommendations (Hegde & Shetty, 2015). kNN is a classification technique that is highly dependent on the quality of the similarity measures applied.

There are different approaches to measure the similarity distance, including the Euclidean distance (Sun et al., 2011); the Minkowski distance, which generalises the Euclidean distance (Merigo & Casanovas, 2011); the Mahalanobis distance (De Maesschalck et al., 2000); cosine similarity (Amatriain et al., 2011); and the Pearson correlation (Melville et al., 2002). Each of these approaches has advantages and disadvantages, but traditionally, RSs have used either the Pearson correlation or cosine similarity (Ricci et al., 2011). In this study, cosine similarity is used to measure the similarity distance between CH objects, users, and users and CH objects. This provides quality measures from the vector space model (Sidorov et al., 2014), which is the knowledge representation model used for this study.

2.7.1.2 Sampling

Sampling is the process of selecting a sub-portion of relevant information from a large portion of information in DM. Sampling can be considered in the pre-processing and interpretation stages. The important issue is finding an appropriate sample containing all of the relevant information of the original data collection.

RSs employ various sampling approaches, but the most popular ones are *random sampling with replacement* and *random sampling without replacement* (Amatriain et al., 2011). In random sampling with replacement, objects selected can be used more than once; they are not removed from the population. In the case of sampling without replacement, once objects are selected, they are removed from the population. The challenge of random sampling is over-specialisation to a certain part of the dataset. C*ross-validation* is used in this study to address the challenge of over-specialisation, as highlighted by Amatriain et al. (2011). Cross-validation is a statistical technique that assesses the statistical analysis results and generalises them to an independent dataset.

2.7.1.3 Reducing dimensionality

Usually, in RSs, there is very sparse knowledge that describes the objects' features. For example, some objects in the collection have limited information that describes their features. In this situation, the distance measurement between the objects is less valuable in a high dimensional space. Thus, dimensionality reduction is required by transforming the original vector into a lower-dimensionality space without interrupting its original values. There are two popular methods of reducing dimensionality: SVD (Sarwar et al., 2002) and principal component analysis (Jolliffe, 2011).

SVD, also known as matrix factorisation, is a technique used to reduce a high-dimensional space without interrupting its original features. SVD has been used as a tool for improving social knowledge representations to produce quality recommendations (Paterek, 2007). Other notable works that used SVD to reduce the dimensionality space for TV programme recommendations include those carried out by Barragáns-Martínez et al. (2010) and Bennett and Lanning (2007), for Netflix. The main drawback of SVD is overfitting, which can be addressed by updating the factorised approximation online, as described by Rendle and Schmidt-Thieme (2008). SVD is adopted in this study to integrate content and social knowledge representations to improve CH recommendations.

2.7.2 Analysis

This section discusses the prediction and description methods for CH recommendations and analysis. The prediction methods are the classification algorithms, and the description methods include association rule mining and clustering algorithms.

Various classification algorithms have been used for RS: kNN (Adeniyi et al., 2016), decision trees (Bouza et al., 2008), ruled-based classifiers (Pappas & Popescu-Belis, 2013), Bayesian network classifiers (Friedman et al., 1997), artificial neural network classifiers (Zurada, 1992), and support vector machine classifiers (Cristianini & Shawe-Taylor, 2000). Each of these algorithms has potentials and limitations, depending on the task at hand. Herlocker et al. (2004) presented a comprehensive review of these algorithms with respect to RSs.

Description methods define the association distances between objects or users. Algorithms in this category include association rule mining, which makes predictions based on objects' occurrences with the objects in the collections (Lin et al., 2000), and k-means clustering, which is a partitioning approach (Li & Kim, 2003).

2.7.3 Result interpretation

The results obtained using classification algorithms and descriptions must be interpreted to produce quality CH recommendations. Result interpretation is the process of confirming whether the produced CH recommendations are good or bad. There are various ways to interpret results for CH recommendations, which include accuracy measurement, recall and precision, F1 score, and receiver operating characteristics curves (Pavlidis, 2018).

Accuracy measurements interpret results through the percentage of average correct predictions provided by a classifier to generate recommendations, as discussed by Cremonesi et al. (2011). Recall and precision are mostly used for results obtained from unbalanced datasets. Recall interprets results by measuring the actual positive prediction generated for CH recommendations, while precision describes the precision of the obtained results (Ma et al., 2015; Pavlidis, 2018). The F1 score and receiver operating characteristics curves, recall, and precision harmonic mean come into play when both recall and precision are need. For the purpose of this study, accuracy measurements, for example, association scores, are used to interpret results for CH recommendations.

2.8 Visual search interface

Visual perception and its capability to provide an interactive visual summary of an entire domain are the key factors in information visualisation. A large proportion of human senses is occupied by visual perception (Nørretranders, 1991). Therefore, it is important to consider human visual perception when building a VSI for exploring large datasets of CH collections. Furthermore, a CH RS is a platform for the search and exploration of CH objects, which demands an interface that encourages search by exploration (Wilson et al., 2010). Some of the features that encourage the exploration of CH materials for new discovery include the visual presentation of the collection's summary, personalised recommendation presentations, and a user feedback interface. In this section, issues surrounding VSI features are discussed.

2.8.1 Domain collection summary

Searching for and exploring new information within a large document collection is challenging due to the lack of a domain summary presentation of the whole collection. Even though CH object search systems like CULTURA (Hampson et al., 2012), SCRABS (Amato et al., 2017), and Europeana (Petras et al., 2017) made digital CH resources available to experts and the wider public, users still struggle to make new discoveries as a result of the absence of a domain summary of the digital collections on these platforms' user interfaces (Amato et al., 2018).

A similar challenge was also highlighted in work presented by Ciocca et al. (2012). Their work provided a search interface for browsing museum image collections on multi-touch displays but lacked a visual summary, posing challenges to users who wish to search by exploration. However, an interactive visual summary is one of the features that this study explores including while building the user interface for the proposed hybrid RS.

2.8.2 Visualisation

Visualisation is one of the key features that search interfaces required for the exploration of information. The Google Art Project presents a VSI for high-resolution images of museum collections from highly recognised museums around the globe (Müller & Winters, 2018). ArtVis is another visual interface that combines visualisation and analysis of artwork collections (Dumas et al., 2014).

However, neither the Google Art Project nor ArtVis provides users with a dynamic approach that could initiate an exploratory search from the interface, which allows room for improvement that this study explores. Wang et al. (2008) provided an interface that describes semantically enriched museum collections, but they did not provide an interactive dynamic visual interface that could allow users to initiate their searches.

2.9 Evaluating recommender systems

The critical aspect of any research study is the *evaluation*. This is the procedure that critically assesses the research findings or systems, which include gathering and breaking down information about the research exercises, qualities, and results. The purpose of the *evaluation* is to assess and judge the research results, enhance its viability, and illuminate the right decisions on the findings. Unlike other research areas, RSs focus on recommendation prediction accuracy (Bobadilla et al., 2013).

Initially, researchers evaluated and ranked RSs based on their prediction powers. Nonetheless, it is currently broadly agreed that accurate prediction is significant, but this is insufficient for providing useful recommendations (Hegde & Shetty, 2015). For example, not many users anticipate recommendations from their exact tastes; rather, they are prepared to discover new items without using their past preferences. Therefore, it is essential to identify the set of variables or properties in the context of the CH domain that could influence the expected discovery.

In this section, the literature related to RS experimental settings and properties is critically reviewed.

2.9.1 Experimental settings

According to Shani and Gunawardana (2011), three possible experiment scenarios can be used to evaluate RSs: offline experiments, user studies, and online experiments. In all of these scenarios, a few guidelines, such as hypotheses, controlling variables, and generalisation power, are vital to follow.

a) Offline experiments: These experiments are relatively low-cost when compared to user studies and online experiments. This experiment process is used to tune algorithms' parameters and advance the best-tuned parameters to the next stage of the experiment, as carried out by Gomez-Uribe and Hunt (2016). A work presented by Cantador et al. (2017) on user preferences and recommendations tuned algorithms' parameters to complete an offline experiment for the system evaluation.

Offline experiments are especially useful for simple user-models (Ricci et al., 2011). A critical advantage of offline experiments is that they can simulate user interactions without engaging user studies and online experiments that are relatively expensive to conduct. However, if the user model is inaccurate, it may lead to an optimising framework with a performance that does not correlate with its real-time performance, as highlighted by Bobadilla et al. (2013). Therefore, in some cases, offline experiments are challenging to conduct.

b) User studies: Unlike offline experiments, user studies provide researchers with an opportunity to measure user behaviours when interacting with the RS (Shani & Gunawardana, 2011). This privilege allows researchers to gather qualitative data and crucial information through questionnaires and interviews to interpret quantitative outcomes. User studies are more convenient than offline experiments for the CH domain (Amato et al., 2018).

However, user studies have some shortfalls. User studies are costly to carry out; they require costly user participation and time consumption. Amato et al. (2018) took 16 months to obtain the participants required for their experiments. Proper guidelines on how to conduct a user studies experiment without difficulties were provided by Knijnenburg and Willemsen (2015).

c) Online experiments: Online experiments leverage user studies. The only difference is that in online experiments, the users are interacting with a fully deployed system. One of the prominent works on online experiments is that by Cheng et al. (2016). They experimented by deploying their work on Google Play, a mobile app store with over one billion active users. Despite their advantage of being Google employees, the experiment took them over two years to reach the required results. Considering the limited time the proposed research can allow for the experiment, an online experiment cannot be a smart move for this research experiment setting.

2.10 Summary and conclusion

This chapter provided an in-depth review of the literature related to the research topic. Key questions that need further research were presented and discussed. This led to the further identification of methodological approaches applied in past studies on the broader research area and the evaluation of CH RSs. The research gap identified is within the domain knowledge sources and representations, in the choice of recommendation techniques, and in the VSI.

From the literature review, it is clear that the broad knowledge representations assembled from various knowledge sources for recommendation are social, individual, and content. The social knowledge representation is suitable for a domain with implicit user interaction but not appropriate for a domain with unstable user preferences. Another issue discussed in this chapter is that social knowledge can be misleading as historical data are unreliable. Unlike social knowledge, individual knowledge has difficulty predicting reliable information on certain items. For example, it is not certain that users with similar tastes in books would also like the same music. Content knowledge representation is suitable for a content-based RSs. The domain of this study lacks user preferences and social tagging, which is the challenge that this study addresses by extracting social tags and users' interests from available domain sources. However, despite their individual challenges, if integrated, hybrid approaches can provide various options for maximum exploitation, for example, integrating content and social representations to improve CH recommendations. Knowledge representations are assembled from the knowledge extracted from various domain sources.

In the CH domain, extracting information that can form a relationship between CH objects is one of the challenges of knowledge discovery for CH recommendations. For example, in paintings, extracting relational information between objects is less possible because the information is not coherent. Even though tools exist to extract knowledge from unstructured sources, several obstacles and cost factors remain for automated approaches. The literature reviewed in this chapter discussed various knowledge representation models' strengths and weaknesses. Meanwhile, for this study, the matrix-based model is adopted for knowledge representations because of its dynamic approach to social and content knowledge sources, which are core knowledge sources that provided the harvested datasets for this research.

This chapter also discussed six recommendation technique approaches. Each approach has limitations, but in some cases, when combined, they complement each others' limitations. For example, a hybrid approach that combines content-based and collaborative approaches addresses the challenge of the cold-start problem in collaborative filtering. A hybrid approach that integrates knowledge representations assembled from static combinations of knowledge sources available has also been developed in the music domain.

However, the study described in this thesis adopts a monolithic hybrid approach. The hybrid representation integrates content and social representations assembled from the CH domain knowledge sources to address the challenges of the cold-start problem, bad recommendations, and out of context recommendations. The content knowledge representation incorporates the semantic relationships of CH objects to augment the semantic gap in the content knowledge representation. Similarly, the social representation includes the CH users' social tagging and interests to bridge the social knowledge gap in the representation to address the challenges of CH recommendations.

To address the problem of bad CH recommendations, the work described in this thesis requires every resource to be mapped to a unique arrangement of vocabulary terms, providing a proposed system to distinguish compatible resources. Such resources are relied upon to have indistinguishable depictions utilising terms from the vocabulary and can, in this manner, be combined. This ascertains the equality among CH objects, which is the essential solution for the CH objects' similarity issues. This builds the 'findability' of already hidden yet related knowledge and information, assisting in new information discovery and hoisting the problem of bad CH recommendations to some degree.

Another aspect discussed in this chapter is VSI. It was found that interactive visual exploratory search interfaces for CH collections revolved mainly around a summary of a collection of CH objects and allowed users to initiate their search from the interface. These are some of the challenges that this research addressed by providing **VISE** for CH collections (Usman & Antonacopoulos, 2019). The design and implementation of VISE are discussed in Chapter Six.

Chapter Three: Proposed Methodology and Experimental Set-up

3.1 Introduction

This chapter describes the new approach and methods proposed in this study and the experimental set-up, including the evaluation measures used for the experiments. The research methodology adopted is similar to the one presented by Peffers et al. (2007). The research process model, experimental set-up, and measures used to evaluate the CH recommendations and VSI are discussed in this chapter.

3.2 The research process model

The research process model combines six phases: a literature review, definition of the solution objectives, design and development, demonstration, evaluation, and communication, as shown in Figure 3.1. The process model provided a novel approach for developing artificial intelligence and information systems, the research domain.

The first and second phases of the research process were discussed in Chapter Two. The first phase provided insight into the current state of the research topic. Experts, critical questions about the topic that needed further research, and methodologies used in past studies of a similar area of research were also discovered and discussed in the first and second phases. During these phases of the research process, the research gap and how the proposed approach could accomplish the overall research aim were identified and discussed. In this section, the design and development and evaluation phases are discussed.

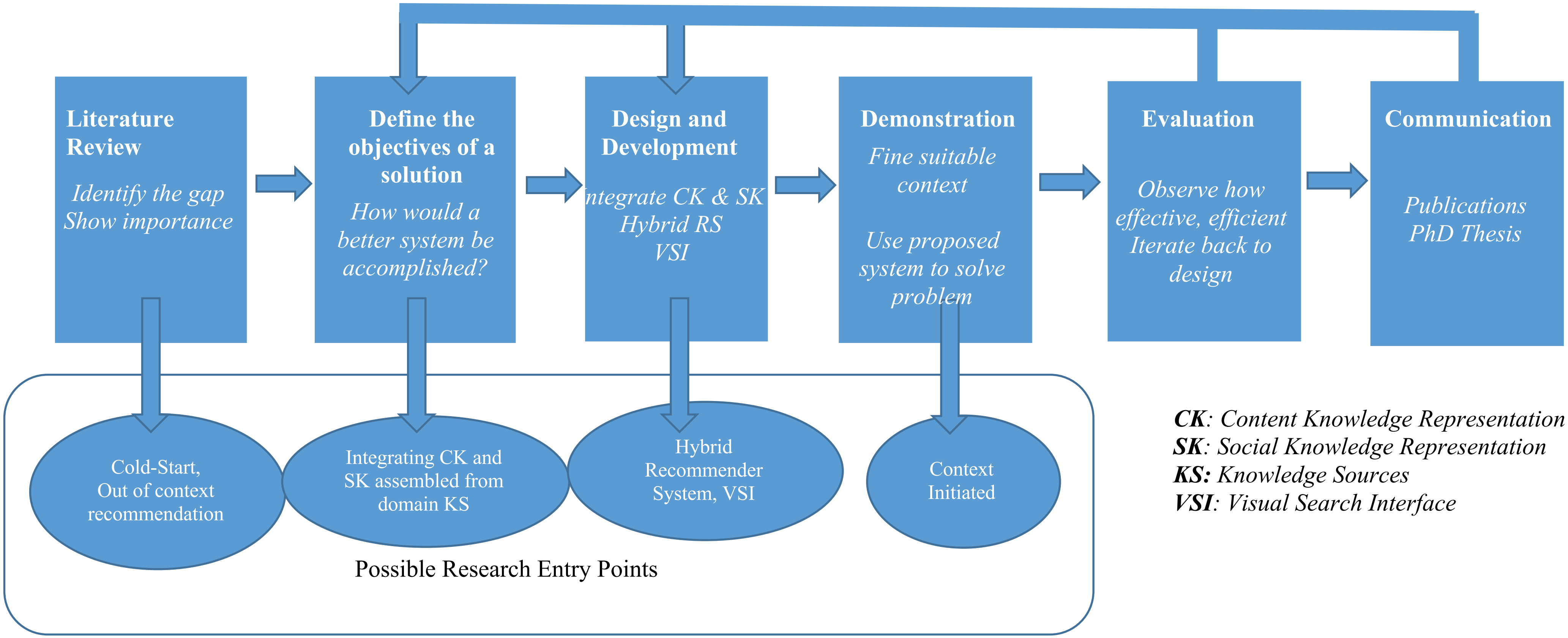


Figure 3.1: Research process model

3.3 Proposed research design and development

The proposed approach consists of four developmental phases:

1. Knowledge extraction and pre-processing,
2. Knowledge representations,
3. Knowledge integration, and
4. Hybrid recommendation and VSI.

The proposed system has a mechanism that dynamically controls the contribution of each knowledge source. This will reflect the anticipated IR and recommendation performance from the search context, as well as a level of personalisation.

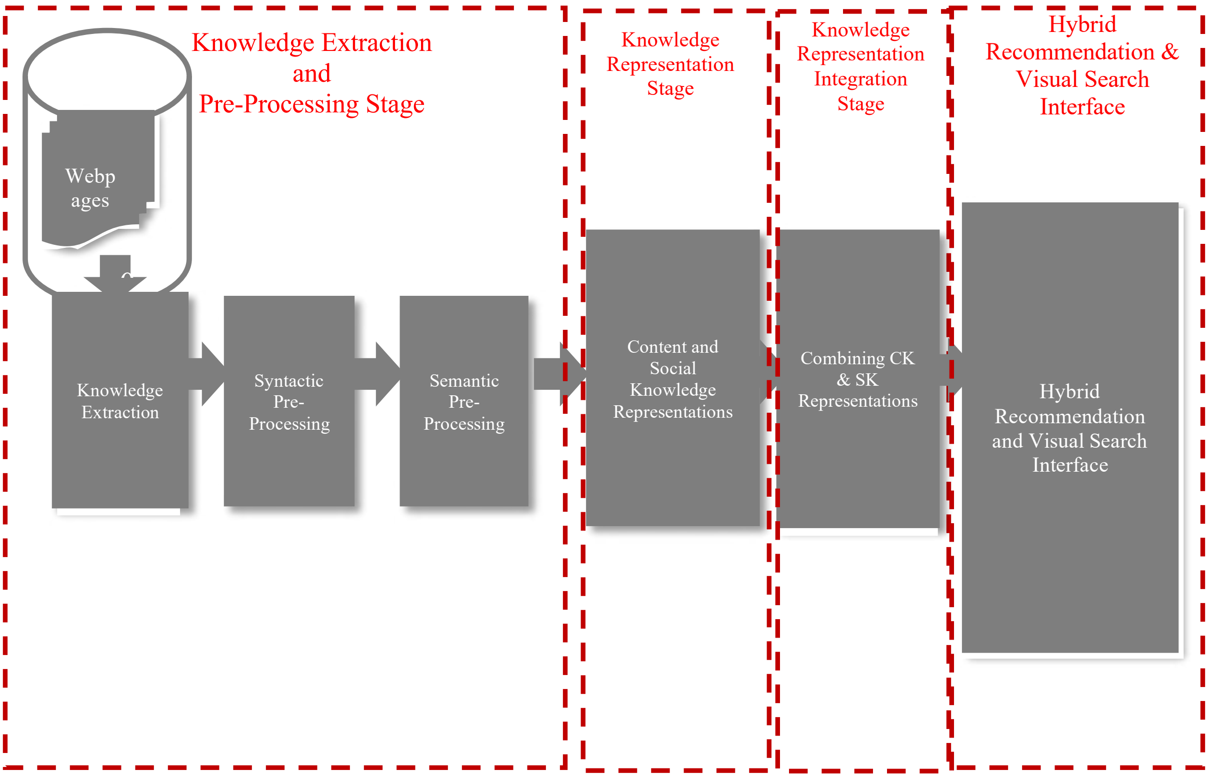


Figure 3.2: Proposed research development phases

The first part of the proposed development phase is the extraction and pre-processing of data (knowledge). At this phase, data to demonstrate the study were collected and pre-processed or transformed into a collection of CH objects, CH users, and users' interests. Knowledge extraction and pre-processing are discussed in Chapter Four (see Section 4.4).

After the first phase, the pre-processed knowledge is represented. The content knowledge that describes CH objects and users and the social knowledge that presents CH objects' social tags and users' interests are represented as vectors. Cosine similarity was used to measure the distance between the CH objects and the users at this phase.

The third part of the developmental phase integrates the content and social knowledge representations to augment the sparse social knowledge and improve the recommendation of CH materials. This is completed by concatenating content and social knowledge vectors; see Chapter Five for more details on how the two knowledge representations – content and social – were integrated to provide recommendations of CH materials.

Hybrid recommendations and the VSI are in the final development phase of the proposed system. Hybrid recommendations dynamically combine social knowledge and integrated knowledge representations to improve the recommendation of CH materials. The influence of each knowledge representation on the production of a CH recommendation varies, depending on the current user and search status. A VSI is an interface that enables a search and recommendation of CH materials. The hybrid CH RS and VSI are discussed in Chapters Six and Seven, respectively. After this research process phase, the next phases are the demonstration and the evaluation of the proposed system, a hybrid CH RS.

3.4 Evaluating a cultural heritage recommender system

One of the core benefits of conducting any research on the recommendation of CH materials is the capacity to evaluate the outlined system's properties, for example, the recommendation quality and recommendation novelty. The quality of CH recommendations should use the new approach to match the CH users' interests since they are the end-users. To validate the results and perform the experiments, CH users should, ideally, be allowed to interact with the proposed system in a live CH setting and provide their feedback on the quality of each CH recommendation produced. However, this evaluation methodology comes with several challenges that deem it impractical. One of the challenges is that it is quite expensive to run a live platform for exploring an extensive collection of CH materials, and such approaches are mostly for actual business cases, not for experimental purposes. Another significant challenge is that users interacting with a live system expect to be provided with consistent, high-quality recommendations. Such consistent quality cannot be assured during the development period of the proposed approach, which is likely to produce poor recommendations, and thus, participative users can easily be dissuaded from future participation in the evaluations.

Standard user studies also face similar challenges to a live system. Even though it is affordable, there is a probability that participants are unlikely to continue participating in the evaluation process. At most, they may only participate for a limited period, for example, a month. Therefore, instead of an online, user-centric method, an offline, system-centric evaluation method is needed – one that can reasonably demonstrate how users will review the CH material recommendations.

System-centric evaluation methods depend on pre-existing user data, such as users' reviews and ratings. Such data can be collected from users' interactions, profiles, and past reviews or ratings of CH materials, which then serve as a ground truth against defined evaluation scores and recommendations. Compared to the user-centric approach, the system-centric approach allows automated evaluations of new recommendation techniques that are impractical for user-centric evaluation methods. The findings generated from a system-centric approach can further be validated with a user study, which satisfies the requirements of both researchers' and participants' interests in evaluating the work.

The system-centric approach to evaluation requires standard and freely available data, which this research domain lacks. While the Europeana dataset (Charles & Isaac, 2015) provides evaluation data, it does not provide the required users' profiles and reviews and social tags. To bridge that gap, a custom CH dataset was harvested from available domain knowledge sources, Europeana webpages, and Facebook; see Chapter Four for more details.

3.4.1 Evaluation measures

A dataset must meet certain requirements to evaluate a CH RS, as discussed in Chapter Four. Once a required dataset is available, the experiment can be run. The remainder of this section presents a brief overview of the measures used to evaluate the new approach.

3.4.1.1 Association score

The association score between CH objects $O_i \ and \ O_j$ can be defined as the proportion of users who agree that there is an association between the CH objects $O_i \ and \ O_j$, calculated as

$$association(O_i, O_j) = \frac{likes(O_i, O_j)}{users\ interaction(O_i, O_j)}, \qquad (3.1)$$

where $users\ interaction(O_i, O_j)$ is the number of CH users that interacted with both $O_i \ and \ O_j$, and $likes(O_i, O_j)$ is the number of CH users who liked both $O_i \ and \ O_j$. This measure is similar to the result found from the user-centric evaluation method; users were asked to explore the given query and CH recommendations and then express their views on whether they liked the recommendation. The users' responses were averaged to provide an association score.

CH objects' popularity properties in the ground truth were normalised using association scores. This was achieved by comparing the number of user interactions with CH object pairs to the number of CH users who liked the CH object pair. The estimated number of CH users is low if both CH objects pairs are unpopular, which requires only a reasonable number of likes to achieve a high association score. However, if both CH objects in the pair are popular, a large number of CH users are estimated to interact with and like both the CH objects to achieve a high association score. The association score values range from 0 to 1 to estimate the quality of CH recommendations. Association scores are later used in Sections 6.4 and 8.4 to measure the CH recommendation quality.

3.4.1.2 Prediction accuracy

It is essential to comprehend how accurate the proposed system makes rating and usage predictions. To evaluate the ratings predictive accuracy, we used the root mean squared error (RMSE), a widespread metric used for predictive accuracy evaluation. The RMSE between the actual user rating and the predicted ratings is given by

$$\mathrm{RMSE} = \sqrt{\frac{1}{|\tau|}} \sum_{(u,i)\epsilon\tau} (\hat{r}_{ui} - r_{ui})^2. \tag{3.2}$$

Alternatively, the mean absolute error (MAE) is given by

$$\mathrm{MAE} = \sqrt{\frac{1}{|\tau|}} \sum_{(u.i)\in\tau} |\hat{r}_{ui} - r_{ui}| \,, \tag{3.3}$$

where $\hat{r}_{ui}$ is the predicted ratings for a test set τ of user u and CH object i, and r_{ui} is the actual ratings from the harvested custom dataset.

RMSE and MAE rely on the errors' magnitudes. Unlike RMSE, MAE is a more natural average error measurement (Willmott & Matsuura, 2005). The result obtained from Equation (3.2) is used to compare the user score and association score.

To measure the usage prediction, a target CH user was selected, and the associated CH objects were removed from the dataset. The system was then asked to predict a set of CH objects. In this case, there were four possible outcomes for the hidden CH objects and the recommendation, as shown in the confusion matrix below.

	Recommended	Not Recommended
Used	True-Positives (tp)	False-Negatives (fn)
Not Used	False-Positives (fp)	True-Negatives (tn)

A confusion matrix presents a classification description performance in table format on a set of known valid values from test data. From the confusion matrix above, the number of examples that fall into the cell can be counted and the following values computed:

$$Precision = \frac{tp}{tp+fp} \tag{3.4}$$

$$Recall\ (True\ Positive\ Rate) = \frac{tp}{tp+tn} \tag{3.5}$$

$$False\ Positive\ Value = \frac{fp}{fp+tn} \tag{3.6}$$

To improve the recall, longer CH recommendation lists are allowed. Note that the most interested measure is Precision at N.

3.4.1.3 User satisfaction level

This metric was used to evaluate the user satisfaction level of the VSI provided for exploring and recommending CH materials. The results obtained to measure the users' satisfaction levels came from the questionnaires given to the users during experiments. Two questionnaires were presented to the participants to express their satisfaction levels using the five-point Likert scale.

Since the results came from the two questionnaires (two samples), a two-sample t-test (Moore, 1957) was performed to measure the users' satisfaction levels of the interface. The two-sample t-test is a statistical measure used to test the difference between two sample population means.

3.5 Conclusion

Exploiting available domain knowledge sources to improve the retrieval and recommendation of CH materials is critical to this study. To achieve this, the current chapter developed a conceptual methodology for integrating content and social knowledge representations assembled from available domain sources and creating a VSI to improve recommendations of CH materials. This chapter also discussed another critical component – the evaluation of CH recommendations.

Constant user interaction with the new CH recommendation approaches is essential during the evaluation stage of the development. Nevertheless, it is not always feasible to continually ask participants to evaluate the small changes made during the evaluation period, so the ability to achieve this without their constant participation is critical. However, to perform such evaluations for the recommendation of CH materials requires an open offline dataset, which is currently not readily available. There are publicly available CH datasets, but they do not contain user interactions. Thus, a custom dataset that incorporated user interactions and social tagging was harvested for system-centric evaluations. This data was used to analyse the evaluation measures for ground truth scores, such as similarity and user preferences.

The evaluation measures for ground truth similarity and user preferences are the association score, the user prediction accuracy, and user satisfaction level of the VSI. The association score and prediction accuracy are used throughout the study as the standard evaluation measures for CH recommendation quality, and the user study is utilised at the end to evaluate the association score and user prediction quality presented by the work.

Chapter Four: Custom Dataset for Cultural Heritage Recommendations

4.1 Introduction

Digital CH materials are of different types and formats; therefore, developing rich knowledge representations from different domain sources for the recommendation of CH materials is critical to this research study. This development can be achieved by having a collection of CH objects and users so that content and social knowledge representations can be assembled directly from the collection. To exploit the collection for knowledge representations, the dataset must contain the important features of CH objects and users' interests. Also, since one of this study's priorities is to bridge the knowledge gap by integrating content and social knowledge representations, two knowledge sources – **Europeana**[9] and **Facebook**[10] – are considered available domain sources for content and social knowledge extraction, respectively. Reasons for choosing these sources over other sources are discussed in Sections 4.2 and 4.3.

In this chapter, Europeana and Facebook are discussed in terms of their knowledge richness and the anticipated challenges over other sources if exploited for CH recommendations. A review of such domain knowledge sources further revealed the need for wider and more users' interests, social tags, and metadata to enrich the dataset. Thus, a CH custom dataset was harvested from these domain knowledge sources. Strengths and weaknesses of Europeana and Facebook as sources of knowledge for this study are first highlighted, and then, the harvesting process of the CH custom dataset for a novel CH RS is discussed. Finally, this chapter discusses the comparative study between the Europeana dataset and the harvested CH custom dataset.

4.2 Europeana and the issues of extracting knowledge from its web content

The primary responsibility of museums across Europe is accumulating items of national and worldwide importance, saving them, translating them, and opening them to as many individuals

[9] https://pro.europeana.eu/resources/apis
[10] https://developers.facebook.com/docs/apis-and-sdks/

as can be reasonably expected (Ciocca et al., 2012). As the online access availability continues to grow for people around the world, these important CH objects can be globally accessible. In the case of Europeana, they are open-licenced sources of content knowledge, which provided an advantage for utilising their web content for this study. Reasons for using Europeana as the source of content and semantic knowledge are stated in Table 4.1.

Table 4.1: Reasons for using Europeana

Reason	**Discussion/Facts**
Rich content information of digital CH objects	Europeana houses over 50 million digital CH materials that are accessible online (Petras et al., 2017).
Open source	No licence is required to access the information.
Metadata aggregation	They have the Europeana Data Model that converges CH materials to metadata standards (Manguinhas, 2016).
Data quality	Data quality is their top priority; Europeana previously organised a series of competitions to choose and support the best thoughts for innovative reuse of digital CH materials (Manguinhas, 2016).

From the Europeana webpages (see Figure 4.1), CH objects' textual descriptions and images are available for knowledge extraction.

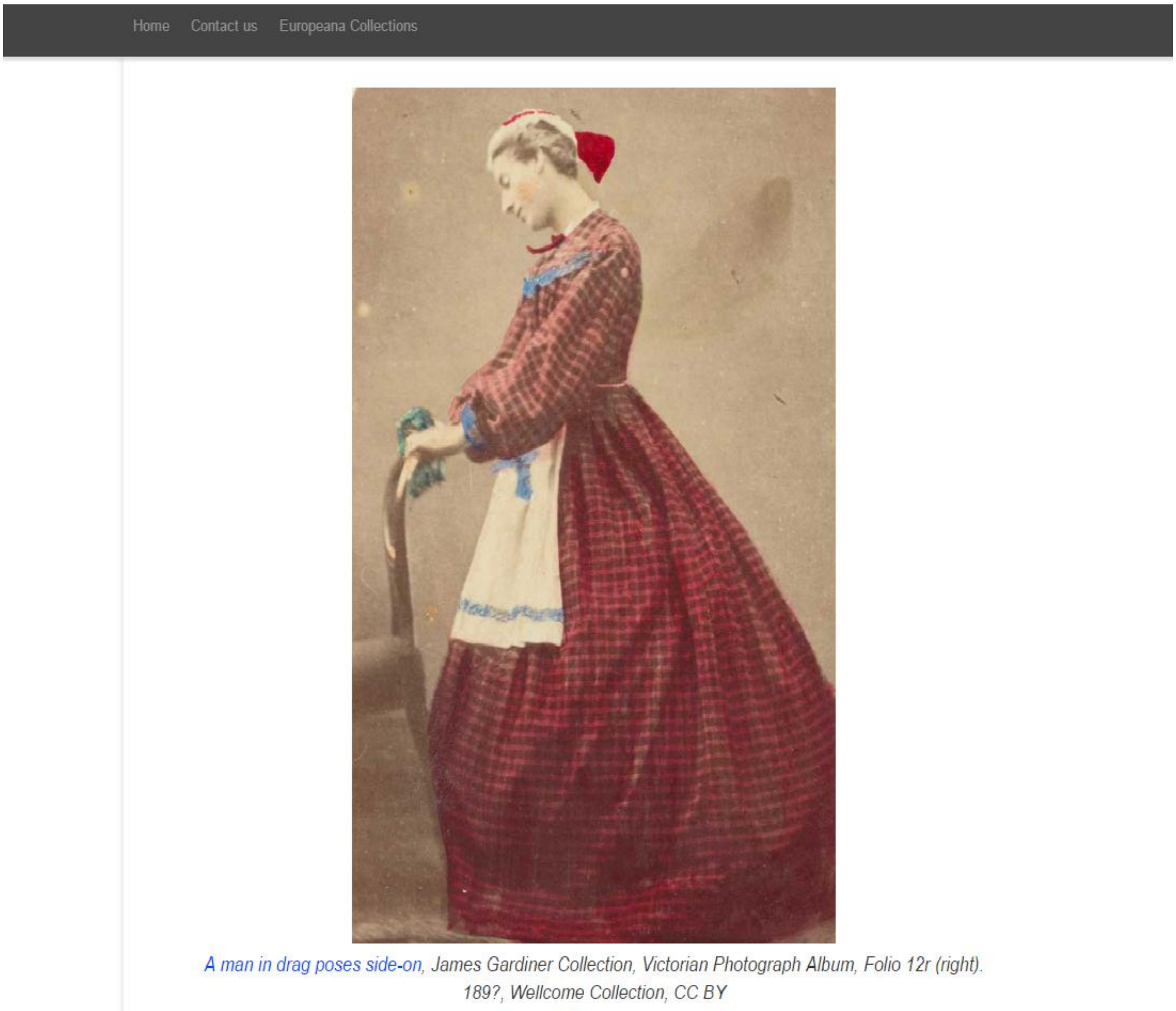


Figure 4.1: Example of a Europeana webpage

Therefore, this study capitalises on the advantages of Europeana's webpages to extract content knowledge using various tools (see Sections 4.4.1 and 4.4.2). Another source of knowledge for this study is Facebook. Issues surrounding Facebook as a source of social knowledge for this study are discussed in the next section.

4.3 Issues with Facebook as a source of social knowledge

Facebook, with over 2.8 billion monthly active users, remains the world's most well-known social network, by far, as shown in Figure 4.2. For example, 73% of internet users in the United Kingdom use Facebook, and 55% of them retain their real-world contacts and share their personal interests on the platform (Matz et al., 2019). Hence, Facebook represents global social knowledge, which is useful to this study for users' interests and social tagging.

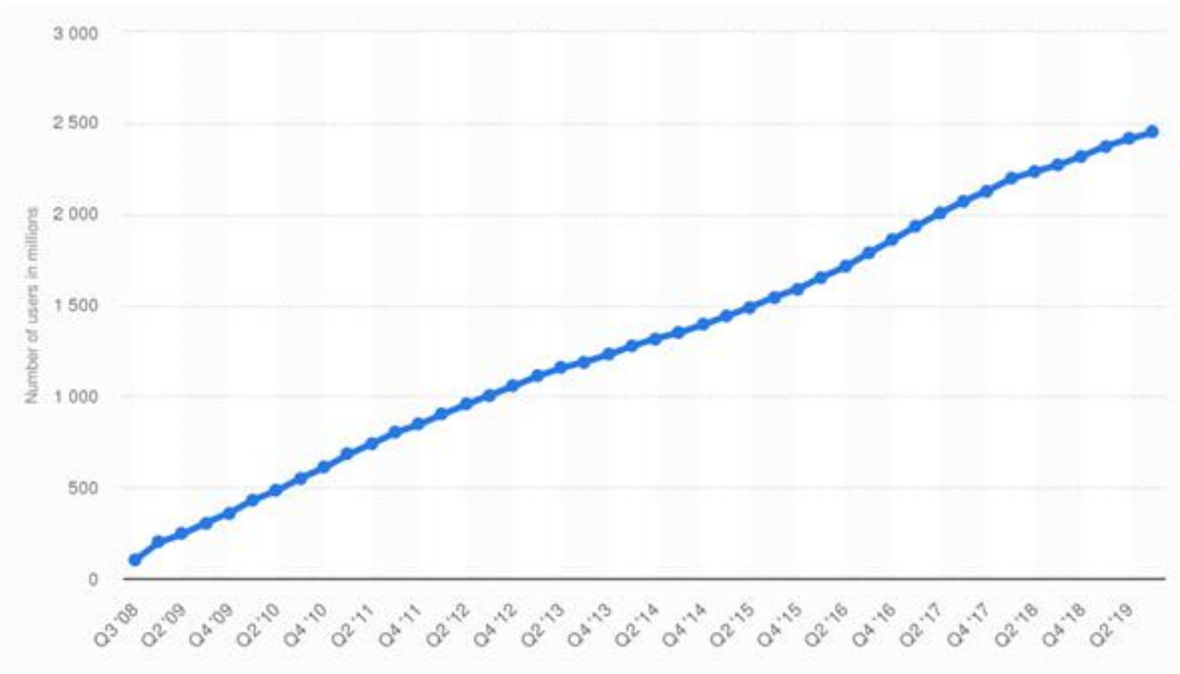


Figure 4.2: Active Facebook users worldwide from 2008 to 2019 (Source: Facebook, 2019)

However, despite these statistics, Facebook is surrounded by controversies and critiques, especially with regards to data protection and privacy. Table 4.2 highlights a brief history of some of these (notable) controversies.

Table 4.2: Summary of data privacy issues surrounding Facebook as a source of knowledge

Year	Controversies and Issues	Responds and Amends
July 2007	The issue of **external privacy** began when an undergraduate student from the University of Virginia named **Adrienne Felt** discovered a loophole that could be used to inject a script to access user profiles on the Facebook platform (Felt & Evans, 2008).	It took Facebook more than two weeks to rectify the issue.

August 2007	After handling the external privacy breach in July, Facebook faced another challenge: **internal privacy threats**. These came as a result of a source code leak by a member of Facebook's engineering team (Atkinson, 2007).	This incident raised concerns over the security of custom data on the platform.
November 2007	Facebook introduced **Beacon**, a system that provides access to third-party websites for information exchange (Perez, 2007).	This action by Facebook raised serious data privacy concerns from users. It was later rectified by requiring user permission before publishing information.
February 2008	An article published by the *New York Times* triggered concerns over **data ownership.** In the article, the author revealed that Facebook did not provide a mechanism for users to **permanently deactivate their account** and, thus, raised an alarm that user data could permanently remain on Facebook's servers (Tufekci, 2008).	This led to the amendment of the Facebook Privacy Policy in March 2008, which allows users to delete their accounts permanently if they so wish.
January 2011	EPIC, a centre for controlling electronic privacy information, filed a lawsuit against Facebook for **sharing vital user information for business purposes** with a third-party, especially users under 18 (Semitsu, 2011). This	Facebook briefly suspended the third-party policy to calm the situation but later upheld the policy.

	vital information included mobile numbers and home addresses.	
March 2018	**Cambridge Analytica Scandal:** In March 2018, it was discovered that Cambridge Analytica, a British political consulting firm that uses data for strategic communication during electoral processes, illegally acquired the personal information of as many as 87 million users from Facebook for commercial use (Common, 2018). This incident raised many issues regarding users' data privacy and protection. Eventually, Mark Zuckerberg, Facebook founder and CEO, was summoned by the United States' Senate and Congress for a public hearing titled 'Facebook, Social Media Privacy, and the Use and Abuse of Data' in April 2018.	Facebook has restricted the investigation firm Strategic Communication Laboratories and its political arm, Cambridge Analytica, for an inability to adhere to standards regarding the treatment of personal information and for what might be among the biggest misuses of personal information in US history. Mark Zuckerberg issued the following statement: 'The good news is that the most important actions, to prevent this from happening again today, we have already taken years ago. But we also made mistakes, there's more to do, and we need to step up and do it.'

The data security and privacy issues discussed in Table 4.2 indicate that Facebook's approach to data protection and privacy are post-active; they act after an incident rather than before an incident. This situation shows that Facebook does not have strong data protection or privacy policies, despite claiming that they do. As this research needed to use Facebook as a source of social knowledge for CH recommendations, this study implemented additional measures with regards to personal data privacy by asking the users' permission and stating the type of data needed prior to using it.

The question is 'why did this research still need to use Facebook as the source of social knowledge despite all of these data protection and privacy issues?' The answers and justification are clearly stated in Table 4.3.

Table 4.3: The answers and justification for using Facebook as a knowledge source

Reasons	**Justifications**
Fame/popularity	Facebook has over 2.8 billion users across the globe, which, without a doubt, makes it the most popular social network platform in the world (Contratres et al., 2018).
Many active users	Facebook has over 1.1 billion daily active users from different knowledge backgrounds. According to Common (2018), 53% of the current youth population accesses their daily news through Facebook, and over 30 million applications are installed daily.
Growth strength	In the last decade, Facebook occupied 30% of the internet users, but after ten years, it reached over 80% of the current internet users, showing over a 50% growth increase (Common, 2018).
Development APIs and tools	In April 2010, through its CEO, Mark Zuckerberg, Facebook introduced new features to the Facebook developers' platform during the F8 developer conference. These features included the ***All Mutual Friends***

	API, Groups API, and the ***Open Graph Protocol.*** These features provided developers access to extract knowledge from Facebook with users' permission for development purposes (Zuckerberg, 2010).

It is important to note that despite all of the issues surrounding Europeana and Facebook, they were the best sources of social knowledge for this study, as shown in Sections 4.2 and 4.3. The harvesting process of the CH custom dataset from these knowledge sources is discussed in Section 4.4.

4.4 Harvesting a custom cultural heritage dataset

A data model that could be used for both services was required to construct a hybrid knowledge representation and then evaluate the RSs. Thus, a personalised dataset was harvested. This dataset was not personal; it was created from the webpages collected from Europeana's website for content knowledge and from Facebook's APIs for social knowledge, such as social tagging and users' interests. The harvesting process had three phases, as illustrated in Figure 4.3.

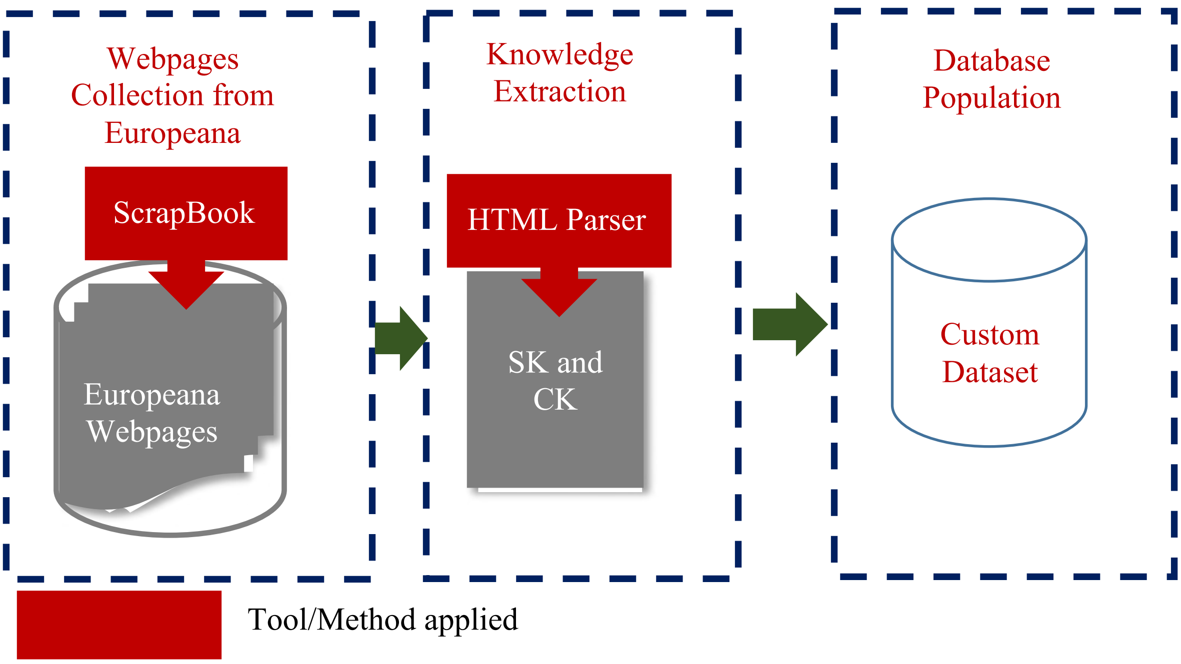


Figure 4.3: Custom dataset harvesting phases

From Figure 4.3, the first phase of the harvesting process, webpages were collected from the Europeana website. The second phase extracted the required content knowledge from the gathered webpages and the social knowledge from Facebook. The extracted knowledge needed to be stored for knowledge pre-processing (see Chapter Five). The data storage is the final phase of the process. The three phases are discussed in Sections 4.4.1 and 4.4.2.

4.4.1 Webpage collection from Europeana

The first phase (see Figure 4.3) of the personalised dataset harvesting process was collecting webpages from Europeana. The challenging tasks for this process phase were transforming webpages to form content knowledge and user information for the CH dataset.

To gather the webpages from the Europeana website, different factors needed to be considered, including the content of the web documents and what to accomplish with the webpages. For this study, approximately 750,000 webpages were collected using the Firefox ScrapBook crawler, a tool that automatically saves and manages webpages. ScrapBook provided an option for the specification requirements, as shown in Figure 4.4. Each Europeana website page contains information about two CH objects, meaning that over 1.5 million CH materials are in the dataset collection.

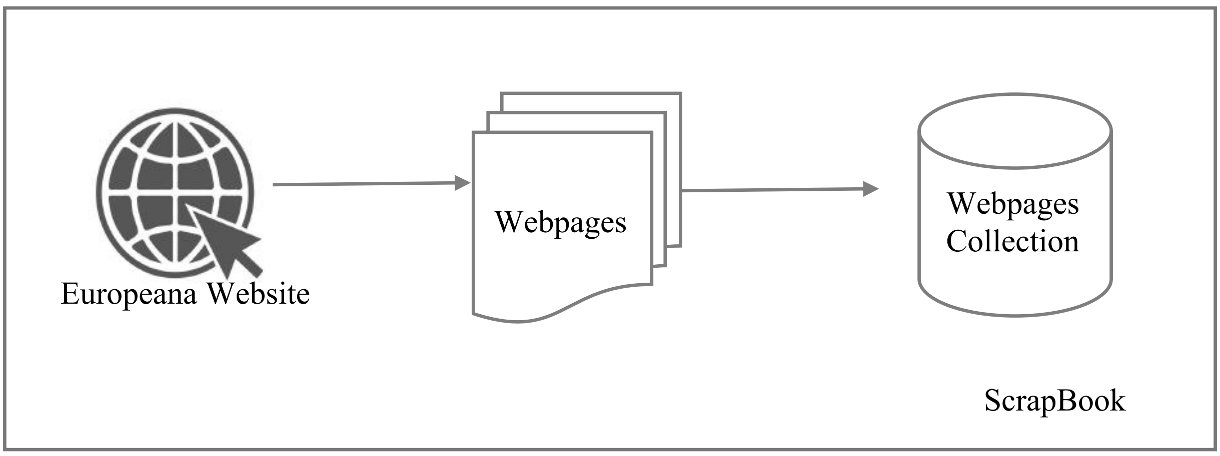


Figure 4.4: ScrapBook webpage crawling process

Generally, webpages are cluttered, with vast amounts of less informative and typically unrelated materials. These include advertisements, navigation information, JavaScripts, CSS, and sponsor hyperlinks. None of these are related to the main content; rather, they make the main content very difficult to locate. Thus, a knowledge extraction application is required.

4.4.2 Knowledge extraction

The second phase of harvesting the personalised dataset for CH recommendation was extracting content and social knowledge from the Europeana website. This phase had two parts: content knowledge extraction and social knowledge extraction.

4.4.2.1 Content knowledge extraction

Generally, webpages are clustered with less informative materials unrelated to the main content, as shown in Figure 4.5. This makes it very difficult to locate the required knowledge. Within these webpages, content knowledge that describes the CH objects was expected to be extracted.

```
class="object-title" property="http://purl.org/dc/elements/1.1/title dc:title">akvarell | Okänd</h2> </div> </div> <div class="da
class="data-group"> <li> <section> <ul class="comma-list data-group"> <li class="split-items"> S/S Hanna av Helsingborg </li> <li
till ”Transitation” och inmättes till 1862 brt - 1093 nrt. Lastförmågan var ca 2 400 ton inkl. bunkers. År 1863 såldes ångaren ti
Sons, Argostoli, Grekland. Februari 1900 förvärvades fartyget av partrederiet N. P. Swensson, Helsingborg. Namnet ändrades till ”
fördes ångaren 1902-06 av sjökapten L. P. Norrman, Mölle, 1906 av sjökapten J. P. Johnsson, 1907-13 av sjökapten E. Winck, Helsin
eller torpederad. Reg.n:r 3586; byggd år 1885 av stål; tontal: brutto 1598, netto 1144; indic. Hkr. 680; redare Rederi AB. Hencke
mars 1915, Nordsjön.<br/>På eftermiddagen fredagen den 12 mars 1915 avgick Hanna från Tyne med last av 1950 ton stenkol, destiner
sydvart hän utefter engelska kusten på 2 à 3 mils avstånd från denna. August Leffler ledde, och Hanna följde efter. Resan fortgic
fartygen befunno sig i närheten av Flamborough Head, skakades Hanna emellertid helt plötsligt av en våldsam explosion i förskeppe
slagsida samt började sjunkas. Strax därefter bröts den av på mitten, och förskeppet försvann under vatten. </li> <li class="spli
alltfort gick före. Sedan de delar av fartyget, som ännu voro över vattnet, en sista gång genomsökts, för att ingen skulle lämnas
båtar som besättning från det förolyckade fartyget. Ej långt därefter sågs även akterskeppet försvinna.<br/>Vid överräknandet av
hava dödats vid explosion. 2 man, som uppehållit sig i styrbords skans, hade däremot sluppit ifrån med livet, ehuru den ene, lätt
dock med egen hjälp lyckats ta sig ned i en av båtarna.<br/>De överlevande medföljde August Leffler till Spurns fyrskepp. Här öve
<br/>Huruvida katastrofen förorsakats av mina eller torped, torde vara svårt att avgöra. 2:dre styrmannen, som vid olyckstillfäll
explosionen. - Besättningen hann icke rädda något av sina tillhörigheter. En del sprungo i båtarna i nattdräkt. <br/>Omkomna: 6 m
f. den 4/5 1886, ogift, hemort: Stockholm.<br/>Nils Wilhelm Hansson, eldare, f. den 21/6 1892, ogift, hemort: Karlskrona.<br/>Kar
1898, ogift, hemort: Stockholm.<br/>Skadad: 1 man. Paul Andersson, lättmatros” </li> <li class="split-items"> På gränsen, 2012-04
class="split-items"> Konst </li> </ul> </section> </li> </ul> </div> </div> <div class="data-section cf" data-section-id="people"
class="data-header">Creator:</h4> <ul class="comma-list data-group"> <li> <a href="/portal/en/search?q=who%3A%28Ok%C3%A4nd%29">Ok
class="subsection-label"> Classifications </h3> <div class="subsection-content"> <ul class="data-group"> <li> <section> <h4 class
</li> <li> <a href="/portal/en/search?q=what%3A%22F%C3%B6rem%C3%A5l%22">Föremål</a> </li> <li> <a href="/portal/en/explore/topics
</h4> <ul class="comma-list data-group"> <li> <a href="/portal/en/search?q=what%3A%22Kulturhistoria%22">Kulturhistoria</a> </li>
</div> <div class="data-section cf" data-section-id="properties"> <h3 class="subsection-label"> Properties </h3> <div class="subs
<li> <a href="/portal/en/search?q=dc_language%3A%28sv%29">sv</a> </li> </ul> </section> </li> </ul> </div> </div> <script id="js-
closed[[/extended_information]]"[[#id]] data-section-id="[[.]]"[[/id]]>
  <h3 class="subsection-label">
    [[ title ]]
    [[#info]]
      <a class="eu-tooltip-anchor" href="[[.]]">
```

Figure 4.5: Webpage cluttered with less informative materials

According to the research conducted by Vosinakis and Tsakonas (2016) on museum users' and visitors' interests, CH users are interested in the age (date) of the object, author or origin, image, and geographical locations. Thus, these attributes that describe CH objects were considered and located during the content knowledge extraction process, as shown in Figure 4.6.

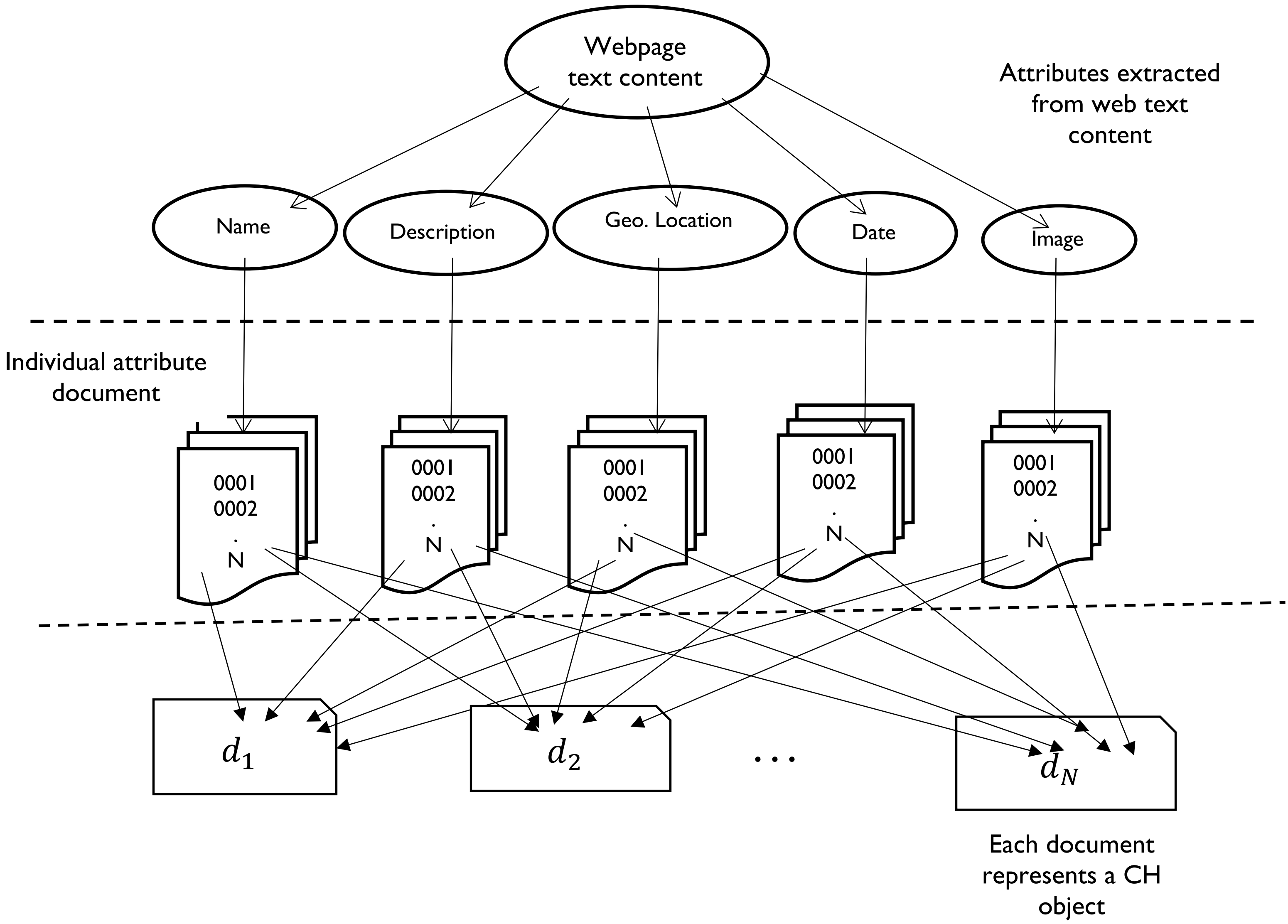


Figure 4.6: Content knowledge extraction

Figure 4.6 describes the content knowledge extraction process, which had three stages. The first stage was extracting important attributes (e.g. name, description, location) to generate individual documents. In this study, the tool used for extracting the important attribute cluttered within the webpages was the 'lxml' library of the Python programming language. The provision of a powerful API has made it the most feature-rich and simple-to-utilise library for handling XML and HTML. The second stage involved using the individual attribute documents to produce other text documents that represented CH objects at the final stage of the extraction process. The important feature attributes extracted were represented as a vector D, as shown in Equation (4.1).

$$D = \{d_1, d_2, d_3 \dots d_N\} \tag{4.1}$$

Each document d_N in D represents the important features of CH object N. The second part of knowledge extraction was the social knowledge extraction from Facebook and its incorporation into the dataset.

4.4.2.2 Social knowledge extraction

For this study, social knowledge, such as social tags and user's interests and locations, were generated using the Facebook Graph API, which obtains user data from Europeana's Facebook platform. The Graph API, also known as the 'social graph' is composed of nodes, edges, and fields. Nodes refer to individual objects (e.g. users, a page, or comments); edges are the connections between the collection and a single object, and fields refer to the information about each object in the collection. The Graph API provided a method, `graph.get_object()`, which acquires all of the user information related to CH objects in the collection. The user information is collected from the users of Europeana's Facebook page (https://www.facebook.com/Europeana/). For this study, the social knowledge of about 200 Facebook users that like the page was extracted.

In order to encode the meaning of extracted knowledge, DBpedia, a universal vocabulary, was used because of its metadata enrichment potential and open accessibility. Note that the ontology used during the semantic pre-processing in this study was the CIDOC conceptual reference model (CRM) ontology refined by DBpedia (see Chapter Five for further details).

DBpedia is an open-source network push to extract organised content from data created in different Wikimedia projects using linked data and semantic web technologies (Lehmann et al., 2015). This is organised as an open knowledge graph, which is accessible over the web and stores information in a machine-readable format, providing a way for data to be gathered, composed, shared, sought, and used. DBpedia contains knowledge from over 100 various language versions of Wikipedia. DBpedia extracted knowledge from the English version of Wikipedia, which is the largest version, consisting of over 400 million facts that describe 3.7 million objects (Lehmann et al., 2015). The universal vocabulary also extracted knowledge from other language versions of Wikipedia, consisting of 1.46 billion facts that describe ten million additional objects (Ismayilov et al., 2018). For example, Europeana uses DBpedia as a vocabulary to enrich and improve its metadata collections by exploiting the semantic relations and translations it provides.

Petras et al. (2017) showed how Europeana utilised DBpedia to enrich their metadata. For this research, a feasibility study was carried out on the possibility of using DBpedia as the vocabulary for Europeana's Facebook user data.

Europeana Facebook user data vs DBpedia: Analysing the possibilities of using DBpedia as the vocabulary for Facebook user data

To observe the feasibility of using DBpedia as a vocabulary tool for describing user data generated from Facebook, a preliminary experiment was set up. In the experiment, data from 186 Facebook user profiles, which comprises users' interests and activities, were evaluated after querying each term from the dataset against DBpedia for successful results. The data model used by DBpedia is the resource description framework (RDF) (Ismayilov et al., 2018). The RDF is a metadata data model for data interchange of web-implemented information.

The user data generated from the 186 Facebook users' profiles comprised the users' interests in **books, paintings, music,** and **movies**. To query each term from the Facebook users' profiles against the DBpedia dataset, SPARQL was applied because of its compatibility with DBpedia. SPARQL is a semantic query language that manipulates and retrieves information stored in an RDF format. However, in Europeana's Facebook user data, it was observed that a single term could have multiple meanings in the DBpedia dataset, leading to ambiguity. For example, the term 'Paris' may refer to many things, such as places or movies, in the DBpedia dataset (see: https://en.wikipedia.org/wiki/Paris_(disambiguation)), which provided the diversity of finding relevant concepts. The term to ambiguity ratio was calculated using the formula (ambiguity percentage function) provided by Gallagher (2013):

$$A_c = \frac{100}{n} \sum_1^n \left(1 - \frac{1}{h_i}\right). \quad (4.2)$$

A_c is the ambiguity percentage for a single user's interest category c; n represents the total number of terms in c, and h is the total number of hits for a single term in c. Figure 4.7 presents the results obtained from the preliminary experiment.

As stated earlier, Europeana's Facebook users' interests were categorised into books, paintings, music, and movies. Therefore, to find the meaning of the extracted terms from Europeana's Facebook users' data, each term was queried individually in DBpedia. The percentage of related users' interest concepts found from the successful queries and the related concept pages established against the queries are presented by the blue bar in Figure 4.7. For example, from the results obtained, 80% of the queries in 'movies' were successful, while approximately 54% of the queries in 'paintings' were successful.

However, for some queries, DBpedia could not return a single concept page. Such results are called ambiguous and are represented by the orange bar. In this case, the ambiguity percentage of each category was higher than that of the successful queries of each category. Therefore, from the results obtained, the ambiguity ratios clearly showed that social data required a semantic pre-processing (see Section 5.5) before being used to avoid misrepresentation of terms, which is one of the content knowledge representations challenges address in this study. The preliminary experiment further revealed that the social data extracted from Europeana's Facebook page could be used for this study.

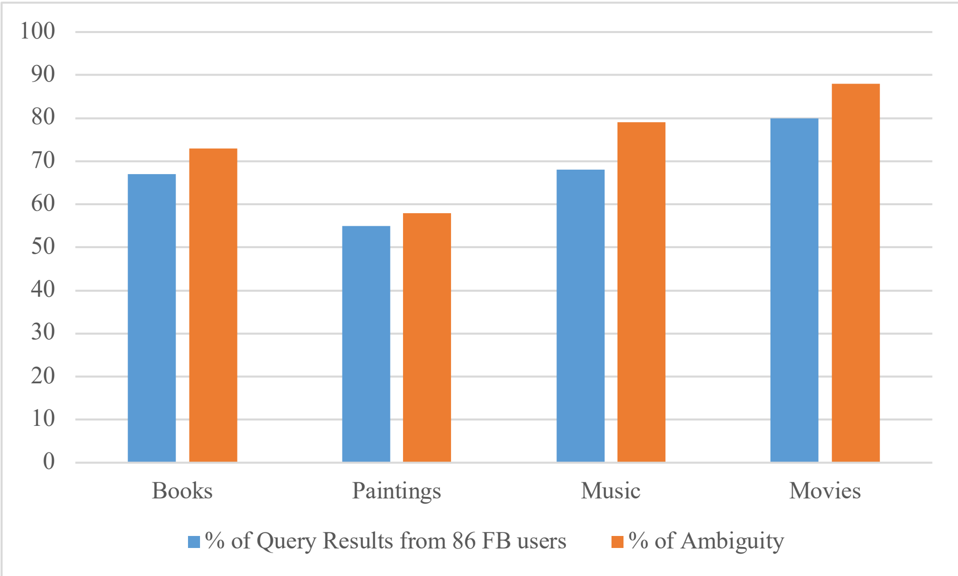


Figure 4.7: The results obtained from the preliminary experiment

4.5 Custom cultural heritage dataset

In this section, the harvested custom CH dataset is discussed. The dataset consists of CH objects, metadata, and social tags from users.

4.5.1. CH objects

There are roughly 680,000 CH objects in the dataset collection; each object has a corresponding text file that describes its content features. The CH objects were divided into five groups: artwork, natural, science and technology, archaeology, and world culture. Artwork comprises paintings and statues. Natural objects include animals and minerals, and science and technology objects are documents that describe ancient technologies and medicine. The CH objects that illustrate different cultures around the globe, e.g. Chinese and African clothing, and historical documents are categorised under world culture. Figure 3.8 presents a summary of the CH object categories in the dataset.

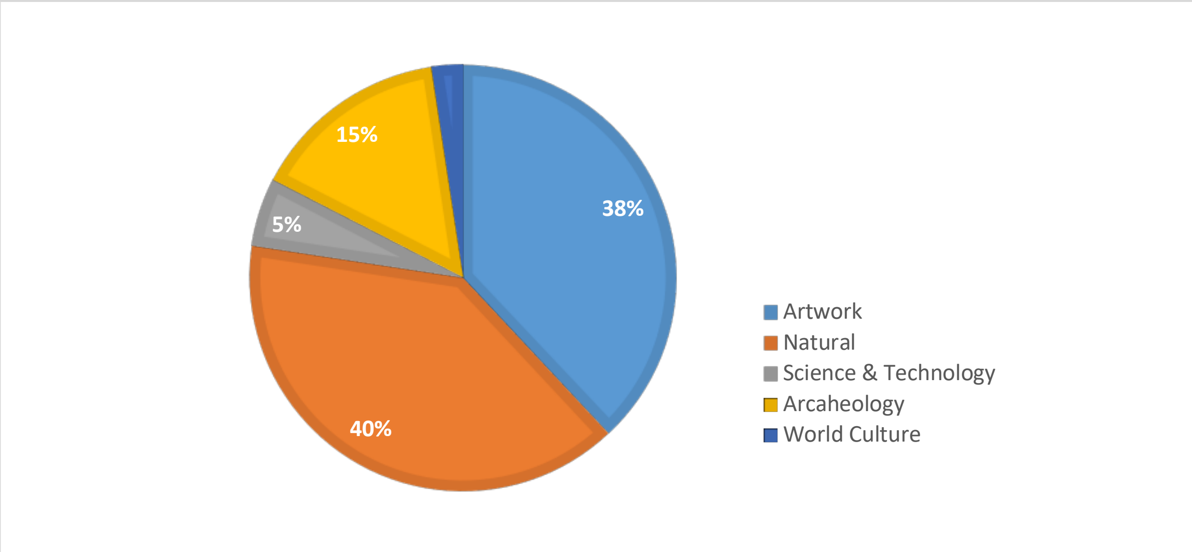


Figure 4.8: Categories of cultural heritage objects in the dataset

In the dataset collection, a majority of the CH objects came from artwork, followed by the science and technology category.

4.5.2. Metadata

Metadata in the dataset was generated from Europeana (see Section 3.4.2), which indicated that the available metadata is rich in the collection. As stated earlier, there are around 680,000 objects in the collection. These objects were created by approximately 358 originators, authors, or artists across various times. Table 4.4 shows the periods when the CH objects originated.

Table 4.4: Cultural heritage objects' origination periods

Centuries	Number of CH objects	% in the dataset collection
20^{th} Century	258,471	38.0
19^{th} Century	267,814	39.4
18^{th} Century	35,401	5.2
Before 18^{th} Century	102,547	15.1

Unknown	15,767	2.3

Most of the CH objects in the dataset collection came from the nineteenth century. Although fewer objects came from the eighteenth century, their creation periods were spread across all the centuries, making them more diverse.

4.5.3. Social tags

Social tags are keywords or tags generated electronically by users as a way to describe CH objects' contents. A total of 4,383 unique social tags were extracted and incorporated into the custom dataset collection. Each CH object had an average of 75 social tags assigned to it. The standard deviation is 0.66, which is a 1.47% error margin. The highest and lowest number of social tags assigned to a single CH object were 73 and 0, respectively. Approximately nine of the CH objects had no social tag assigned to them.

4.5.4 Cultural heritage users and their interests

The custom dataset has roughly 130,000 users with their interests on over 600,000 CH objects. Each user is presented along with their CH objects of interest. Nearly 80% of the users have at least one CH object that they expressed interest in. Less than 15% of the users in the dataset show no interest in any of the CH objects.

4.6 Comparison between Europeana and the custom cultural heritage dataset

There are over 50 million records in the Europeana collections. These records, which include CH objects such as artefacts, arts, books, audio clips, and newspapers, are represented in 271 different datasets (e.g. medical illustration, artwork, and painting datasets) by Europeana.[11] These datasets are free and open-licenced to use for research. For this study, five different datasets from Europeana were merged to form a single dataset named '**Europeana Dataset**', which contained around 701,000 CH objects. In this section, a comparative study between the

[11] https://pro.europeana.eu/resources/datasets

Europeana dataset and the harvested custom CH dataset is evaluated and discussed in terms of their social data richness.

Both the custom CH and Europeana datasets are clean and structured. These strengths give room for a comparative experiment. Figure 4.9 presents the comparative results between the two datasets in terms of the social tagging distribution. The blue graph area shows the social tagging distribution of the customly constructed CH dataset, while the red graph area presents that of the Europeana dataset.

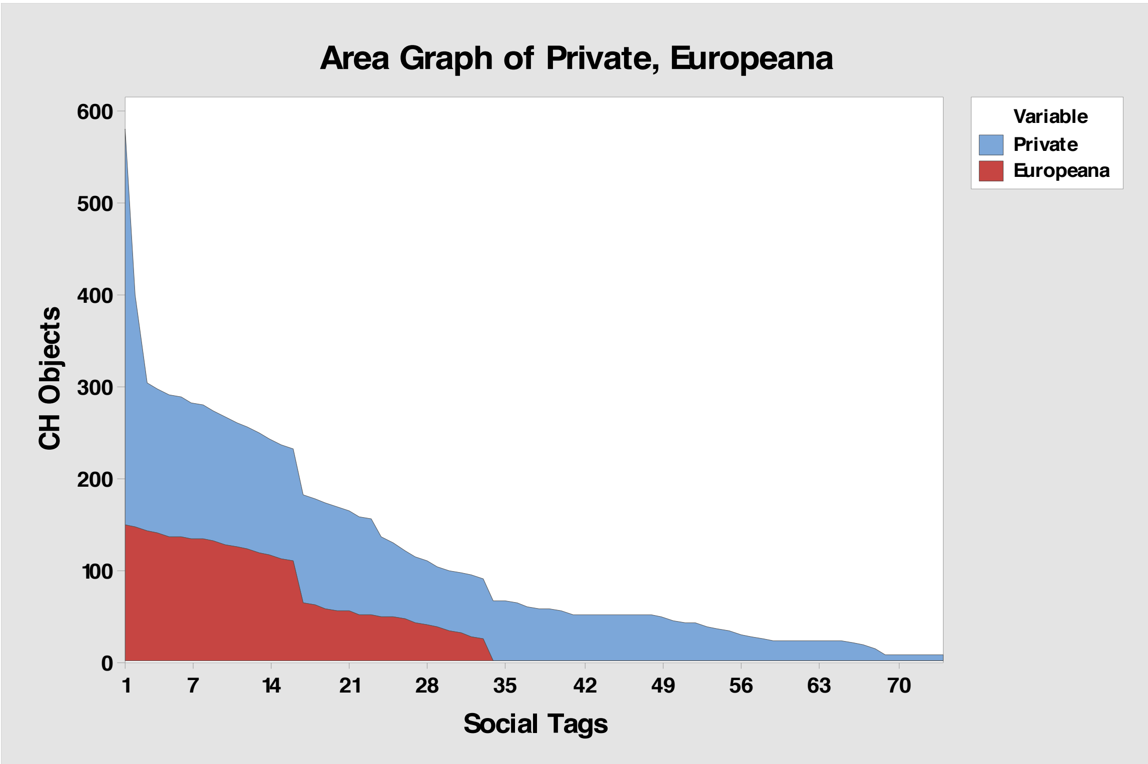


Figure 4.9: Social tags distribution comparison between the Europeana and custom cultural heritage datasets

From the results presented, it can be seen that the Europeana dataset has a maximum number of around 35 social tags per CH object, while the custom dataset has a maximum of about 75 tags, which shows that Europeana has limited number of social tags when compared with the custom dataset. It also shows that as the number of social tags increases, the total number of CH objects that have that many social tags decreases, which is a normal trend. This is because most of the data generated by online users are tagged toward objects' popularity. From the results obtained, it can be concluded that the custom dataset is richer than Europeana's in terms of social data.

4.7 Conclusion

The dataset is one of the critical aspects of running RS algorithm training and experiments. To produce CH recommendations, the required dataset must be rich with content and social knowledge. The performance of the recommendation techniques also depends on the richness of the dataset. This chapter discussed the importance of building a dataset that contains the required knowledge to produce quality CH recommendations. Therefore, a custom dataset was harvested from two available knowledge sources, Facebook and Europeana. Reasons for choosing Europeana and Facebook were detailed in this chapter; see Tables 4.1 and 4.3.

Europeana was chosen as the source of content knowledge because of its content information richness and the data model that convergences CH materials to metadata standards and because it is open-source. Facebook has an edge over other sources of social knowledge because of its popularity and large number of active users, growth strength, and development tools and APIs available to build CH recommendations. The custom dataset used in this study was developed using Europeana and Facebook.

The custom dataset harvested was compared with the Europeana dataset (see Section 4.6), and the results show that it outperformed the Europeana dataset in terms of its richness in social data. Thus, the custom dataset has more social knowledge, such as social tags, because of the additional social information acquired from Facebook.

Chapter Five: Integrating Content and Social Knowledge Representations

5.1 Introduction

The challenges of CH recommendations that this study addresses are the cold-start problem, out of context recommendations, and similar objects but bad recommendations. These Challenges were discussed in Chapter two (see Section 2.4.1, 2.4.2, and 2.4.3).

The CH recommendation challenges are addressed by providing multiple novel knowledge representations assembled from various available domain sources. When a cold-start problem exists, content knowledge representation can be used instead of social knowledge representation to the address the problem (Lika et al., 2014b). However, it is clear from the literature that social knowledge representations produce better CH material recommendations than content knowledge presentations (Wei et al., 2017).

Therefore, integrating content and social knowledge representations provided the opportunity to achieve both cold-start discovery and high-quality recommendations simultaneously (Aslanian et al., 2016). The challenge was retaining the content knowledge representation strength for a cold-start discovery together with the quality recommendations of the social knowledge representation, as shown in Figure 5.1. The cold-start discovery was measured by the number of user interactions; CH objects with few or no user interactions were discovered as cold-start objects.

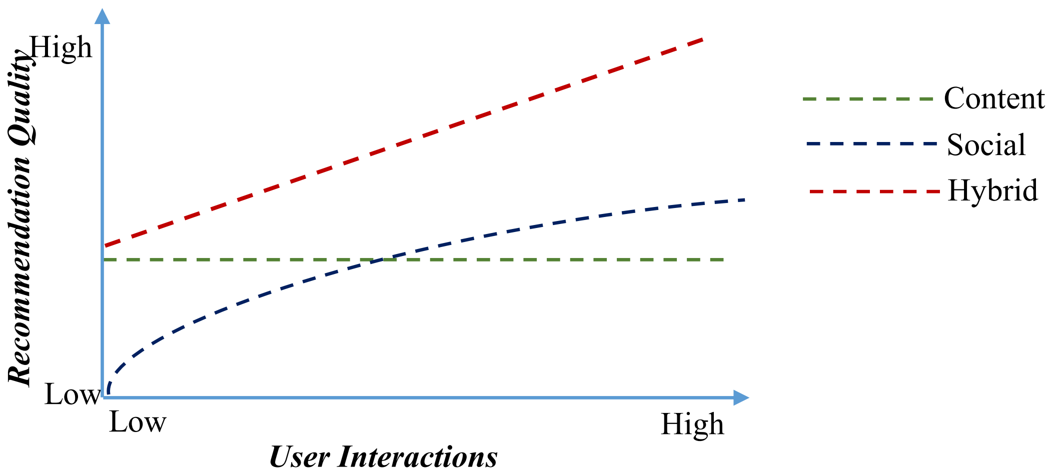


Figure 5.1: Balancing recommendation quality and cold-start discovery

This chapter discusses the characteristics of each knowledge representation and how they were successfully combined to bridge the semantic and social knowledge gaps for CH recommendations. Quality recommendation evaluations of each knowledge representation are discussed later in this chapter.

5.2 Knowledge representation

To combine content and social knowledge representations, it is important to discuss the features of each knowledge representation. The process of knowledge representation consists of three steps: *knowledge pre-processing, knowledge analysis, and result interpretation* (Amatriain et al., 2011). These three steps are carried out in succession, similar to the approach discussed in Chapter Two (Section 2.4.4).

This section discusses the two knowledge representations (content and social) assembled from domain sources (Europeana and Facebook), according to the three processing steps of knowledge representations.

5.2.1 Content knowledge representation

Content knowledge representations describe the important features of CH objects, such as origin (name), description, age (date), geographical location, and image. From Chapter Four (see Section 4.4.2), the output of knowledge extraction is D, a set of documents (d_1 $d_{2,\ldots}$ d_n). Note that each document represents a CH object's feature description. For content knowledge representation, the dataset D needed to be ***pre-processed*** before conducting the ***knowledge analysis***, using appropriate techniques and methods for ***interpretation*** to produce CH recommendations.

5.2.1.1 Knowledge pre-processing

At this stage, before distance measurements and dimensional reductions, the syntactic and semantic pre-processing was conducted to generate a bag of terms (T) from *D* and group the CH objects in the collection according to their semantic relations.

Syntactic pre-processing includes tokenisation, stop-word removal, and stemming. However, in this study, stemming was skipped to avoid poor semantic representation during the semantic pre-processing. The stemming algorithms generate a stemmed form of knowledge with vocabularies that cannot be understood during the semantic pre-processing. Tokenisation is the process of converting the documents in *D* into a collection of tokens. Each token represents a *term* (*t*) belonging to T. In general, tokenisation excludes punctuation, digits, and non-alphabetical characters, but for this research, digits were included as part of the tokens because a CH object's date (digit type) is one of the important attributes considered. The tokens generated from tokenisation contain unwanted terms called *stop-words*. Stop-words include a short function and frequent terms, such as *is, what, when, which, and so on*. The method `Stop_word()` was applied to remove stop-words from the tokens collection and produce T.

$$\mathrm{T} = \{t_1 t_2 t_3 \ldots t_m\} \tag{5.1}$$

In this study, the knowledge representation model used for content representation was the vector space model, a matrix-based model, as shown in Table 5.1. The reasons for that decision were discussed in Chapter Two (see Section 2.2.2).

Table 5.1: Vector space model

	t_1	t_2	.....	t_m
d_1	w_{11}	w_{12}	.....	w_{1m}
d_2	w_{21}	w_{22}	.....	w_{2m}
. . .	. . .	. . .	. . .	. . .
d_n	w_{n1}	w_{n2}	.....	w_{nm}

In Table 5.1, each document *d* is a vector in n-dimensional space. Thus, a matrix *W* is generated to present the weight of terms in the CH object document, as shown in Equation (5.4). The weight term is calculated using *term frequency-inverse document frequency* (TF-IDF) to avoid bias towards occurrences of a term in documents (Lops et al., 2011). TF-IDF is the statistical measure that reflects how imperative a *term* is to a CH object's document in the collection.

$$TF - IDF(t_m, d_n) = TF(t_m, d_n). \log \frac{N}{n_k} \tag{5.2}$$

N and n_k are the number of a CH object's documents in the collection and the number of a CH object's documents in which term t_m has occurred at least once, respectively.

$$TF(t_m, d_n) = \frac{f_{m,n}}{max f_{m,n}} \tag{5.3}$$

$f_{m,n}$ denotes the frequencies of all of the terms t_m that occur in the CH object's document d_n . Thus, *W* is generated as

$$W = \begin{bmatrix} w_{11} & w_{12} & w_{13} \dots w_{1m} \\ \vdots & \vdots & \vdots \quad \vdots \\ w_{n1} & w_{n2} & w_{n3} \dots w_{nm} \end{bmatrix}, \tag{5.4}$$

$$\text{where } w_{nm} = \frac{TF\text{-}IDF(t_m, d_n)}{\sqrt{\sum_{s=1}^{|T|} TF\text{-}IDF(t_s, d_n)^2}} \quad . \tag{5.5}$$

Equation (5.5) is the cosine normalisation assumption of Equation (5.2). This is for the w_{nm} to fall within the [0,1] interval.

Semantic pre-processing is one of the vital stages of knowledge pre-processing in this study, to produce CH objects relationship matrix (Equation [5.8]),which is when semantic relationships between words are identified. This is important in order to allow matching of terms to take into account semantic relatedness beyond the exact text. For example, from Figure 5.3, the terms 'John Alexander' and 'Edwin Austin Abbey' should match because they are semantically related; they are both 'painters/artists' from the '18th century', even though they are textually different.

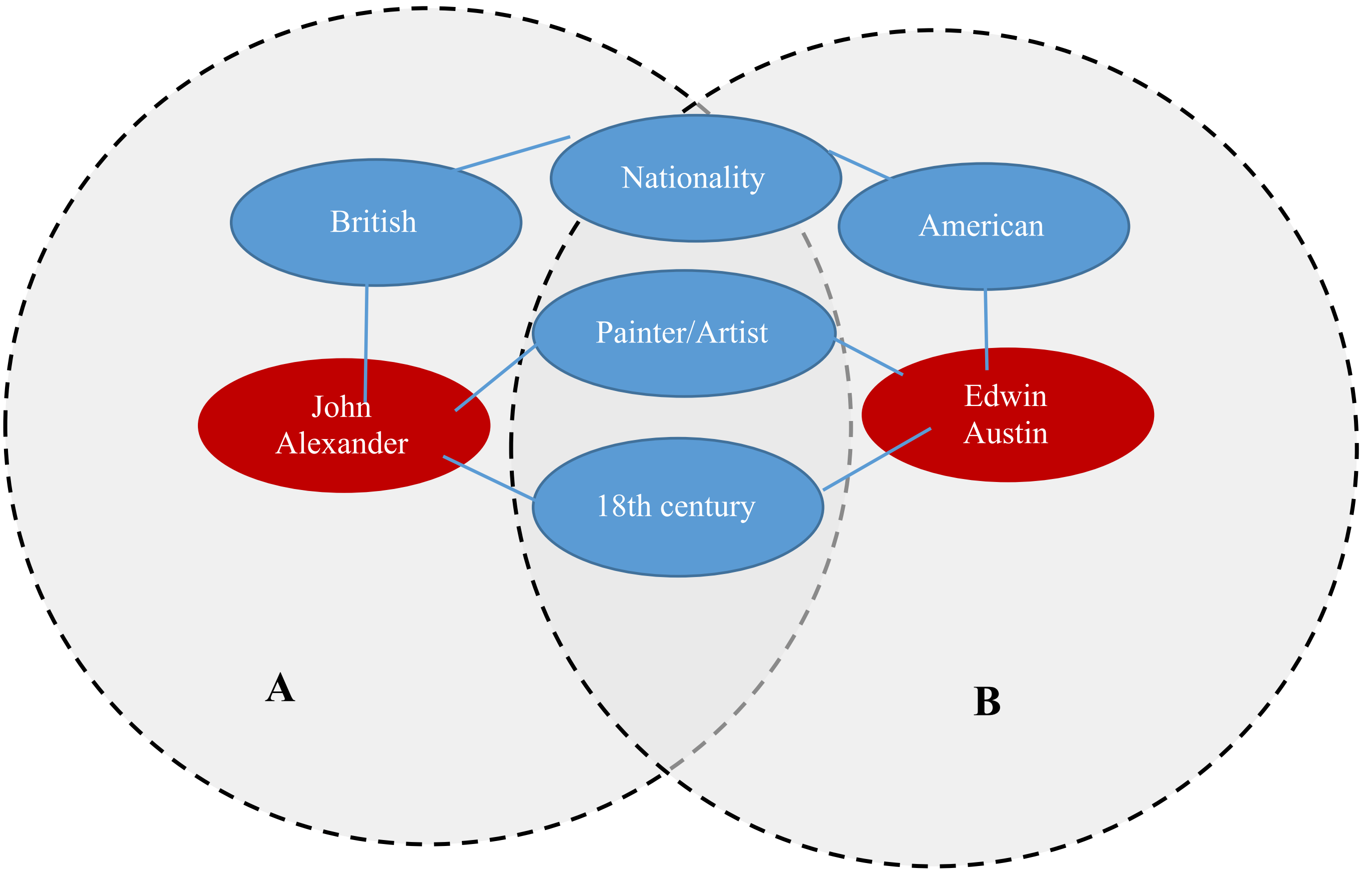


Figure 5.3: Venn diagram of cultural heritage objects' description properties

Figure 5.3 presented the venn diagram of two CH objects description property sets, say A and B. The red oval represent name or origin of the object, while the blue oval represent the other properties describing the CH objects.

Relatedness was constructed using the CIDOC CRM ontology refined by DBpedia to group the terms that shared a semantic relationship. The CIDOC CRM provided definitions and a formal structure for depicting the implicit and explicit ideas, as well as the connections used in the documentation of digital CH objects. The relatedness was expressed as a semantic proximity set of terms *TX* from different assignments, as described in Equation (5.6). Note that $TX \in T$, such that tx_c is a subset of T that belongs to c category group of CH.

$$TX = [tx_1\, , tx_2\, , \dots tx_c] \tag{5.6}$$

In case of this study, there are five category group of CH objects (see Section 4.5.1). Therefore, c = 5

From D and TX, a semantic category *nxc* group matrix, *G*, was generated, as shown in Table 5.2 and described in Equation (5.7).

Table 5.2: Presentation of a semantic category grouping

	tx_1	tx_2	.....	tx_c
d_1	g_{11}	g_{12}	.....	g_{1c}
d_2	g_{21}	g_{22}	.....	g_{2c}
. . .	. . .	. . .	. . .	. . .
d_n	g_{n1}	g_{n2}	.....	g_{nc}

$$G = \begin{bmatrix} g_{11} & g_{12} & g_{13} \cdots g_{1c} \\ \vdots & \vdots & \vdots \quad \vdots \\ g_{n1} & g_{n2} & g_{n3} \cdots g_{nc} \end{bmatrix} \tag{5.7}$$

where $g_{nc} = \begin{cases} 1 \; if \; d_n \in tx_c \\ 0 \; otherwise \end{cases}$

From Equation (5.7), another matrix *R* was generated to present the semantic relations between a CH object's documents. In this case, a probabilistic method was used to predict to which category a particular CH object belonged.

$$R = \begin{bmatrix} r_{11} & r_{12} & r_{13} \cdots r_{1c} \\ \vdots & \vdots & \vdots \quad \vdots \\ r_{n1} & r_{n2} & r_{n3} \cdots r_{nc} \end{bmatrix}, \tag{5.8}$$

where $r_{nc} = \frac{\sum(tx_c : g_{nc} = 1)}{(\sum_1^c tx_c : g_{nc} = 1)}$; c is the number of CH object semantic category groups.

To have a complete content knowledge representation for this study, W and R were transformed into a single knowledge representation, *C*:

$$C = \left(\begin{array}{cccc|cccc} w_{11} & w_{12} & w_{13} \cdots w_{1m} & r_{11} & r_{12} & r_{13} \cdots r_{1c} \\ \vdots & \vdots & \vdots \quad \vdots & \vdots & \vdots & \vdots \quad \vdots \\ w_{n1} & w_{n2} & w_{n3} \cdots w_{nm} & r_{n1} & r_{n2} & r_{n3} \cdots r_{nc} \end{array} \right). \tag{5.9}$$

The first part of the content knowledge presents the term's weight of a CH object's documents in the collection, while the last part presents the semantic relationship's weight across CH objects in the collection. The similarity distance between the CH objects was calculated using cosine similarity. Thus, similarity distance is presented as

$$sim(d_i , d_j) = \frac{1}{2}\left(\frac{\sum_m w_{mi} \cdot w_{mj}}{\sqrt{\sum_m w_{mi}^2} \cdot \sqrt{\sum_m w_{mj}^2}} + \frac{\sum_c r_{ci} \cdot r_{cj}}{\sqrt{\sum_c r_{ci}^2} \cdot \sqrt{\sum_c r_{cj}^2}} \right). \tag{5.10}$$

To illustrate this further with a toy example, for example, given three CH objects d_1 , d_2, and d_3 with dimension *m = 2*, their similarity distance can be calculated as follows:

	t_1	t_2	r_1	r_2
d_1	0.92	0.26	0.2	0.4
d_2	0.41	0.52	0.5	0.6
d_3	0.54	0.21	0.7	0.2

$$sim(d_1\,, d_2) = \frac{1}{2}\left(\frac{((0.92*0.41)+(0.26*0.52))}{\left(\sqrt{(0.92)^2*(0.26)^2}\right)*(\sqrt{(0.41)^2*(0.52)^\wedge 2})} + \frac{((0.2*0.5)+(0.4*0.6))}{\left(\sqrt{(0.2)^2*(0.4)^2}\right)*(\sqrt{(0.5)^2*(0.6)^2})}\right),$$

where $sim(d_1\,, d_2) =$ 0.63, and similarly, $sim(d_1\,, d_3) = 0.91$. Therefore, from the example given above, the CH object d_1 is more similar to d_3 than d_2.

5.2.2 Social knowledge representation

The social knowledge representation is the representation of knowledge that describes (i) CH users' preferences obtained from social tagging and (ii) users' interests gathered from their social activities on social networks (Europeana's Facebook page). The social knowledge representation was developed from user interactions applied to a CH object. Therefore, from the custom dataset harvested in Chapter four, a user-interaction *nxt* matrix, UI, was generated, as shown in Table 5.3 and described in Equation (5.11).

Table 5.2: Presentation of a user interactions

	U_1	U_2	U_3	…	U_t
O_1	ui_{11}	ui_{12}	ui_{13}	…	ui_{1t}
O_2	ui_{21}	ui_{22}	ui_{23}	…	ui_{2t}
O_3	ui_{31}	ui_{32}	ui_{33}	…	ui_{3t}
. . .	. . .	. . .	. . .	. . .	. . .
O_n	ui_{n1}	ui_{n2}	ui_{n3}	…	ui_{nt}

$$UI = \begin{bmatrix} ui_{11} & ui_{12} & ui_{13} \ldots ui_{1t} \\ \vdots & \vdots & \vdots \quad \vdots \\ ui_{n1} & ui_{n2} & ui_{n3} \ldots ui_{nt} \end{bmatrix} \tag{5.11}$$

where ui_{nt} is the number of interactions of user U_t on CH object O_n.

From Equation (5.11), user interaction was aggregated across all users, such that each social context had a weight when applied to a CH object, represented as a Euclidean vector *S*.

$$S = [s_1 s_2 \cdots s_n] \qquad (5.12)$$

where $s_n = \sum_{i=1}^{t} u_{nt}$

Unlike content knowledge representation, each CH object's social context vector *s* contained a sparsity of social data because it had a single dimension for every CH object's social context. Thus, the number of CH objects was the number of vector *s* dimensions.

To illustrate this further with a toy example, for example, given three CH objects and three users. The social knowledge representation will be:

	U_1	U_2	U_3
O_1	3	0	4
O_2	5	1	0
O_3	0	0	0

$$S = [2.3,2.0,0.0]$$

5.3 Integrating content and social knowledge representations

In Chapter Two (see Section 2.2), different recommendation techniques were discussed on the provision of quality recommendations. Each technique has its own individual limitations, hence the idea of a hybrid approach to mitigate these limitations. However, the existing hybrid techniques group the content knowledge representations with respect to their similarity distances and then integrate them with a social knowledge representation. These approaches presume that each content representation of CH objects is meaningful, but that is not the case in the CH domain because of its diversity.

The content and social knowledge representations describe the CH objects' content features and semantic relations and their social relationships with the users. If these representations are combined, there is a high likelihood that the social relationships will be ignored when producing CH recommendations (Ge & Persia, 2017; Pavlidis, 2018). For example, it is possible for a CH object with an *r*-value greater than 0.5 (see Equation [5.8]) to not be categorised within the designated category group; as a result, its social relationship might not be captured. In order to avoid such a scenario, this study integrated content and social knowledge representations in such a way that these relationships would not be ignored.

To integrate the available knowledge representations, content representation C and social representation S were concatenated to form a new vector *I*:

$$I = [w_1 \dots w_m \; r_1 \dots r_c \;\; s_1 \dots s_t] \, , \qquad (5.12)$$

where m, c, and t represent the dimensions of each CH object's content weight, semantic relations, and social relations vectors, respectively. This integration provided the opportunity to measure the correlations between a CH object's content and its social relationships with users. It is crucial to identify these correlations to reduce the sparsity challenge. This can be done by generalising the vector *I*. For this study, the Latent Semantic Analysis (LSA) technique was applied to generalise vector *I*. The reason for this decision was discussed in Chapter Two, Sub-section 2.4.4.2.

5.3.1 Generalisation by latent semantic analysis

Generalisation refers to your model's capacity to adjust appropriately to new, previously unseen knowledge, drawn from the same collection as the one used to make the model. There are different generalisation techniques such as LSA. LSA is an effective generalisation technique for analysing relationships between a set of CH objects and the knowledge they contain by developing a group of concepts related to the CH objects and knowledge representations, providing room to uncover and exploit the relationships between the knowledge representations.

Therefore, LSA was used to generalise the integrated knowledge representation, as described in Equation (5.12), without hindering the social relationships between CH objects and users. The vector *I* was transformed into CH object matrix *I*:

$$I = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1N} \\ \vdots & \vdots & & \vdots \\ I_{m1} & I_{m2} & \vdots & I_{mN} \\ I_{(m+1)1} & I_{(m+1)2} & \vdots & I_{(m+1)N} \\ \vdots & \vdots & \vdots & \vdots \\ I_{(m+c)1} & I_{(m+c)2} & \cdots & I_{(m+c)N} \\ I_{(m+c+1)1} & I_{(m+c+1)2} & \vdots & I_{(m+c+1)N} \\ \vdots & \vdots & \cdots & \vdots \\ I_{(m+c+t)1} & I_{(m+c+t)2} & & I_{(m+c+t)N} \end{bmatrix}, \qquad (5.13)$$

where N is the total number of CH objects in the collection. It is important to note that each column in Equation (5.13) represents the integrated knowledge (combination of content and social knowledge) vector *I* for an individual CH object. In order to ensure that no knowledge representation overpowered the others and to maintain the social relations, each CH object's integrated knowledge vector was normalised to a unit vector using SVD (Amatriain et al., 2011). Normalisation is the process of rearranging knowledge so that there is no data redundancy, so all related knowledge are stored together, and so their dependencies are logical. For example, whenever a particular CH object is dependent on another, the knowledge of two CH objects should be kept within the same proximity. For this study, SVD was used as the basis of LSA to normalise the integrated knowledge vector.

SVD is a powerful matrix factorisation technique that lowers the dimensional feature space without interrupting the actual semantic and social concepts in the input matrix. Using the SVD approach, matrix *I* (see Equation [4.11]) was decomposed into $I = U\lambda V^T$ such that the columns U and V were the eigenvectors of II^T and I^TI, respectively. The λ represents the diagonal matrix that contains a positive singular value. Therefore, II^T and I^TI were computed before computing the SVD of I:

$$A = II^T \tag{5.14}$$

$$B = I^TI \tag{5.15}$$

Equation (5.14) presents the correlations with respect to CH objects, while Equation (5.15) contains CH object correlations with respect to knowledge representations. The key issue of decomposing I using SVD is computing the eigenvectors and eigenvalues of the correlation matrices A and B. Therefore, integrated knowledge I is decomposed and normalised into $A\lambda B^T$, where λ is a diagonal matrix of positive singular value.

5.3.2 The impact of generalisation

To understand whether generalising the integrated content and social knowledge representation using SVD has an impact when a CH object's vector is projected, both normalised and un-normalised input matrices I were considered. For example, in Figure 5.4, which presents the input matrix describing the content, semantic relations, and social tags vectors of each CH object, CH objects O_3 and O_5 suffer cold-start; the weight of all social knowledge for the CH objects are zeros. It can also be observed that CH objects O_1, O_2, and O_4 are represented by both content and social knowledge.

O_1	O_2	O_3	O_4	O_5
0.234	0.176	0.554	0.334	0.239
0.563	0.422	0.235	0.694	0.574
0.235	0.176	0.741	0.741	0.24
0.427	0.32	0.653	0.652	0.436
0.802	0	0.441	0.634	0.818
0.631	0.473	0.257	0.257	0.644
0.551	0.413	0	0.476	0
0	0.514	0	0.637	0
0.574	0.431	0	0.585	0
0.002	0.002	0	0.002	0
0.365	0.274	0	0.372	0
0.631	0.473	0	0.644	0

Figure 5.4: Example of integrated content and social knowledge

To learn the correlation across each dimension, the sample matrix was generalised using LSA. After the generalisation, the weights of the matrix were represented by learning the subspaces of each unprotected CH object vector using SVD:

$$\hat{O}_n = O_n \lambda V^T, \tag{5.16}$$

where O_n is the column vector of CH object n from the input matrix in Figure 5.4.

The result obtained is presented in Figure 5.5. The social knowledge dimensions corresponding to the CH objects affected by cold-start are highlighted in bold and yellow. The figure clearly indicates that after LSA, social knowledge was proliferated into the knowledge representation for the CH objects affected by cold-start. This happens as a result of the content knowledge representation coordinating a strong concept of social knowledge identified by LSA.

O_1	O_2	O_3	O_4	O_5
0.631	0.754	0.256	0.476	0.622
0.249	0.375	0.629	0.159	0.241
0.629	0.754	0.130	0.130	0.620
0.370	0.495	0.185	0.186	0.361
0.096	**0.187**	0.356	0.198	0.087
0.200	0.325	0.590	0.590	0.191
0.259	0.384	**0.492**	0.322	**0.492**
0.462	0.289	**0.708**	0.196	**0.708**
0.241	0.366	**0.633**	0.233	**0.633**
0.621	0.351	**0.490**	0.324	**0.490**
0.438	0.562	**0.367**	0.429	**0.367**
0.200	0.325	**0.719**	0.191	**0.719**

Figure 5.5: Generalised matrix of integrated content and social knowledge

5.4 Evaluation of integrated content and social knowledge representation

The main purpose of integrating content and social knowledge representation is to improve the recommendation quality and increase the discovery rate of cold-start CH objects. Three knowledge representations were evaluated:

- Content – the knowledge representation described in Section 5.2.1 used for the provision of CH recommendations
- Social – the social tags and user activities' knowledge representation extracted from Facebook and described in Section 5.2.2 that provided a yardstick recommendation quality
- Integrated – the combination of content and social knowledge representation defined in Section 5.3

The recommendation quality for this evaluation was measured from the similarity measure of CH objects. The approach used for the similarity measure was *cosine similarity* (Ahn, 2008). This approach considers the CH objects' vectors O_n and computes their association distances as the cosine angle that they form:

$$\cos(O_i, O_j) = \frac{(O_i \cdot O_j)}{\|O_i\|\|O_j\|}, \tag{5.17}$$

where $\cdot, \| O_i \|, and \ \| O_j \|$ are the dot product and the norms of vectors O_i and O_j, respectively.

5.4.1 Evaluation results and discussion

From the results shown in Figure 5.6, the integrated and social knowledge representations outperform the content knowledge representation. This comes as a result of the strong social knowledge present in the representations, which produces a better-quality recommendation. However, for the first two CH recommendations made, there was not much difference between the integrated and social knowledge representations in terms of the CH recommendation quality. This was also true when more than 12 recommendations were produced.

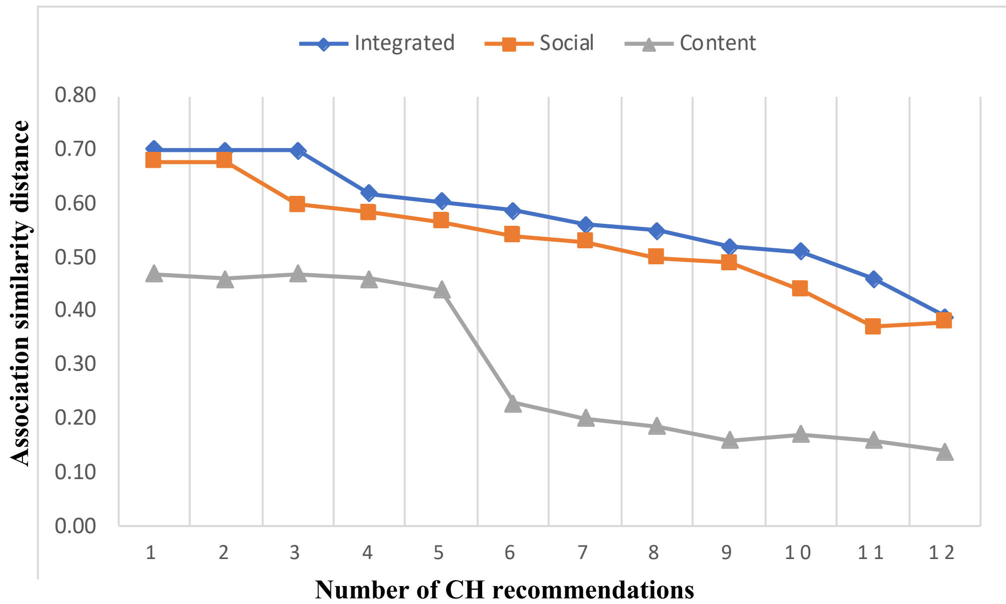


Figure 5.6: Evaluation results without cold-start cultural heritage objects

It is important to note that only around 3% of the CH objects in the dataset collection used for this evaluation were cold-start CH objects, which is why the integrated and content knowledge representations are closely similar. However, when there is a higher rate of cold-start CH objects present, the integrated and social representations' similarities might differ. To evaluate that, social knowledge was removed randomly across 25% of the CH objects in the collection, and the result is presented in Figure 5.7.

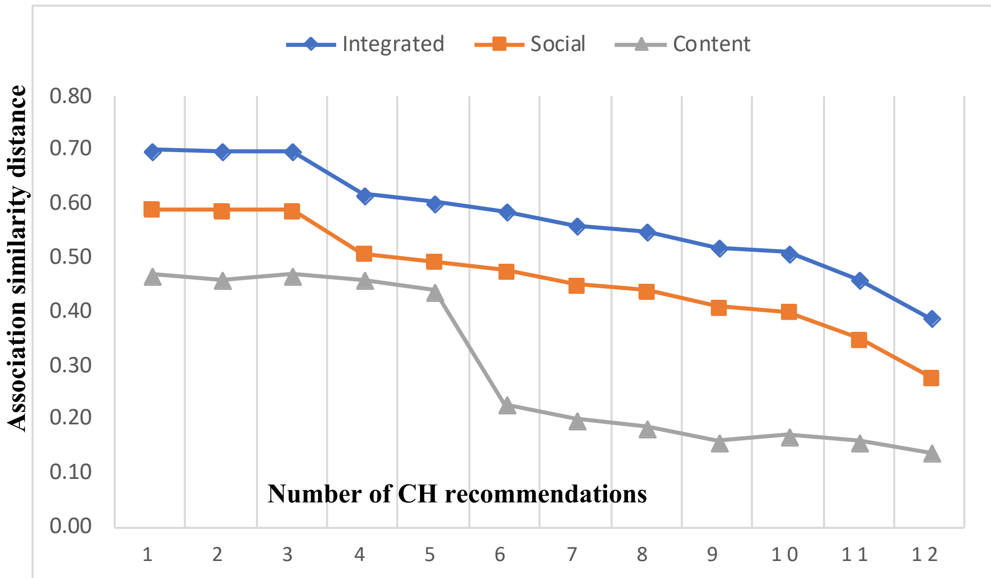


Figure 5.7: Evaluation result with cold-start cultural heritage objects injected

From Figure 5.7, it can be observed that integrated and social knowledge representations continue to outperform the content knowledge representation. However, when comparing the integrated and social knowledge representations, there is a significant difference in terms of CH recommendation quality, which comes as a result of the cold-start CH objects present in the collection. This concludes that the cold-start problem can significantly affect the CH recommendation quality of the social knowledge representation. However, the integrated knowledge representation augments the social knowledge by identifying the lower concepts within the CH objects and learning from the objects with strong social knowledge to propagate the knowledge into the representation that lacks social knowledge. This shows that the integrated knowledge representation provides better quality recommendations than the content and social knowledge representations.

5.5 Summary

In general, social knowledge representation provides better quality CH recommendations than content knowledge representation. However, social representations suffer sparsity of social knowledge in some cases, leading to the problem of cold-start CH objects, which lack social knowledge representations.

To address the challenge of social knowledge sparsity within the representations, both content and social knowledge were integrated into a single knowledge representation to augment the limitations affecting both representations. LSA was applied to generalise the integrated knowledge representation, which provided a new concept area to accommodate both content and social knowledge in a single representation. After the investigation, it was also revealed that the generalisation of these representations affected the overall CH recommendation performance.

Chapter Six: Dynamic Hybrid Cultural Heritage Recommendations

6.1 Introduction

This chapter is an extension of Chapter Five, which discussed the CH object recommendation challenges and how integrating the content and social knowledge representations addressed these challenges. However, it is clear from the literature that weak representations of social knowledge reduce the quality of CH object recommendations. The social knowledge representation produces a better recommendation quality than the content knowledge representation, but it suffers from cold-start. Thus, a dynamic hybrid approach was proposed to further address the drawback of social knowledge representations so that the influence of each knowledge representation differed, depending on the current user status.

The hybrid approach, as discussed in Chapter Five, assembled knowledge representations from a static combination of knowledge. However, this chapter extends the discussion to a dynamic hybrid approach that combines the social knowledge representation and integrated knowledge representation (discussed in Chapter Five) for CH recommendations so that the influence of each knowledge representation depends on the current user status. It is important to note that integrated knowledge representation is the combination of the content and social knowledge representations, while a hybrid representation is the further combination of the integrated and social knowledge representations to produce dynamic hybrid CH recommendations.

This chapter discusses the motivation for using integrated and social knowledge representations to develop the dynamic hybrid representation. The proposed dynamic hybrid approach is then evaluated by comparing the CH object recommendation quality with that of the social and integrated knowledge representations. This demonstrates the effect that each knowledge representation has on the discovery of cold-start.

6.2 Motivation for using integrated and social knowledge representations in dynamic hybrid cultural heritage recommendations

Many of the existing hybrid methods incorporate content into a social knowledge representation to produce static hybrid recommendations, as presented in the previous chapter. In some domains, this approach has an advantage because the content knowledge that describes an item is always available and meaningful. This creates an implicit link between social and content representations and, thus, reduces cold-start.

However, in the CH domain, the specific knowledge that describes the content and social knowledge shows that incorporating content into social knowledge may not be an ideal approach for a dynamic hybrid recommendation of CH materials. CH object content features describe CH types, authors, locations, and dates, while the social representation features high-level knowledge, for example, CH users' opinions, social tags, and metadata. This apparent difference between the knowledge representations limits the ability to produce dynamic hybrid CH recommendations.

Integrated knowledge representation, as discussed in Chapter Five, provides an alternative way to produce dynamic hybrid CH recommendations. This is because there is a link between social and integrated knowledge representations; they share similar knowledge properties that can be introduced to the definition process of dynamic hybrid CH recommendations. Learning integrated knowledge representations can also overcome the problem introduced by content representations for processing dynamic hybrid CH recommendations.

Therefore, the dynamic hybrid CH recommendations presented in this chapter combine the learned integrated knowledge for each CH object with the object's social knowledge. These knowledge representations are always available to address cold-start problems and improve the recommendation quality of CH materials. To develop dynamic hybrid CH recommendations, two development stages are discussed: the method of selecting which knowledge representation to include in the hybrid system and the combination of two knowledge representations into a hybrid CH recommendation.

6.3 Selecting the best integrated knowledge

The integrated knowledge representation described in Chapter Five combines the content knowledge of each CH object and the social knowledge that occurs as a result of user interactions with the object. The content-based approach retrieves the top *K* most similar CH objects that are associated with the social knowledge representation to a query as CH recommendations. The similarity is measured using cosine similarity (Equation [5.17]). However, this similarity varies depending on the user query. Therefore, a consistent weighting is required after every user query. The CH object recommendations – CH objects most similar to the user query – can be presented as a rank list, such that the nearest neighbour to the CH object is $k = 1$. The weight of social knowledge of each CH object in the ranked list is calculated as

$$w_k = \frac{1}{k}, \qquad (6.1)$$

where w_k is the weight of *k*th CH object in the ranked list, ranging from 1 to *K*.

Therefore, the integrated knowledge representation *I* for a given user query is calculated as

$$Q_j = \sum_{i=1}^{K} w_k \, O_j(k) \, , \qquad (6.2)$$

where $O_j(k)$ is *j*th CH object's weight in the integrated knowledge vector (see Figure 5.5), at position *k* in the ranked lists. Using Equation (6.2), each CH object's integrated knowledge was measured according to its position and original strength in the nearest neighbours ranked list.

To produce CH recommendations, cosine similarity was used. Cosine similarity measures the similarity based on the CH social knowledge strength assigned to each CH object. However, since social and integrated knowledge are included in the hybrid CH recommendations, each knowledge representation's non-zero-dimension differences can lead to a low recommendation quality. For CH objects that are rich in social knowledge, the power to produce strong CH material recommendations is desirable, but introducing too much integrated knowledge may reduce such power.

To overcome the challenge of weakening power to produce strong CH recommendations by introducing integrated knowledge, the size of integrated knowledge to be included in a hybrid recommendation was restricted. Each integrated knowledge had a corresponding strength that could decide which to include in the hybrid CH recommendation.

6.4 Weighting social and integrated knowledge

The previous section revealed that the richness of social knowledge on CH objects influences which knowledge representations to include in a hybrid CH recommendation. This can be achieved by introducing weighting parameters on each representation. Therefore, the vector *h* of a dynamic hybrid representation is defined as

$$h_j = \alpha I_j + (1 - \alpha) S_j \qquad (6.3)$$

$$such\ that\colon \alpha = \begin{cases} 0, if\ x\ \geq y \\ \neq 0, otherwise \end{cases},$$

where *i* and s are the vectors of integrated and social knowledge, respectively. Note that both *i* and s are normalised before computing the hybrid vector h. The value $\propto$ controls the influence of both integrated and social knowledge, where x and y are the average numbers of user interactions in integrated and social knowledge representations, respectively.

To determine how the influence of social and integrated knowledge changes as the number of user interactions increases, another experiment was conducted to set α values for hybrid representations, and the result is presented in Figure 6.1. The quality of recommendations achieved by integrated and social representations was calculated for different values of user interactions.

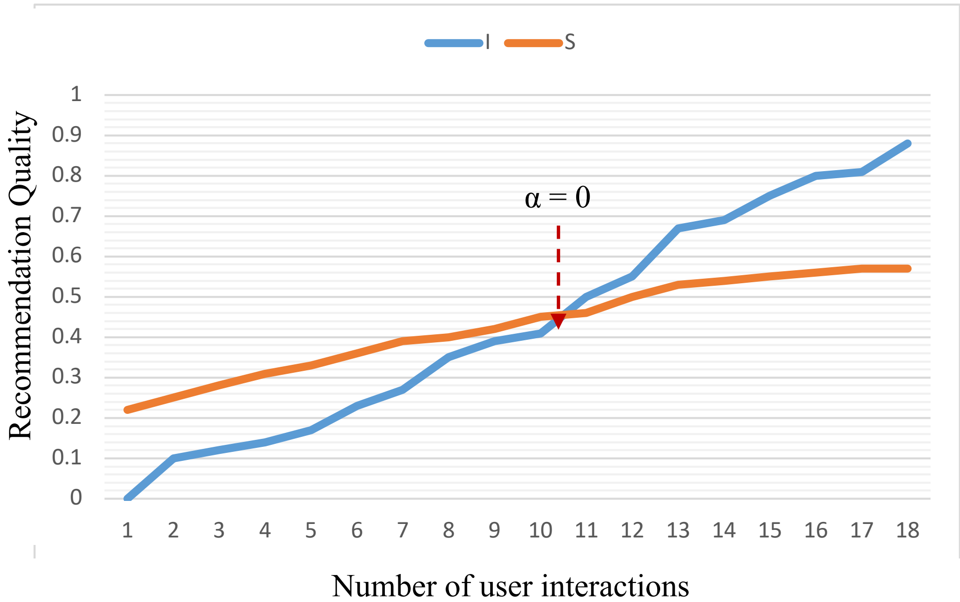


Figure 6.1: Recommendation quality of integrated and social knowledge representations from different values of user interactions

In Figure 6.1, the values of α are set from 0 to 0.5 at an interval of 0.1 to produce a hybrid representation. The values 0 and 0.5 were selected as the lower and upper boundaries, respectively, to guarantee at least an equal weight of user interaction between social and integrated knowledge representations since social representation is a stronger representation for CH recommendations if there is a sufficient number of user interactions.

6.5 Evaluation of the hybrid representation for cultural heritage recommendations

To evaluate the hybrid representation, its recommendation quality was compared with the quality of CH recommendations when the integrated knowledge representation was used. The result obtained provided further insights into the effects of hybrid representation on CH objects with cold-start and a significant number of user interactions. The recommendation quality was measured using the association score (see Equation [3.1]).

6.5.1 Recommendation quality

The recommendation quality achieved against the number of CH recommendations by using the integrated and hybrid representations is presented in Figure 6.2. The blue represents the recommendation quality achieved using the integrated knowledge representation, while the red line represents the recommendation quality produced using the hybrid knowledge representation. The hybrid representation produced better quality recommendations across all of the CH recommendations than those produced by the integrated knowledge representation.

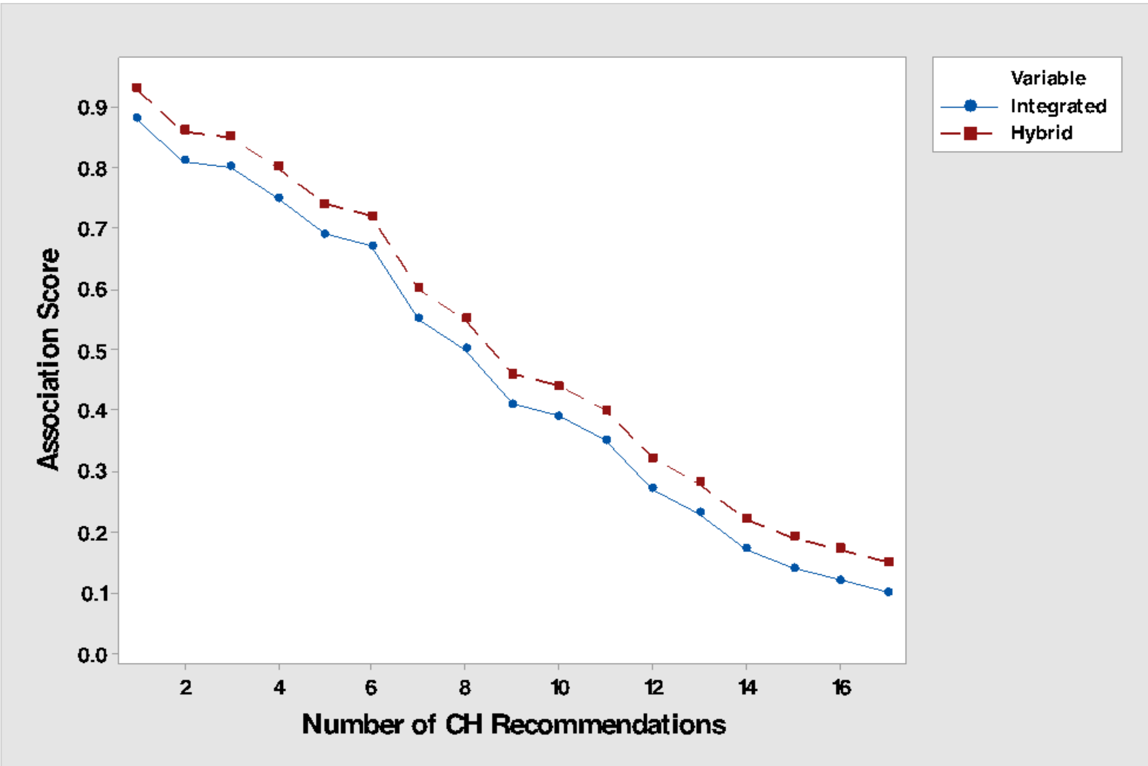


Figure 6.2: Quality of the recommendations produced by the hybrid representation

The better recommendation quality achieved by the hybrid representation over the integrated representation comes as a result of the new social knowledge introduced into the knowledge representation via integrated knowledge. For CH objects that do not require user interactions, a hybrid representation provides a controlled approach to recommending CH objects by exploiting other CH objects' social knowledge.

6.5.2 Effects on the cold-start discovery

In the previous section, the quality improvement of CH object recommendation was achieved by introducing integrated knowledge representation into the hybrid representation. To have a better understanding of how CH objects' user interactions affected the recommendation quality, the effect on cold-start discovery via hybrid representation was examined. Cold-start CH objects (objects that have no user interaction) were injected into the collection, giving approximately 25% of CH objects in the collection no user interaction. The results obtained are presented in Figure 6.3.

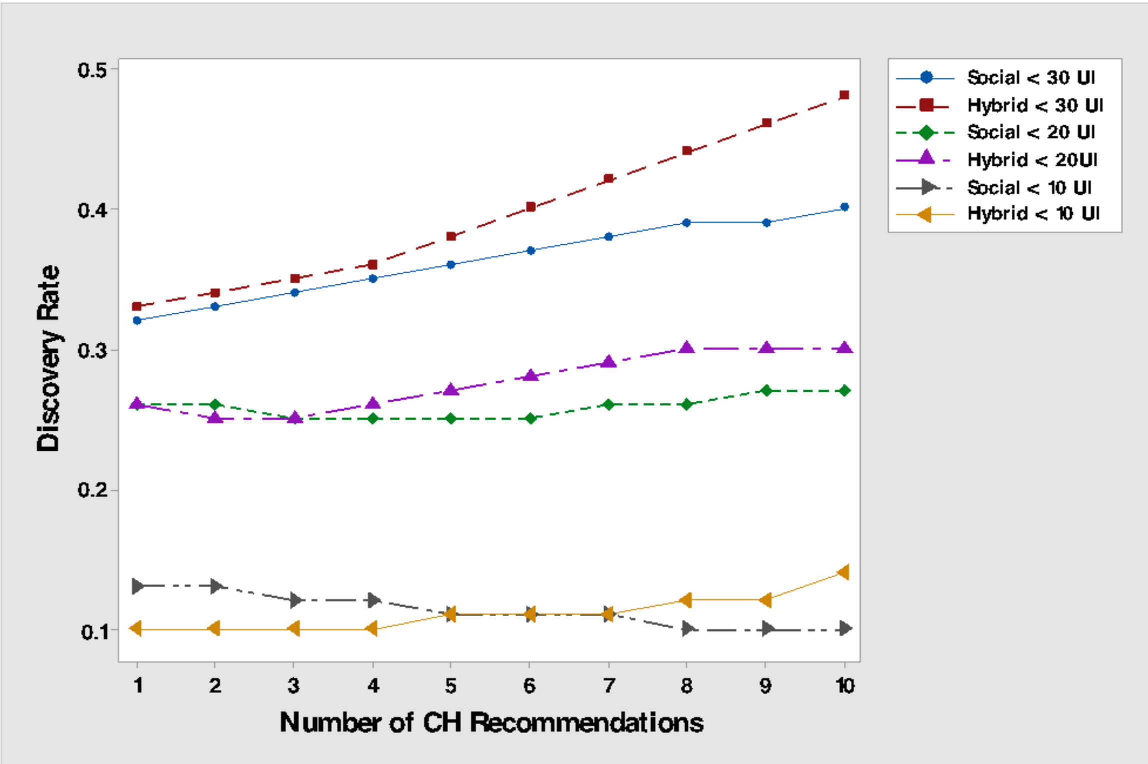


Figure 6.3: Rate of cultural heritage recommendation discovery with respect to the number of cultural heritage objects' user interactions (UI)

The results presented in Figure 6.3 show the cold-start discovery rate of CH objects using social and hybrid representations. The upper segment of the results shows the rate of discovery of CH objects with more than 20 but fewer than 30 user interactions. In this case, the hybrid representation improves more in discovery rate than the social knowledge representation as the number of CH recommendations increases. When the user interactions are fewer than 20 but more than ten, there are no significant changes between the two methods when three CH recommendations are measured, but the discovery rate of the hybrid representation begins improving from four CH recommendations. However, in the case of CH objects with fewer than ten user interactions, the curves fluctuate; there are no significant changes between the two methods of social and hybrid representations.

6.6 Summary

To a certain extent, the cold-start problem affects the recommendation quality in the CH domain. An increased number of user interactions on CH objects increases the recommendation quality. Thus, the ability to augment the sparsity of user interactions within a social knowledge representation opens up the opportunity to improve the CH recommendation quality for all CH objects, which the hybrid representation provided.

This chapter described an approach to further augment the sparsity within a social knowledge representation without directly including the content of CH objects. The content knowledge representation describes the features of CH objects, but the social knowledge representation often presents users' interactions and opinions. Thus, the content knowledge representation is incompatible with being directly included in a dynamic hybrid representation for CH object recommendations. Instead, the integrated knowledge representation, discussed in Chapter Five, was used to indirectly ingest the content into a dynamic hybrid representation.

The content knowledge incorporated into the integrated knowledge representation gave an edge to the hybrid representation. This approach also eased the hybrid recommendation process, for example, by not including generalisation but focusing instead on combining the two similar knowledge representations – social and integrated – into a single hybrid.

The recommendations of CH materials produced by the hybrid representation were of better quality than those produced from the social knowledge representation. With the hybrid representation, strong CH object recommendations could be made even for CH objects that had a low number of user interactions since they could be reinforced by integrated knowledge. The hybrid approach has the ability to improve the quality of CH recommendations, regardless of a CH object’s level of social knowledge available.

Chapter Seven: VISE: An Interface for Visual Search and Exploration of Cultural Heritage Collections

7.1 Introduction

One of the research questions of this study is how a VSI can help CH users with no domain knowledge to explore a large collection of CH materials for new information discovery. The current chapter addresses this question by presenting VISE, an interface that enables a visual search and exploration across a large collection of CH materials. VISE provides an interactive visual summary of information relating to CH materials to address the challenges faced by online users with no domain knowledge when exploring large CH collections.

Knowledge assembled from CH domain sources usually includes abstract terms that are not familiar to CH users, especially to those that have no domain knowledge. Therefore, CH users find it challenging to know which terms to use to explore CH collections for new discoveries. For example, a CH user can have an idea of which query to perform but may not know the specific terms to use. In this chapter, a visual summary of the whole CH collection as a tag cloud is presented to the user to initiate the search process. A tag cloud is a visual presentation of user-generated tags attached to CH objects' contents to represent the prominence of the tags depicted.

Chapter Two (see Section 2.1) established that a VSI is one solution that addresses the challenges of exploring a large collection of CH objects. Thus, this chapter presents VISE. It provides a diverse searching strategy, which is especially useful when users are unaware of the full details of their tasks or do not have domain knowledge, allowing users to initiate their search by selecting terms from an interactive visual interface. In contrast to alternative search interfaces, VISE recommends terms that are specific to objects in digital CH collections and provides insights based on semantic and social relationships between CH objects and users, as discussed in Chapter Five (see Section 5.2).

The current chapter also discusses the evaluation of the VSI satisfaction level of users after using the interface, especially those with no domain knowledge. The work presented in this chapter has been published in ACM's *Journal on Computing and Cultural Heritage* (Usman & Antonacopoulos, 2019).

7.2 Functional requirements of VISE

VISE has the following functional requirements:

I. VISE should provide a visual summary of the entire collection of CH materials to support the search and recommendation of CH objects for new discoveries. VISE uses a tag cloud to present such a visual summary. Tag clouds on VISE should provide CH users with the ability to interactively initiate their search to find relevant CH materials.

II. The search result and CH recommendations presented on VISE should be in a faceted classification. This will encourage CH users to refine their search results in the context of the current user's search status. For example, CH users can improve their search by applying user-specific (e.g. geographical location) constraints to the search results.

III. CH users with no domain knowledge should be able to explore VISE and make new information discoveries.

7.3 Design and implementation of VISE

The creation of VISE involved two main stages: knowledge extraction and representation and creation of the VSI. A large portion of the first design and implementation stage was discussed in Chapters Four and Five. Therefore, this section highlights some of the key points relating to the implementation of the VSI that were not discussed previously.

7.3.1 Knowledge extraction and representation

It is important to note that before building VISE, the knowledge and information used (text and images) were extracted from unstructured and noisy sources (see Section 4.4). As a large portion of the information is encoded in natural language, Goerz and Scholz (2010) described the need to extract knowledge from unstructured sources. Tools and techniques were outlined to extract semantically meaningful insights from unstructured data. The dataset used to demonstrate this study represents digital CH collections with information about CH objects and users harvested from Europeana and Facebook (see Chapter Four).

From the webpages collected, important feature attributes and relevant information were extracted for building VISE. The attributes extracted from those pages and their representations were explicitly explained in Chapter Four. These knowledge representations provided a convenient application for indexing and providing quality CH recommendations. An index optimises the performance and, consequently, the speed of an IR system in finding relevant information. The visual interface provided for the exploration of a CH collection generated from these knowledge representations.

7.3.2 Visual search interface

The key purpose of the VSI was to provide an alternative to a keyword search that would encourage users to search for CH information by exploration. To achieve this, an interactive visual summary was created, which presented the whole CH collection in the form of a *tag cloud* (Castella & Sutton, 2014).

The tag cloud is a visualisation technique that provides an interactive image display comprising information from a designated set of collection documents. This allows important terms from the collection to be presented in a visual schema that encourages search by exploration. The design and implementation of the VSI are composed of three pre-processing stages, as shown in Figure 7.1.

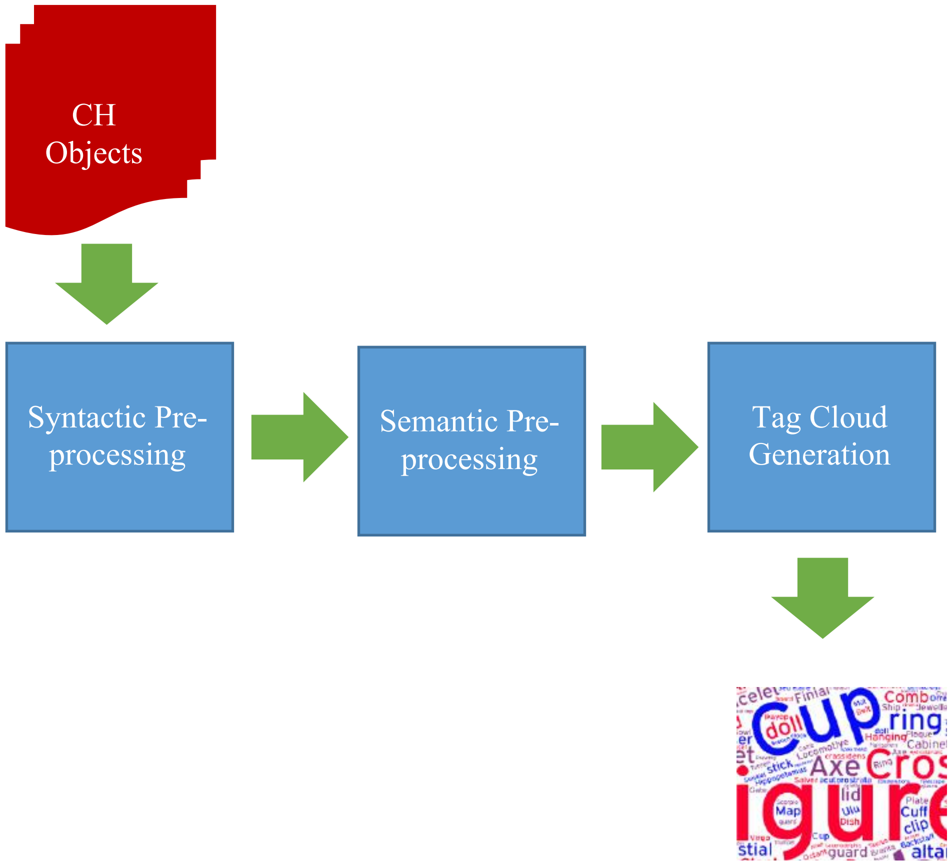


Figure 7.1: VISE design processing stage

CH objects from webpages collected needed to be pre-processed before generating the tag cloud. Tag clouds were created from the bag of terms generated from the content knowledge extracted (see sub-Section 5.1.1.1). The syntactic and semantic pre-processing stages were discussed in Chapters Four and Five, respectively. Briefly, syntactic pre-processing includes tokenisation, stemming, and stop-word removal. It is important to note that in this case, stemming was omitted because the stemming algorithms generate stemmed forms not included in most electronic dictionaries, which can introduce a setback during the semantic pre-processing. Semantic pre-processing uses the CIDOC CRM ontology refined by DBpedia to group the terms that are similar in meaning and nature by assigning a single term, called a *root*, to represent them. CIDOC-CRM provides definitions and a formal structure for depicting the implicit and explicit ideas, as well as the connections used in the documentation of digital CH objects. The root provides a short and broad description and records semantic relations between CH objects.

Tag cloud generation uses the *root* processed from the semantic pre-processing stage to generate a visual summary of the entire CH object collection as a tag cloud. The root refers to the term representing a group of similar terms, for example, the terms ‘football’, ‘hockey’, and ‘tennis’ can have ‘sport’ as their *root*. After extracting and presenting the terms, the *VISE User Interface* was built, providing two options for searching for CH objects: keyword search option and visual search option, as shown in Figure 7.2.

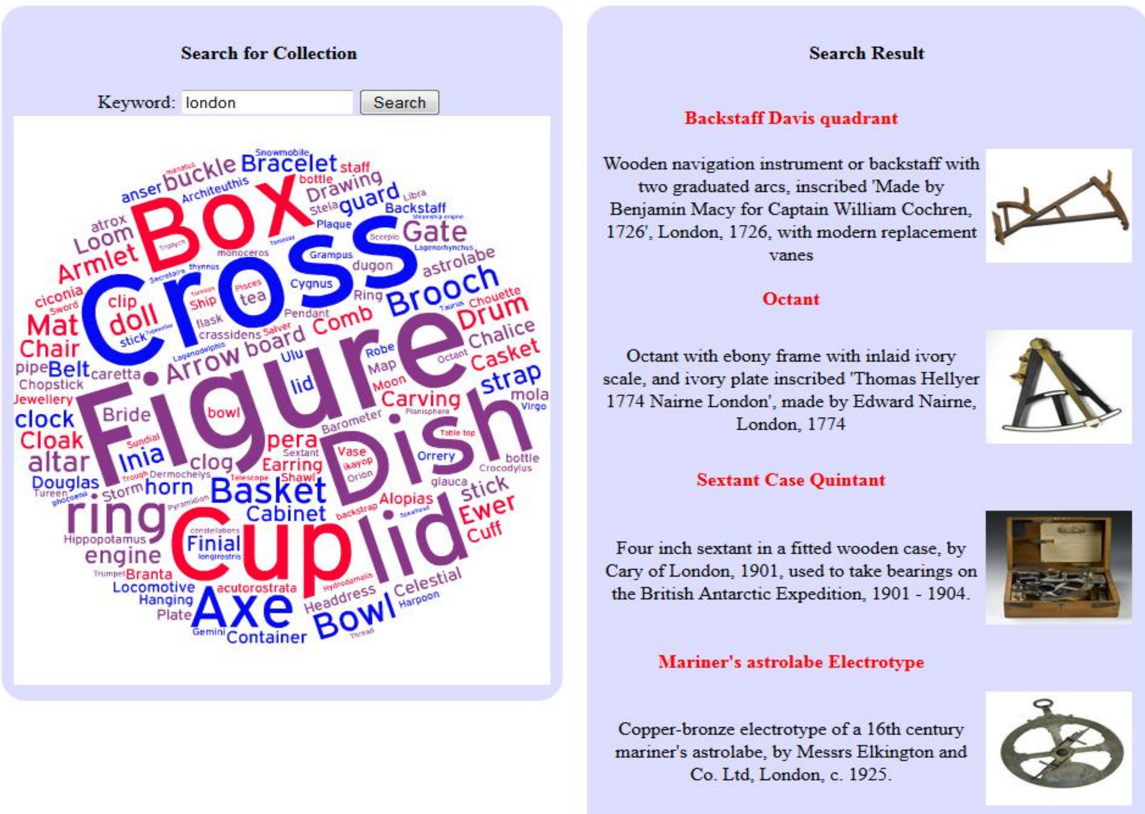


Figure 7.2: VISE user interface

From the interface, CH users can key-in or select the recommended terms from the tag cloud to populate the search box, as illustrated in Figure 7.3. This provides an opportunity for CH users, especially those with little or no domain knowledge, to explore a large collection of CH materials.

Figure 7.3: Search box populated with the recommended term selected from the tag cloud

The CH objects returned from the search result are listed in relevance order on the VISE search result area. The result displays the CH object’s name, description, and image, as shown in Figure 7.4.

Figure 7.4: Cultural heritage object from a search result on VISE

7.4 User satisfaction evaluation of the visual search interface

While in the IR community, the quality of algorithms that match users' queries and the quality of the indexing methods are paramount, one of the objectives of this study was to provide an interactive VSI that would encourage the exploration of CH collections. Thus, the evaluation presented in this chapter is that of the user satisfaction of the proposed VSI compared to a system with no visualisation.

7.4.1 Participants

The user group for this experiment included 50 participants aged between 27 and 38 years; 24 were female, and 26 were male. Out of the participants, 19 had domain knowledge, while the remaining 31 had little or none.

The evaluation was conducted through a USE questionnaire – Usefulness, Satisfaction, and Ease of use (Lund, 2001). It should be noted that the evaluation was not carried out in a museum but rather in a laboratory as it was not feasible to install VISE in a real museum setting. The target application scenario is an online search of the CH objects; therefore, the location of the users was not a factor in the experiment. The evaluation was carried out in six laboratory sessions within a three-month period. In the first two sessions, the participants were experts (with domain knowledge background), while non-experts attended the remaining sessions.

7.4.2 Procedure

Two systems, VISE (denoted as a system with visualisation [SWV]) and a system with no visual interface (SWNV), were provided to the participants to search for CH objects. Each participant was instructed to explore a CH object collection for 30 minutes. No specific task was given to the participants so as not to influence their overall satisfaction. After the experiment, two questionnaires were handed to the participants to express their levels of satisfaction with the two systems to test the following hypotheses:

H₀: *The mean user satisfaction level between the SWV and SWNV does not depend on the VSI.*

H₁: *Such dependency does exist.*

The participants completed the first questionnaire after using the SWV, and they completed the second questionnaire after using the SWNV. In the questionnaires, participants could express their satisfaction through a five-point Likert scale, from 1 – strongly disagree to 5 – strongly agree. The users evaluated VISE along three dimensions (ease of use, satisfaction, and usefulness), as suggested by Lund (2001); each dimension had a series of questions to answer. Participants were briefed on how the two systems worked before using them. Figure 7.3 presents each participant's satisfaction level in percentages.

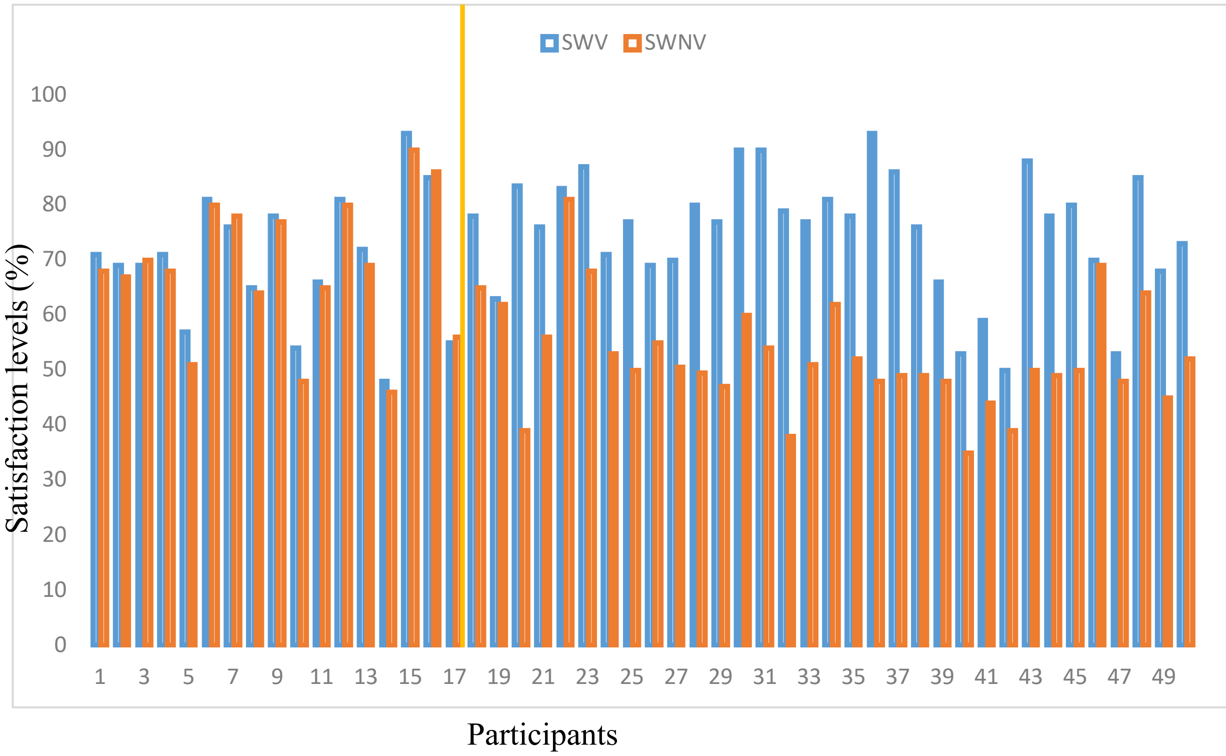


Figure 7.3: Participants' satisfaction levels (%)

It can be observed in Figure 7.3 that there are no significant differences in satisfaction levels between the two systems among the group of users that participated in the first two sessions of the experiments (users with domain knowledge). The difference in satisfaction, however, among the participants of the later sessions (non-expert users) is significant between the two systems, as discussed next section.

7.4.3 Results and discussion

It is important to note that the data collected throughout the experiment are ordinal, and the samples are independent. Therefore, a non-parametric test, a two-sample t-test, was performed with a 95% confidence interval, and the result is shown in Figure 7.4

Estimation for Difference

Difference	95% CI for Difference
15.67	(10.77, 20.57)

Test

Null hypothesis H_0: $\mu_1 - \mu_2 = 0$
Alternative hypothesis H_1: $\mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
6.35	96	0.000

Figure 7.4: Two-sample t-test result

From the results obtained, the p-value < 0.05, which proves that the participants' satisfaction levels depended on the visual interface, rejecting the null hypothesis. Overall, participants were more satisfied with the proposed search interface, VISE, than with SWNV. More specifically, in terms of the domain background knowledge, there was no significant difference observed in satisfaction levels for participants with domain knowledge, while there was for those with no domain knowledge, as shown in Figure 7.5. This is because users with domain knowledge know the specific keyword to initiate a search because of their knowledge background, while users with no domain knowledge do not know the keyword to use.

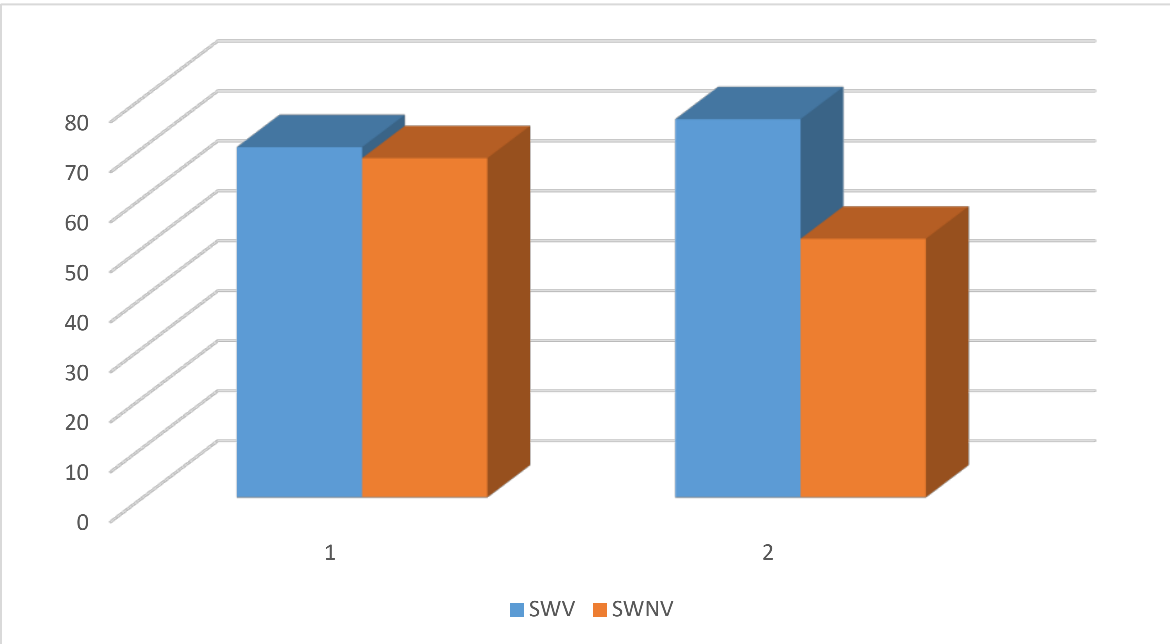


Figure 7.5: Satisfaction level based on domain background knowledge (1 – with domain knowledge and 2 – without domain knowledge)

7.5 Summary

The development of a VSI that can enable the exploration of a large collection of CH materials for new discoveries is one of the objectives of this study. To achieve that objective, VISE was developed. This chapter presented the functional requirements of VISE, which included the provision of a visual summary of the whole collection. The techniques involved in designing and developing VISE were also discussed. Finally, the user satisfaction evaluation of VISE was discussed.

The evaluation of VISE revealed that users were more satisfied and attracted to explore CH collections via VISE than via the SWNV. Users with little or no domain knowledge found it easier to explore collections and find CH objects of interest while using VISE, in contrast to the SWNV.

Chapter Eight: User Evaluation

8.1 Introduction

The main goal of every RS is providing quality recommendations to users. Therefore, it is important to involve users in evaluating the CH recommendations. For the work described in this Thesis, three recommendation approaches were evaluated through user study:

- Content – as discussed in Chapter Five (see Section 5.2.1)
- Social – as discussed in Chapter Five (see Section 5.2.2)
- Hybrid – as discussed in Chapter Six

In addition to recommendation quality, the recommendation novelty was also measured via the user evaluation. Novel recommendations are CH recommendations that users are unaware of. Data on demographics and user satisfaction levels were collected to further analyse the results in full context and the performance of each recommendation approach.

The current chapter starts by discussing the evaluation design, including what parameters are likely to be measured and how to measure them for quality CH recommendations. Then, the user participation, such as the demographics and domain knowledge background of the users who participated in the evaluation process, is discussed. The recommendation quality results obtained from the experiments are later presented, leading to some unanticipated observations. The last sections further discuss how these observations affect the performance and novelty of each approach. Finally, the chapter's summary and conclusion are provided.

8.2 Evaluation design

The main aim of the user evaluation for this study was to assess the three recommendation approaches in terms of their novelty, serendipity, and recommendation quality. To achieve that aim, it was important to consider some key issues: (i) an unbiased recommendation presentation for the users so as not to influence their feedback; (ii) an assessment of the recommendation quality, serendipity, and novelty; and (iii) what type of data is more important to collect.

8.2.1 Presenting recommendations to the users

The CH recommendations were presented to the users through a web interface for evaluation; the web interface presents CH recommendations from different approaches – content, social, and hybrid. To avoid bias, the recommendation approaches used were not indicated on the interface. The list of the top ten CH recommendations produced from these approaches was presented to the users. The recommendations were presented randomly to hide the actual recommendation order from the users and avoid giving an edge to the users that constantly review recommendations according to their presentation order.

Users could evaluate as many CH recommendations as possible within the tasks given to them. In order to avoid technically minded users from accessing information that could influence their decisions and interactions with the CH objects presented, all of the required parameters and recommendation orders were stored on the server side.

8.2.2 Capturing recommendation quality

To capture the CH recommendation quality, users were instructed to evaluate the quality of the CH recommendations presented to them. To achieve that, users interacted with the first set of CH recommendations presented to them and expressed their opinions of the CH object. This answered the question, 'Given these CH objects, how good are these CH recommendations?' The users' opinions were expressed by rating the recommendations on a scale of one star to five stars.

8.2.3 Capturing novelty

To capture the recommendation novelty, it was important to identify which CH recommendations presented were known by the users and which were not known. This was achieved by asking users the following questions:

1. Do you know the CH object's author (origin) or name and geographic location?
2. Do you know the CH object's author or object's name but not the geographic location?
3. Do you not know either the CH object's name or location?

The knowledge gathered from these questions further revealed the proportion of novel CH recommendations that each approach produced.

8.2.4 Demographic questionnaire

Different groups of users are likely to have different CH recommendation perceptions. To understand that and to have knowledge of the users that participated in the study, each user was issued a demographic questionnaire to complete at the beginning of the evaluation process. The questions presented in the questionnaire are shown in Table 8.1.

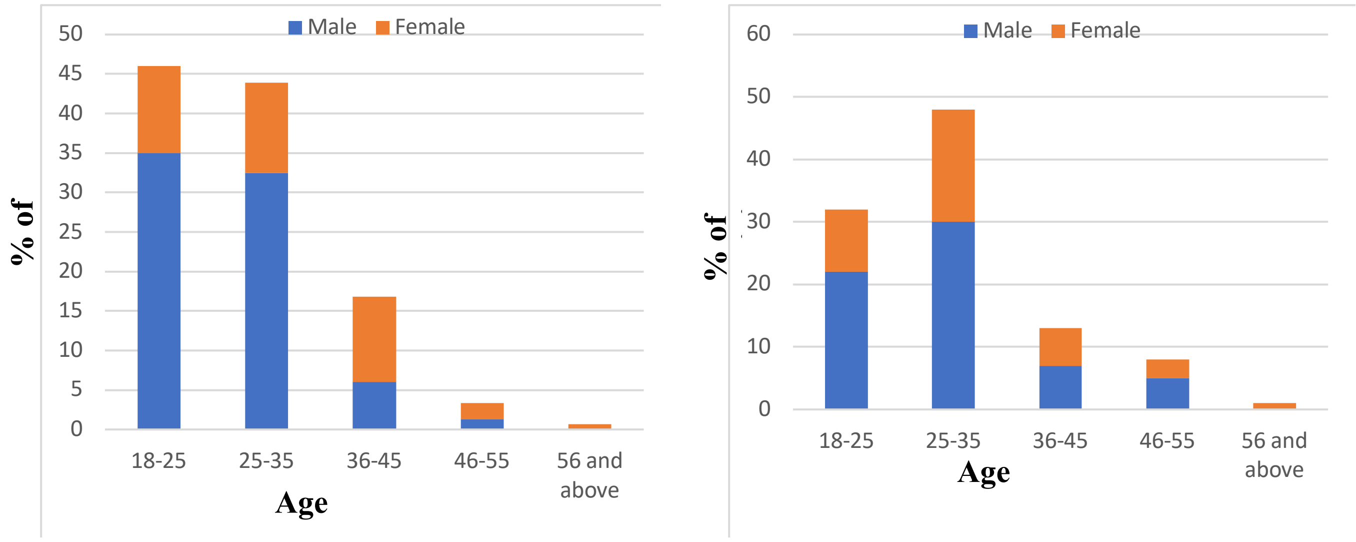
Table 8.1: User demographic questionnaire

Question	**Input Type**	**Options**
Gender	Single Selection	Female Male
Age Group	Single Selection	18–25 26–35 36–45 46–55 56 and above
Cultural Heritage Knowledge Background	Single Selection	None – no particular knowledge in CH-related topics Basic – learned at school, read from web or magazine Advanced – often visit museums, work with CH materials

		Professional – librarian, archaeologist, etc.
How often did you visit museums and art galleries from a web interface?	Single selection	Very often Once in a week Once in a month Never
Type of CH interests	Multiple selections	Tangible Culture, e.g. artworks, paintings Intangible Culture, e.g. artist expression Natural Heritage, e.g. geological elements None

8.3 User participation

The study had 148 participants. Most participants were post-graduate students at the University of Salford in the United Kingdom and ATB University in Nigeria. Some of the participants were users of Europeana's Facebook page. The proposed system was deployed online for 30 days, and the weblink was distributed to participants through emails and social media. The participants evaluated approximately 2,359 CH recommendations; 36% of the recommendations came from initial recommendations presented to the users at the start of the experiment, while the remaining recommendations came from the users' queries. A demographic questionnaire was provided to the participants at the beginning of the study. The demographic questionnaire (see Table 8.1) captures the demographic and domain knowledge background information of the participants to show the participants' diversity.



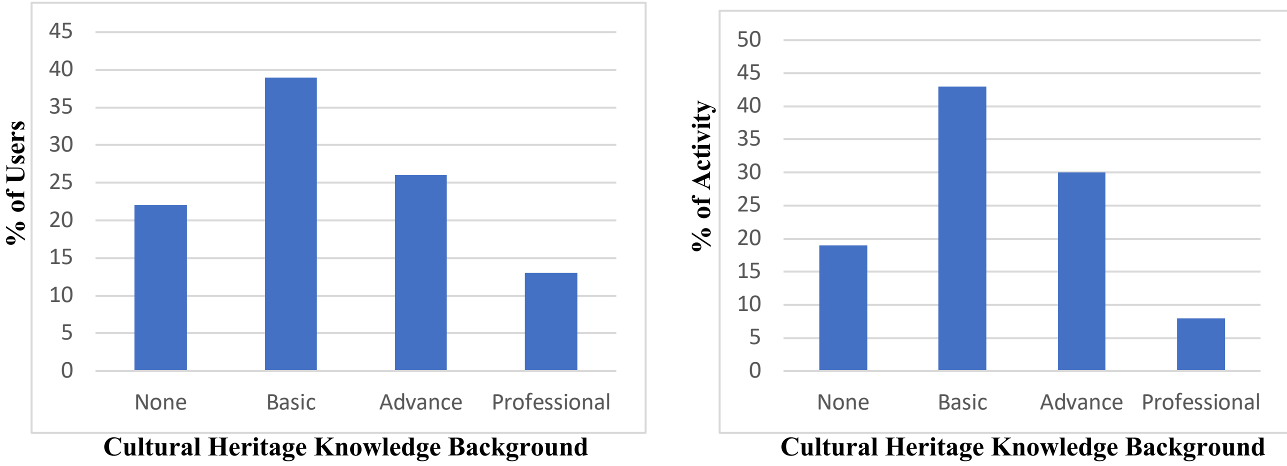
(b) Activity

(a) Users

Figure 8.2: Users' gender by age range

The demographic data captured were collated and evaluated. Figure 8.2 shows the gender and age ranges of the users who participated in the experiment. From the results obtained, 65% of the participants were male, and 35% were female. The majority of participants were between 18 and 25 years, followed by the age group of 26 to 35. Figure 8.2b shows users' engagement (activity) in the study. The results show that the 18 to 35 age range engaged with more activities than other age ranges during the study, and male participants between 26 and 35 years engaged more in the study when compared with other user groups.

Another important piece of data obtained was the users' domain knowledge backgrounds. Figure 8.3 presents the users' background knowledge and engagement. From the results, it seems there is not much correlation between users' CH knowledge backgrounds and their engagement in the study. Also, approximately 78% of the participants had a minimum of a basic knowledge background; for example, they had either learned from school or read from books and on the web.

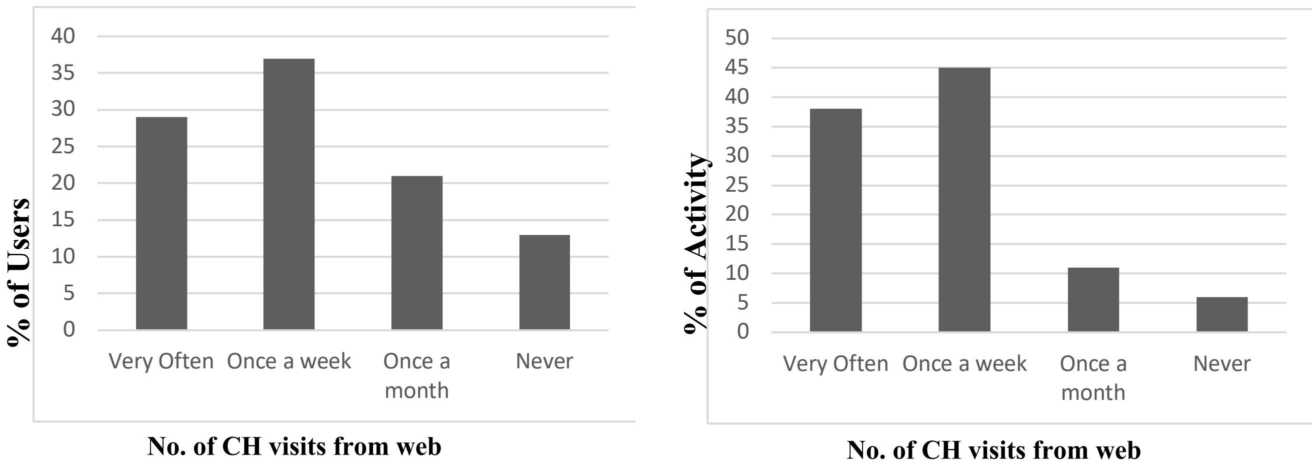


(a) Users

(b) Activity

Figure 8.3: Cultural heritage knowledge background

The information obtained on how often users explored CH objects through a web interface provided insight into users' activeness. From Figure 8.4, it is observed that around 93% of the users had explored CH materials through a web interface at least once. Unlike domain knowledge background, there was a correlation between users and activity in terms of exploring CH materials through a web interface. This is due to the majority of the activity in the study coming from users that either explored CH materials very often or at least once a week.



(b) Users

(a) Activity

Figure 8.4: Visiting periods of cultural heritage materials from a web interface

Users might be interested in three types of CH materials: tangible culture, intangible culture, and natural heritage. Tangible culture comprises physical CH objects, such as artwork, clothing, books, and other artefacts that are preserved for the future. Intangible culture includes non-physical CH objects that are often maintained by different societies during a specific period of history, for example, aesthetic and spiritual beliefs and social values. Natural heritage includes the natural environment that attracts tourists, for example, cultural landscapes.

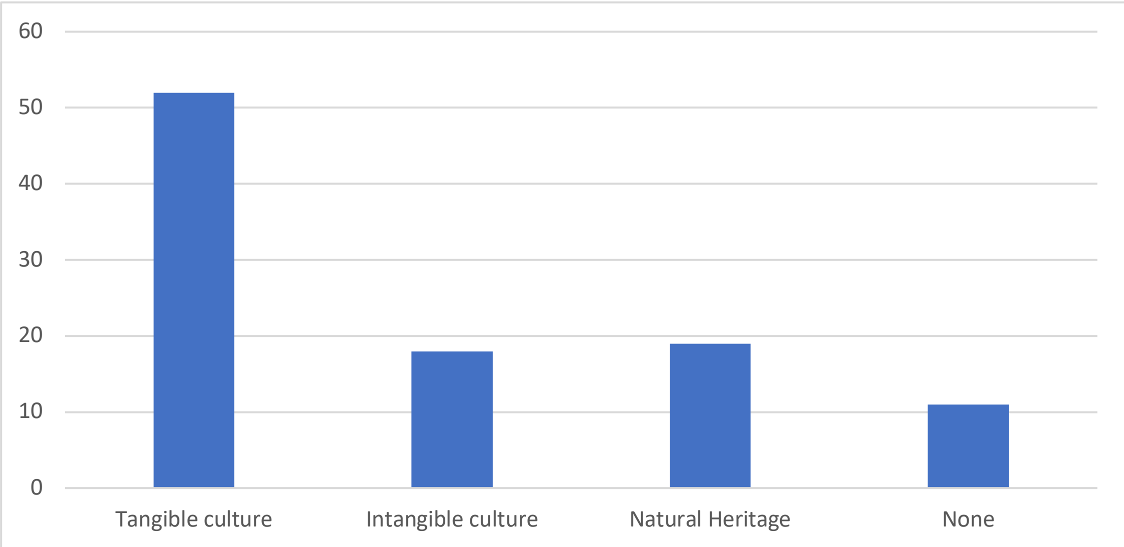


Figure 8.5: Users' interests on types of cultural heritage

Participants provided the information on CH objects in which they were interested. This information was used to trigger the initial CH recommendations presented to the users during the study. The users' interests are illustrated in Figure 8.5. The figure clearly shows that approximately 52% of the users were interested in tangible culture; these objects occupy a large portion of the dataset.

8.4 User score and association score comparison results

The information gathered from the users during the user study provided the opportunity to investigate the prediction of CH recommendations when compared with the results obtained in Chapter Six (see Sections 6.4.1 and 6.4.2).

To compare the association score – the proportion of CH users who agreed that there was an association between two or more CH objects (see Equation [3.1] in Sub-section 3.4.1.1) – with the user score, the mean average score (MAS) of each query presented was calculated as

$$MAS(Q) = \frac{1}{|Q|}\sum_{q\in Q}\frac{1}{K}\sum_{k=1}^{K} score(q, R_k) , \tag{8.1}$$

where Q is a set of an individual CH object queries; an instance $q \in Q$ is a query evaluated by a single CH user. The number of CH recommendations is K; for this study, this was fixed at ten. The k^{th} CH recommendation for q is R_k, and the individual CH user score is $score(q, R_k)$. The results obtained from the experiment are presented in Figure 8.6. The results show the average MAS score across all queries. The bars in Figure 8.6 represent the MAS of the hybrid, social, and content approaches. The first group of bars indicates the MAS for users, the score provided by CH users, while second group of bars indicates the `association score (see Equation [5.17]). Throughout the evaluation, the proposed hybrid approach outperformed the other approaches, followed by the social representation approach.

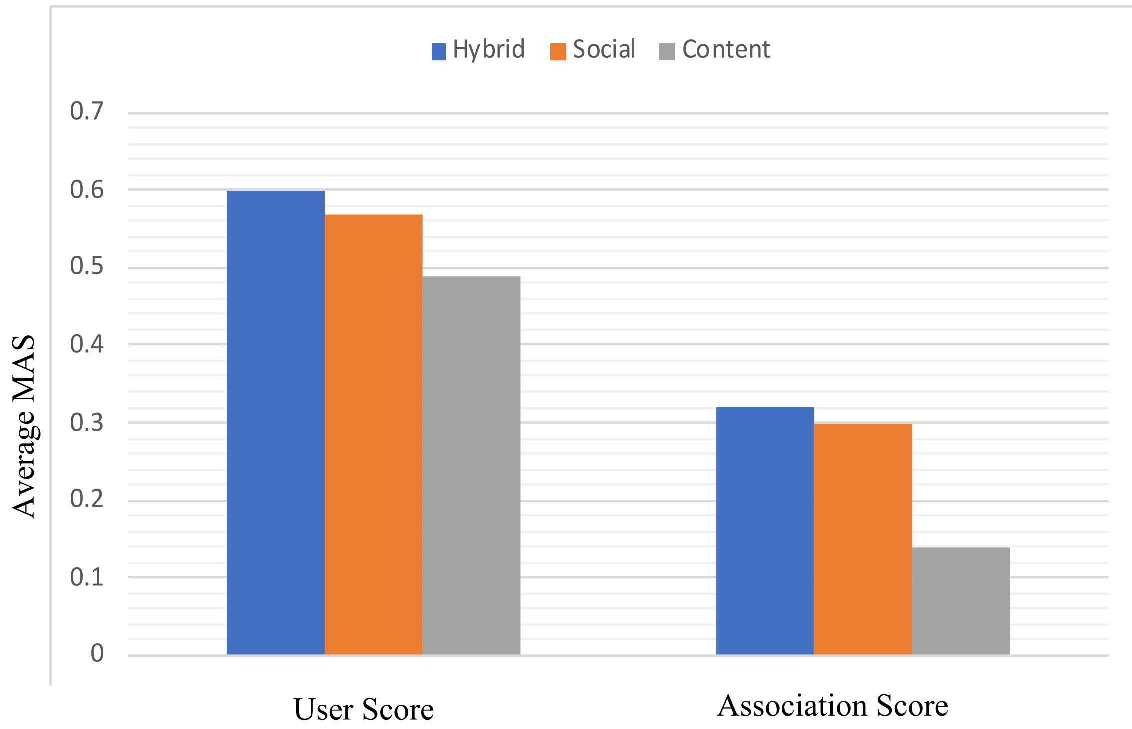


Figure 8.6: User and Association scores comparison

The results generated for the user score also followed the same recommendation quality pattern for each approach. In this case, it was the interpretation of the result values that was important, not the exact values generated. Also, the user score indicated that the hybrid approach performed better when compared with the social and content approaches.

In order to have further insight into how the scores correlated, the MAS for each individual query's user score was calculated for all approaches, as shown in Figure 8.7. The red line indicates the best fit across all the queries. From the result, it can be observed that the association score values range between 0 and 1, while the user score values are between 0.25 and 0.80. This indicates that it is likely that the user scores are more moderate than the association scores; hence, user scores would not produce extremely bad or good CH recommendations.

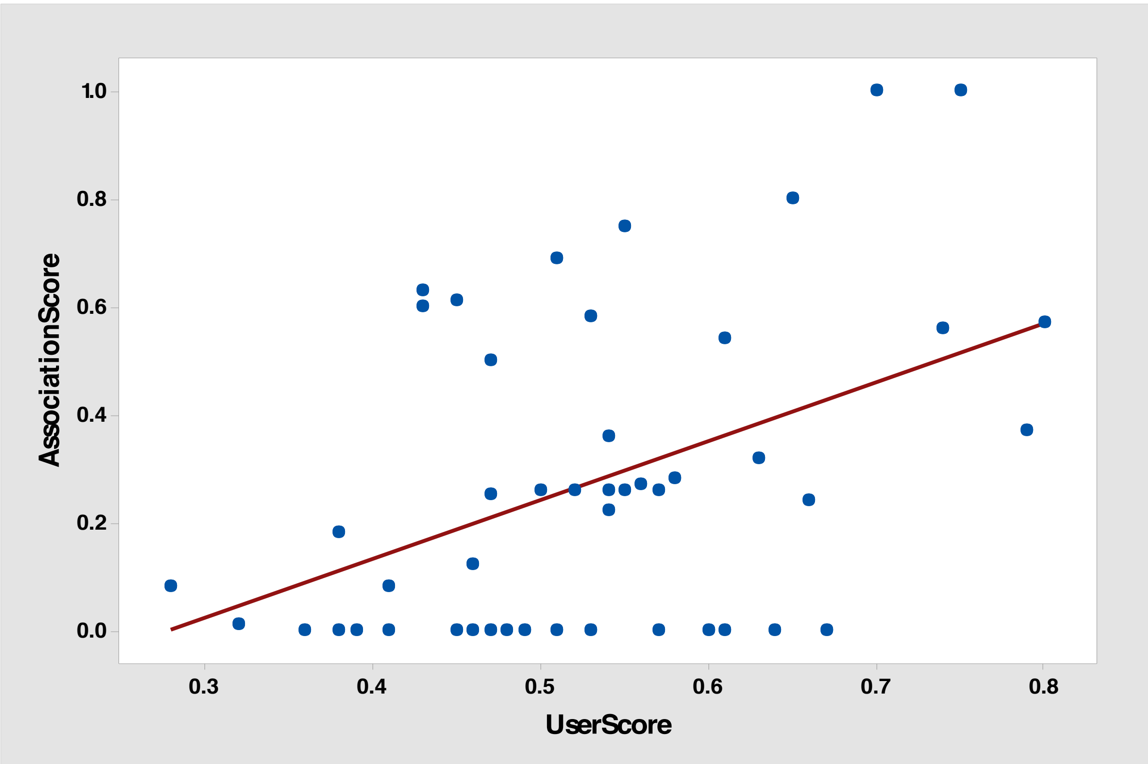


Figure 8.7: User score and association score comparison

8.5 Recommendation quality results

From the previous sections, it is observed that the association scores strongly correlate with the measured values provided by users during the user study experiments. This section examines a detailed analysis of the user opinions obtained during the user study regarding the performance of each recommendation approach.

To evaluate the recommendation quality of each recommendation approach, the average MAS across all queries were calculated, and the results are presented in Figure 8.8. The users evaluated an average of 380 queries during the experiments. In Figure 8.8, the blue, red, and green lines show the quality of CH recommendations achieved by the hybrid, social, and content knowledge representation approaches, respectively.

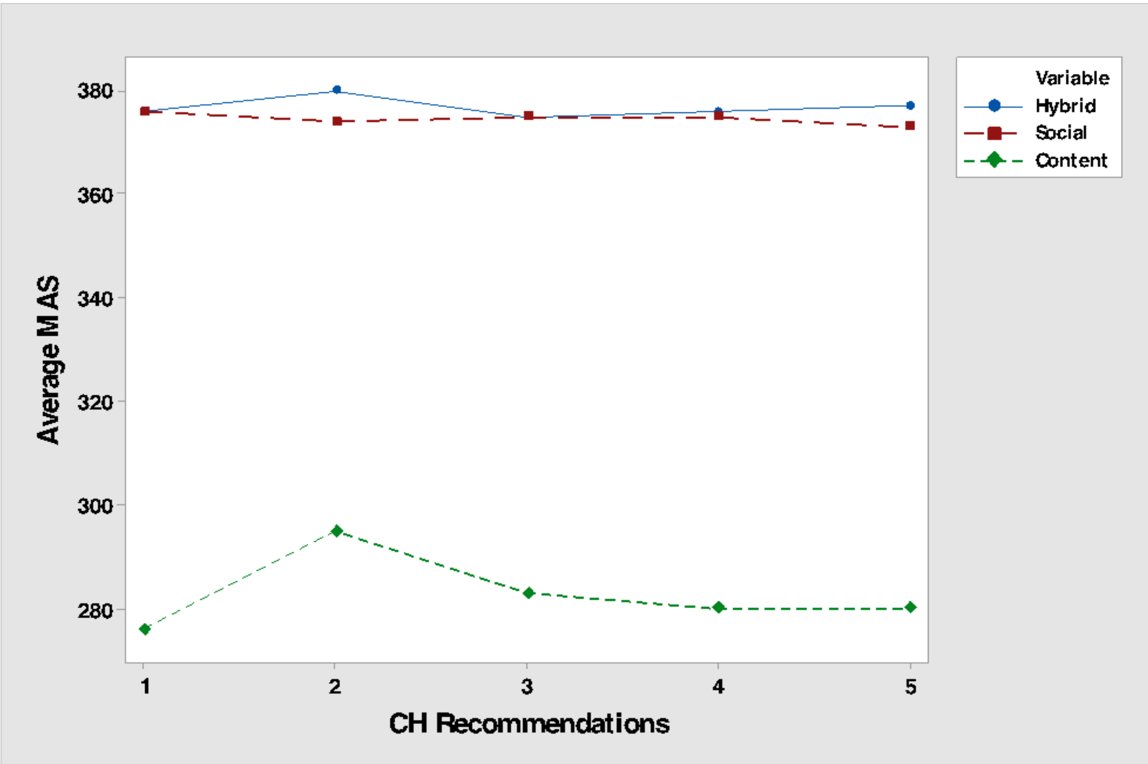


Figure 8.8: Recommendation quality for cultural heritage objects

It can be observed that the hybrid approach also outperformed the content and social knowledge representation approaches in terms of recommendation quality. However, the recommendation quality was not much different between the hybrid and social representation approaches.

8.6 Recommendation novelty results

The recommendation novelty is the ability to recommend unknown CH objects to users. Novelty is measured as the rate at which unknown but exciting CH recommendations are produced for users from known queries. The novelty results of each CH recommendation over five recommendations is presented in Figure 8.9.

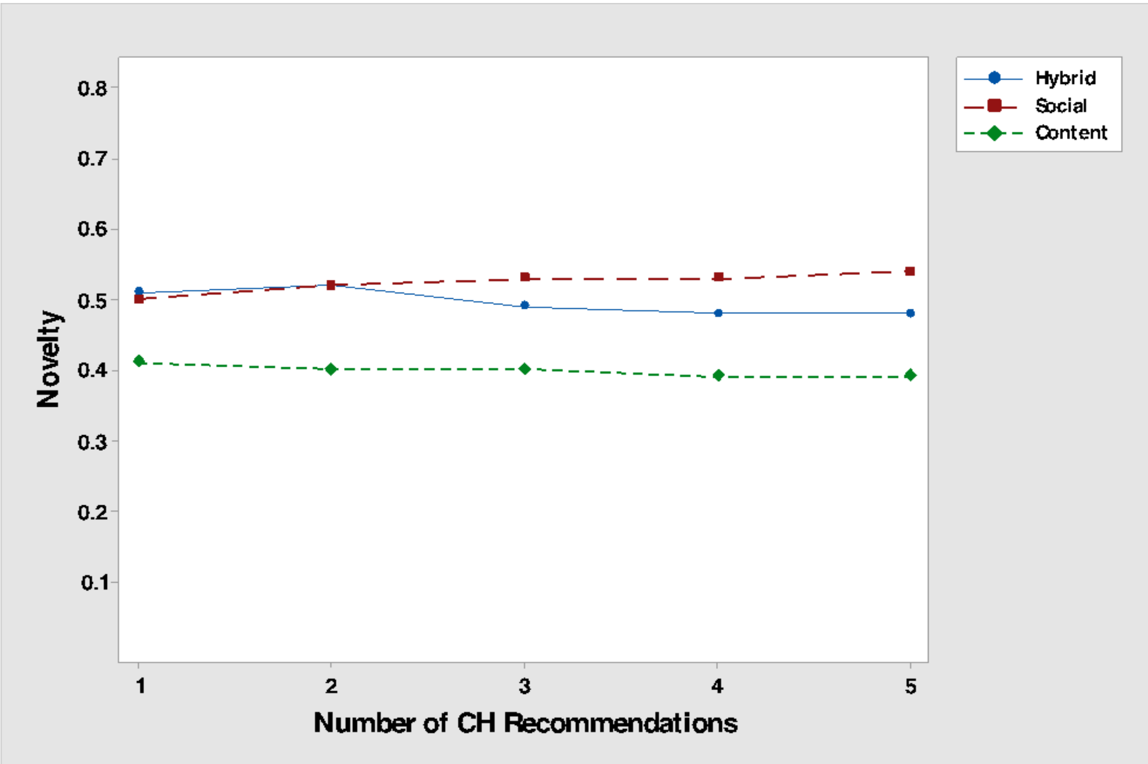


Figure 8.9: Cultural heritage recommendation novelty

Unlike the results obtained for recommendation quality, the social knowledge representation approach provided better recommendation novelty than the hybrid and content approaches. For example, 55% of the five CH recommendations made via the social representation approach were novel. This is because the social knowledge representation does not have a single content knowledge in the representation.

8.7 Summary

A user study was conducted to evaluate the quality and novelty of the CH recommendations made from the content, social, and hybrid approaches. During the study, users were first presented with a questionnaire to complete to provide their demographic information and knowledge background. Later, each participant was presented with queries; each query had at least five CH recommendations made from each approach to be evaluated.

A significant amount of participation was achieved during the four weeks of the evaluation period. Around 150 users participated and evaluated over a thousand queries. The data obtained from the questionnaire show that the participation was spread across users' genders, ages, and CH objects of interest.

The results obtained from the user study show that the hybrid representation produced a better quality of CH recommendations when compared with the content and social knowledge representations in isolation. Even though the social representation did not provide a higher-level recommendation quality than the hybrid representation, it outperformed the hybrid representation in recommending novel CH objects because it produced CH recommendations with only social knowledge, without introducing content knowledge.

Chapter Nine: Conclusions and Future Work

9.1 Introduction

The work presented in this Thesis revisits the research questions raised in the introductory chapter: (i) How can a hybrid approach address the challenges of CH recommendations, such as cold-start, out of context recommendations, and bad recommendations? (ii) How can a VSI help CH users with no domain knowledge to explore large CH collections for new discoveries? Research objectives were set to address these questions (see Section 1.2).

This concluding chapter discusses the summary of the main findings from the research objectives and the potential for future work.

9.2 Findings and contributions

In this section, the emrging findings and contributions from the set research objectives are discussed.

9.2.1 Findings

To discuss the findings and achievements made from this research, it is important to revisit the initial research objectives provided in Section 1.2.

I. Bridge the knowledge gap in the knowledge representations assembled from the available domain knowledge sources

The integration of content and social knowledge representations to bridge the semantic and social knowledge gaps and address the cold-start problem is one objective achieved by this study. As discussed in the previous chapters, it was discovered that most hybrid approaches cluster content and social knowledge to integrate the knowledge representations, but this study's approach provided a concept that allowed the social knowledge representation to learn directly from the content knowledge representation. This established correlations between CH objects' contents and their corresponding social knowledge, user interactions, and social tagging and, thus, bridged the semantic and social knowledge gap in the knowledge representations for CH recommendations. The hybrid approach presented in this work also contributed to the discovery rate of cold-start CH objects, recommending CH objects that lacked user interactions. This work has been submitted for publication in the *IEEE Access Journal*.

II. Develop a dynamic hybrid approach and a visual search interface

The dynamic hybrid approach is an extension of integrated knowledge representations; it dynamically combines content and social knowledge representations to address CH recommendation challenges. In this case, weights were assigned to both content and social representations so that the influence of each knowledge representation varied, depending on the current user search status. This approach also has the ability to learn indirectly from the content knowledge representation for better CH material recommendations.

Also discussed in this thesis, VISE provides a visual summary of information relating to the CH domain to help online users with no domain knowledge to explore large CH collections for new information discovery. User satisfaction evaluations were conducted, and it was revealed that CH users were more satisfied with VISE when compared with the non-VSI. This work has been published in ACM's *Journal of Computing and Cultural Heritage* (Usman & Antonacopoulos, 2019).

III. Build a custom dataset for cultural heritage recommendations

As clearly stated in the literature review, producing quality CH recommendations requires strong knowledge representations rich in the knowledge required to generate the recommendations. However, the available public datasets lack the required content and social knowledge, such as user interactions and social tagging, to generate CH recommendations. Thus, a custom CH dataset that includes all of the required knowledge for CH recommendations needed to be built.

To achieve this objective, the work presented in this thesis built a custom dataset for CH recommendations and a VSI from different domain knowledge sources, Europeana webpages and Facebook. The custom dataset consists of approximately 700,000 CH objects and their corresponding social knowledge, for example, user-interaction and social tagging, as provided by online users. The custom dataset was compared with one of the available public datasets, Europeana; the custom dataset was richer than Europeana's in terms of social knowledge distribution.

9.2.2 Contributions

The major contribution derived from this work is the integration of content and social knowledge representations to bridge the knowledge gap and provide quality CH recommendations. This approach addresses the issues of the cold-start problem, out of context recommendations, and similar objects but bad recommendations. A dynamic hybrid approach, a further dynamic combination of integrated and social knowledge representations assembled from the available domain knowledge sources, is another contribution derived from this work.

Another contribution derived from this work is the design and development of an interface for the VIsual Search and Exploration (VISE) of CH collections to enable users with little or no domain knowledge to quickly discover new CH materials from extensive CH collections. This contribution has been published in the Association for Computing Machinery's (ACM) *Journal Of Computing and Cultural Heritage* (Usman & Antonacopoulos, 2019).

The harvest of a custom CH dataset from the available domain knowledge sources is another contribution derived from this work. The current available open datasets of CH materials, for example, Europeana datasets, do not have all of the data required for recommendation techniques to make CH recommendations.

9.3 Conclusion and Future work

This study focused on addressing the research questions raised in Chapter One (see Section 1.2). The questions were answered by achieving the research objectives. During this study, the challenges of CH recommendations, such as the cold-start problem, out of context recommendations, and bad recommendations, were discovered and addressed.

The major contributions this research included building a knowledge gap bridge between the content and social knowledge representations to improve CH recommendations and developing a VSI to help users with no domain knowledge to search large CH collections for new information discoveries. The limitations of knowledge representations and user profiles assembled from various domain sources were also identified. Some of these challenges were addressed in this thesis; other challenges could be addressed in future work:

- Considering other knowledge sources, such as multimedia, to solve the sparsity problem
- Incorporating context-aware knowledge (e.g. time or location) into a CH RS using deep learning and recurrent neural networks
- Developing a personalised VSI

References

Adeniyi, D., Wei, Z., & Yongquan, Y. (2016). Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Applied Computing and Informatics*, *12*(1), 90–108.

Adomavicius, G., & Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge & Data Engineering*, (6), 734–749.

Adomavicius, G., & Tuzhilin, A. (2011). Context-aware recommender systems. In *Recommender systems handbook* (pp. 217–253). Springer.

Aggarwal, C. C. (2016). *Recommender systems*. Springer.

Ahn, H. J. (2008). A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem. *Information Sciences*, *178*(1), 37–51.

Al-Shamri, M. Y. H. (2016). User profiling approaches for demographic recommender systems. *Knowledge-Based Systems*, *100*, 175–187.

Amato, F., Moscato, V., Picariello, A., Colace, F., Santo, M. D., Schreiber, F. A., & Tanca, L. (2017). Big data meets digital cultural heritage: Design and implementation of SCRABS, a Smart Context-awaRe Browsing Assistant for Cultural EnvironmentS. *J. Comput. Cult. Herit.*, *10*(1), 1–23. https://doi.org/10.1145/3012286

Amato, F., Moscato, V., Picariello, A., & Sperlí, G. (2018). A recommender system for multimedia art collections. In G. De Pietro, L. Gallo, R. J. Howlett, & L. C. Jain (Eds.), *Intelligent interactive multimedia systems and services 2017* (pp. 200–209). Cham: Springer International Publishing.

Amatriain, X., Jaimes, A., Oliver, N., & Pujol, J. M. (2011). Data mining methods for recommender systems. In *Recommender systems handbook* (pp. 39–71). Springer.

Aslanian, E., Radmanesh, M., & Jalili, M. (2016). Hybrid recommender systems based on content feature relationship. *IEEE Transactions on Industrial Informatics*.

Atkinson, S. (2007). Risk reduction through technological control of personal information.

Bandyopadhyay, S., & Maulik, U. (2005). Knowledge discovery and data mining. In *Advanced methods for knowledge discovery from complex data* (pp. 3–42). Springer London.

Bar, A., Rokach, L., Shani, G., Shapira, B., & Schclar, A. (2012). Boosting simple collaborative filtering models using ensemble methods. *arXiv preprint arXiv:1211.2891*.

Bar, Y., Levy, N., & Wolf, L. (2015). *Classification of artistic styles using binarized features derived from a deep neural network*. Cham.

Barkan, O., & Koenigstein, N. (2016). *Item2vec: Neural item embedding for collaborative filtering* [Paper presentation]. 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP).

Barragáns-Martínez, A. B., Costa-Montenegro, E., Burguillo, J. C., Rey-López, M., Mikic-Fonte, F. A., & Peleteiro, A. (2010). A hybrid content-based and item-based collaborative filtering approach to recommend TV programs enhanced with singular value decomposition. *Information Sciences*, *180*(22), 4290–4311.

Bench-Capon, T. J. (2014). *Knowledge representation: An approach to artificial intelligence* (Vol. 32). Elsevier.

Bennett, J., & Lanning, S. (2007). *The Netflix prize* [Paper presentation]. Proceedings of KDD Cup and Workshop.

Billsus, D., & Pazzani, M. J. (2000). User modeling for adaptive news access. *User Modeling and User-Adapted Interaction*, *10*(2–3), 147–180.

Bird, S. (2006). *NLTK: The natural language toolkit* [Paper presentatoin]. Proceedings of the COLING/ACL on Interactive Presentation Sessions.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-Based Systems*, *46*, 109–132.

Bouza, A., Reif, G., Bernstein, A., & Gall, H. (2008). *Semtree: Ontology-based decision tree algorithm for recommender systems* [Paper presentation]. Proceedings of the 2007 International Conference on Posters and Demonstrations-Volume 401.

Brand, M. (2003). *Fast online SVD revisions for lightweight recommender systems* [Paper presentation]. SDM.

Breese, J. S., Heckerman, D., & Kadie, C. (1998). *Empirical analysis of predictive algorithms for collaborative filtering* [Paper presentation]. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence.

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

Burke, R. (2007a). Hybrid web recommender systems. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 377–408). Springer Berlin Heidelberg.

Burke, R. (2007b). Hybrid web recommender systems. In *The adaptive web* (pp. 377–408). Springer.

Cacheda, F., #237, Carneiro, C., Fern, D., #225, ndez, & Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Trans. Web*, *5*(1), 1–33. https://doi.org/10.1145/1921591.1921593

Cantador, I., Bellogín, A., Cortés-Cediel, M. E., & Gil, O. (2017). *Personalized recommendations in e-participation: Offline experiments for the 'Decide Madrid' platform* [Paper presentation]. Proceedings of the International Workshop on Recommender Systems for Citizens.

Castella, Q., & Sutton, C. (2014). *Word storms: Multiples of word clouds for visual comparison of documents* [Paper presentation]. Proceedings of the 23rd International Conference on World Wide Web.

Charles, V., & Isaac, A. (2015). *Enhancing the Europeana Data Model (EDM)* [White paper].

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Shah, H. (2016). *Wide & deep learning for recommender systems* [Paper presentation]. Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA.

Ciocca, G., Olivo, P., & Schettini, R. (2012). Browsing museum image collections on a multi-touch table. *Information Systems*, *37*(2), 169–182.

Clough, P., Hill, T., Paramita, M. L., & Goodale, P. (2017). *Europeana: What users search for and why* [Paper presentation]. International Conference on Theory and Practice of Digital Libraries.

Common, M. F. (2018). Facebook and Cambridge Analytica: Let this be the high-water mark for impunity. *LSE Business Review*.

Contratres, F. G., Alves-Souza, S. N., Filgueiras, L. V. L., & DeSouza, L. S. (2018). *Sentiment analysis of social network data for cold-start relief in recommender systems* [Paper presentation]. World Conference on Information Systems and Technologies.

Covington, P., Adams, J., & Sargin, E. (2016). *Deep neural networks for YouTube recommendations* [Paper presentation]. Proceedings of the 10th ACM Conference on Recommender Systems.

Cremonesi, P., Garzotto, F., Negro, S., Papadopoulos, A., & Turrin, R. (2011). Comparative evaluation of recommender system quality. In *CHI'11 extended abstracts on human factors in computing systems* (pp. 1927–1932).

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press.

Cunningham, H., Maynard, D., & Bontcheva, K. (2011). *Text processing with gate*. Gateway Press CA.

Cunningham, P., Bergmann, R., Schmitt, S., Traphoner, R., Breen, S., & Smyth, B. (2001). Websell: Intelligent sales assistants for the world wide web. *Artificial Intelligence*, *15*(1), 28–32.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The Mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, *50*(1), 1–18.

Dumas, B., Moerman, B., Trullemans, S., & Signer, B. (2014). *ArtVis: Combining advanced visualisation and tangible interaction for the exploration, analysis and browsing of digital artwork collections* [Paper presentation]. Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces.

Felfernig, A., Friedrich, G., Jannach, D., & Zanker, M. (2011). Developing constraint-based recommenders. In *Recommender systems handbook* (pp. 187–215). Springer.

Felfernig, A., Isak, K., Szabo, K., & Zachar, P. (2007). *The VITA financial services sales support environment* [Paper presentation]. Proceedings of the National Conference on Artificial Intelligence.

Felt, A., & Evans, D. (2008). *Privacy protection for social networking platforms*.

Friedman, N., Geiger, D., & Goldszmidt, M. (1997). Bayesian network classifiers. *Machine learning*, *29*(2–3), 131–163.

Gäde, M. (2014). *Country and language level differences in multilingual digital libraries*.

Gallagher, E. (2013). Measuring semantic ambiguity. In.

Ge, M., & Persia, F. (2017). A survey of multimedia recommender systems: Challenges and opportunities. *International Journal of Semantic Computing*, *11*(03), 411–428.

Goerz, G., & Scholz, M. (2010). Adaptation of nlp techniques to cultural heritage research and documentation. *Journal of Computing and Information Technology*, *18*(4), 317–324.

Goldberg, D., Nichols, D., Oki, B. M., & Terry, D. (1992). Using collaborative filtering to weave an information tapestry. *Communications of the ACM*, *35*(12), 61–70.

Gomez-Uribe, C. A., & Hunt, N. (2016). The Netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, *6*(4), 13.

Hampson, C., Agosti, M., Orio, N., Bailey, E., Lawless, S., Conlan, O., & Wade, V. (2012). The CULTURA Project: Supporting next generation interaction with digital cultural heritage collections. In M. Ioannides, D. Fritsch, J. Leissner, R. Davies, F. Remondino, & R. Caffo (Eds.), *Progress in cultural heritage preservation: 4th international conference, EuroMed 2012, Limassol, Cyprus, October 29 – November 3, 2012. Proceedings* (pp. 668–675). Springer Berlin Heidelberg.

Hegde, A., & Shetty, S. K. (2015). Collaborative filtering recommender system. *International Journal of Emerging Trends in Science and Technology*, *2*(07).

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 5–53.

Hofmann, T. (2004). Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, *22*(1), 89–115.

Holzer, S. J. J., Kar, A., Trevor, A. J. B., Kalogiros, P., Spanos, I., & Rusu, R. B. (2018). Visual search using multi-view interactive digital media representations. In: Google Patents.

Hong, M., Jung, J. J., Piccialli, F., & Chianese, A. (2017). Social recommendation service for cultural heritage. *Personal and Ubiquitous Computing*, *21*(2), 191–201.

Hyvönen, E. (2009). Semantic portals for cultural heritage. In *Handbook on ontologies* (pp. 757–778). Springer.

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2018). Wikidata through the Eyes of DBpedia. *Semantic Web* (Preprint), 1–11.

Jain, S., Grover, A., Thakur, P. S., & Choudhary, S. K. (2015). *Trends, problems and solutions of recommender system* [Paper presentation]. International Conference on Computing, Communication & Automation.

Jiang, Z., Liu, H., Fu, B., Wu, Z., & Zhang, T. (2018). *Recommendation in heterogeneous information networks based on generalized random walk model and Bayesian personalized ranking* [Paper presentation]. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining.

Jolliffe, I. (2011). *Principal component analysis*. Springer.

Knijnenburg, B. P., & Willemsen, M. C. (2015). Evaluating recommender systems with user experiments. In *Recommender systems handbook* (pp. 309–352). Springer.

Laramée, F. D. (2019). How to extract good knowledge from bad data: An experiment with eighteenth century French texts. *Digital Studies/Le champ numérique*, *9*(1).

Lehmann, J., Isele, R., Jakob, M., Jentzsch, A., Kontokostas, D., Mendes, P. N., Auer, S. (2015). DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, *6*(2), 167–195.

Li, D., Chen, C., Lv, Q., Shang, L., Zhao, Y., Lu, T., & Gu, N. (2016). An algorithm for efficient privacy-preserving item-based collaborative filtering. *Future Generation Computer Systems*, *55*, 311–320.

Li, Q., & Kim, B. M. (2003). *Clustering approach for hybrid recommender system* [Paper presentation]. Proceedings IEEE/WIC International Conference on Web Intelligence (WI 2003).

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014a). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4), 2065–2073.

Lika, B., Kolomvatsos, K., & Hadjiefthymiades, S. (2014b). Facing the cold start problem in recommender systems. *Expert Systems with Applications*, *41*(4, Part 2), 2065–2073. https://doi.org/10.1016/j.eswa.2013.09.005

Lin, W., Alvarez, S. A., & Ruiz, C. (2000). Collaborative recommendation via adaptive association rule mining. *Data Mining and Knowledge Discovery*, *6*, 83–105.

Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook* (pp. 73–105). Springer.

Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability Interface*, *8*(2), 3–6.

Ma, T., Zhou, J., Tang, M., Tian, Y., Al-Dhelaan, A., Al-Rodhaan, M., & Lee, S. (2015). Social network and tag sources based augmenting collaborative recommender system. *IEICE Transactions on Information and Systems*, *98*(4), 902–910.

Mahmood, T., & Ricci, F. (2007). *Towards learning user-adaptive state models in a conversational recommender system* [Paper presentation]. LWA.

Manguinhas, H. (2016). Europeana semantic enrichment framework. Documentation, Europeana. In.

Matz, S. C., Menges, J. I., Stillwell, D. J., & Schwartz, H. A. (2019). Predicting individual-level income from Facebook profiles. *PloS One*, *14*(3), e0214369.

Melville, P., Mooney, R. J., & Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. *Aaai/iaai*, *23*, 187–192.

Mensink, T., & Van Gemert, J. (2014). *The Rijksmuseum challenge: Museum-centered visual recognition* [Paper presentation]. Proceedings of International Conference on Multimedia Retrieval.

Merigo, J. M., & Casanovas, M. (2011). A new Minkowski distance based on induced aggregation operators. *International Journal of Computational Intelligence Systems*, *4*(2), 123–133.

Mirzadeh, N., Ricci, F., & Bansal, M. (2005). *Feature selection methods for conversational recommender systems* [Paper presentation]. 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service.

Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142–151.

Montaner, M., López, B., & De La Rosa, J. L. (2003). A taxonomy of recommender agents on the internet. *Artificial Intelligence Review*, *19*(4), 285–330.

Moore, P. G. (1957). The two-sample t-test based on range. *Biometrika*, *44*(3/4), 482–489.

Müller, T. F., & Winters, J. (2018). Compression in cultural evolution: Homogeneity and structure in the emergence and evolution of a large-scale online collaborative art project. *PloS One*, *13*(9), e0202019.

Nguyen, T. T., Hui, P.-M., Harper, F. M., Terveen, L., & Konstan, J. A. (2014). *Exploring the filter bubble: The effect of using recommender systems on content diversity* [Paper presentation]. Proceedings of the 23rd International Conference on World Wide Web.

Nørretranders, T. (1991). *The user illusion: Cutting consciousness down to size*. Viking.

Palopoli, L., Rosaci, D., & Sarnè, G. M. (2013). Introducing specialization in e-commerce recommender systems. *Concurrent Engineering*, 1063293X13493915.

Panniello, U., Tuzhilin, A., Gorgoglione, M., Palmisano, C., & Pedone, A. (2009). *Experimental comparison of pre- vs. post-filtering approaches in context-aware recommender systems* [Paper presentation]. Proceedings of the Third ACM Conference on Recommender Systems.

Pappas, N., & Popescu-Belis, A. (2013). *Sentiment analysis of user comments for one-class collaborative filtering over TED talks* [Paper presentation]. Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval.

Paterek, A. (2007). *Improving regularized singular value decomposition for collaborative filtering* [Paper presentation]. Proceedings of KDD Cup and Workshop.

Pavlidis, G. (2018). Recommender systems, cultural heritage applications, and the way forward. *Journal of Cultural Heritage*.

Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325–341). Springer.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Pereira, N., & Varma, S. (2016). Survey on content based recommendation system. *Int. J. Comput. Sci. Inf. Technol*, *7*(1), 281–284.

Perez, J. C. (2007). Facebook's Beacon more intrusive than previously thought. *PC World, 30*.

Petras, V., Hill, T., Stiller, J., & Gäde, M. (2017). Europeana – a search engine for digitised cultural Heritage Material. *Datenbank-Spektrum*, *17*(1), 41–46. https://doi.org/10.1007/s13222-016-0238-1

Rendle, S., & Schmidt-Thieme, L. (2008). *Online-updating regularized kernel matrix factorization models for large-scale recommender systems* [Paper presentation]. Proceedings of the 2008 ACM Conference on Recommender Systems.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). *GroupLens: An open architecture for collaborative filtering of netnews* [Paper presentation]. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work.

Ricci, F., Rokach, L., & Shapira, B. (2011). Introduction to recommender systems handbook. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 1–35). Springer.

Ristoski, P., & Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Web Semantics: Science, Services and Agents on the World Wide Web*, *36*, 1–22.

Saat, N. I. Y., Noah, S. A. M., & Mohd, M. (2018). Towards serendipity for content-based recommender systems. *International Journal on Advanced Science, Engineering and Information Technology*, *8*(4–2), 1762–1769.

Salakhutdinov, R., Mnih, A., & Hinton, G. (2007). *Restricted Boltzmann machines for collaborative filtering* [Paper presentation]. Proceedings of the 24th International Conference on Machine Learning.

Sansonetti, G., Gasparetti, F., Micarelli, A., Cena, F., & Gena, C. (2019). Enhancing cultural recommendations through social and linked open data. *User Modeling and User-Adapted Interaction*, 1–39.

Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2002). *Incremental singular value decomposition algorithms for highly scalable recommender systems* [Paper presentation]. Fifth International Conference on Computer and Information Science.

Sarwar, B. M., Karypis, G., Konstan, J. A., & Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. *Www*, *1*, 285–295.

Schafer, J. B., Frankowski, D., Herlocker, J., & Sen, S. (2007). Collaborative filtering recommender systems. In *The adaptive web* (pp. 291–324). Springer.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2001). E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, *5*(1–2), 115–153.

Schafer, J. B., Konstan, J. A., & Riedl, J. (2002). *Meta-recommendation systems: User-controlled integration of diverse recommendations* [Paper presentation]. Proceedings of the Eleventh International Conference on Information and Knowledge Management.

Schein, A. I., Popescul, A., Ungar, L. H., & Pennock, D. M. (2002). *Methods and metrics for cold-start recommendations* [Paper presentation]. Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.

Semitsu, J. P. (2011). From Facebook to mug shot: How the dearth of social networking privacy rights revolutionized online government surveillance. *Pace L. Rev.*, *31*, 291.

Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), *Recommender systems handbook* (pp. 257–297). Springer.

Shardanand, U., & Maes, P. (1995). *Social information filtering: Algorithms for automating "word of mouth"* [Paper presentation]. The Chi.

Sidorov, G., Gelbukh, A., Gómez-Adorno, H., & Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, *18*(3), 491–504.

Smirnov, A. V., Kashevnik, A. M., & Ponomarev, A. (2017). Context-based infomobility system for cultural heritage recommendation: Tourist Assistant—TAIS. *Personal and Ubiquitous Computing*, *21*(2), 297–311.

Smyth, B. (2007). Case-based recommendation. In P. Brusilovsky, A. Kobsa, & W. Nejdl (Eds.), *The adaptive web: Methods and strategies of web personalization* (pp. 342–376). Springer Berlin Heidelberg.

Sun, H., Peng, Y., Chen, J., Liu, C., & Sun, Y. (2011). A new similarity measure based on adjusted Euclidean distance for memory-based collaborative filtering. *JSW*, *6*(6), 993–1000.

Takács, G., Pilászy, I., Németh, B., & Tikk, D. (2008). *Matrix factorization and neighbor based algorithms for the Netflix prize problem* [Paper presentation]. Proceedings of the 2008 ACM Conference on Recommender Systems.

Thompson, C. A., Goker, M. H., & Langley, P. (2004). A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research*, *21*, 393–428.

Thorat, P. B., Goudar, R., & Barve, S. (2015). Survey on collaborative filtering, content-based filtering and hybrid recommendation system. *International Journal of Computer Applications*, *110*(4), 31–36.

Trewin, S. (2000). Knowledge-based recommender systems. *Encyclopedia of Library and Information Science*, *69*(Supplement 32), 180.

Tufekci, Z. (2008). Can you see me now? Audience and disclosure regulation in online social network sites. *Bulletin of Science, Technology & Society*, *28*(1), 20–36.

Unbehauen, J., Hellmann, S., Auer, S., & Stadler, C. (2012). Knowledge extraction from structured sources. In *Search computing* (pp. 34–52). Springer.

Usman, M. A., & Antonacopoulos, A. (2019). VISE: An interface for visual search and exploration of museum collections. *Journal on Computing and Cultural Heritage (JOCCH)*, *12*(4), 25.

Vesanto, A., Nivala, A., Salakoski, T., Salmi, H., & Ginter, F. (2017). *A system for identifying and exploring text repetition in large historical document corpora* [Paper presentation]. Proceedings of the 21st Nordic Conference on Computational Linguistics, NoDaLiDa, 22–24 May 2017, Gothenburg, Sweden.

Vosinakis, S., & Tsakonas, Y. (2016). Visitor experience in Google Art Project and in second life-based virtual museums: A comparative study. *Mediterranean Archaeology & Archaeometry*, *16*(5).

Walsh, D., Hall, M. M., Clough, P., & Foster, J. (2018). Characterising online museum users: A study of the National Museums Liverpool museum website. *International Journal on Digital Libraries*, 1–13.

Wang, Y., Aroyo, L. M., Stash, N., & Rutledge, L. (2007). *Interactive user modeling for personalized access to museum collections: The Rijksmuseum case study*. Heidelberg.

Wang, Y., Stash, N., Aroyo, L., Gorgels, P., Rutledge, L., & Schreiber, G. (2008). Recommendations based on semantically enriched museum collections. *Web Semantics: Science, Services and Agents on the World Wide Web*, *6*(4), 283–290.

Wei, J., He, J., Chen, K., Zhou, Y., & Tang, Z. (2017). Collaborative filtering and deep learning based recommendation system for cold start items. *Expert Systems with Applications*, *69*, 29–39. https://doi.org/10.1016/j.eswa.2016.09.040

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.

Wilson, M. L., Kules, B., & Shneiderman, B. (2010). From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends® in Web Science*, *2*(1), 1–97.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Wu, M.-C., Lo, Y.-F., & Hsu, S.-H. (2008). A fuzzy CBR technique for generating product ideas. *Expert Systems with Applications*, *34*(1), 530–540.

Yang, Z., Li, D., Lin, R., Tang, Y., Li, W., & Liu, H. (2018). *An academic social network friend recommendation algorithm based on decision tree* [Paper presentation]. 2018 IEEE

SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI).

Zanker, M., & Jessenitschnig, M. (2009). *Collaborative feature-combination recommender exploiting explicit and implicit user feedback* [Paper presentation]. Commerce and Enterprise Computing, 2009. CEC'09. IEEE Conference on.

Zhang, M., Tang, J., Zhang, X., & Xue, X. (2014). *Addressing cold start in recommender systems: A semi-supervised co-training algorithm* [Paper presentation]. The Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval.

Zhang, Z.-K., Liu, C., Zhang, Y.-C., & Zhou, T. (2010). Solving the cold-start problem in recommender systems with social tags. *EPL (Europhysics Letters)*, *92*(2), 28002.

Zhao, Z.-D., & Shang, M.-S. (2010). *User-based collaborative-filtering recommendation algorithms on hadoop* [Paper presentation]. 2010 Third International Conference on Knowledge Discovery and Data Mining.

Zheng, E., Kondo, G. Y., Zilora, S., & Yu, Q. (2018). Tag-aware dynamic music recommendation. *Expert Systems with Applications*, *106*, 244–251.

Zheng, V. W., Zheng, Y., Xie, X., & Yang, Q. (2010). *Collaborative location and activity recommendations with GPS history data* [Paper presentation]. The Proceedings of the 19th International Conference on World Wide Web.

Zuckerberg, M. (2010). Facebook F8 Developer Conference. In: April.

Zurada, J. M. (1992). Introduction to artificial neural systems. West, St. *Paul, Minn*.

APENDICES

VISE: An interface for VIsual Search and Exploration of museum collections

MAHMUD AHMED USMAN and APOSTOLOS ANTONACOPOULOS, Pattern Recognition and Image Analysis (PRImA) Research Lab, University of Salford, Greater Manchester, United Kingdom

This paper presents VISE, an interface that enables VIsual Search and Exploration across collections of approximately 836,000 museum objects extracted from the websites of the National Museums Scotland and the Rijksmuseum in the Netherlands. VISE provides an interactive visual summary of information relating to the museum to address the online users with no domain knowledge challenges of exploring large museum collection. User satisfaction evaluation was conducted to measure the user satisfaction level for using VISE as the interface for search and exploration of information from large museum collection when compared to non-visual search interface. The evaluation of the visual interface revealed that users are more satisfied and attracted to explore museum objects via VISE than via the system with no visual search interface. Users with little or no domain knowledge find it easier to explore collections and find objects of interest while using VISE in contrast to the system with no visual interface.

CCS Concepts: • **Information System**; • **Information Retrieval**; • **Users and interactive retrieval**; • **Search interfaces**;

Additional Key Words and Phrases: Visual Search, User interface, knowledge representation, Cultural heritage

ACM Reference Format:
Mahmud Ahmed Usman and Apostolos Antonacopoulos. 2019. VISE: An interface for VIsual Search and Exploration of museum collections. *ACM J. Comput. Cult. Herit.* 12, 3, Article 13 (June 2019), 9 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Exploration of large collections of Digital Cultural Heritage (DCH) objects contained in museum collections presents significant challenges. Information about hundreds of millions of objects from museum collections is available online leading to information overload. Users, especially those with no domain knowledge, find it difficult to explore collections and find objects of interest [Villa et al. 2013].

Locating DCH objects of interest and finding information about them typically involves keyword search where users enter search terms and are presented with ranked lists of relevant objects. Keyword search is efficient if users have reasonable domain knowledge [Clough et al. 2017], but many online museum visitors generally access the internet nowadays either for research or leisure [Walsh et al. 2018] without specific domain knowledge. VISE, an interface for VIsual Search and Exploration of museum objects, has been developed to address the challenges that the online users face when exploring museum objects for new discoveries.

VISE provides a diverse searching strategy, which is especially useful when users are unaware of the full details of their tasks or do not have domain knowledge. This allows users to initiate their search by selecting terms from an interactive visual interface. In contrast to a keyword search interface, VISE suggests terms that are specific to museum digital objects and it provides insights based on semantic relationships between museum objects.

Authors' address: Mahmud Ahmed Usman, m.a.usman@edu.salford.ac.uk; Apostolos Antonacopoulos, a.antonacopoulos@primaresearch.orgPattern Recognition and Image Analysis (PRImA) Research Lab, University of Salford, Greater Manchester, United Kingdom.

XXXX-XXXX/2019/6-ART13 $15.00
https://doi.org/10.1145/1122445.1122456

The provision of a visual search interface that encourages exploration of about 836,000 museum objects and the extraction of museum objects'important feature attributes from the web pages of the National Museums Scotland and the Rijksmuseum in The Netherlands, are the major contributions of this presented work.

The evaluation of user satisfaction of VISE is presented and discussed. The goal is to determine whether the visual search interface provides higher user satisfaction levels compared to a search interface without visualisation. VISE combines an interactive visual summary of museum collections with information visualisation via interactive surface as described by Keim [Keim 2002].

2 RELATED WORK

Visual perception and its capabilities in providing an interactive visual summary of a whole domain are the key factors in information visualisation. In fact, a large proportion of human senses is occupied by visual perception [Nørretranders 1991]. Therefore, it is important to consider human visual perception when building a visual search interface for the exploration of large datasets of objects from museum collections. Furthermore, a visual search system as an alternative to keyword search, demands an interface that will encourage search by exploration as presented by Wilson et al [Wilson et al. 2010].

Even though systems like CULTURA [Hampson et al. 2012], SCRABS [Amato et al. 2017], and Europeana [Petras et al. 2017] made DCH resources available to experts and to the wider public to explore, users still struggle to make new discoveries as a result of the absence of recommendations and/or of an interactive visual summary of the digital collections on these platforms [Amato et al. 2018].

Interactive visual exploratory search interfaces in the context of museum collections revolved mainly around a summary of a collection of museum objects and allowing users to initiate their search from the interface. Similar interface for browsing museum image collections on multi-touch display is presented by [Ciocca et al. 2012]. Google Art represents museum image collections in high resolution from highly recognised Museums around the globe. ArtVis [Dumas et al. 2014] is another visual interface that combines visualisation and analysis of artwork collections. But all those visual interfaces do not provide users with a dynamic approach that could initiate exploratory search from the interface. Wang et al [Wang et al. 2008] describe a semantically enriched museum collections recommender system. However, it does not provide an interactive dynamic visual interface that allows users to initiate their search.

In terms of representation, the information in VISE is structured in a similar way to the work of McCallum and Nigam [McCallum et al. 2000].

3 DESIGN AND IMPLEMENTATION OF VISE

The creation of VISE involved two main stages: (a) Knowledge Extraction and Representation, and (b) Creation of the Visual Search Interface.

3.1 Knowledge Extraction and Representation

It is important to note that before building VISE, the knowledge and information used (text and images) used were extracted from unstructured and noisy sources. As a large portion of information is encoded in natural language, Goerz et al [Goerz and Scholz 2010] described the need to extract knowledge from unstructured sources. Tools and techniques were outlined to extract semantically meaningful insights from unstructured data.

The dataset used to demonstrate this research represents museum collections with information about objects harvested from about 300,000 web pages from the websites of the National Museums Scotland (NMS) and the Rijksmuseum in the Netherlands. Each web page presents knowledge on two or more museum objects, resulting in a total of about 836,000 museum objects in the dataset.

Generally, web pages are cluttered with less informative materials that are not related to the main content, making it very difficult to locate important information. According to Falk and Dierking [Falk and Dierking 2016], museum users are mainly interested in Age, Author, Origin or Geographical location, and Image of the museum object while exploring collections. These key feature attributes of the museum object are located within web page content collected from the NMS and the Rijksmuseum.

From the web pages collected, important feature attributes and relevant information were extracted for building VISE. The attributes extracted from those pages and their representation are shown in Figure 1. This representation provided a convenient application for indexing. An index optimises the performance and consequently the speed of an Information Retrieval (IR) system in finding relevant information. The visual interface provided for the exploration of museum collection was generated from these feature attributes.

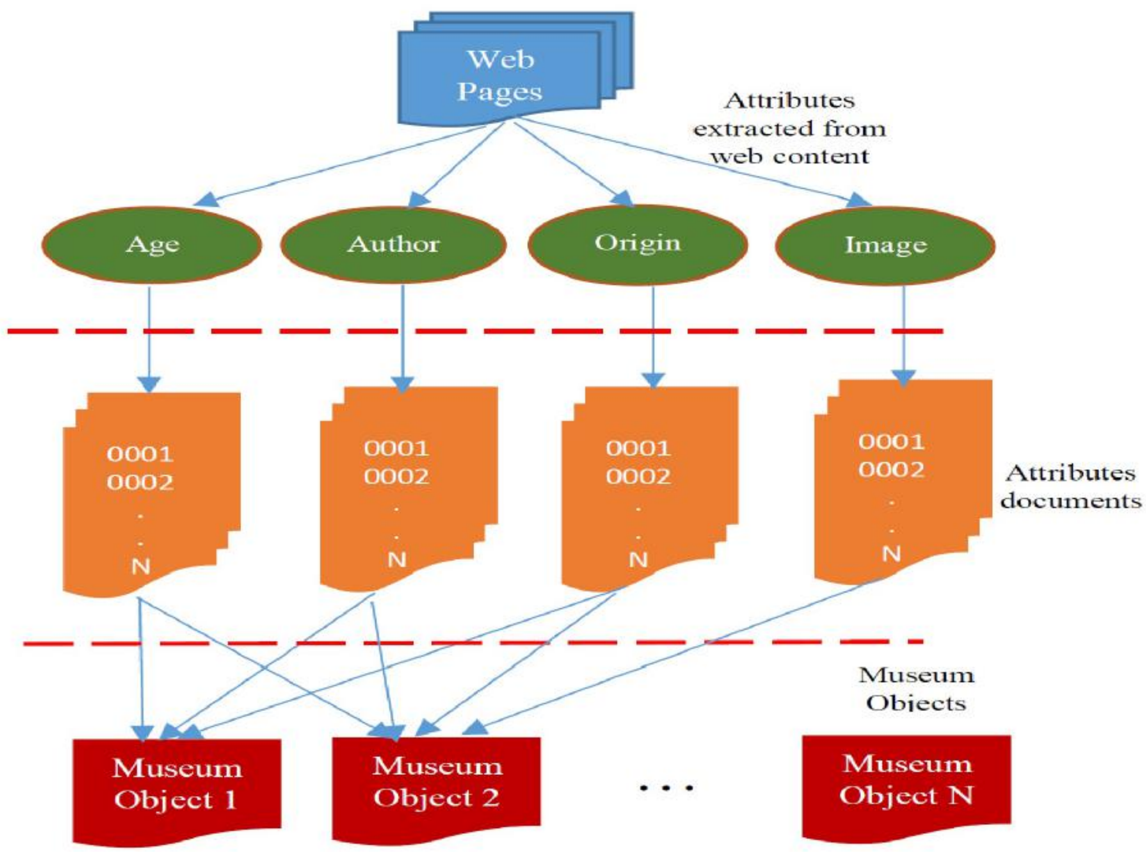


Fig. 1. Museum object attributes extraction and representation

3.2 Visual search Interface

Main objective of this research was to provide an alternative to keyword search that will encourage users to search museum collections by exploration. To achieve that, an interactive visual summary was created which presents the whole museum collections in the form of a Tag Cloud [Castella and Sutton 2014].

The Tag Cloud is one of the visualisation techniques that provides an interactive image display comprising information from the designated set of collection documents. This allows the important terms from the collections to be presented in visual schema that encourages search by exploration. The design and implementation of VISE is composed of three pre-processing stages as shown in Figure 2.

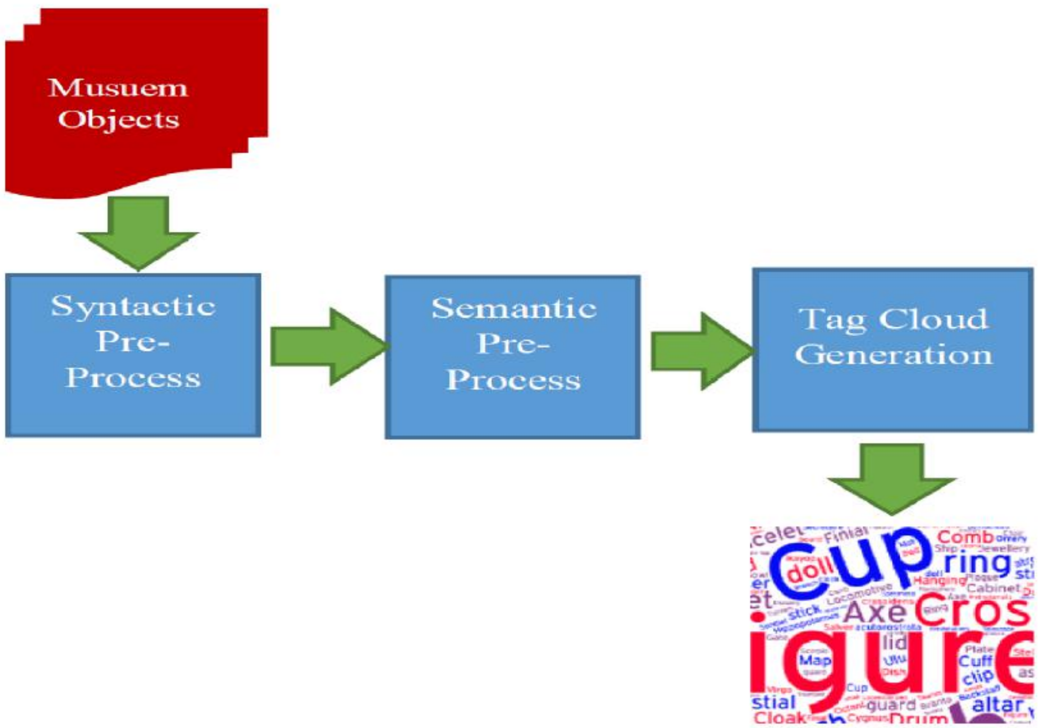


Fig. 2. VISE design processing stage

Museum objects from web pages collected need to be pre-processed before generating the tag cloud:

1. Syntactic Pre-Processing includes tokenisation and stop-word removal. It is important to note that stemming is omitted at this pre-processing stage because the stemming algorithms generate stemmed forms not included in most electronic dictionaries, and this can introduce a setback during semantic pre-processing.

2. Semantic Pre-Processing, using the CIDOC Conceptual Reference Model (CRM) ontology refined by DBpedia to group the terms that are similar in meaning and nature by assigning a single term called *root* to represent them. CIDOC-CRM provides definitions and a formal structure for depicting the implicit and explicit ideas as well as the connections used in documentation of DCH. The root provides a short and broad description, and records semantic relations between those museum objects.

2. Tag Cloud Generation uses the root processed from semantic pre-processing stage to generate a visual summary of the whole museum collections as tag cloud. We use *term frequency-inverse document frequency* (tf-idf) to determine which term will be presented on the tag cloud in order to avoid bias.

After the extraction of the terms, we built the *VISE User Interface* providing two options for searching for museum objects: keyword search option and visual search option. The users can initiate their search by selecting terms from the visualisation as shown in figure 3.

Fig. 3. VISE user interface

4 USER SATISFACTION EVALUATION

While, in the IR community the quality of algorithms that match usersâĂŹ queries and the quality of the indexing methods are paramount, the main aim of this research was to provide an interactive visual search interface that will encourage the exploration of museum collections. Thus, the evaluation presented in this paper is that of user satisfaction of the proposed visual search interface compared to a system with no visualisation.

4.1 Participants

Our user group include (50) participants aged between 27 and 38 years. Twenty-four (24) were female and twenty-six (26) were male. Out of the participants, 38% of them had domain knowledge while the remaining 62% had little or none.

The data collection was done through a USE questionnaire [Lund 2001]. It should be noted that the evaluation was not carried out in a museum but rather in a laboratory as it was out of our control to install VISE in real museum setting. In fact, the target application scenario is online search of the objects, therefore the location of the users should not be factor in the experiment. The evaluation was carried out in six laboratory sessions within a three-month period. In the first two sessions, the participants were experts (with domain knowledge background) while the remaining sessions were attended by non-experts.

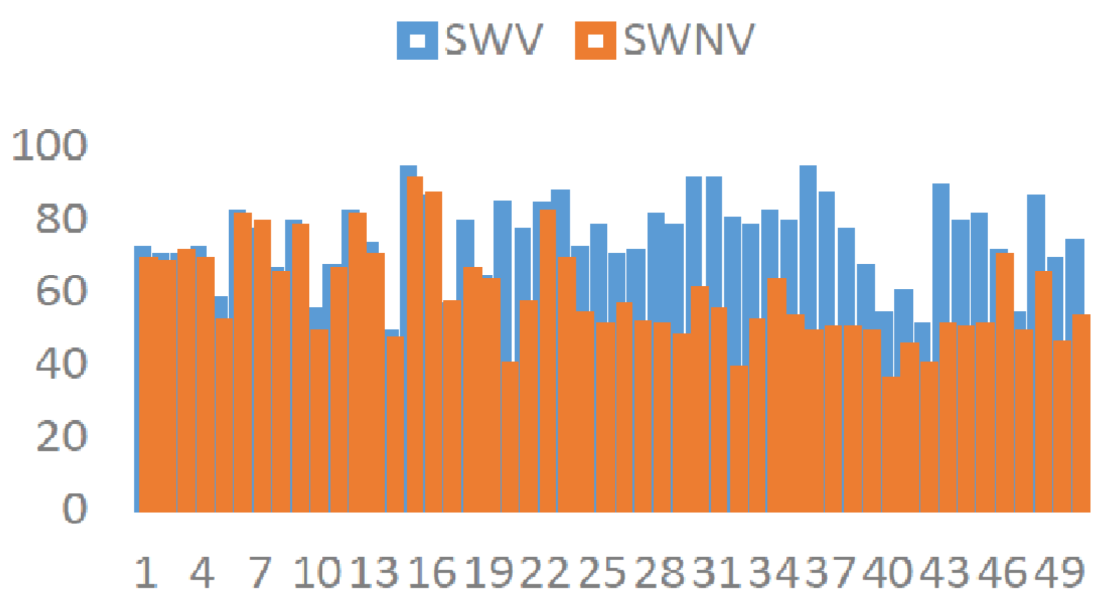


Fig. 4. Participants satisfaction level in percentage

4.2 Procedure

Two systems, VISE (denoted as System With Visualisation (SWV)) and a System With No Visual interface (denoted as SWNV), were provided to the participants to search for museum objects. Each participant was instructed to explore the museumsâĂŹ collections dataset extracted from NMS and Rijksmuseum for a period of thirty minutes. No specific task was given to the participants in order not to influence their overall satisfaction. After

Estimation for Difference

Difference	95% CI for Difference
15.67	(10.77, 20.57)

Test

Null hypothesis	$H_0: \mu_1 - \mu_2 = 0$
Alternative hypothesis	$H_1: \mu_1 - \mu_2 \neq 0$

T-Value	DF	P-Value
6.35	96	0.000

Fig. 5. 2-sample t-Test result

the experiment, two different questionnaires (see below) were handed to the participants to express their level of satisfaction about the two systems in order to test the following hypotheses:

H0: *The mean user satisfaction level between SWV and SWNV does not depend on visual search interface.*

H1: *Such dependency does exist.*

The first questionnaire was completed by the participants after using the VISE (SWV) while the second questionnaire was completed after using the system without visual interface (SWNV). In the questionnaires, participants could express their satisfaction through a 5-point Likert scale, from 1 —strongly disagree to 5 —strongly agree. The users evaluated VISE along three dimensions (Ease of use, satisfaction, and usefulness) as in Lund [Lund 2001], each dimension has a series of questions to answer. Participants were briefed on how the two systems work before using the systems. Figure 4 presents each participant's satisfaction level in percentage.

It can be observed in Figure 4, that there are no significant differences in satisfaction level between the two systems among the group of users that participated in the first two sessions of the experiments (users with domain knowledge). The difference in satisfaction, however, among the participants of the later sessions (non-expert users) is significant between the two systems, as will be discussed next.

4.3 Result and Discussion

It is important to note that the data collected throughout the experiment are ordinal and the samples are also independent. Therefore, a non-parametric test, a 2-sample t-test is performed with 95% Confidence Interval (CI) and the result is shown in Figure 5

Fig. 6. Satisfaction level based on domain Background knowledge (1 âĂŞ with domain Knowledge and 2 âĂŞ without domain Knowledge).

From the result obtained, p-value < 0.05, which proves that the participants'satisfaction level depends on the visual interface. Thus, rejecting the null hypothesis. Overall, participants are more satisfied with the proposed system, VISE than with SWNV.

More specifically, in terms of the domain background knowledge, there is no significant difference observed in satisfaction level for participants with domain knowledge, while there is for those with no domain knowledge, as shown in Figure 6. It is important to note that the data collected throughout the experiment are ordinal and the samples are also independent. Therefore, a non-parametric test, a 2-sample t-test is performed with 95% Confidence Interval (CI) and the result is shown in Figure 5

5 FUTURE WORK

The idea of proposing and creating an interactive visual summary of about 836,000 museum objects is to address and overcome the information overload in the domain of digital cultural heritage âĂŞ museum collections. The users can interactively select terms from the visual interface to initiate their search and make new discoveries.

VISE currently presents the visual summary of museum collections from a static combination of knowledge representations that is assembled from the NMS and the Rijksmuseum webpages. In the future, we plan to present VISE based on a dynamic combination of knowledge representations from any available knowledge sources, for example from usersâĂŹ social interactions via social networks such as Twitter and Facebook, so that the influence of each representation will vary according to the current user and search characteristics. Also, ongoing research work is planned to incorporate the visual appearance feature of the museum objects into the dataset in order to present personalised recommendations through the visual interface.

The evaluation presented in this paper was performed in a laboratory setting but in the near future, we are planning to perform a further online experiment in a museum setting that will give access to real-time users'interactions from which to further refine the underlying knowledge representation.

REFERENCES

Flora Amato, Vincenzo Moscato, Antonio Picariello, Francesco Colace, Massimo De Santo, Fabio A Schreiber, and Letizia Tanca. 2017. Big data meets digital cultural heritage: Design and implementation of scrabs, a smart context-aware browsing assistant for cultural environments. *Journal on Computing and Cultural Heritage (JOCCH)* 10, 1 (2017), 6.

Flora Amato, Vincenzo Moscato, Antonio Picariello, and Giancarlo Sperlí. 2018. A Recommender System for Multimedia Art Collections. In *International Conference on Intelligent Interactive Multimedia Systems and Services*. Springer, 200–209.

Quim Castella and Charles Sutton. 2014. Word storms: Multiples of word clouds for visual comparison of documents. In *Proceedings of the 23rd international conference on World wide web*. ACM, 665–676.

Gianluigi Ciocca, Paolo Olivo, and Raimondo Schettini. 2012. Browsing museum image collections on a multi-touch table. *Information systems* 37, 2 (2012), 169–182.

Paul Clough, Timothy Hill, Monica Lestari Paramita, and Paula Goodale. 2017. Europeana: What users search for and why. In *International Conference on Theory and Practice of Digital Libraries*. Springer, 207–219.

Bruno Dumas, Bram Moerman, Sandra Trullemans, and Beat Signer. 2014. ArtVis: combining advanced visualisation and tangible interaction for the exploration, analysis and browsing of digital artwork collections. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM, 65–72.

John H Falk and Lynn D Dierking. 2016. *The museum experience revisited*. Routledge.

Guenther Goerz and Martin Scholz. 2010. Adaptation of nlp techniques to cultural heritage research and documentation. *Journal of computing and information technology* 18, 4 (2010), 317–324.

Cormac Hampson, Maristella Agosti, Nicola Orio, Eoin Bailey, Seamus Lawless, Owen Conlan, and Vincent Wade. 2012. The CULTURA project: supporting next generation interaction with digital cultural heritage collections. In *Euro-Mediterranean Conference*. Springer, 668–675.

Daniel A Keim. 2002. Information visualization and visual data mining. *IEEE transactions on Visualization and Computer Graphics* 8, 1 (2002), 1–8.

Arnold M Lund. 2001. Measuring usability with the use questionnaire12. *Usability interface* 8, 2 (2001), 3–6.

Andrew McCallum, Kamal Nigam, and Lyle H Ungar. 2000. Efficient clustering of high-dimensional data sets with application to reference matching. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*. Citeseer, 169–178.

Tor Nørretranders. 1991. *The user illusion: Cutting consciousness down to size*. Viking.

Vivien Petras, Timothy Hill, Juliane Stiller, and Maria Gäde. 2017. Europeana–a Search Engine for Digitised Cultural Heritage Material. *Datenbank-Spektrum* 17, 1 (2017), 41–46.

Robert Villa, Paul D Clough, Mark M Hall, and Sophie A Rutter. 2013. Search or Browse? Casual Information Access to a Cultural Heritage Collection.. In *EuroHCIR*. Citeseer, 19–22.

David Walsh, Mark M Hall, Paul Clough, and Jonathan Foster. 2018. Characterising online museum users: a study of the National Museums Liverpool museum website. *International Journal on Digital Libraries* (2018), 1–13.

Yiwen Wang, Natalia Stash, Lora Aroyo, Peter Gorgels, Lloyd Rutledge, and Guus Schreiber. 2008. Recommendations based on semantically enriched museum collections. *Web Semantics: Science, Services and Agents on the World Wide Web* 6, 4 (2008), 283–290.

Max L Wilson, Bill Kules, Ben Shneiderman, et al. 2010. From keyword search to exploration: Designing future search interfaces for the web. *Foundations and Trends® in Web Science* 2, 1 (2010), 1–97.