VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

A RELATION/TOPIC-BASED VISUALISATION TO AID EXPLORATORY SEARCH IN LARGE COLLECTIONS

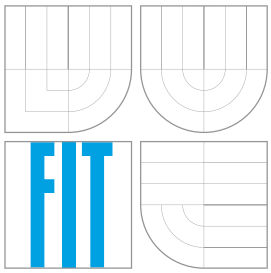
DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE
AUTHOR

Bc. DRAHOMÍRA HERRMANNOVÁ

BRNO 2012

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA INFORMAČNÍCH TECHNOLOGIÍ
ÚSTAV POČÍTAČOVÉ GRAFIKY A MULTIMÉDIÍ

FACULTY OF INFORMATION TECHNOLOGY
DEPARTMENT OF COMPUTER GRAPHICS AND MULTIMEDIA

VIZUALIZACE ZALOŽENÁ NA VZTAZÍCH/TÉMATECH K PODPOŘE PRŮZKUMNÉHO HLEDÁNÍ

A RELATION/TOPIC-BASED VISUALISATION TO AID EXPLORATORY SEARCH IN LARGE COLLECTIONS

DIPLOMOVÁ PRÁCE
MASTER'S THESIS

AUTOR PRÁCE — Bc. DRAHOMÍRA HERRMANNOVÁ
AUTHOR

VEDOUCÍ PRÁCE — Doc. RNDr. PAVEL SMRŽ, Ph.D.
SUPERVISOR

BRNO 2012

Abstrakt

V posledních letech bylo vyvinuto mnoho nových přístupů a metod k vizualizaci a prohlížení obsahu kolekcí dokumentů. Tyto nové metody jsou zaměřené na problémy spojené s rostoucím množstvím dostupných dat a s měnícímy se způsoby, jak lidé s těmito daty pracují. Uživatelé nyní vyžadují lepší podporu takzvaného Exploratory Search (do češtiny by se tento termín dal preložit jako "objevné hledání") pro podporu vyhledávání vazeb mezi dokumenty a pro porovnávání dokumentů. Ačkoliv vizualizace mají potenciál v tomto směru zlepšit prohledávání kolekcí dokumentů (ve srovnání s klasickým textovým vyhledáváním), nestaly se zatím mezi uživateli příliš populárními. Důvodů pro to může být celá řada, od návrhu těchto vizualizací až po jejich implementaci a zpusob použití. Tato práce zkoumá tyto důvody a také faktory, které mohou zlepšit a zpříjemnit použití vizualizací pro hledání. Následně, po zvážení všech těchto faktorů, je navrženo a vyvinuto uživatelské rozhraní pro vizuální prohledávání kolekcí dokumentů, použití tohoto rozhraní je demonstrováno na dvou odlišných kolekcích a v závěru práce je rozhraní vyhodnoceno.

Abstract

In recent years a number of new approaches for visualising and browsing document collections have been developed. These approaches try to address the problems associated with the growing amounts of content available and the changing patterns in the way people interact with information. Users now demand better support for exploring document collections to discover connections, compare and contrast information. Although visual search interfaces have the potential to improve the user experience in exploring document collections compared to textual search interfaces, they have not yet become as popular among users. The reasons for this range from the design of such visual interfaces to the way these interfaces are implemented and used. This work studies these reasons and determines the factors that contribute to an improved visual browsing experience. Consequently, by taking these factors into account, a novel visual search interface that improves exploratory search and the discovery of document relations is designed, implemented and evaluated.

Klíčová slova

Vizualizace informací, prohledávání obsahu, průzkumné hledání

Keywords

Information Visualisation, Content Exploration, Exploratory Search

Citace

Drahomíra Herrmannová: A Relation/Topic-Based Visualisation to Aid Exploratory Search in Large Collections, diplomová práce, Brno, FIT VUT v Brně, 2012

A Relation/Topic-Based Visualisation to Aid Exploratory Search in Large Collections

Prohlášení

Prohlašuji, že jsem tento diplomový projekt vypracovala samostatně pod vedením pana Doc. RNDr. Pavla Smrže, PhD., a že jsem uvedla všechny literární prameny a publikace, ze kterých jsem čerpala.

.....
Drahomíra Herrmannová
July 30, 2012

Poděkování

During the whole academic year I was working as an Erasmus exchange student at The Open University in the United Kingdom where I also worked on this thesis under the guidance and supervision of a researcher Petr Knoth from the university. During the work on this project, I managed to write a paper about the topic, which was accepted at the International Workshop on Mining Scientific Publications held as part of the Joint Conference on Digital Libraries 2012 in Washington, D.C. [16]. This project is based on the paper. I would like to thank my supervisor from BUT FIT, Doc. RNDr. Pavel Smrž, PhD., for his pedagogical and professional guidance, which helped me to successfully fnish my project. I would also like to thank my supervisor from The Open University in the United Kingdom, Petr Knoth, for his constant comments of my work and for his professional advice, which apart from helping me in finishing my project also helped me with developing new skills and learning many new things. His help was very important during all the stages of development of my project.

Contents

1 Introduction **4**

2 Project aims **6**
2.1 Considered types of collections 6
2.2 Visual search interface 6

3 Exploratory Search and Visualisation **8**
3.1 Motivation 8
3.1.1 Example 1: News Search 9
3.1.2 Example 2: Cultural Heritage Exploration 9
3.1.3 Example 3: Document Collection Exploration 10
3.2 Background 11
3.2.1 Data, Information, Knowledge 11
3.2.2 Exploratory Search 11
3.2.3 Current search interfaces 13
3.2.4 Information Visualisation 14

4 Related Work **15**
4.1 Collection level visualisations 16
4.2 Document level visualisaitons 18
4.3 Browsing and query focused visualisations 20
4.4 Selected approach 20

5 Design principles of visual search interfaces **22**
5.1 Added value 22
5.2 Simplicity 23
5.3 Visual legibility 23
5.4 Use of colours 23
5.5 Dimension 23
5.6 Fixed spatial location 23

6 Analysis **24**
6.1 Task 24
6.2 Exploration of scientific publications 24
6.3 Exploration of cultural heritage content 25
6.4 Problems 26
6.4.1 Metadata 26
6.4.2 Visualised set of documents 27

6.4.3 Simplicity and usability . . . 27
6.4.4 Added value . . . 27
6.4.5 Applicability in different collections . . . 27

7 Design **28**
7.1 Considered types of document collections . . . 28
7.1.1 Links between documents . . . 28
7.1.2 Comparing multiple documents . . . 29
7.2 Objectives . . . 29
7.3 Functionality . . . 30
7.3.1 Exploring document relations . . . 31
7.3.2 Discovering interesting connections across dimensions . . . 31
7.3.3 Comparing and contrasting documents in the document stack . . . 32
7.4 Application in cultural heritage collection . . . 32

8 Implementation **35**
8.1 Preparation . . . 35
8.1.1 Basic data set . . . 35
8.1.2 Visualisation tools . . . 35
8.2 Layers of the application . . . 37
8.2.1 Server-side . . . 37
8.2.2 Client-side . . . 37

9 Testing and Evaluation **40**
9.1 Testing . . . 40
9.2 Evaluation . . . 40
9.3 Results . . . 41
9.3.1 Knowledge and skills of participants . . . 41
9.3.2 Comparison of the visual and textual search interfaces of CORE . . . 41
9.3.3 Features of the visual search interface . . . 42
9.3.4 Overall rating . . . 43
9.4 Conclusion . . . 43

10 Discussion **44**
10.1 Application in other domains . . . 44
10.2 Project contribution . . . 44
10.3 Support of exploratory search . . . 45
10.4 Natural language processing tools for generating metadata . . . 45

11 Conclusion **46**

A CD content **50**

B Usage manual **51**

List of Figures

3.1 Types of search activities by G. Marchionini [22] 11
3.2 Two levels of information as seen by J. Zhang [35] 12

4.1 TIARA visualisation showing visualisation of a field "cause of injury." The x-axis is showing time, while the y-axis is showing number of documents belonging to different topics. 15
4.2 FacetAtlas visualisation showing visualisation of a query word "diabetes" . 16
4.3 Overview of Wikipedia topics 17
4.4 InfoSky visualisation 17
4.5 A ThinkPedia visualisation showing articles related to "Semantic Web" . . 18
4.6 A Hopara visualisation showing topics and articles related to article about "Tacoma Narrows Bridge" 19
4.7 A Wivi visualisation 19
4.8 A NVSS visualisation showing citations of articles from years 1991–1992 . . 20
4.9 The Apolo visualisation showing citation network around a selected article . 21
4.10 A WikiVis visualisation 21

6.1 Visualisation of similar documents from CORE 25

7.1 Preview of the visual search interface, showing one document in the document stack and its relations 29
7.2 Discovering interesting connections across dimensions by selecting a relevant document 30
7.3 Comparing and contrasting documents 31
7.4 Comparison of union and intersection mode 32
7.5 Visualisation of content related to one historic event 33
7.6 Visualisation of content related to one historic event, filtered by one selected document and by fine-tuning location settings 33
7.7 Visualisation of the intersection mode of multiple events 34

8.1 Client-server architecture of the visualisation 36
8.2 Class diagram of the server-side of the application 39

Chapter 1

Introduction

Search has been for a long time an integral part of many applications and systems. Nowadays, it has become a daily activity for almost everyone and it is a common way of accessing data and information. Unfortunately, search can be often a complex and a time-consuming task [22]. Among the main reasons are *information overload* and the so-called *"lost in hyperspace"* problem. Information overload comes with the incredible (and growing) speed with which content is generated. This term addresses the fact that with the growing amount of content it becomes harder (or even impossible) to comprehend it. "Lost in hyperspace" refers to the problem of navigating in large quantities of virtual (typically hypertext) content. While following links and relationships, people might easily lose track of how they got to their current "position".

Over the last 20 years, search has become an essential activity of our lives and the way people search and what they require from search interfaces has changed. Gary Marchionini [22] divides search tasks in two basic types — *lookup search* tasks and *exploratory search* tasks. The names of these two concepts already suggest how search has evolved from single-step "fact retrieval" or "question answering" to complex activity that incorporates analysing, comparing and evaluating the content.

While exploratory searches constitute a significant proportion of all searches [26], current search interfaces do not sufficiently support them. This issue has been addressed by a number of researchers by exploring the use of *information visualisation*. Visual search interfaces make use of our visual skills in order to help us to navigate through content. An important aspect of visualisations is that they make it easier to communicate structure, organisation and relations in content. They can also be well utilised to improve search experience, by depicting more information than a typical text search interface using the same space, and they can simplify the process of finding relevant information and can provide graphical aid in results diversification.

In this context, this project aims to create a visual search interface to aid exploratory search in document collections. Document collection visualisations typically project content along one or more selected dimensions — this might be time or other properties of documents in the collection. In contrast, this project addresses the problem by exploring generally applicable principles without considering a specific document collection. With these principles in mind, a novel visual interface, that can work with any type of dimension and any number of dimensions, is designed and implemented and its usability is demonstrated on the domain of research publications and the domain of cultural heritage artifacts.

This thesis can be divided into two main parts. First part focuses on theoretical background of the topic. It introduces the fields of *exploratory search* and *information visualisa-*

tion and explains how visualisation can be used to aid content exploration. It also explores related work in this field and examines the general design principles which can be considered when designing a visualisation. Second part of the project is practical. It describes the analysis of the problem, design of the visualisation and following implementation. In this part the most attention is paid to the design and implementation of the visual search interface.

The remainder of this text is organised as follows. Chapter 2 briefly introduces the goal of this project and explains what were the main goals, tasks, requirements and challenges. Chapter 3 explores the background of the topic. Most importantly, it creates a motivation for using visualisations to aid exploratory search. It points out some issues of current search and explains how these issues could be addressed with visualisation. Chapter 4 reviews the current work in the field of visualising document collections and search results. Based on the related work, this chapter aims to roughly divide this field to categories according to the specialisation and purpose of the visualisations. The visualisation developed as a part of this project is then classified according to this division.

Chapter 5 is dedicated to studying some important design principles for creating visualisations that are also applicable in the field of visual search interfaces. These design principles are based on the study of the related work. Chapter 6, talks about problems and challenges that had to be addressed during the design of the interface. The interface is then described in Chapter 7. Following that, the Chapters 8 and 9 talk about development, testing and evaluation of the visual search interface. Finally, I discuss the contribution of the approach in the Chapter 10 and outline the future work 11.

Chapter 2

Project aims

The main goal of this project was to explore the possibilities of utilising visualisation to aid content exploration and exploratory search in large collections of documents. The task was to design and develop a visual search interface and demonstrate its usability by applying this interface in two different fields.

2.1 Considered types of collections

One of the main requirements for the visual search interface was to create an interface that could work with any document collection, regardless types of properties of the documents in the collection. By a document collection we can understand anything from well structured and hierarchically organised collection like Wikipedia to a collection of documents with very little or without any organisation (for example, we might have a collection of text files or PDFs without metadata and we might want to build a visualisation upon this collection in order to help its exploration).

The two collections that were utilised for the designed visual search interface had, with regard to structure and organisation, very different properties, one of them being a collection of very well arranged and complete metadata, while the other was often missing convenient metadata which could be used to build a visualisation. This situation was one of the challenges during the development of this project because it was necessary to develop a visualisation that would suit both cases.

2.2 Visual search interface

During the development of the visual search interface attention was payed to its usability. The use of visualisations for content exploration has been already explored by many researchers, however despite that visualisations still didn't become a common way for browsing document collections. For example among web search engines the most popular ones typically offer a only textual interface (although some web search engines aimed to present ways of visualising search results[1,2] and Google recently introduced their Knowledge Graph[3]).

[1] http://askken.heroku.com/

[2] http://search-cube.com/

[3] http://www.zdnet.com/google-search-gets-semantic-with-knowledge-graph-3040155245/

Because of these reasons I examined the related work in this field and I created a list of design principles (with focus on visual search interfaces) that are in my opinion significant for designing visual search interfaces and visualisations in general. Following these principles I esigned and implemented a visual search interface which could be used to depict content of any document collection. This approach differentiates this work from many current approaches which usually focus on a specific data set and create a visualisation based on the attributes of the specific collection.

Chapter 3

Exploratory Search and Visualisation

The following chapter constitutes a brief introduction to the fields of exploratory search and information visualisation. First part of the chapter explains in which situations the current search interfaces might be insufficient and proposes visualisation as a way of addressing some issues of current search interfaces. This chapter also explains the concepts of exploratory search and information visualisation.

3.1 Motivation

Information overload has in recent years become an ubiquitous problem. Most of the information is accessible through Internet, a huge database of interlinked resources, but also in electronic libraries and repositories. About ten years ago researches estimated that about 1 exabyte of data is generated every year with more than 99.9% available digitally [18]. Growth rate of Internet users world-wide since that time was more than 480% [13] and it's not just Internet users that help to generate data – a big portion of data is generated thanks to various monitoring systems, sensors, cameras, etc [18]. With this rate of growth it's impossible for an individual to process all information available. And the problem of being unable to grasp the available information is not the only problem that emerges with this growth, it also complicates the navigation in this content. These two issues already have names and I mentioned them in the Introduction, it is the problem of *information overload* and the *"lost in hyperspace"* problem.

The problem of efficient exploration of the available content starts with specifying the request or the search query. The user might not have a deep understanding of the topic he is interested in, on the contrary he might want to learn something about it. In this case the search query might be very general or even ambiguous (it has been also observed that search queries often are ambiguous [30]). In such case it might be difficult for the user to understand how search results relate to his original request and to find interesting information in the results.

1	Viewpoint: Fukushima makes case for renewable energy	4 April 2011
2	Japan earthquake: Nuclear evacuation centre in Fukushima	14 March 2011
3	Japan quake: Power line laid to Fukushima nuclear plant	17 March 2011
4	Japan's Fukushima plant opened to media	12 November 2011
5	Japan earthquake: Concern for Fukushima residents	15 March 2011
6	IAEA chief visits Fukushima nuclear plant	26 July 2011
7	Japan earthquake: Radiation tests in Fukushima schools	5 April 2011
8	Fukushima governor says Japan earthquake victims need help	16 March 2011
9	Japan bans Fukushima rice shipment due to contamination	18 November 2011
10	Tokyo radiation hotspot 'not linked to Fukushima'	13 October 2011

Table 3.1: Most relevant search results for query "Fukushima" retrieved from http://www.bbc.co.uk/news/

3.1.1 Example 1: News Search

Imagine a person searching for news about a specific event. This person could use some personal favourite news service, such as BBC News[1]. Example query could be a keyword *Fukushima*, a place of recent nuclear accident. First 10 most relevant results for this query (as of 27th December 2011) are shown in table 3.1. The columns show order, title and the date when the document was published. BBC search results page also shows first one or two sentences from the document under the document title.

Lets now assume that this person has none or very little knowledge of this event and wants to learn about it. A list of resources like the one shown in the table doesn't tell the user anything about how the resources are related to the even or among each other. By going through the result list article by article user can eventually find the relevant information, but this process requires examining the articles one by one. By looking at the result list we can see that the results are very diverse, for instance the 9th retrieved article obviously relates to business while the next (10th) might relate to environment and safety.

3.1.2 Example 2: Cultural Heritage Exploration

Another example could be exploration of cultural heritage artifacts collection. Such collection could consist of various artifacts from different collections and sources and could contain information about paintings, sculptures and other objects but also about historic events, about buildings and their architecture or information about significant people in history. Objects in such collection could be connected by historic period, architectural style, by artist, author or by an event. Naturally we could display results of search in such collection as a list or alongside a time axis. However the interesting connections that could tell us more about the collection items would remain hidden or might be difficult to see.

[1] http://www.bbc.co.uk/news/

1	Tree-Based Inference for Dirichlet Process Mixtures
2	R/BHC:Fast Bayesian Hierarchical Clustering for Microarray Data
3	R/BHC: Fast bayesian Hierarchical clustering for microarray data
4	Discovering Non-binary Hierarchical Structures with Bayesian Rose Trees
5	Bayesian Rose Trees
6	Bayesian rose trees
7	Robust methods in data mining
8	Unsupervised Learning
9	Tree-Structured Stick Breaking Processes for Hierarchical Modeling
10	Time-Sensitive Dirichlet Process Mixture Models

Table 3.2: Most similar articles to article *Bayesian Hierarchical Clustering* by *Katherine Heller and Zoubin Ghahramani* retrieved by http://core.kmi.open.ac.uk/ as of 28th December 2011

In this case some form of visual representation of the links and relations in the content in could help to understand the narrative and connections between the collection objects.

3.1.3 Example 3: Document Collection Exploration

Another example could be a simple document search, for instance in a digital library or a repository. Documents contained in these collections can be research articles and papers, popular literature or medical records. Such documents usually contain information about authors, date of the publication of the document, publisher and place of publication etc. A typical approach of displaying documents in such collection is (like in the two previous examples) displaying them as a list.

One example of such collection is *CORE*[2]. CORE collects scientific papers from Open Access Repositories and analyzes them – calculates similarities between them. The user can search this collection and explore documents by their similarity. The table 3.2 is showing documents most closely related to the document *Bayesian Hierarchical Clustering*[3] by *Katherine Heller and Zoubin Ghahramani* (this list was retrieved on 28th December 2011).

By looking at the list we can immediately see that documents 2 and 3 and documents 5 and 6 link to the same document. Documents 4, 5 and 6 will probably talk about similar topics, just as documents 1 and 10. Document 7 relates to data mining, while document 8 will most likely talk about machine learning. There is a clear connection between some of the documents, the small groups of documents seem to be cohesive. However this information can only be obtained by going through the list item by item.

[2] http://core.kmi.open.ac.uk/

[3] http://core.kmi.open.ac.uk/display/22012

3.2 Background

In this section I would like to introduce the basic concepts behind the task and to explain how visualisation can be used to address some issues of current search interfaces.

3.2.1 Data, Information, Knowledge

Three most basic concepts that relate to the field are *data*, *information* and *knowledge*. Data is stored facts, it carries no meaning by itself. Data are the metadata of documents without any futher organisation. Information is data that was somehow processed, interpreted. This interpretation might be classification, organisation or correction of the metadata. Knowledge is gained by a person through processing data and information (by comparing, contrasting, analysing and evaluating data and information), a knowledge of some topic means that person is familiar with it. Sometimes we also talk about *wisdom* [25], which stands above knowledge and might be characterised as the ability to utilise knowledge to make wiser decisions.

3.2.2 Exploratory Search

The three examples described in the beginning of this chapter were mentioned in order show how search is usually conducted and to point to some drawbacks of this approach. As I will mention also in in Section 3.2.3, probably the majority of current search interfaces present search results as a ranked list. This approach might be well suited for some search tasks but unsatisfactory for others. In the Chapter 1 I talked about a division of search tasks by Gary Marchionini [22]. He divides search tasks into *lookup*, *learn* and *investigate* tasks. *Learn* and *investigate* tasks according to Marchionini constitute *exploratory search*. His division is shown in Figure 3.1.

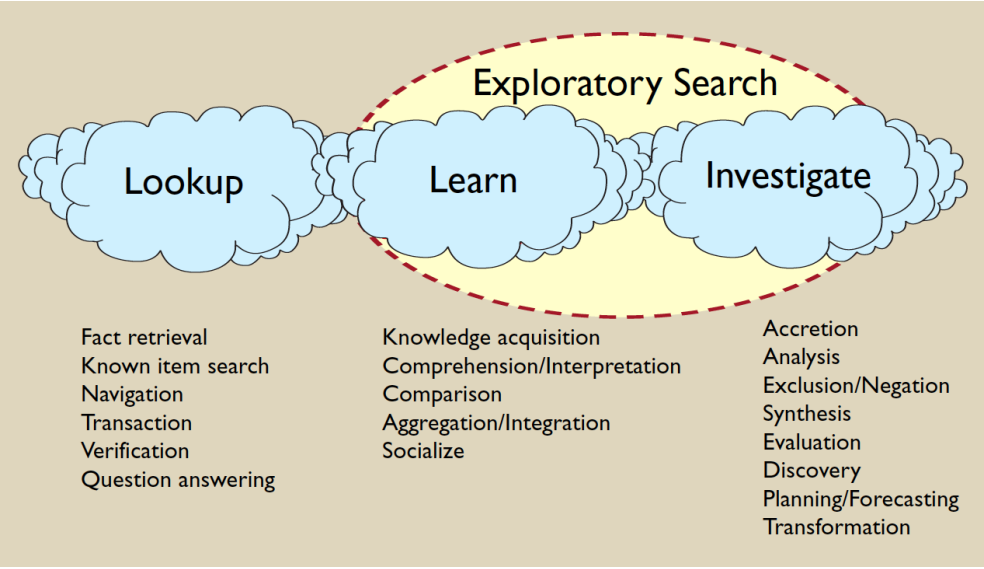


Figure 3.1: Types of search activities by G. Marchionini [22]

What this division shows us is the fact that user goals and reasons for searching can vary. In some cases the users might be looking for a very specific information, they might want to lookup and answer to a question or to retrieve well defined and structured data

from a database. In other cases the search query might be very general or even ambiguous. In such cases the users might need to analyze the results of search and reformulate the query in order to get more specific results.

According to [33], in exploratory search users typically combine browsing and querying. Zhang [35] describes browsing as a kind of activity where results are evaluated by user, while querying results are retrieved and evaluated by the system. We could also say that querying means retrieving a set of facts with no meaning and no relations while browsing means extracting information from data and using this information to compare retrieved objects and to assign them some weight. This situation is well shown in Figure 3.2. It shows two levels of information as seen by Zhang [35]. Micro level of information stands for single documents and their metadata with no relations. On the macro level documents have some meaning and relate to each other.

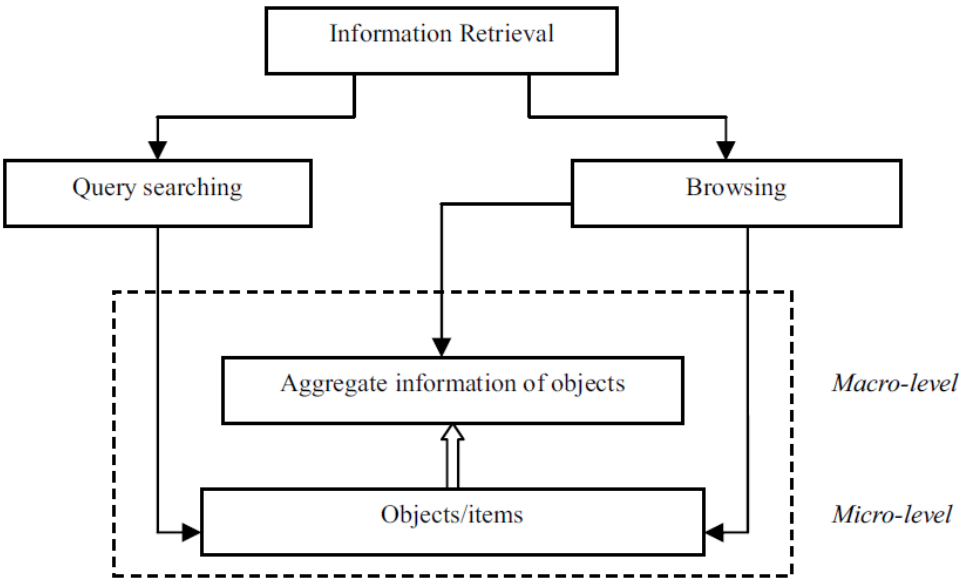


Figure 3.2: Two levels of information as seen by J. Zhang [35]

A good search system with support for exploratory search should clearly access not just the micro-level — retrieve just a set of documents without any further specification of their relations or meaning. However many current search applications provide data or information often only or mostly on the micro-level. This is happening even though the exploratory search tasks constitute a significant portion of all searches.

Regarding the user goals and types of search tasks I would like to mention a study by Andrei Broder [4] in which he conducted an analysis of web searches. He divides search activities into three distinct types:

- **Navigational.** The user wants to navigate to a specific page.
- **Informational.** The user wants to acquire some information on a certain topic. In this case the information might be present on multiple pages.
- **Transactional.** The user wants to perform some transaction, this might be opening a video or shopping on a website

According to the study [4] the informational queries represent the largest category. Other studies came to similar results [21, 26].

Many different methods have been explored in order to support content exploration and exploratory search. As pointed out in the introduction one of such methods is *information visualisation*. The related work done in this field will be presented in Chapter 4. This project aims to study this field and how it can be used to support exploration. As a goal of this project I would like to design and implement a visual search interface which would demonstrate the advantages of presenting search results visually. This visual search interface should allow to utilise the existing relations between the items in the search result list and graphically present these relations to the user.

3.2.3 Current search interfaces

A common search process could be divided in two steps: a query consisting of one or more keywords is entered into the search system and subsequently the system responds with a list of relevant items (documents). This list might or might not be sufficient depending on the type of task. As I already mentioned, a typical example of exploratory search task is a situation when person has none or very little knowledge of the topic. In this case the search query would be probably very general or even ambiguous. After receiving the result list the person will typically examine documents in the list and reformulate the search query in order to get more relevant or more specific results. This approach runs into several problems.

- **Result list size.** For instance a single search done using one of current popular web search engines[4,5,6] can result in a list consisting of even hundreds of millions of documents. At the same time one person will probably examine only first few documents of the list.
- **Ranking of search results.** In order to present the person with the most relevant results first, the documents in the list are typically ranked according to various criteria. One of such criteria might be their interconnection by links — one algorithm utilising this approach is called PageRank (PageRank is used by well-known web search engine Google[4]). This algorithm ranks documents based not on their content, but purely on their "location in the Web's graph structure" [24]. From a certain point of view we could really say that documents with more links pointing to them would appear more important. Trouble with such approach however is that it doesn't provide any information about "hidden" connections between documents like connection by their topics, similarity or shared attributes.
- **Search query ambiguity and results diversification.** Typically the search query is built using only few keywords which often have many interpretations [1]. This fact is addressed using various *results diversification* methods — these methods aim to avoiding situations when the search results are too homogeneous and contain representation of only some facets the search query might have.
- **Complicated browsing of search history.** Current search tools usually don't provide any means of browsing search history. Every time a new search is done a new

[4] https://www.google.co.uk/
[5] http://www.bing.com/
[6] http://uk.yahoo.com/

result list is returned. That means if person wants to review results of some previous query he would have to reenter the query in the system and carry out new search.

Information visualisation provides means how to address some of these issues. Visualisation allows to display more information that a typical search interface using the same space. Allowing the user to see more search results inside of one screen might be a way how to deal with the result list size.

Visualisation are often used because of their ability to communicate information faster than just with text. This ability can help the user to faster understand the meaning of the visualised information and to direct him to interesting information. It can also help in search results diversification.

Visualisations can also depict any number of documents, even from previous searches. In Chapter 4 I will mention one visualisation [20] utilising search history to improve the visualisation.

Finally, visualisation can also help to address the question of how to rank the result list in order to provide more information about the content of the result list. Visualisation can depict more types of metadata, more "dimensions" of the documents at once. The visualisation developed during this project provides such feature. It can depict multiple dimensions corresponding to different meta-information of the documents and each of the dimensions contains a ranked list of results. This is one of the most important features of the developed visual search interface.

3.2.4 Information Visualisation

Simply put *information visualization* is a way how information is presented to the user. It is a way of spatially organizing information in order to show links, relationships and other properties of the content. Visualization can be a way how more information can be communicated than with textual representation, how information can be explored and understood more easily. Visual perception is a sense which we use to navigate the world so it would seem logical to use visualization for navigating through content [23].

The use of information visualisation for supporting exploratory search has been studied by many researchers and is a popular topic among researchers. In Chapter 4 I will present several such visual interfaces. However even though the field of information visualisations has been fruitful there aren't many visual search interfaces which would be preferred to textual search interfaces or be used on daily basis. I tried to explore reasons behind this before designing the search interface. Several researchers have conducted evaluations of existing visual search interfaces and information visualisations. I examined these studies and created a list of design principles which I believe should be considered when designing a visualisation. These principles are listed in Chapter 5.

Chapter 4

Related Work

Current approaches to visualising document collections can be divided according to the granularity of information they provide about the collection into the following groups:

1. *Collection level* — visualise attributes of the collection. These visualisations typically aim at providing a general overview of the collection content.

2. *Document level* — visualise attributes of the collection items, their mutual links and relations.

3. *Intra-document level* — visualise the internal structure of a document, such as the distribution of topics within the document.

In this paper, I am concerned with document level visualisations, however certain concepts from collection level are also applicable.

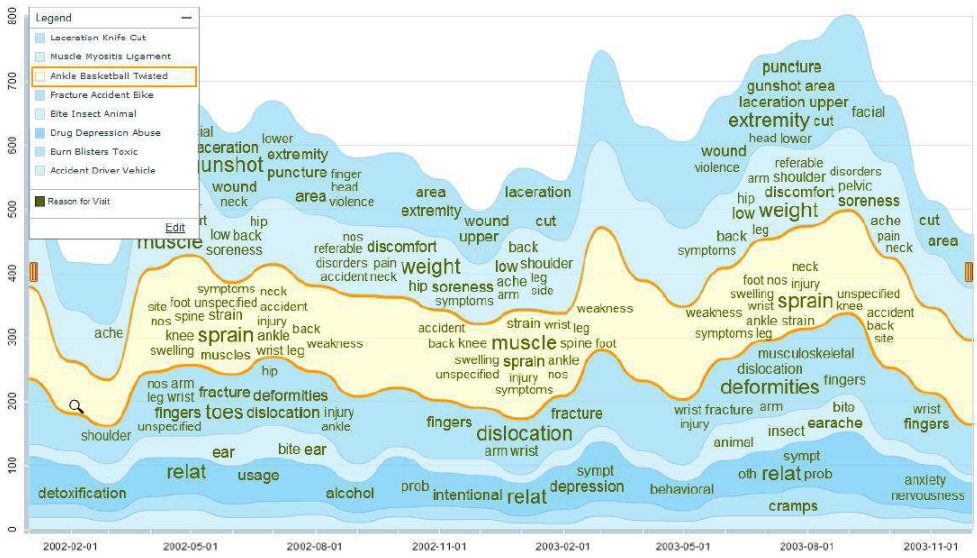


Figure 4.1: TIARA visualisation showing visualisation of a field "cause of injury." The x-axis is showing time, while the y-axis is showing number of documents belonging to different topics.

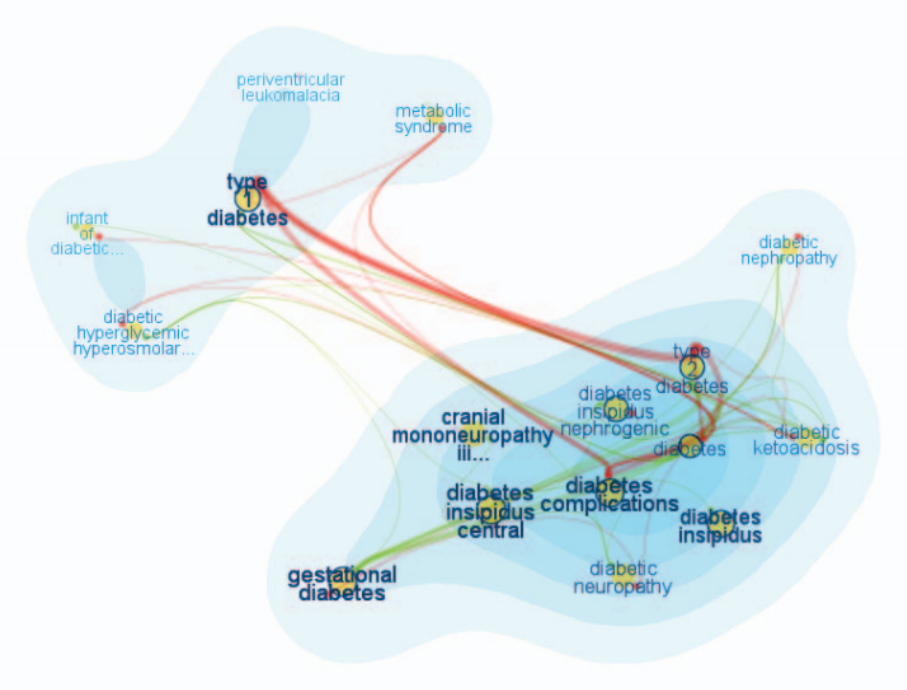


Figure 4.2: FacetAtlas visualisation showing visualisation of a query word "diabetes"

4.1 Collection level visualisations

A well-known example of the first (collection level based) type of visualisations are tag clouds [14] that visually (using attributes like font size and colour) communicate statistical information (such as word, tag or topic frequency) about the collection.

A considerable number of collection-level visualisations focus on depicting topics or themes contained in the collection. While the ThemeRiver [15] and the TIARA [32] (Figure 4.1) visualisations both show changes of themes in the collection over time, [6] use visualisation to reveal theme structure of a collection (for example, an overview of Wikipedia topics from their visualisation can be seen in Figure 4.3).

The FacetAtlas project [5] (Figure 4.2) focuses on multi-faceted documents and keywords, and combines search with a visualisation depicting the keyword senses and different relations between documents based on these senses. Collection level visualisations can also be used for visualising document clusters in a collection. Galaxies [34] or InfoSky [12] (Figure 4.4) are good examples of document cluster visualisations.

In the field of visualising research papers, we can also find a number of tools that aim to create collection overviews. The GRIDL [29] is one such visualisation which purpose is visualising search results in digital libraries. Other tools from this field include the ASE [11] and NVSS [28] tools (Figure 4.8), which use citation networks.

Visualisations focused on collection level information are well suited for analytical and statistical tasks. They can help in the exploration of the collection by providing an overview of the collection content, like in the FacetAtlas [5] (Figure 4.2) or in the TIARA [32] (Figure 4.1). In this case, the exploration happens at the collection level which provides the user with a general overview of the collection's characteristics.

Figure 4.3: Overview of Wikipedia topics from [6]

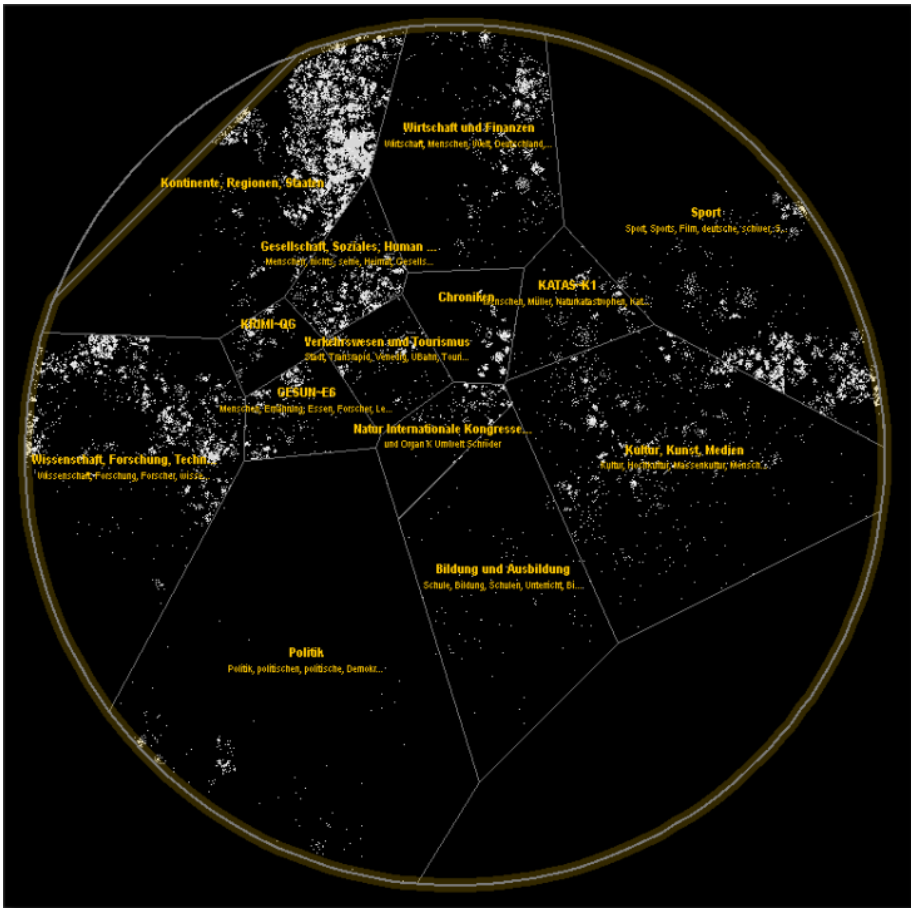


Figure 4.4: InfoSky visualisation

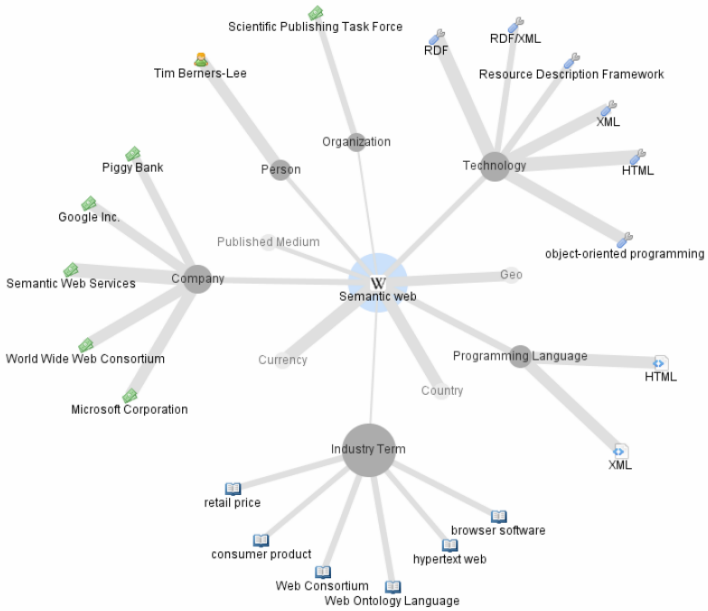


Figure 4.5: A ThinkPedia visualisation showing articles related to "Semantic Web"

4.2 Document level visualisaitons

The second group of visualisations focuses on visualising document level properties. In this paper, I am interested mainly in their use to aid information discovery and content exploration. Regarding this area, a growing number of researchers have been interested in various collections and networks which emerged on the Web in recent years, one such collection being Wikipedia. Data from Wikipedia are hierarchically organised and highly interlinked, which provides good foundation for visualisations.

Hirsch et. al. in [17] created two visualisations, one of them built upon Freebase (a collection similar to Wikipedia) and the other upon Wikipedia (Figure 4.5). Both visualisations present the user with articles related to the currently browsed article and with types of connections between these articles. This way of visualising related articles helps users to quickly explore relevant topics (information about places, people, etc.).

Milne and Witten in [23] (Figure 4.6) chose a slightly different approach. They utilised suggestion of related articles and their clustering, thanks to which they could increase legibility of the visualisation. This is an important quality which can influence whether the user will use or abandon the visualisation.

The Wivi visualisation (Figure 4.7) created by [20] uses a different approach for suggesting relevant articles. It builds a graph of already visited articles and suggests relevant unvisited articles based on relevance to all articles in the browsing history. Relevance of unvisited articles is indicated using a varying distance of articles in the visualisation.

Suggestion of relevant items based on multiple interesting documents (instead of one) is a useful feature which might help to narrow the selection of relevant items. In the visual search interface developed in this project, I utilise a similar approach. The user is given the possibility to choose and add to the visualisation any documents and any number of documents.

Regarding document level visualisations of collections of research papers I would like to

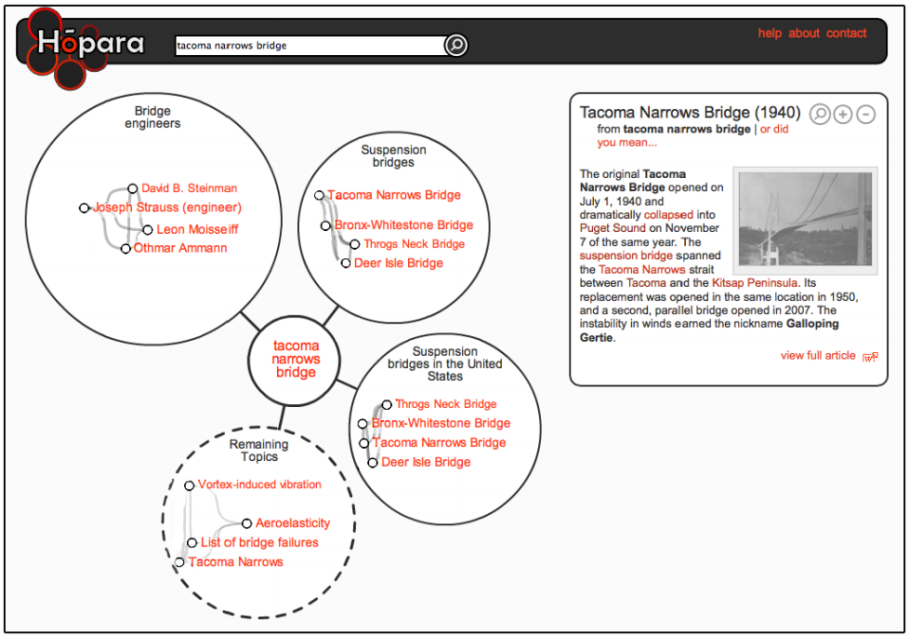


Figure 4.6: A Hopara visualisation showing topics and articles related to article about "Tacoma Narrows Bridge"

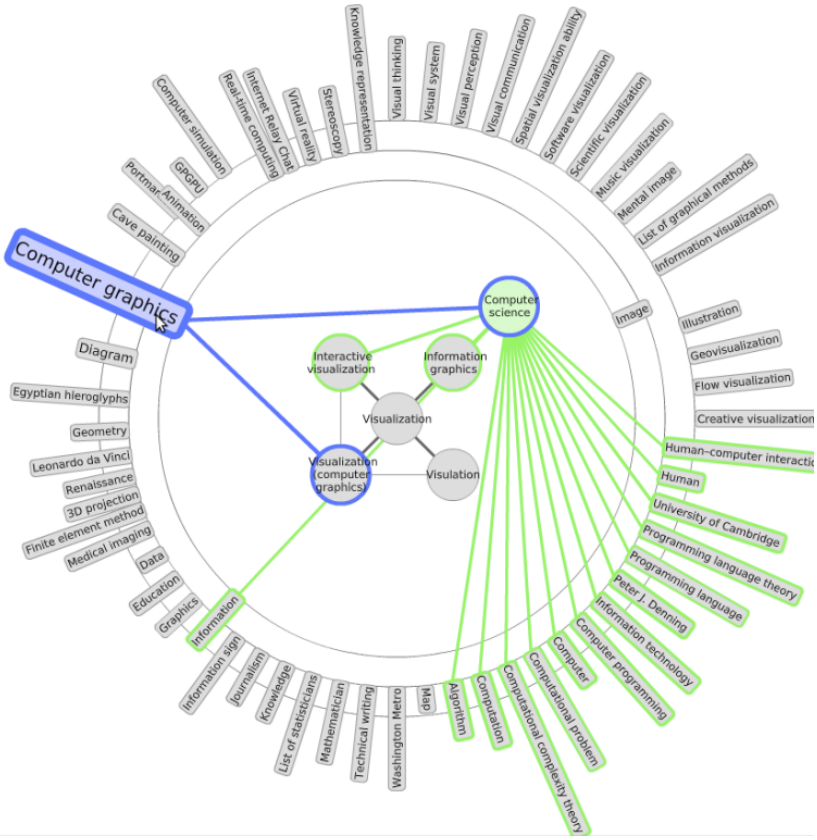


Figure 4.7: A Wivi visualisation

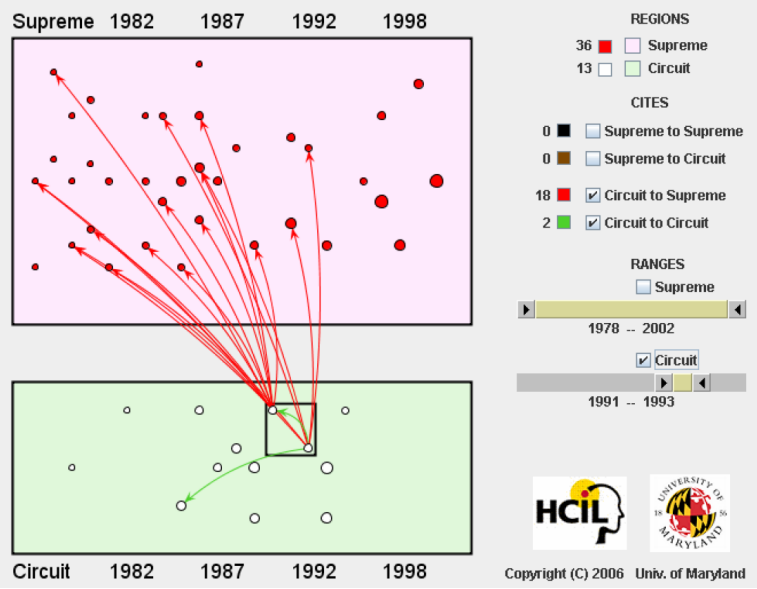


Figure 4.8: A NVSS visualisation showing citations of articles from years 1991–1992

mention [31] (Figure 4.8) and [7] (Figure 4.9) tools which both provide a visualisation of the local subgraph surrounding a specified document.

4.3 Browsing and query focused visualisations

Visual search interfaces can also be divided according to the way the exploration is carried out. As I mentioned in section 3.2.2 J. Zhang [35] divides search tasks into the following two groups (Figure 3.2) that are also applicable to visual search interfaces:

- *Browsing-focused* — The user starts exploration at a specific point in the collection (typically a root document or a topic; usually the same point is used every time) from which the user navigates through the collection.
- *Query-focused* — The user starts with a query, which determines the entry point from which the exploration starts.

As in textual search interfaces, one way to visually explore document collections is to start with an initial point and browse through the collection by navigating from this initial point. The starting point might be, for example, an overview of the whole collection like in [6] (Figure 4.3) and [12] (Figure 4.4) or it might be a root element of a hierarchy as in the category view of the WikiVis visualisation described in [3] (Figure 4.10).

In contrast to this way of exploring the collection, the query-based search interfaces start with the user specifying a query and building a visualisation based on one ([17], Figure 4.5 and [23], Figure 4.6) or multiple ([20], Figure 4.7) documents from this result list.

4.4 Selected approach

Regarding the two previously mentioned divisions, the visualisation described in this thesis could be categorized as document and query based. It aims to visualise articles, related to

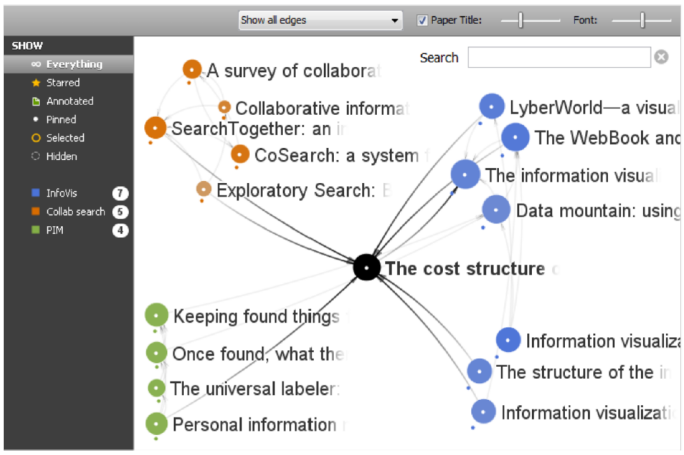


Figure 4.9: The Apolo visualisation showing citation network around a selected article

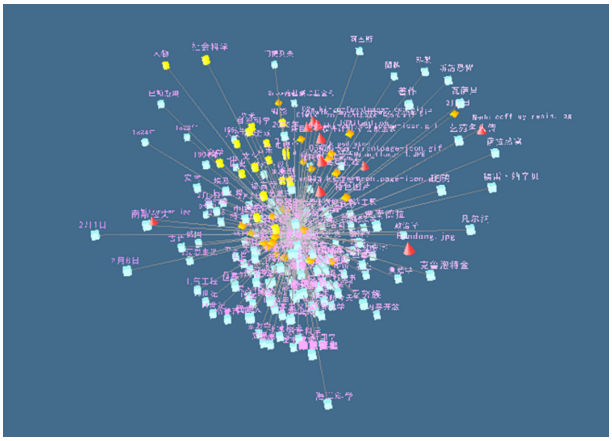


Figure 4.10: A WikiVis visualisation

a user query and through showing relations between these articles to help user to decide how and where to focus his further search and exploration.

Chapter 5

Design principles of visual search interfaces

In this chapter, some general design principles for creating document visualisations are studied. I selected those that are relevant for visual search interfaces and I provide examples of their use.

An empirical study of 6 visual interfaces was carried out by [8]. They concluded that users typically perform better (in terms of efficiency and accuracy) with simpler visual interfaces, regardless of their cognitive abilities. A similar study was conducted by Sebrechts et al. [27] who performed a comparative evaluation of textual, 2D and 3D versions of the NIRVE search interface.

The study pointed out that visual interfaces, in contrast to classical textual interfaces, should simplify the process of accessing information. According to the authors, the usability of visual interfaces is dependent on three factors: the visual interface, the task being performed using this interface and the user performing the task. This means that visual interfaces might be better suited for some information seeking tasks than others (for example, visual search interfaces are probably better suited for exploratory tasks than for lookup tasks).

Sebrechts et al. also observed several factors that affected the usability of the visual interfaces: the use of colours, number of documents in the visualisation, fixed spatial location of the visualisation and the difference between 2D and 3D interfaces. I have analysed these factors, discovered interesting examples of their use in the design of visual search interfaces and organised them into the following list of design principles.

5.1 Added value

First principle I would like to mention is added value[1] with respect to a textual solution. Every visual interface should provide an advantage over a textual interface. The visual interface can assist in the discovery of different information that might otherwise be difficult to see, it might increase the speed of communicating the information, it might help to organise the information more clearly, etc.

According to [27, 2] the visual interface should reduce the mental workload of the user. When document collection exploration is considered, relations between documents might be

[1] Added value stays at a different level of abstraction than the remaining design principles. It refers more to the overall concept of the visual interface rather than how the visual interface is presented.

easier to comprehend when using visual representation rather than textual. FacetAltas [5] (Figure 4.2) is a good example of a visualisation which manages to graphically communicate relations (in this case relations and connections between items based on different facets) that would be difficult to present textually.

5.2 Simplicity

One of the main reasons why textual interfaces are often preferred over visual interfaces is that they can be often used without almost any previous knowledge. This is due to their simplicity and the fact their design mostly follows standard patterns. Visual interfaces that are simple and do not require any learning curve have been found generally better than more complex ones [8]. I am not aware of any popular visual search interface that would be preferred over a textual one for its simplicity.

5.3 Visual legibility

Visual legibility strongly influences user experience with the search interface. Hardly readable text labels, overlapping items or too many items in the view may be a reason for the user to prefer a textual interface even if the visual interface conveys more information. For example, Hopara search interface [23] (Figure 4.6) accomplishes legibility by the use of document clustering and by suggestion of relevant topics.

5.4 Use of colours

Use of colours is a simple but a very powerful tool. Colours can help to immediately identify a shared feature, the type of a relation, a membership in a group, etc. The study [27] pointed out that colours helped to immediately identify groups of articles (regardless of the type or dimension of the interface). [20] (Figure 4.7) and [5] (Figure 4.2) show how colours can be used in visual search interfaces.

5.5 Dimension

Dimension of the visualisation projection. 3D interfaces might be useful and legible in some cases but inconvenient in other cases. A disadvantage of 3D interfaces is that not all parts of the visualisation might be visible in a single view (as in WikiVis visualisation presented in [3], Figure 4.10) — this reduces the legibility and makes the navigation more difficult.

5.6 Fixed spatial location

Fixed spatial location of the visualisation. Sebrechts et al. [27] point out that once users started to rotate the 3D visual interface, they lost track of relations that were no longer visible. This might apply also to 2D interfaces which require zooming. As a result, it is important to consider the use of features, such as rotation and zooming, and what effect they have on navigation.

Chapter 6

Analysis

This chapter aims to analyse the task of the project in detail and to explain which problems and challenges had to be addressed during the design and development of the interface. It also describes the two document collections which were used for demonstrating the functionality of the visualisation.

6.1 Task

The main aim of the project was briefly outlined in chapter 2. The goal was to develop a visual interface for visualising a set of objects (like documents or records) relevant to a specific query. The main reason behind the requirement for developing a visual search interface was to support exploratory search in collections of objects. This interface had to be easily applicable in different domains. Initially the interface had to be applied in the following two fields:

1. Exploration of scientific publications.
2. Exploration of cultural heritage content.

6.2 Exploration of scientific publications

One of the fields where the visualisation had to be applied was exploration of a collection of scientific publications. By scientific publications we can understand any documents like research papers and reports, theses and dissertations. A collection of this kind might be for example a digital library or a university repository.

Documents in such collection are typically described according to basic metadata like author, publisher, date of publication, type of the publication, etc. These types of metadata are explicit (and typically are entered into the repository by a responsible person). The repository then usually provides a search interface that offers search in this metadata and/or in the full text of the publication.

These types of documents also carry other types of metadata, which are not explicit and might need to extracted from the documents. This implicit metadata might be semantic similarity of documents, automatically extracted entities and concepts or a citation network created from these documents.

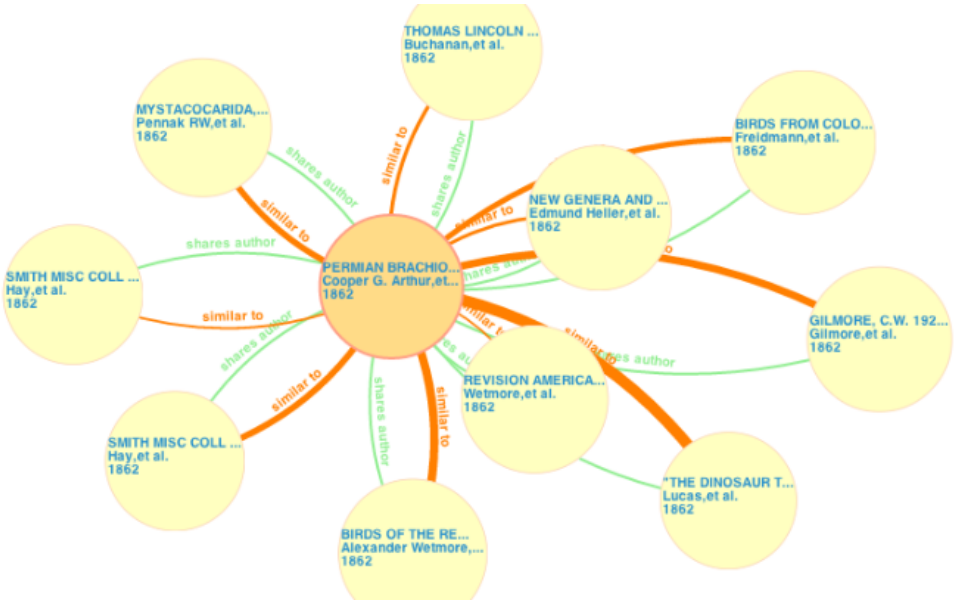


Figure 6.1: Visualisation of similar documents from CORE

A particular document collection which was used for developing the visualisation is CORE[1] which was briefly introduced in Section 3.1.3. CORE is a service that aims to provide access to documents from many Open Access repositories[2]. Apart of harvesting documents, CORE aims to also improve search and navigation in the collection by analysing the documents and extracting some additional metadata. Particularly, CORE provides information about semantic similarity of documents and extracts citations from documents. The collection can be accessed through a textual search interface, through an API or through a mobile application (which also provides a textual search interface).

The main motivation for this project was in improving content exploration and exploratory search in the CORE collection. CORE currently provides a visualisation of similar documents (Figure 6.1), however this visualisation only depicts a limited number of documents and only alongside the dimension of similar documents.

6.3 Exploration of cultural heritage content

Second field where the visual search interface had to be applied was exploration of cultural heritage collection. This field was briefly introduced in Section 3.1.2. By a cultural heritage collection we can understand a collection of objects in a museum, gallery or a historical archive. A typical museum collection will probably contain information about objects like paintings, sculptures and other artistic objects, information about archaeological objects or about antiquities, information about people, places and events, etc., depending on the type of the collection.

Documents in such collection are (as in the collection of scientific publications) typically described according to some explicit metadata, like style, historical period, type of item,

[1] http://core.kmi.open.ac.uk/

[2] CORE currently — as of 20th July — contains well over 8 million documents of which about 400 000 documents are full text documents.

artist or owner, etc. Each item in the collection can be described by a set of such properties. This type of collection will probably contain many links and relations between the collection items. Some of these connections are created through a shared property — through an artist who created the item, through a historical perid or through a shared style or type of item. Other types of connections might be created for example through an event — one artist might have influenced another, two items might have been mentioned in the same book, etc.

One such collection of cultural heritage objects is Decipher[3]. As part of this project it was required to demonstrate the use of the visualisation on this collection. Decipher represents information about objects, people or places as events instead of representing each object as a set of properties, which differentiates it from typical museums or gallery collections.

For example a painting "Lady with an Ermine" can be described by an event "Leonardo da Vinci painted 'Lady with an Ermine' between the years 1489 and 1490." Other events that describe this painting would be for example "Cecilia Gallerani was portrayed by Leonardo da Vinci in 1489" and "The 'Lady with an Ermine' was acquired by Prince Adam Jerzy Czartoryski in 1798". Each of these events is described by a set of properties that correspond to specific dimensions. These dimensions can be for example 'a type of object' (painting, "Lady with an Ermine"), 'agent' (Leonardo da Vinci, Cecilia Gallerani, Prince Adam Jerzy Czartoryski) or 'activity' (paint, acquire). Using this form of description any object can be characterised.

Such detailed and complete metadata as in the Decipher collection are well suited for search and for creating visualisations. Decipher already provides timeline and map views for presenting the data visually, however it doesn't provide any general visualisation that could be utilised for exploratory search. This functionality was required from the visual search search interface that was developed in this project.

6.4 Problems

The following section deals with some problems and challenges that had to be addressed during the project.

6.4.1 Metadata

One of the problems during the design and development of the interface was incomplete or missing metadata. In the Chapter 4 I presented several visualisations that were utilising well structured and interlinked collections like Wikipedia[4] or Freebase[5]. However many real world document collections might not be so well interlinked, classified and categorised. Previously in this chapter I introduced two document collections which were utilised in this project. Documents from the Decipher collection were very well structured and organised. On the other hand, documents from CORE contained basic metadata, information about similarity of documents and in some cases also citations and extracted concepts, however there was no hierarchical organisation that would structure documents by topic, field or category. One of the challenges during the design and development of the interface was

[3] http://decipher.open.ac.uk/project/
[4] http://www.wikipedia.org/
[5] http://www.freebase.com/

how to visually structure the collection and which features and information to use in the visualisation.

6.4.2 Visualised set of documents

In Chapter 3 I pointed out that one of the advantages of using visual interfaces in contrast to textual ones is in the ability to show more documents at the same time. This ability of visualisations could help to reduce one of the issues of current search interfaces — too big search result list. However with the ability of showing more documents at once comes the question of how many documents can be visualised without loosing legibility and also how many documents can the user still comprehend. The question is also which documents from the result set should be selected for visualisation as the best representation of the result set.

6.4.3 Simplicity and usability

Following the design principles listed in Chapter 5 I was also concerned about the simplicity and usability of the visualisation. For example, the CORE project already offers a textual search interface which is used by hundreds of users every day[6]. The textual interface uses standard design or pattern used in the majority of current search engines and interfaces, therefore the users are able to work with the interface almost immediately. On the other hand, if the users start to use the visual interface, it will probably take them some time to learn to work with the interface and to explore its features. The interface should simplify and shorten this process as much as possible. The use of visualisation also brings some additional requirements for hardware and software equipment. These requirements should preferably also be minimised.

6.4.4 Added value

The visual interface should bring some added value compared to textual search interface. It should show some relations and connections that aren't immediately visible when using the textual interface. The CORE project currently already provides a simple visualisation of documents similar to the specified document. This visualisation is shown in Figure 6.1. It depicts the selected document in the center of the visualisation and surrounding are documents similar to this document. Links between the nodes show how similar the documents are (stronger link stands for higher similarity) and if the documents share an author (link with a different color-coding). The requirement for the visualisation was to provide an improvement over the current visualisation — to depict more documents and/or more information and to provide an interface that would better support exploration of related documents.

6.4.5 Applicability in different collections

In the beginning of this chapter I presented two different document collections that were utilised in the project. It was required to provide a visualisation that would be applicable in these two collections but also in any other document collection, regardless of the types of the documents contained in the collection.

[6] http://core.kmi.open.ac.uk/search

Chapter 7

Design

This chapter describes the design of the visual search interface. The design of the visualisation is based on the design principles presented in Chapter 5 and takes in account the problems outlined in Chapter 6.

7.1 Considered types of document collections

One of the main requirements for the visualisation was to provide added value not just in contrast to classical textual search interfaces but also to the existing visualisation, which is part of the CORE application 6.4.4. In order to achieve this I looked at the documents as to a set of dimensions (sometimes these are called facets). Every document in a collection is defined according to this set. The dimensions are typically of different types. Each document can be described by a set of properties each of which expresses the value of a corresponding dimension.

Although the specific dimensions are dependent on the document collection domain, they are in a real-world document collection always present. For example, an article in a news collection can be described by the properties corresponding to dimensions, such as time, themes, locations, relations to other articles. Documents describing cultural heritage artifacts can be characterised by artifact type, historical period, style, material, etc. Similarly, research articles can be represented by citations, authors, concepts, similarities with other research articles, etc. Each dimension offers a different point of view on the specified document and on documents that relate to it.

These document properties are either explicit or implicit. Explicit properties relate to user defined properties, typically citations, authors, location. Implicit properties refer to properties, such as document similarity, which usually need to be discovered. Many visual interfaces are tailored to specific domains and their dimensions. In this project, all types of document collections are considered.

7.1.1 Links between documents

These dimensions, which define documents in the collection, can have both internal links and links among the dimension. For example, we can have a scientific article which is defined by its authors, similar documents and the topics it talks about. The documents similar to this article might be similar between each other — these are internal links. The similar documents might also share an author or a topic with the specified article — these are links among dimensions. In the visualisation I aimed to reveal these links and to allow

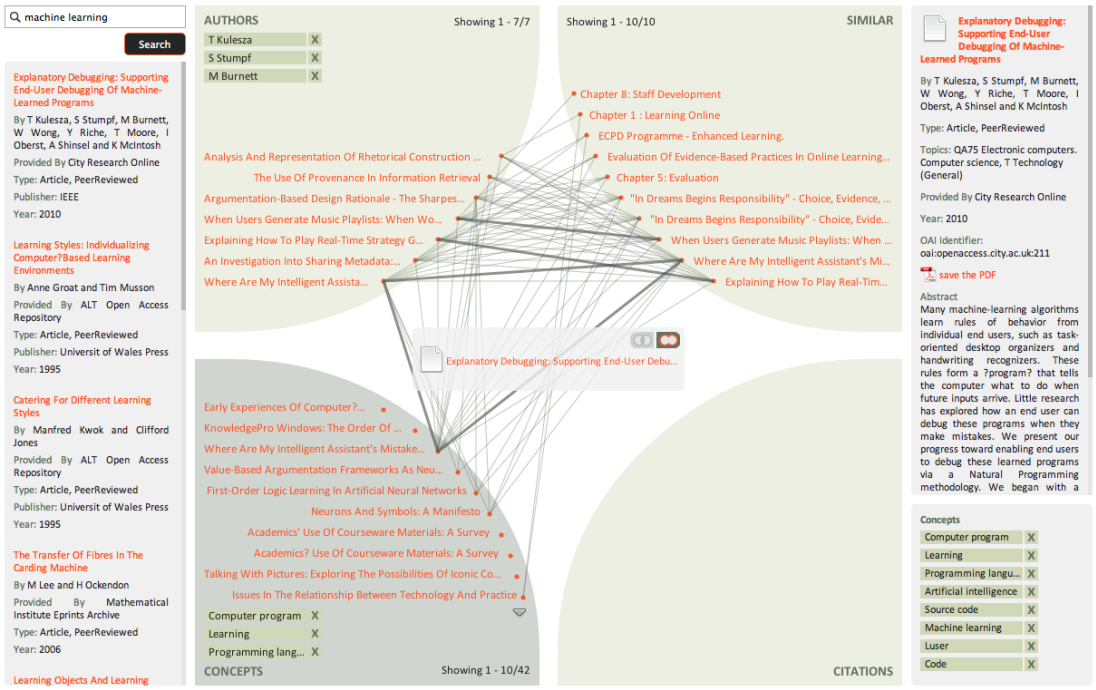


Figure 7.1: Preview of the visual search interface, showing one document in the document stack and its relations

the user to compare and contrast them. This revealed relations should then help the user to understand better the relevance and relation of the specified document to other documents and to decide how to focus further search.

7.1.2 Comparing multiple documents

To better aid the exploration of interesting content the visualisation also allows the comparison of multiple documents at once. This approach can help to reveal common attributes of the selected documents and to reveal their mutual connections. I believe the ability to visualise links inside and between dimensions and to compare documents presents the added value of the visualisation. This feature can also help to reduce the problem of selecting relevant documents for the visualisation. If the visualisation could show multiple dimensions at once, it should help the user to faster understand what is the meaning of the visualised documents while allowing him to see more documents than textual search, using the same space.

7.2 Objectives

To summarize the previous section, the visual search interface is based on the combination of the following principles which differentiate this approach from previous work:

- *Support for comparing and contrasting content.* The search interface should offer the means for comparing and contrasting properties of multiple documents.

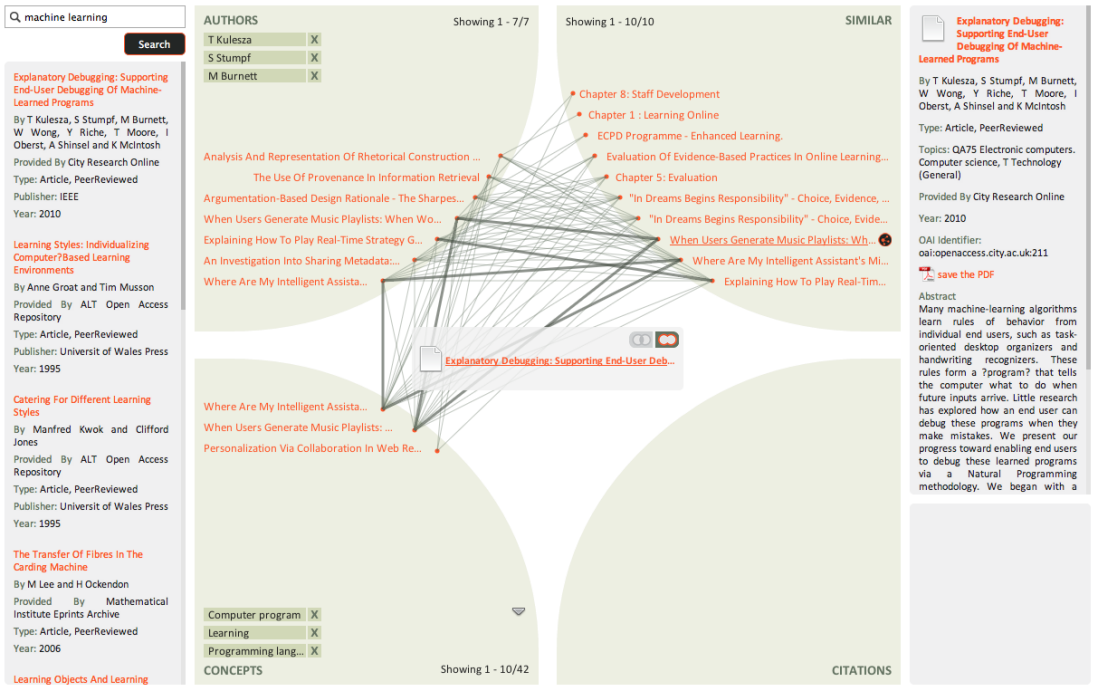


Figure 7.2: Discovering interesting connections across dimensions by selecting a relevant document

- *Support for exploration across dimensions.* The search interface should help assist in the discovery of interesting relationships across dimensions by taking into account multiple aspects simultaneously.
- *Universal approach to the visualised dimensions.* The visual search interface can be adapted to any document collection.

While the first two principles are difficult to realise in textual search interfaces, satisfying the third one is a challenge for visual interfaces. The contribution of this project is in addressing these principles at once.

7.3 Functionality

The proposed visual search interface consists of a visualisation area which is supported by a left and right sidebar. The left sidebar features a search box, which is the starting point of visual search, and an area for the search results. In the first step, the user enters an initial query into the search box and a list of relevant documents will be displayed. The user can select one of the documents and see its details in the right sidebar. Any of these documents can be dragged into the visualisation area, which initialises the visualisation. The visualisation enables the user to perform the following activities: *exploring document relations*, *discovering interesting connections across dimensions*, *comparing and contrasting documents.*

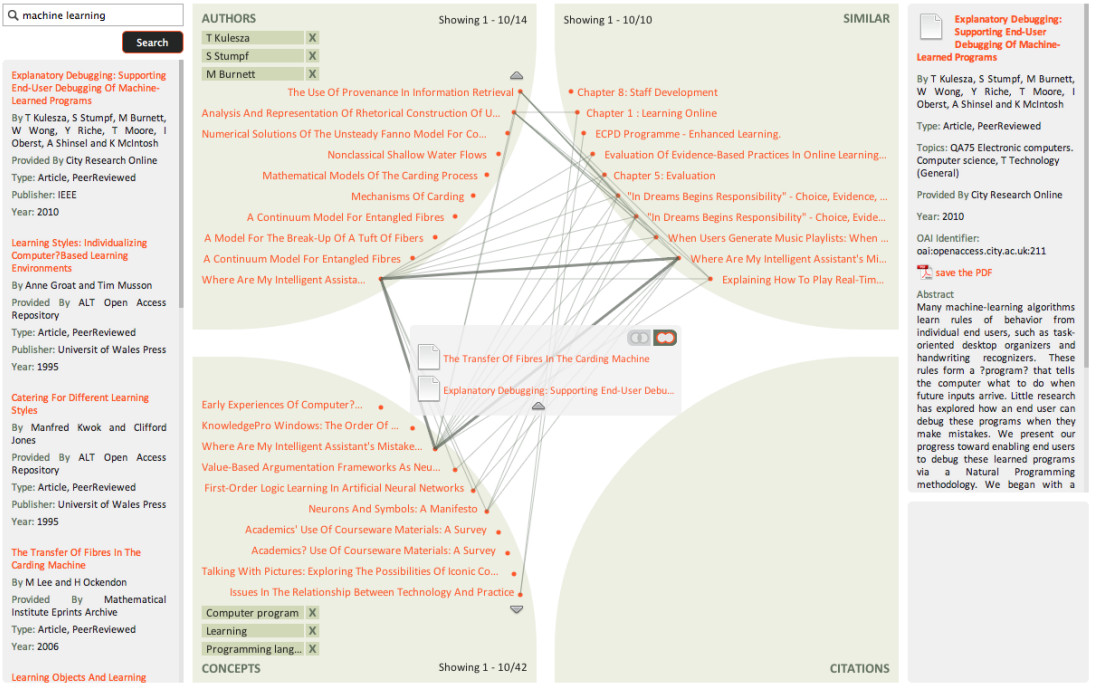


Figure 7.3: Comparing and contrasting documents

7.3.1 Exploring document relations

The visualisation itself shows the selected document in the centre of the screen in an area which is called the *document stack*. Any number of documents can be added to the document stack. This area is surrounded by a set of predefined dimensions that are suitable for the visualised document collection. Figure 7.1 shows the dimensions relevant to the domain of scientific articles (this is a version of the visualisation integrated into the CORE application 3.1.3). For this example, I chose document authors, concepts (the document topics or themes), similar documents and citations. In a typical collection, there will be many documents related to the content of the document stack and the user can scroll through them. Each of the dimensions offers a different view on the related documents. For example, the area showing document authors might reveal other documents from the same authors. Similarly, the concepts area enables the user to explore documents discussing the same topics. Some of the areas can be customised to further specify the relatedness criteria. This can be achieved by modifying the dimension settings that appear in the bottom right sidebar. For instance, the visualisation allows deselecting any of the concepts in the concepts view and consequently fine-tuning the list of the relevant documents.

7.3.2 Discovering interesting connections across dimensions

Just like the documents in the stack, the related documents are also described by the same set of properties — authors, concepts, similar documents and citations. These documents relate not only to the document stack, but also to one another, across the dimensions. For example, one of the cited documents can share an author with a document in the stack. The cited document will appear in two views — in the authors view and in the citations view. The visualisation displays these connections using thin lines. If the same document

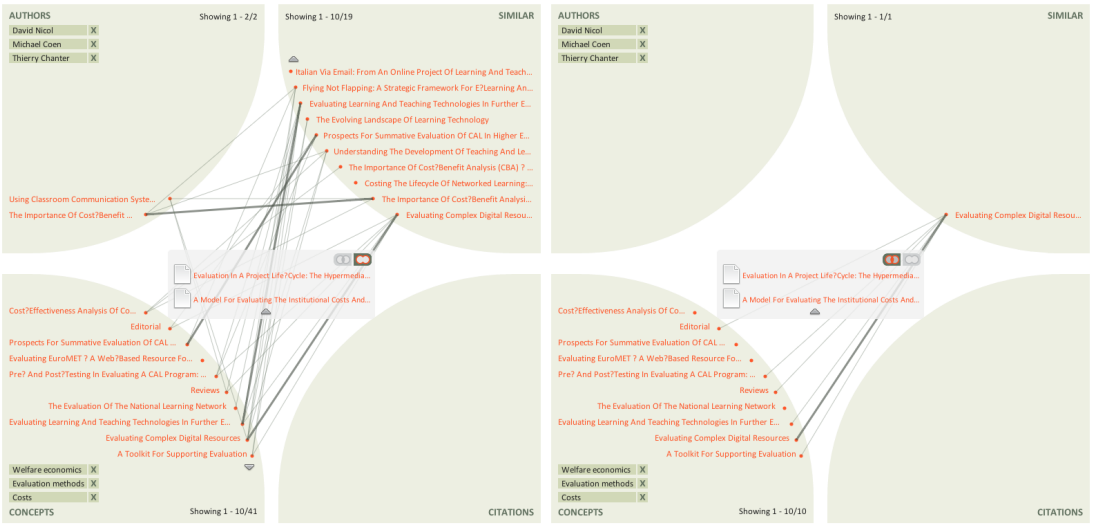


Figure 7.4: Comparison of union and intersection mode

appears in multiple views, it will be connected by a thicker connecting line.

Any of the related documents can be selected in order to reveal connections across dimensions (using a small *reveal connections* icon which appears after hovering over the related document). This is used to highlight only the documents that relate to both the document stack and the selected document. As shown in Figure 7.2, the interface adjusts the content displayed and hides documents that are not related to the selected document.

7.3.3 Comparing and contrasting documents in the document stack

At any time, the user can drag more documents displayed in the visualisation area or in the left sidebar to the document stack. This allows the comparing and contrasting of their properties and relations. The documents in the stack can share any properties. The user can switch between the union and intersection mode as shown in Figure 7.3 to see all the properties and relations of the documents in the stack or only the shared properties and relations. Figure 7.4 is showing comparison of union and intersection mode of two documents in document stack. Visualised documents can be removed from the document stack by clicking at a minus icon that appears after hovering over a document in the stack.

7.4 Application in cultural heritage collection

The previous section demonstrated how the visual search interface can be applied in the domain of research papers. As part of the project it was required to apply the visual search interface also in the field of cultural heritage content. Unfortunately at the time of writing this text the project Decipher, which was planned to be used for the visualisation 6.3, was still under development and its data collection was not accessible. Therefore in this case the demonstration of usage of the visual search interface was done as a mockup on a very small set of sample data.

For the field of cultural heritage content I chose to visualise the following dimensions: agents, objects, activities and locations. These dimensions were described in Chapter Anal-

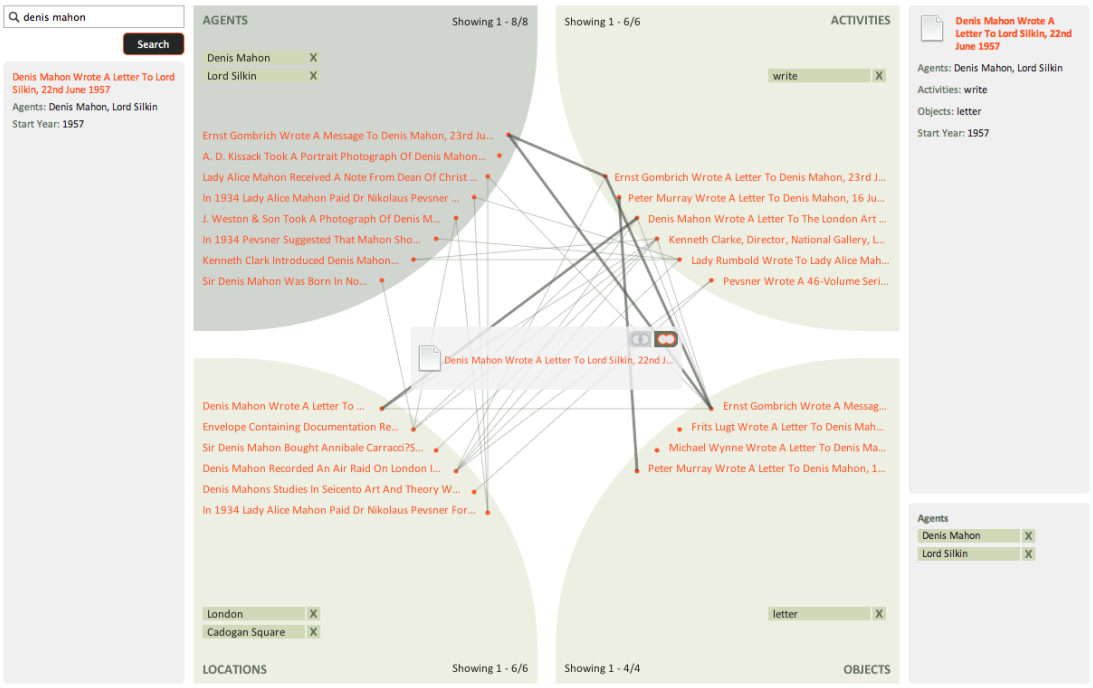


Figure 7.5: Visualisation of content related to one historic event

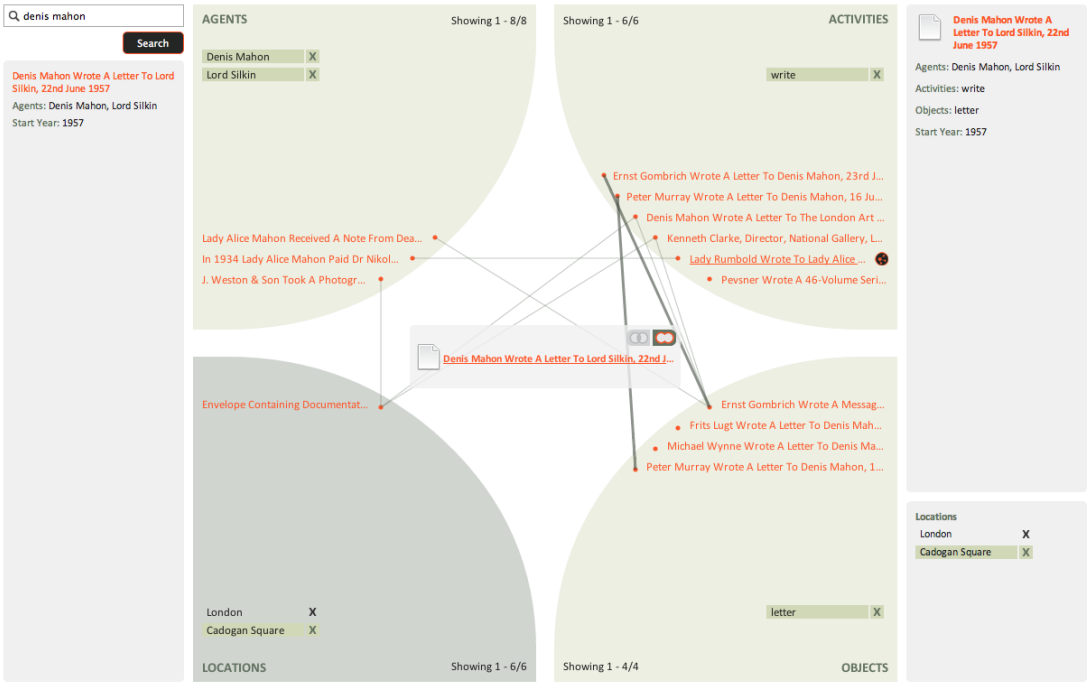


Figure 7.6: Visualisation of content related to one historic event, filtered by one selected document and by fine-tuning location settings

Figure 7.7: Visualisation of the intersection mode of multiple events

ysis 6.3. The Figure 7.5 is showing a visualisation of content related to an event "Denis Mahon Wrote A Letter To Lord Silkin, 22nd June 1957." The Figure shows that in the case of cultural heritage content, each of the dimensions offers additional settings which can help to better specify the results. For example the locations dimensions offers a selection of a specific place or places, which is show in Figure 7.6. Figure 7.7 depicts the content related to two selected documents, in this case the visual search interface is showing the intersection mode.

Chapter 8

Implementation

This chapter describes the implementation part of this project — the development of the designed visual search interface. It explains which technologies were used for the visualisation, the architecture of the application and implementation process.

8.1 Preparation

8.1.1 Basic data set

As I mentioned in Chapter 7.4 the project Decipher was at the time of writing this thesis still at development and its data collection isn't yet available. Therefore the CORE collection (described in Chapter 6) was selected as the basic data set for the development of the project. It provides a big set of documents with basic metadata, with information about document similarity and in case of documents will full text also information about citations.

As described in chapter 7 I decided to visualise the documents along found dimensions: dimension of authors, dimension of similar documents, dimension of cited documents and dimension of document concepts (topics). Unfortunately in the case of citations only few of them link to another document from CORE, most citations are stored only as a title. As a result, in the final visualisation the dimension of citations is often empty. Also not all documents from CORE contain information about concepts.

8.1.2 Visualisation tools

The visualisation was implemented as a web browser application. For creating the visualisation itself I needed to choose a visualisation tool. At the beginning of the development of the application I experimented with multiple visualisation libraries and tools that can help create visualisations for the web. Following is an overview of some popular visualisation tools:

- *Arbor*[1] Arbor is a visualization tool which provides algorithm for creating force-directed layouts. It uses jQuery and webworkers. This tool is used by CORE project to create their visualization (figure 6.1).
- *Prefuse Flare*[2] "is an ActionScript library for creating visualizations that run in the Adobe Flash Player" [19]. It supports many different types of graphs from simple

[1] http://arborjs.org/
[2] http://flare.prefuse.org/

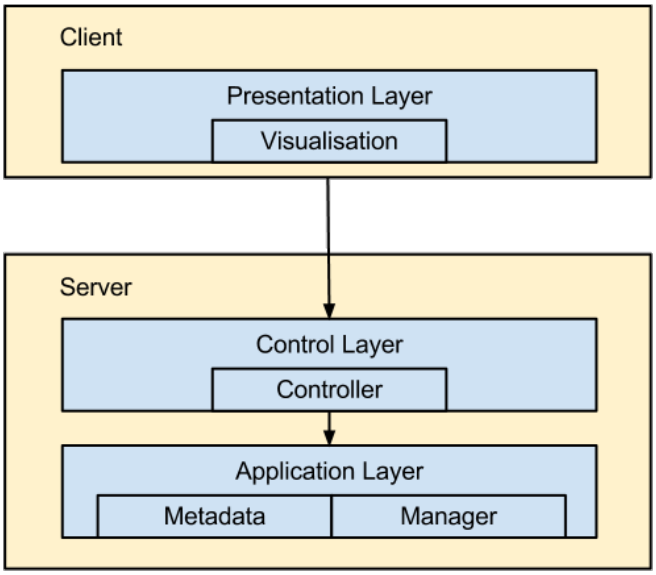


Figure 8.1: Client-server architecture of the visualisation

to very complex ones. It is used by two visualizations I mentioned in section 4 – by Hopara (figure 4.6) and Wivi (figure 4.7).

- *Processing.js*[3] is a JavaScript library for creating visualizations using Processing visual language. Visualizations can be created either with Processing language or using JavaScript. Processing.js is used for instance by visual web browser *Ask Ken*[4].
- *InfoVis Toolkit*[5] is another JavaScript library. Just as the two previously mentioned libraries, InfoVis supports different types of graphs and visualizations such as force-directed layout or tree graphs.
- *Protovis*[6] *and D3.js*[7] Protovis uses JavaScript and SVG (SVG is a language for describing 2D vector graphics using XML). Protovis is no longer in development, however it was replaced by a very similar tool called D3. D3 uses JavaScript and works with DOM (Document Object Model) to create visualizations.
- *HTML5*[8] The latest version of HTML provides canvas element for 2D drawing which can be used in combination with JavaScript to create visualisations.

After testing the available tools the choice fell on the combination of HTML5, CSS3 and JavaScript. The elements of the visualisation are styled and positioned using CSS3. HTML5 is a relatively new technology (currently, as of July 2012, it is still under development [9]), however its basic features are already supported by the major web browsers (including web browsers for mobile devices) [10]. This and the fact I wanted to familiarise myself with a

[3] http://processingjs.org/
[4] http://askken.heroku.com/
[5] http://thejit.org/
[6] http://mbostock.github.com/protovis/
[7] http://mbostock.github.com/d3/
[8] http://www.w3.org/TR/html5/

popular new technology were among the main reasons why I chose HTML5 for creating the visualisation.

8.2 Layers of the application

The visualisation (the client) communicates with the server using REST (Representational State Transfer) architecture. The client requests metadata from the server and builds the visualisation upon this metadata. The communication between the client and the server requires three methods and a specific data format, therefore it was also necessary to develop the server side services. The CORE application is created in Java language and is using Spring Framework MVC architecture, so the choice of the server-side technologies was simple. The basic architecture of the visualisation is shown in Figure 8.1.

8.2.1 Server-side

The visualisation requires access to three REST resources, which are listed in table 8.1. The communication between the server and the client is done via JSON format. The required format of the response can be found in the application documentation.

The server side code of the application is using existing CORE classes for accessing index with documents and for accessing database, which stores information about document similarities. The requests from client are processed by `DocVisController` class. This class receives requests from the client, calls the appropriate classes and sends responses back to the client. For converting the responses to the JSON format I utilised the Google GSON library[9]. Search is done using `DocumentSearcher` class from CORE. For loading the dimensions of a document and links between documents in these dimensions I created a class `MetadataManager`. This class uses `DocumentSearcher` and `SimilaritySearcher` classes from CORE for loading the related documents. After all documents are loaded, the class compares all documents in all dimensions and searches for connections to other documents. The `DocumentMetadata` and the `MetadataWrapper` classes are used to build the list of dimensions and documents within the dimensions. The `MetadataWrapper` class is then converted to JSON using the Google Gson library[9]. Class diagram of the server-side of the application is shown in Figure 8.2.

8.2.2 Client-side

The client side of the application is using HTML5, CSS3 and JavaScript. The code of the web page with the visualisation is created using JSP, which is the technology CORE uses for creating the web site, however it can be simply converted to pure HTML. The elements of the visualisation are created using HTML5 and are positioned and styled using CSS3. The page contains the HTML `canvas` element, which is used for drawing connections between articles. The functionality of the visualisation is created using JavaScript. For simplifying work with JavaScript I utilised jQuery library[10,11] and several jQuery plugings, particularly jCanvas[12] for simpler work with the HTML5 canvas, jQuery Templates[13] and jScrollPane[14].

[9] http://code.google.com/p/google-gson/
[10] http://jquery.com/
[11] http://jqueryui.com/
[12] http://calebevans.me/projects/jcanvas/index.php
[13] http://api.jquery.com/jquery.tmpl/
[14] http://jscrollpane.kelvinluck.com/

Resource URI	Method	Response format	Description
`/metadata/{documentId}`	GET	JSON	Retrieve metadata (dimensions) of a document specified by ID.
`/detail/{documentId}`	GET	JSON	Retrieve details of a document specified by ID.
`/search/{searchCriteria}`	GET	JSON	Search for specified criteria and return list of articles.

Table 8.1: REST resources requested by the visualisation

When the visualisation is loaded, the JavaScript code sends AJAX requests to the server (listed in 8.1), fills the page with the received data and creates the visualisation (draws connections between documents and provides the filtering, scrolling and other functionality of the visualisation) upon the received data. Detailed description of the client-side functions can be found in the application documentation.

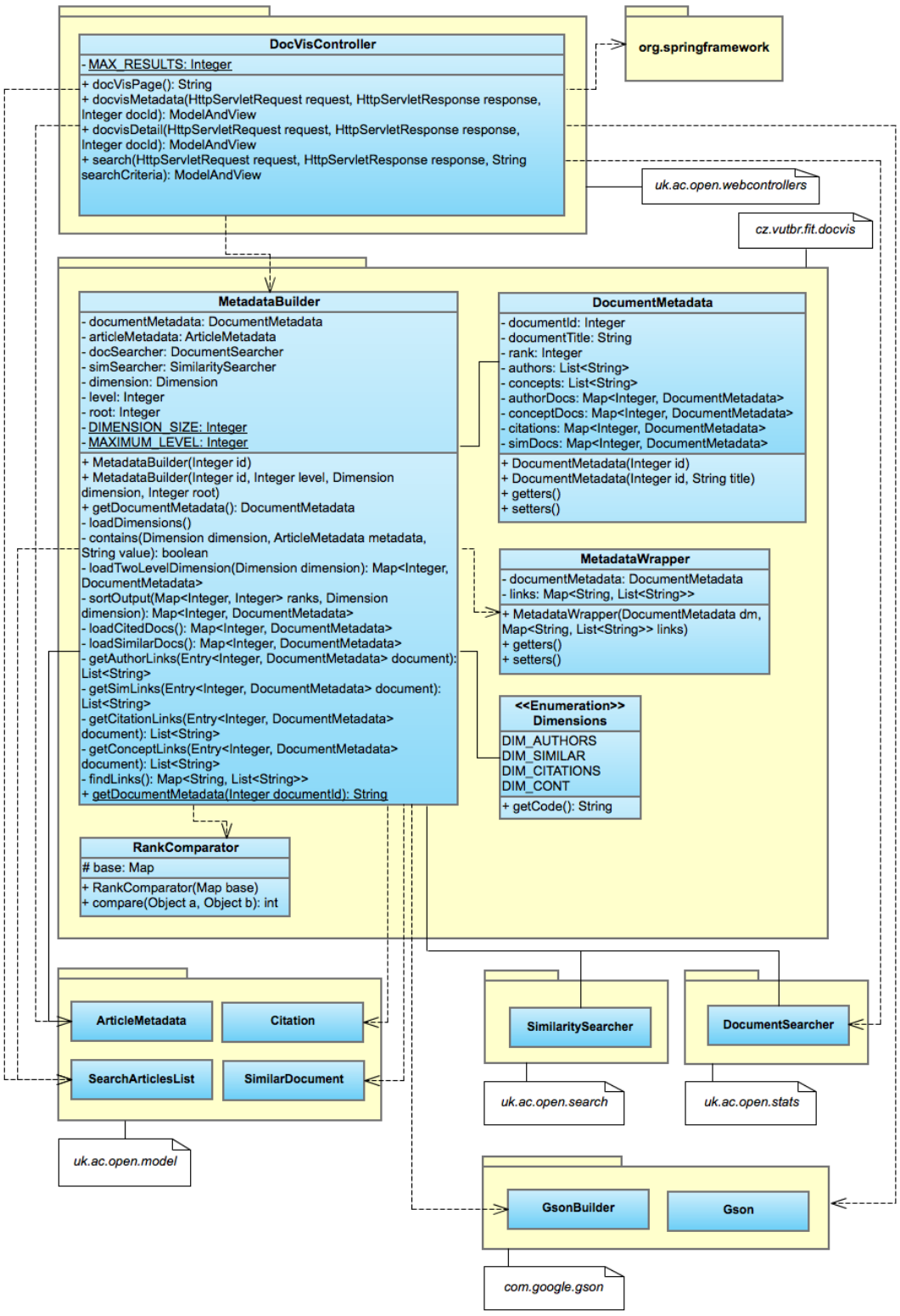


Figure 8.2: Class diagram of the server-side of the application

Chapter 9

Testing and Evaluation

This chapter describes the process of testing and evaluating the application.

9.1 Testing

The visualisation was tested using operating systems Windows 7, Mac OS X Lion and Ubuntu Linux and web browsers listed in table 9.1. The table shows which browser version was used in each system. Visualisation works correctly in all listed browsers.

	Windows 7	Mac OS X Lion	Ubuntu Linux
Google Chrome	20.0	20.0	18.0
Mozilla Firefox	14.0	12.0	11.0
Opera	12.00	12.00	12.00
Safari	-	5.1.7	-

Table 9.1: List of operating systems and web browsers used for testing the visualisation

9.2 Evaluation

After implementing the visualisation I conducted a qualitative evaluation through an anonymous user study to gain data on how the users benefit from the visualisation. The evaluation was done using document collection from the CORE. I focused on comparing how the visualisation supported exploration in contrast to the textual interface that CORE provides. The visualisation was tested and evaluated by six participants, two of them were women. The participants were asked to browse some subject they were interested in and to try to search for some interesting documents that relate to their selected subject. They were asked to perform this action using the textual search interface and the visual search interface. Following that the participants were asked to fill in a questionnaire which consisted of the following questions:

- **Knowledge and skill of participants**
  - How often do you use CORE?
  - Do you use the search feature of CORE?

  - How would you rate your computer skills?
  - How well did you know the subject you were searching for?

- **Comparison of the visual and textual search interfaces of CORE**
  - Would you say that the textual search interface of CORE helped you to find interesting content related to your subject?
  - How many interesting documents you found using the textual search interface?
  - Would you say that the visual search interface of CORE helped you to find interesting content related to your subject?
  - How many interesting documents you found using the visual search interface?
  - How much did the visual search interface help you with search in contrast to the textual representation?

- **Features of the visual search interface**
  - How much did the ability to see related documents divided into dimensions help you with search?
  - How much did the ability to contrast multiple documents help you with search?
  - How much did the ability to switch between union and intersection mode help you with search?

- **Overall rating**
  - How do you rate the usefulness of the graphical representation in general?
  - Please rate the textual and the visual search interface.
  - Would you use the visual search interface in the future, if it was implemented in other document collections, like Wikipedia?

9.3 Results

9.3.1 Knowledge and skills of participants

The first set of questions was aimed at gathering information about the computer skills of the participants, about their previous experience with CORE and with its search feature and their knowledge of the topic they chose to explore. None of the participants was using CORE for searching for documents and majority weren't familiar with the system at all. Only one of the participants stated he was using the CORE search feature. Majority of the participants said their computer skills were above average. The fourth question aimed at determining how well did the participants know the topic which they chose. All of the participants chose a topic they were at least aware of, majority chose a topic they were about moderately familiar with.

9.3.2 Comparison of the visual and textual search interfaces of CORE

The second set of questions was aimed at comparing the textual and visual search interfaces of CORE. It was focused at comparing how do both interfaces support exploration of interesting content. It was interesting to see that only those participants who were familiar

with the system felt that the textual search interface of CORE helped them to find relevant content. This might suggest that the textual search interface is missing features that would support exploratory search. On the contrary all of the participants felt that the visual search interface helped them with exploration and they all managed to find more interesting articles using the visual search interface than using the textual search interface. The last question of this set received the most interesting responses. Some of the responses pointed out that using the textual search interface one must clearly specify the search query in order to receive satisfactory results:

> "...through the CORE textual interface the only possibility to find relevant documents is to put the right title in the query or by browsing through similar documents."

> "I would use classic search for quick searching or in case I know the name of paper (so I know exactly what I want)."

With the visual search interface all participants seemed to agree it can aid to explore interesting content. One of the participants even wrote the visual search is more enjoyable.

> "I got definitely better results with the visual search. It helped me to find related articles."

> "The visual search gives you nice overview on one page and shows the results more clearly."

> "In the visual interface there is in addition a possibility to see documents sharing the same author and concepts, which is helpful. And browsing the documents is much more fun!"

9.3.3 Features of the visual search interface

The third set of questions was focused on discovering what benefit did the features of the visual search interface bring. In general, the participants felt the ability to see different dimensions of a document as well as the ability to compare multiple documents helped them to reveal interesting content. The most positive responses were towards the ability to see related documents divided into dimensions:

> "It made it (the search) quicker and clearer."

> "This was the most helpful feature for me."

Also the ability to compare and contrast multiple documents at once was in general regarded helpful. One of the participants felt it would be more significant on a larger collection of documents. The participants were a little more sceptical about the union and intersection modes, from the responses it seemed this wasn't an often used feature.

> "...if the collection was larger this could help to refine the results."

> "I tried it once, but the intersection mode seems quite useful."

9.3.4 Overall rating

The overall rating of usefulness of the visual search interface was mostly positive. It was felt the visual search interface has the potential of helping to discover more relevant content — one participant used the words "...when I don't know the exact name of what I am looking for..." However it seemed that some of the participants, especially if they had no previous experience with the CORE, took some time to learn how to work with the visualisation. One of the participants expressed this with the following words:

> "I think one needs to get used to this kind of interface to get the most out of it. Using of some features is not very intuitive and can be confusing at the beginning, but after little introduction it turns into a useful tool."

All participants agreed they would be interested in using the visual search interface in other collections.

9.4 Conclusion

The results of the evaluation has shown the interest of the participants in using the visual search interface. The study suggests that some form of visualising search results in order to show relations in the content would be a welcomed feature. Unfortunately most participants rated their computer skills as high so the evaluation is missing an opinion of a technically less skilled person. As some of the participants expressed, the visual interface had a learning curve, which might be in case of less skilled person even slower. This could be a major drawback of the visual interface and should be in the future addressed by possibly providing the novice users with a manual on how to start working with the interface or by providing them with a simpler version of the interface (maybe with some features disabled).

It would be also interesting to apply the visual search interface in more document collections, preferably in a collection with many users. Because most of the participants were not familiar with the CORE it took them a while to discover what types of documents does the collection offer and what could they search for.

The overall rating of the visual search interface seemed positive, especially in comparison with the textual search interface of CORE. It seemed the participants preferred the visual search interface for its ability to show different dimensions of a document and to show connections between documents.

Chapter 10

Discussion

10.1 Application in other domains

In Chapter 7, I have described the functionality of the visual search interface and demonstrated how it can be applied in the domain of research papers. The functionality and the interface design are universal and can be used in any document collection. For example, the interface could be applied on a collection of news articles. The dimensions in this case might be time, location, topic, author, links to other news articles, etc. Even though the use of the interface was demonstrated on a domain with four dimensions, the principles and the functionality are the same. The only difference is in the number of dimensions. The maximum number of dimensions is in theory not restricted, the only restrictions being the size and resolution of the screen and the limitations of human perception. If more views than what can fit on the screen need to be visualised, the interface should allow the user to select the desired combination, but should not allow visualising more than the maximum number to keep the interface simple and legible.

10.2 Project contribution

In Chapter Related Work 4, the document collection visualisations were divided according to the granularity of information they provide about the collection. I mentioned some collection-level visualisation tools for visualising collecitons of research papers. Particularly I mentioned the ASE [11], NVSS [28] (Figure 4.8) and GRIDL [29] tools. While these tools can be classified as *collection level* visualisations, the visualisation developed by this project provide a *document level* visualisation.

I also mentioned some document level visualisations which provide a visualisation of the local subgraph surrounding a specified document. In contrast, the developed tool provides a view on multiple dimensions of a specified document (or a set of documents) and relations between these dimensions. Another difference is that the designed visual search interface allows search results in these dimensions to be ranked, ordered according to their relevance and paginated, preserving a key feature of traditional search interfaces. This feature is difficult to provide in visualisations of the local subgraph surrounding a specified document and I am not aware of any such interface that would support it.

10.3 Support of exploratory search

In the design of the presented visual search interface, I aimed at addressing some of the main issues of current search interfaces. The presented interface addresses the two problems mentioned in the introduction: *information overload* and "*lost in hyperspace.*" The interface mitigates *information overload* in two ways. It (a) helps the user to identify different types of connections between documents and (b) it also helps to explain their meaning. I believe this makes it easier for the user to find important information and comprehend it. It also prevents the users from "getting lost" in the document space by allowing them to add new documents into the stack without the necessity to leave the current position.

The connections in the visual interface correspond to correlations between dimensions. Current search engines typically evaluate the relevance of a user query with respect to all these dimensions at once, which might make it more difficult for the user to discover these correlations. While this behaviour of search engines is often desirable (as it hides complexity), I believe it is not always the case when exploratory search is needed. For example, in the domain of research publications, if there is a large number of connections between similar documents and citations (relevant to the documents in the stack) indicating a strong correlation between these two dimensions, it probably means that the citations used in the research papers cover well the visualised domain. If this correlation appears between authors and citations, but does not appears between other dimensions, it might indicate that authors do not refer to similar work but rather cite their own papers. These correlations are difficult to spot when using traditional search engines.

10.4 Natural language processing tools for generating metadata

In Chapter 6 I mentioned some of the problems I had to face during the design and development of the visual search interface. One of the main problems were incomplete or even missing metadata. We might only have a collection of documents without any metadata at all. In such cases in order to build a visualisation or to be able to search the collection it is necessary to extract the metadata from the documents. Various natural language processing methods and tools exist that can help with this task, from online tools to libraries for different languages. Implicit metadata like title of the document, author or year of publication doesn't typically change and therefore can be extracted and stored before the visualisation is used. However we might also require to use implicit metadata like semantic similarity of documents or information about document clusters and which cluster does the document belong to. This information typically changes with the changing content of the collection. For example, if the collection is constantly growing and new documents are being added, we might discover new clusters or new similar documents. In such case it is necessary to re-calculate the similarities or the clusters every time new documents are added or every time the visualisation is used. Calculating this metadata every time new documents are added might be very time consuming if the documents are added to the collection very often. On the other hand if this metadata is calculated when the visualisation is requested it can slow down the visualisation and worsen the user experience. The type of collection and the way it is used should be the factors for deciding how the metadata should be extracted and/or calculated.

Chapter 11

Conclusion

The task of this project was to create a visualisation of documents in a collection that would support content exploration and exploratory search in this collection. This visual search interface was required to be general and applicable in any document collection, regardless of the type of documents in the collection.

I analysed common design principles of document visualisations and, based on these principles, I managed to design and develop a novel document level query focused visual search interface and to demonstrate its application in two different document collections — in a collection of scientific publications and in cultural heritage collection. I also analysed common design principles for creating visualisation and listed these principles in the thesis. The contribution of my approach is in the combination of the following aspects: support for comparing and contrasting content, support for the discovery and exploration of content across dimensions, and adaptability of the visual interface to different domains.

Following the implementation of the visual search interface I also conducted a qualitative evaluation. The evaluation has shown that the designed visual search interface helped exploratory search, which was the main aim and focus of the visualisation. According to results of the evaluation the users would be interested in using the visual search interface in the future.

As a future work, I would like to provide the technical support for reusing the visual search interface in other domains by the means of an API. This API would make it possible to use this search interface in different types of document collections. The user would provide a definition of dimensions (a description of the types of information shown in each dimension) and a response for each method listed in Table 8.1. The API would then build a visualisation on top of the metadata received from the server.

In Section 9 I presented results of a qualitative evaluation done on the domain of scientific publications. I would also like to perform a user study of the visualisation implemented to a different collection. The results of the user study should help to fine-tune the approach and to demonstrate the usability of the application in different collections.

Bibliography

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 5–14, New York, NY, USA, 2009. ACM.

[2] Ricardo A. Baeza-Yates and Berthier Ribeiro-Neto. *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[3] Robert P. Biuk-Aghai. Visualizing co-authorship networks in online wikipedia. In *Communications and Information Technologies, 2006. ISCIT '06. International Symposium on*, pages 737 –742, 18 2006-sept. 20 2006.

[4] Andrei Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, September 2002.

[5] Nan Cao, Jimeng Sun, Yu-Ru Lin, David Gotz, Shixia Liu, and Huamin Qu. Facetatlas: Multifaceted visualization for rich text corpora. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):1172–1181, November 2010.

[6] Allison J.B. Chaney and David M. Blei. Visualizing topic models, 2012. Department of Computer Science Princeton University, Princeton, NJ USA.

[7] Duen Horng Chau, Aniket Kittur, Jason I. Hong, and Christos Faloutsos. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 167–176, New York, NY, USA, 2011. ACM.

[8] Chaomei Chen and Yue Yu. Empirical studies of information visualization: a meta-analysis. *Int. J. Hum.-Comput. Stud.*, 53(5):851–866, November 2000.

[9] World Wide Web Consortium. Html5: A vocabulary and associated apis for html and xhtml, July 2012.

[10] Alexis Deveria. Compatibility tables for support of html5, css3, svg and more in desktop and mobile browsers, July 2012.

[11] R. Gove, C. Dunne, B. Shneiderman, J. Klavans, and B. Dorr. Evaluating visual and statistical exploration of scientific literature networks. In *Visual Languages and Human-Centric Computing (VL/HCC), 2011 IEEE Symposium on*, pages 217 –224, sept. 2011.

[12] Michael Granitzer, Wolfgang Kienreich, Vedran Sabol, Keith Andrews, and Werner Klieber. Evaluating a system for interactive exploration of large, hierarchically structured document repositories. In *Proceedings of the IEEE Symposium on Information Visualization*, INFOVIS '04, pages 127–134, Washington, DC, USA, 2004. IEEE Computer Society.

[13] Miniwatts Marketing Group. World internet users and population stats, March 2011. Accessed: 2012/01/03.

[14] Yusef Hassan-Montero and Victor Herrero-Solana. Improving tag-clouds as visual information retrieval interfaces. In *Merida, InScit2006 Conference*, October 2006.

[15] Susan Havre, Elizabeth Hetzler, Paul Whitney, and Lucy Nowell. Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization and Computer Graphics*, 8(1):9–20, January 2002.

[16] Drahomira Herrmannova and Petr Knoth. Visual search for supporting content exploration in large document collections. *D-Lib Magazine*, 18(7/8), July/August 2012.

[17] Christian Hirsch, John Hosking, and John Grundy. Interactive visualization tools for exploring the semantic graph of large knowledge spaces, 2009.

[18] Daniel A. Keim. Visual exploration of large data sets. *Commun. ACM*, 44:38–44, August 2001.

[19] UC Berkeley Visualization Lab. Prefuse flare, January 2012.

[20] Simon Lehmann, Ulrich Schwanecke, and Ralf Dörner. Interactive visualization for opportunistic exploration of large document collections. *Inf. Syst.*, 35(2):260–269, April 2010.

[21] Dirk Lewandowski. Query types and search topics of german web search engine users. *Information Services and Use*, 26(4/2006), 2006.

[22] Gary Marchionini. Exploratory search: from finding to understanding. *Commun. ACM*, 49(4):41–46, April 2006.

[23] David N. Milne and Ian H. Witten. A link-based visual search engine for wikipedia. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, JCDL '11, pages 223–226, New York, NY, USA, 2011. ACM.

[24] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[25] Edward W. Christensen Paul E. Bierly III, Eric H. Kessler. Organizational learning, knowledge and wisdom. *Journal of Organizational Change Management*, 13(6), 2000.

[26] Daniel E. Rose and Danny Levinson. Understanding user goals in web search. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.

[27] Marc M. Sebrechts, John V. Cugini, Sharon J. Laskowski, Joanna Vasilakis, and Michael S. Miller. Visualization of search results: a comparative evaluation of text, 2d, and 3d interfaces. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pages 3–10, New York, NY, USA, 1999. ACM.

[28] B. Shneiderman and A. Aris. Network visualization by semantic substrates. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):733 –740, sept.-oct. 2006.

[29] Ben Shneiderman, David Feldman, Anne Rose, and Xavier Ferré Grau. Visualizing digital library search results with categorical and hierarchical axes. In *Proceedings of the fifth ACM conference on Digital libraries*, DL '00, pages 57–66, New York, NY, USA, 2000. ACM.

[30] Ruihua Song, Zhenxiao Luo, Jian-Yun Nie, Yong Yu, and Hsiao-Wuen Hon. Identification of ambiguous queries in web search. *Inf. Process. Manage.*, 45(2):216–229, March 2009.

[31] F. van Ham and A. Perer. „search, show context, expand on demand“: Supporting large graph exploration with degree-of-interest. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):953 –960, nov.-dec. 2009.

[32] Furu Wei, Shixia Liu, Yangqiu Song, Shimei Pan, Michelle X. Zhou, Weihong Qian, Lei Shi, Li Tan, and Qiang Zhang. Tiara: a visual exploratory text analytic system. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 153–162, New York, NY, USA, 2010. ACM.

[33] Ryen W. White, Bill Kules, Steven M. Drucker, and M. C. Schraefel. Supporting exploratory search, introduction. *Communications of the ACM*, 49:36–39, April 2006.

[34] James A. Wise, James J. Thomas, Kelly Pennock, David Lantrip, Marc Pottier, Anne Schur, and Vern Crow. Readings in information visualization. pages 442–450, 1999.

[35] Jin Zhang. *Visualization for Information Retrieval.* The Information Retrieval Series. Springer, January 2008.

Appendix A

CD content

This thesis comes with a DVD containing all data of the project. The final version of this text can be found in a root directory of this DVD in PDF format. There are also two subdirectories in this directory:

`doc/` directory

- `doc/client`
  Software documentation of client-side code in HTML format.
- `doc/server`
  Software documentation of server-side code in HTML format.
- `doc/thesis`
  Source codes of this thesis, in LaTEX format.

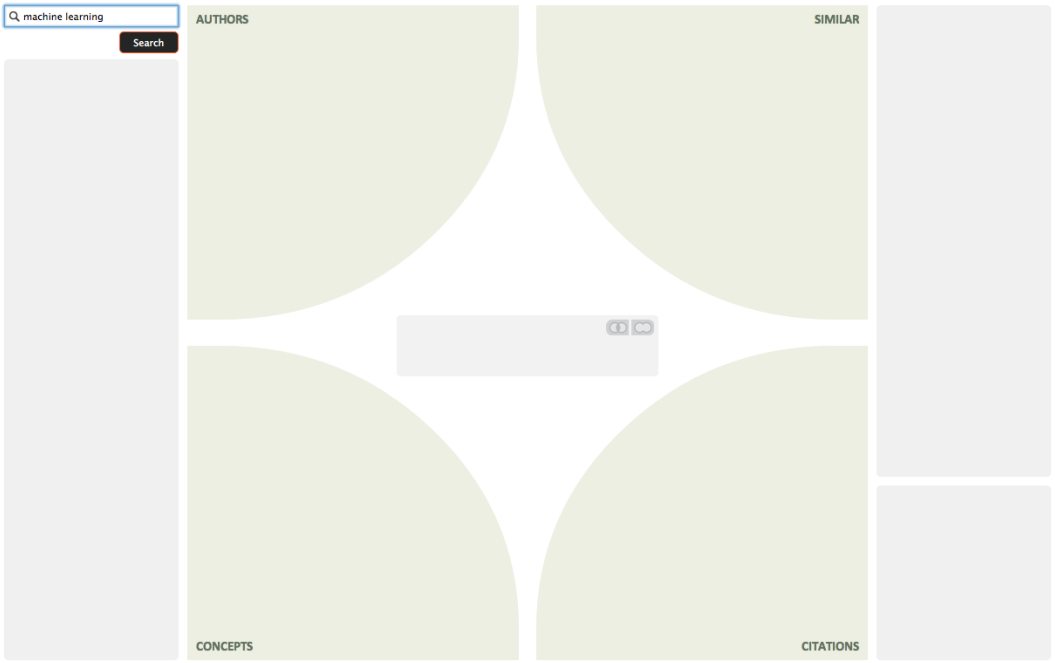
`src/` directory

- `src/client`
  Client-side source codes.
- `src/server`
  Server-side source codes.

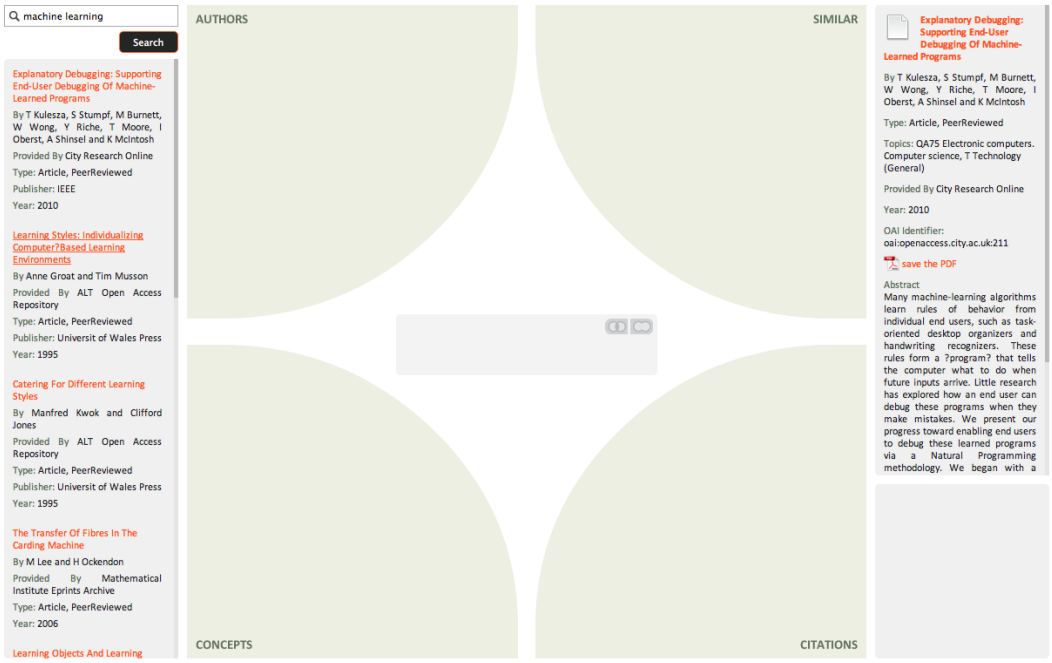
Appendix B

Usage manual

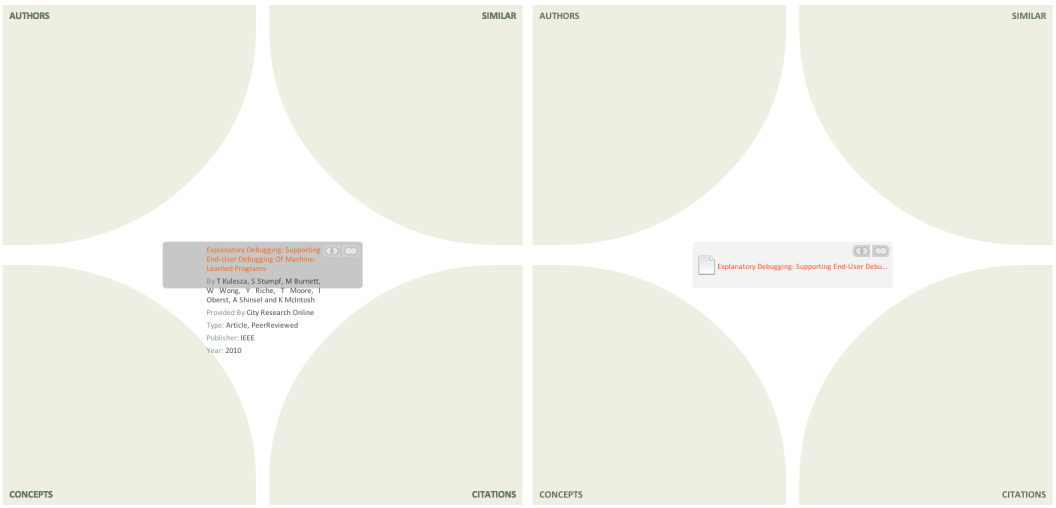
The following chapter constitutes a graphical manual explaining in steps the usage of the visualisation.



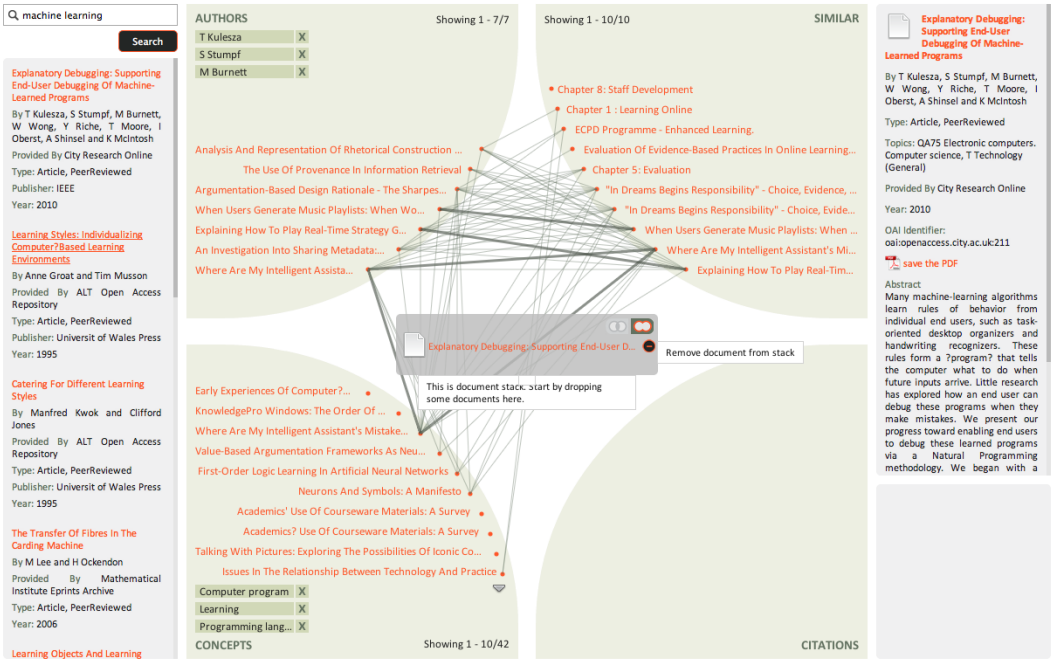
(1) The empty visualisation. The left column of the visualisation features a search box.



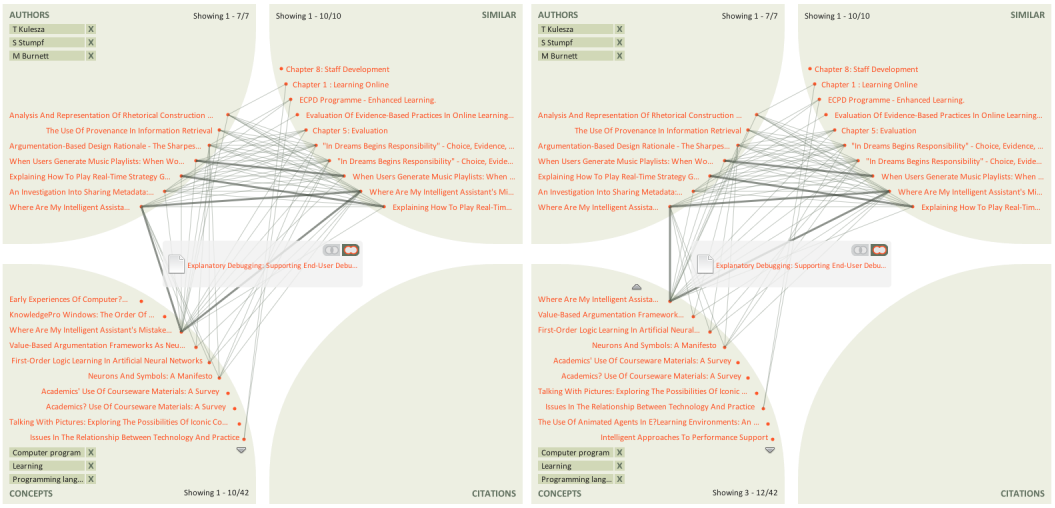
(2) After searching the left column will display the list of results. A header of any document from the result list can be clicked in order to view details of this document in the right column.



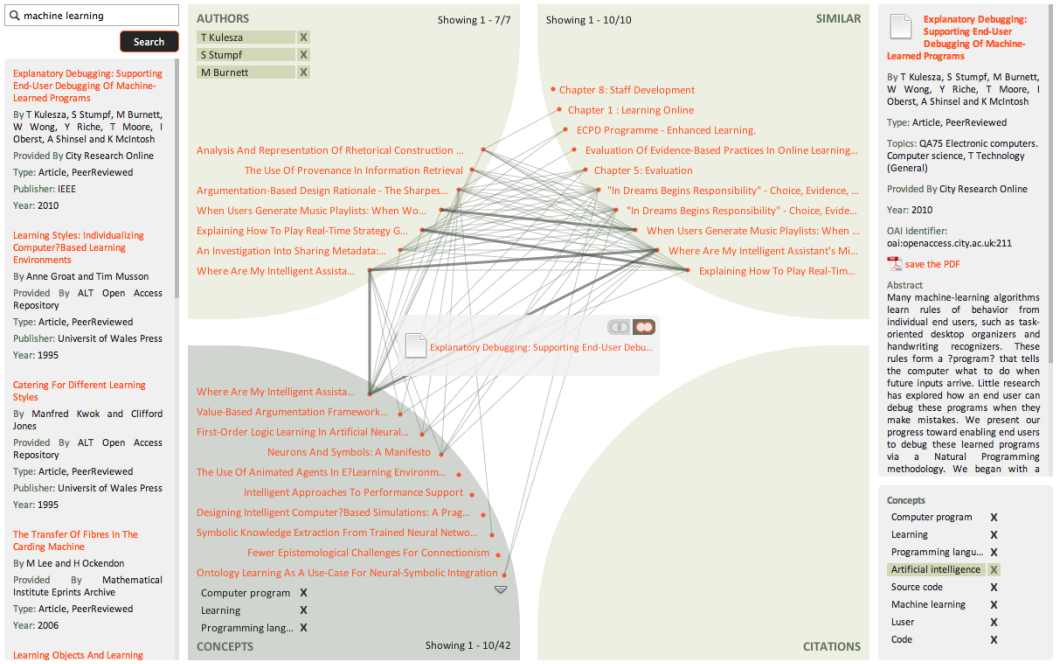
(3) Any document from the left column with search result can be dragged and dropped to the central part of the visualisation called document stack.



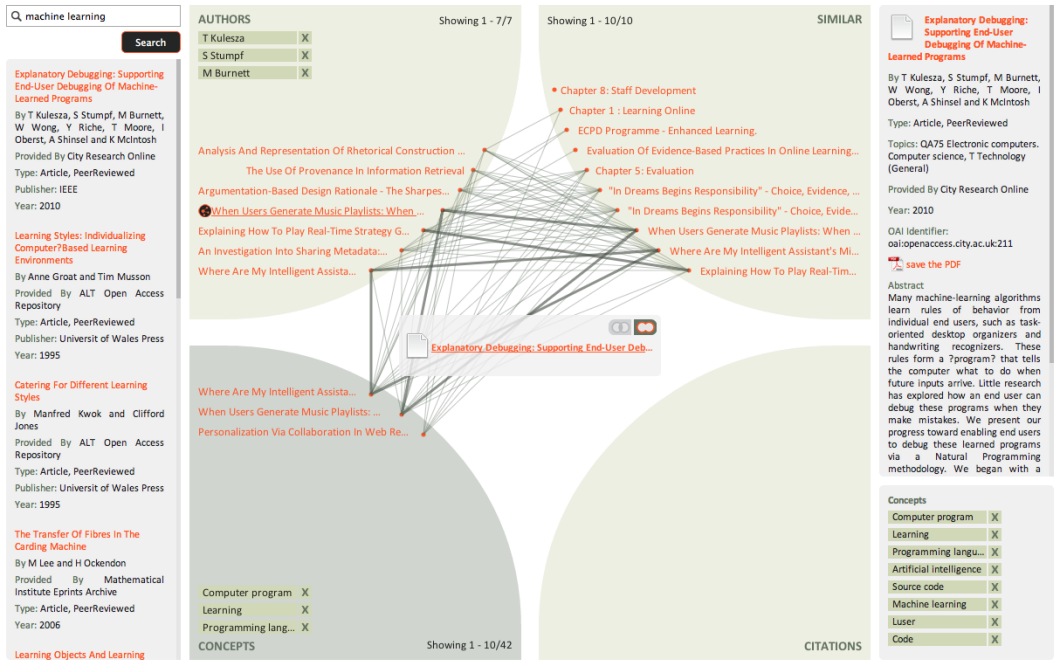
(4) Dropping a document into document stack will initialise the visualisation. Any document can be removed from stack.



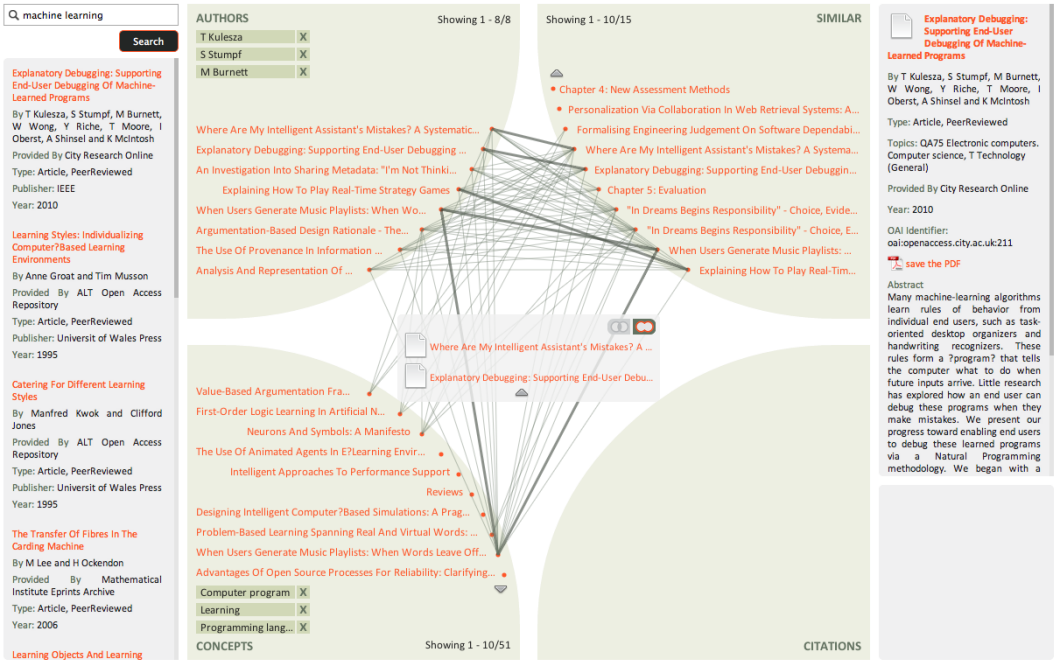
(5) Dimensions, which contain more documents that can be displayed, can be scrolled using the small arrows above and below the documents in the dimension.



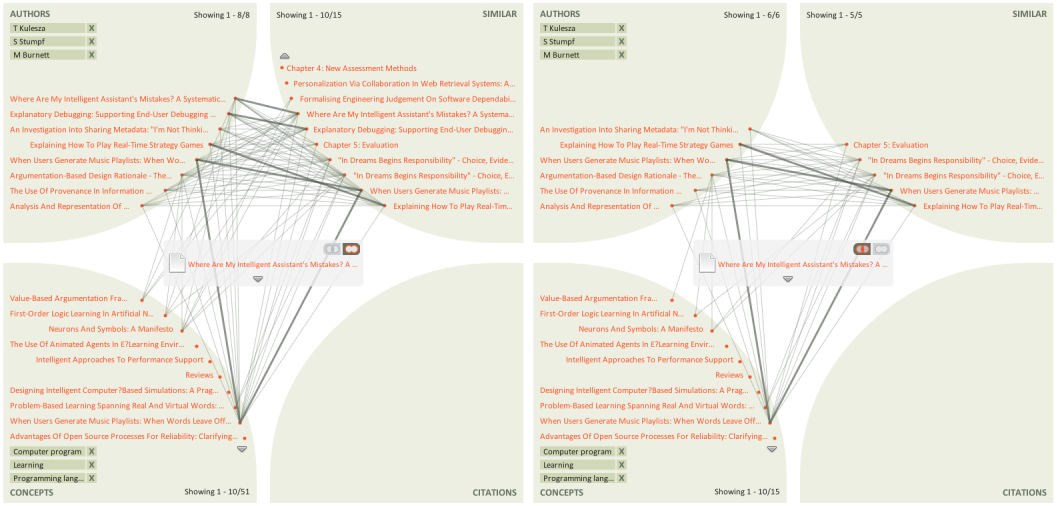
(6) Some dimensions offer additional settings. These settings can be displayed in the right bottom corner of the visualisation by clicking on a dimension. Each of these settings can be switched off or on by clicking on a small cross next to the title. This will fine-tune the results in the dimension.



(7) Dimensions can be filtered also according to shared properties of documents in stack and of any document in any dimension. This filter can be activated by clicking on a small icon which is revealed after hovering over the document.



(8) Multiple documents can be added to stack, either from the search result list or from any dimension.



(9) With more documents in stack it is possible to switch between union and intersection mode which will display either only the related documents shared by all documents in stack or all related documents.