Otto-von-Guericke
University Magdeburg

Faculty of Computer Science

Data and Knowledge Engineering Group

Bachelor's Thesis

Interactive Visualization of Large Concept Lattices for Data Exploration

Author:
Johannes Filter

Advisors:
Prof. Dr. Andreas Nürnberger
Otto-von-Guericke University Magdeburg

Prof. Dr. Ana García-Serrano
Universidad Nacional de Educación a Distancia

July 28, 2015

Abstract

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Inhaltsangabe

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Acknowledgements

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

Acronyms

DT Dynamic Taxonomy. viii, 11, 20–22

FCA Formal Concept Analysis. viii, ix, 1, 3, 6, 7, 11–13, 16–23, 30, 32, 39

IR Information Retrieval. viii, 3, 6, 7, 11, 13–15

1. Introduction

The digital revolution is affecting every part of our life. Also the humanities scholars witness a change in their work life when analog collections are digitized. They have to apply computer science methods to organize and analyze huge amount of data. The area "digital humanities" evolved during the last ten years and it can be defined as an "intersection between the humanities and information technology" [46].

The computer science department of the Universidad Nacional de Educación a Distancia (UNED) in Madrid (Spain) cooperates with human scholars to conduct research in the digital humanities. In this project[1], there are historical maps which have been digitized and annotated. To extract knowledge from the collection, the research group advocates for the use of Formal Concept Analysis (FCA) for topic organization [12, 13]. With this method, the maps are organized on their annotation.

They successfully implemented a FCA algorithm but lack an interactive user interface which will be developed in this thesis. The interface will be evaluated with an user study.

While FCA is a mathematically well-funded principle, the resulting visualization of large concept lattices are a problem. Large concept lattices arose when you apply FCA to large amount of entities (details will be explained in section 2.1). When applying FCA to a document collection, you are likely to encounter huge amount of entities. That is why alternative visualization techniques are needed.

This thesis focuses on concept lattices that were built from document collection. In this scenario, FCA is especially valuable when the user does

[1] `http://linhd.uned.es/p/proyecto-dimh`

not know anything about the collection and wants to browse and explore the data.

This user interface will run in modern web browser to avoid dealing with a heterogeneous operating systems environment. Because of the fast-changing environment of the web, it is important to keep with the latest technologies and techniques to not fall apart. The software utilizes a large amount of frameworks to efficiently create a pleasant user interface.

The remainder of this thesis is structured as follows: The background of Formal Concept Analysis will be presented in Chapter 2 and the background of User Search Interfaces in Chapter 3. In Chapter 4 I will present my (first) approach and the implementation, which will be evaluated in Chapter 5. Built on the Evaluation, I will adjust my work and present an updated version of my work in Chapter 6. I conclude in Chapter 7 and give ideas for future work.

2. Background

Before we can analyze already existing work in this area, draw our conclusions and develop our own system, we have give background information. This chapter gives an introduction into formal concept analysis, followed by an introduction into user interface design principles.

2.1 Formal Concept Analysis

Formal Concept Analysis (FCA) is a mathematically well-funded technique to analyze data. FCA creates relationships among objects specified by attributes. It derives from old philosophical ideas and was formalized by Rudolf Wille [24].

First, we describe the formal definitions and explain them with examples. Second, we give examples for the use of FCA in Information Retrieval (IR).

2.1.1 Definition

FCA is constructed from a formal context. A *formal context* is defined as as a tripple $K = (G, M, I)$ where G is a set of objects, M is a set of attributes and I is a binary relation $I \subseteq G \times M$. I specifies whether an object has an attribute or not. (G and M come from the German words 'Gegenstand' and 'Merkmal'.)

Table 2.1 illustrates an example (from David Eppstein [22]) where G comprises the integers from 1 to 10 and M comprises the attributes composite, even, odd, prime and square.

Table 2.1: Formal context, integers 1 to 10 as objects and attributes

	composite	even	odd	prime	square
1			×		×
2		×		×	
3			×	×	
4	×	×			×
5			×	×	
6	×	×			
7			×	×	
8	×	×			
9	×		×		×
10	×	×			

Let the operator $'$ for $A \subseteq G$ be defined as following:

$$A' = \{m \subseteq M \mid I(g, m)\ \forall g \in A\}$$

A' is the set of those attributes that are present in all objects from given A.

Let the operator $'$ for $B \subseteq M$ be defined as following:

$$B' = \{g \subseteq G \mid I(g, m)\ \forall m \in B\}$$

B' is the set of objects that have at least the attributes given in B.

If for $A \subseteq G$ such that $A = A''$, then A is called *closed*. The same is true for $B \subseteq M$ and $B = B''$.

For example, let a set of objects be defined as $A_1 = \{1, 4\} \subseteq G$. This results into: $A_1' = \{square\}$ and $A_1'' = \{1, 4, 9\}$. A_1 is not closed but $A_2 = \{1, 4, 9\} \subseteq G$ is called close because $A_2 = A_2''$.

A *formal concept* is a pair of (A, B) where $A \subseteq G$ and $B \subseteq M$ and $A = B' \wedge B = A'$. Informally, all objects in A share exactly the same attributes in B. A is a set of objects called the *extent* of a formal concept. B a set of attributes called the *intent* of a formal concept. The extent and the intent of all formal concepts are always closed.

From the example in 2.1, we can derive several formal concepts. Three randomly chosen concepts are shown in table 2.2.

Table 2.2: Three formal concepts from the formal context in table 2.1

Concept	Extent	Intent
C_1	{4,6,8,10}	{composite, even}
C_2	{2,4,6,8,10}	{even}
C_3	{9}	{composite, odd, square}

It is always possible to define an order relation on the formal concepts. Let us introduce the relation $\leq$ as follows:

$$(A_i, B_i) \leq (A_j, B_j) \Longleftrightarrow A_i \subseteq A_j$$

With the help of $\leq$, we can derive relationships from the the concepts in table 2.2. We see that $C_1 \leq C_2$. This means that C_1 is more specific than C_2 and C_2 is more general than C_1. We can also see that C_3 is unrelated to C_1, and that C_3 is unrelated to C_2.

A formal context with $\leq$ is called a *concept lattice* of the context. It can be shown, that for two formal concepts C_i and C_j, there always exists a formal concept C_x such that $C_i \leq C_x \wedge C_j \leq C_x$. That means that there is always a formal concept wich is 'above' in the hierarchy and also related to the two formal concepts C_i and C_j.

The interested reader is advices to read "The Basic Theorem on Concept Lattices" for instance in [9] for an formal explanation.

In the next section we will take a look at the static visualization of concept lattices.

2.1.2 Static Visualization

It is often said that a picture is worth a thousand words. To convey the abstract insights of a concept lattice, it can be visually represented in a *Hasse diagram*. Figure 2.1 shows the Hasse diagram of the concept lattice derived from the formal context described in table 2.1.

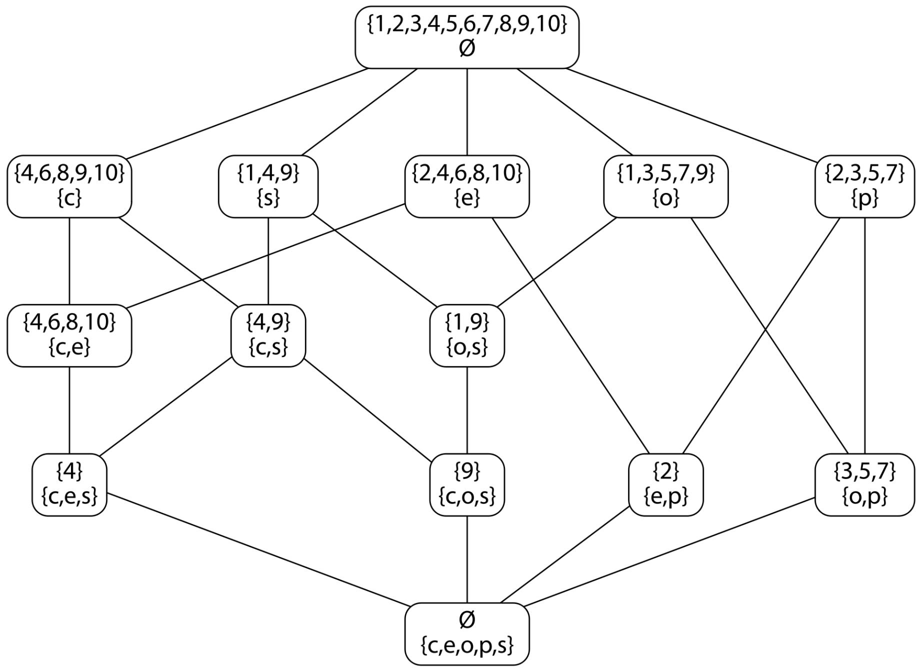

Figure 2.1: Hasse diagram, with the integers 1 to 10 as objects and attributes square (s), prime (p), composite (c), even (e), and odd (o)

A Hasse diagram is a graph where the vertices represent formal concepts and edges represent the relation $\leq$ among the formal concepts. An edge between formal concepts C_i and C_j is drawn, when $C_i \leq C_j$ and there does not exist a formal concept C_x such as $C_i \leq C_x \leq C_j$. To increase the readability, the nodes are ordered in layers. The concepts in top are more general, the concepts in the bottom are the more specific. There are two special concepts: supremum and infimum. The supremum is vertex node in the top and the attributes in its intent are those which are present in all objects. In most cases its intent is empty. The infimum is the vertex in the bottom and the objects in its extent are those which have all attributes. In most cases its extent is empty.

After this general introduction, we will describe in the next section how we can apply FCA to IR.

2.1.3 Application of FCA in IR

The examples depicted in table 2.1 and figure 2.1 are only exemplary. How did FCA affected the real world? According to Poelmans et al. has been FCA "applied in many disciplines such as software engineering, knowledge discovery and IR" [40] and they did two comprehensive surveys on the application of FCA [40, 41].

Carpineto et al.[11] describe the start of FCA in IR:

> In the 80's, basic ideas were put forth - essentially that a concept can be seen as a query (the intent) with a set of retrieved documents (the extent) and that neighbor concepts can be seen as minimal query changes."

This means that the objects are documents and the documents are treated as a set of words. An attribute means that a word occur in this documents. In this simple form, this has a major drawback because it is not possible to we weight terms or for instance to rank among documents with same attributes.

The different algorithms to create concept lattes or an in-depth analysis of FCA for text mining are not covered in this thesis. The interested reader is guided to study the work from Carpineto and Romano [9] for a detailed investigation.

The visual representation of of those IR systems is discussed in the related work section 3. Before we do this, let us get to know some principles of interface design.

2.2 Interface Design

The interaction from the user with the system is what exactly matters to the user. The interaction of humans with computers has its own research area, human-computer interaction, and one of its pioneers is Ben Shneiderman. In the following, two principles from him will be presented: The "Eight Golden Rules of Interface Design" and the "Visual Information Seeking Mantra".

2.2.1 Eight Golden Rules of Interface Design

This rules are genreal advices for user interface designers which should apply to all interfaces. Ben Shneiderman et al. present this rules in their book [44]. The rules are names and explained with my own remarks.

- Strive for consistency: Use similar actions in similar situations. Use identical terminology, colors, fonts etc. throughout the system.
- Cater to universal usability: Design for the needs of a diverse user group (skill level, age, gender and others)
- Offer informative feedback: Give system feedback for every action.
- Design dialogs to yield closure: Sequences of actions should be grouped. Give feedback on completion of a group.
- Prevent errors: Design the system that the user cannot even do errors in the first place. But if she does some, offer instructions how to recover.
- Permit easy reversal of actions: Actions should be undone. This gives the user confidence to explore the system.
- Support internal locus of control: The user should think that she is in charge of control.
- Reduce short-term memory load: Reduce the number of things the user has to keep in mind while using the system.

As you can see, the rules are open to interpretation. There exits alternative principles for instance: Donald Norman's Design Principles [39] or Jakob Nielsen's "10 Usability Heuristics for User Interface Design" [38].

This principles can be applied to all user interfaces. In the next section, design principles will be presented where the user views a large collection of items.

2.2.2 Visual Information Seeking Mantra

The visual information seeking mantra (the Mantra) was introduced by Ben Shneiderman [43] and are based on his experience with past projects. Albeit the Mantra was intended to be a "descriptive and explanatory" [6], "in effect, the Mantra has become a prescriptive principle for many information visualization designers", write Craft and Cairns [15].

The Mantra describes user interface design principles for systems when the "users are viewing collections of items, where items have multiple attributes" [43]. The starting principles are: overview first, zoom and filter, and then details on demand. They will be explained below and added by three other principles.

- Overview: Gain an overview of the entire collection.
- Zoom: Zoom in on items of interest.
- Filter: Filter out uninteresting items.
- Details-on-demand: Select an item or group and get details when needed.
- Relate: View relationships among items.
- History: Keep a history of actions to support undo, replay, and progressive refinement.
- Extract: Allow extraction of sub-collections and of the query parameters.

Some task needs more explanation which I will give in the following with help from related literature.

Zoom and Filter

This task are responsible for reducing the complexity of the data collection. 'Zoom' means that the user focuses on items she wants to see. 'Filter' means that she can hide items which are not interesting for her.

History

It is important to give the user the possibility to easily recover from mistakes or loss of interest. In addition, "it is rare that a single user action produces the desired outcome. Information exploration is inherently a process with many steps, so keeping the history of actions and allowing users to retrace their steps is important." writes Shneiderman [43].

Extract

Once interesting objects are found, the user should have the possibility to extract them from the system. Shneiderman describes printing, emailing or saving the item to the disk as 'extraction'.

2.2.3 Final Remarks

The presented ideas are based mostly on the experience of one person: Ben Shneiderman. The huge number of citations show that his work is influential for a lot of people. But this is not real science. Craft and Cairns [15] are calling for empirical justification of the Mantra. This is an indication that human-computer interaction is only at the start point - there is a still a lot of research to do.

In addition, you have to interpret this rules and adapt them to the current situation. Furthermore, there overall humans are involved, a very complex system which is hardly explored.

3. Related Work

After introducing formal concept analysis in section 2.1, let us review and discuss related work. In the first three sections, we go over a lot of different FCA-based approaches. Eventually, we evaluate one FCA-based real world application in detail in regard to the interface design principles described in section 2.2. In section 3.4, a non-FCA based approach is shown which is very related to FCA: DT and Faceted Search.

3.1 Full Hasse Diagrams

The traditional, static visualization of concept lattices are Hasse diagrams as describes in section 2.1. They are also named *line diagrams*. Eklund et al. [20] conducted user studies and proclaim that non-FCA-experts can read Hasse Diagrams if you fine-tune the Hasse diagram. For instance by choosing appropriate colors, making us of symbols and carefully positioning the vertices in layer.

But in the domain of IR on documents you get formal contexts with a lot of objects. Those Hasse diagrams scale bad for large concepts lattices. Kuznetsov et al. [33] describe this resulting visualization: "Representing concept lattices constructed from large contexts often results in heavy, complex diagrams that can be impractical to handle and, eventually, to make sense of." Especially the high connectivity of the graph results in enormous edge crossing. The image 3.1 shows the first results from data in the digital humanities project. The software XX was used.

The visualization is useless. What can be done to improve the situation? The Hasse can be pruned by reducing the number of vertices. The different techniques are discussed in the next section.

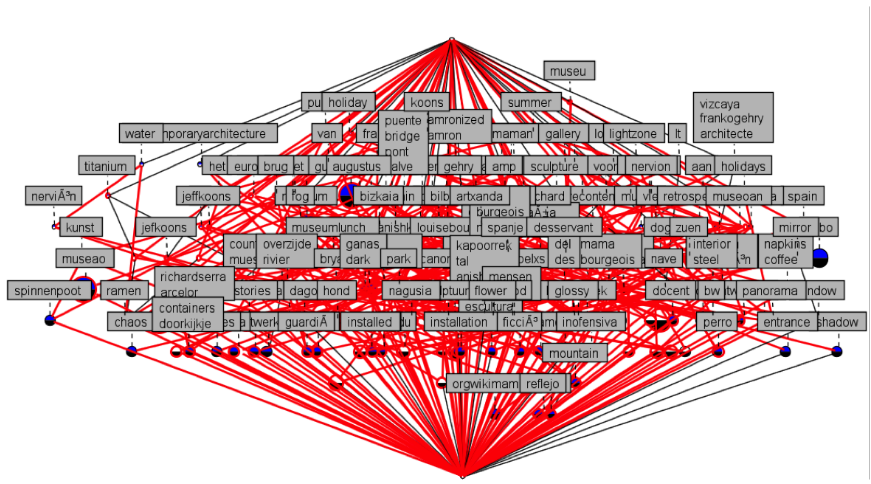

Figure 3.1: First visualization of digital humanist data with traditional FCA software

3.2 Pruned Hasse Diagrams

3.2.1 Reduce Number of Formal Concepts

One way to reduce the number of vertices is to compute the *iceberg lattices* as described by Stumme et al.[45]. They result after the application of a data mining technique "frequent item-set mining" from Agrawal et al.[1]. Only formal concepts are selected which are considered 'frequent'. A formal concept is frequent if its intent, the set of attributes, is frequent. Let B be the intent and $minSupport \in [0,1]$, then B is frequent if $|B'|/|G| \geq minSupport$. This means the attribute set has to specify a high portion of objects; at least $minSupport$.

This approach has some drawbacks as Kuznetsov et al. [33] point out: "One should be careful not to overlook small but interesting groups, for example, "exotic" or 'emergent' groups not yet represented by a large number of objects, or, groups that contain objects who are not members of any other group." They propose to only select 'stable' concepts and write "A concept is stable if its intent does not depend much on each particular object of the extent." [33]. It is also possible to apply traditional cluster techniques like fuzzy K-Means clustering to FCA [2].

While all these techniques undoubtable reduces the number of formal concepts, it is to question if the results are any helpful. In our case of IR, we apply FCA to explore the data and get insights about the lattice structure. When pruning the nodes, you are losing many data relationships, many formal concepts and, consequently, the "power" of FCA as exploratory technique is significantly reduced. When we deal with large concepts lattices, the number of nodes has to be very low if we want to represent them with Hasse diagrams. In the year 2015, the question is not how can I visualize and analyze 16 formal concepts as in figure 2.1 - it is more how can I analyze 160000 formal concepts. For this tasks, this approach is not appropriate.

But this reduction techniques can be useful to reduce the clutter in the lattice. They can be used in combination with our visualizing techniques as described in the upcoming section 3.3.

3.2.2 Nest Formal Concepts

Another idea are *nested line diagrams* - line diagrams are another name for Hasse diagrams. All attributes of a formal context are partitioned into layers. For instance if you just have two layers. For the first layer: built up the Hasse diagram with the attributes of the first layer. For each vertex in the resulting Hasse diagram, built up a Hasse diagrams *inside* the vertex. This secondary Hasse diagrams are built from the objects in each vertex (extent of the formal concept). This can be done for an arbitrary number of layers. An example from Carpineto and Romano [9] is shown in figure 3.2. The general idea should be clear without explaining the context - if not Carpineto and Romano [9] describe it more in detail.

But how to partition the attributes? You have to select them manually. If you have an taxonomy of the attributes you can make use of them. In the case of text mining we do not have an taxonomy and thereby cannot make use of it. Furthermore: FCA is good at building relationships with arbitrary attributes. If you have a taxonomy, there are way better techniques one will be presented in section 3.4: "Dynamic Taxonomies and Faceted Search".

3.3 Local View

Instead of showing the full Hasse diagram, the user can see a local view on the lattice. We will give an overview about the basic idea and applications

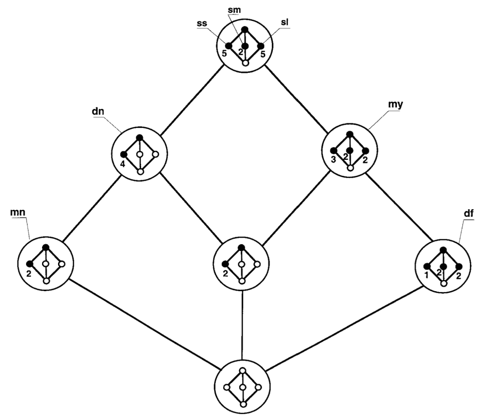

Figure 3.2: Nested line diagram

before we review one real-world application in detail.

One could argue that you just have to visualize everything and then allow to zoom on nodes. This techniques is common among network visualizations [28]. But because of the high connectivity of the graph, this is not helpful to Hasse diagrams. You can see this at the tool 'FCART' presented by Neznanov and Parinov [37]. In figure 3.3 they visualize a concept lattices comprising more than 20000 concepts. Without going into detail and without arguing the general appearance of the user interface we can see, that this kind of approach is doomed to fail. Furthermore, this procedure takes a lot of computation power.

Let us review other, more successful approaches.

3.3.1 Introduction

There exist a lot of different names for a similar approach. Eklund et al. name it *conceptual neighborhood*[19, 21] and Carpineto and Romano name it *hybrid navigation* [8]. But there share the same ideas: the interface is always focused on exactly *one* formal concept. The user can navigate through the lattice by going up (becoming more general) or going down (becoming more special). In the context of IR this means removing terms or adding terms. They also offer the possibility to query the system with a search. In most cases, the user would start with a search and if there exist a corresponding

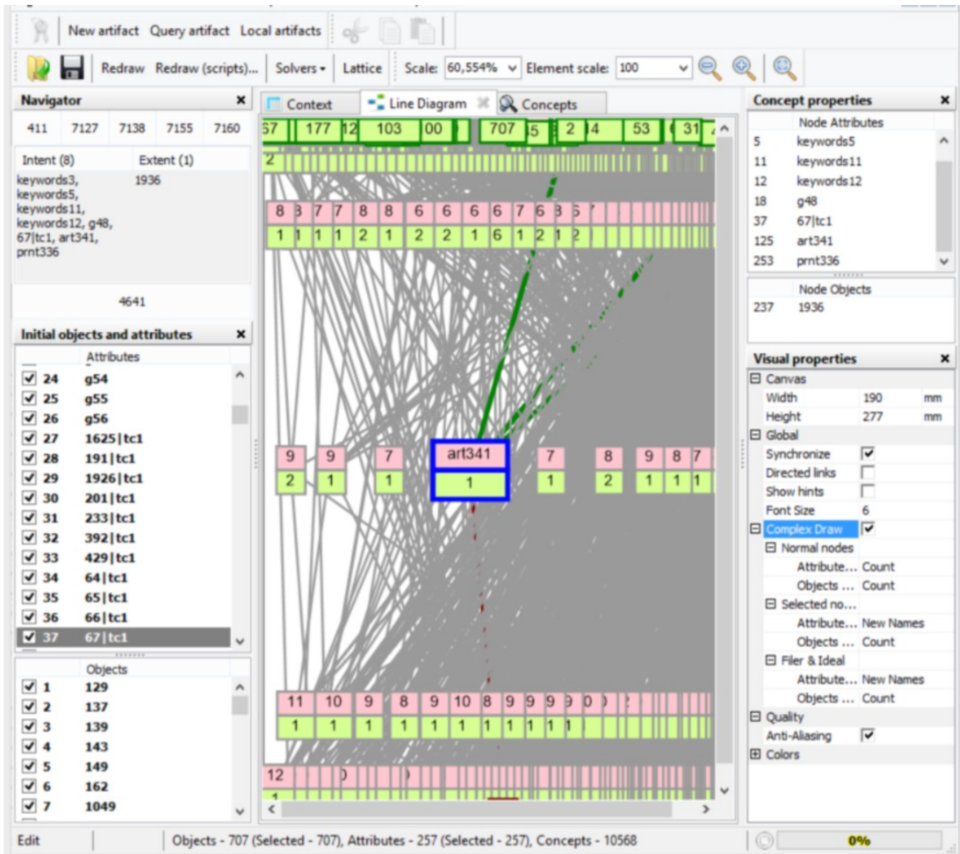

Figure 3.3: FCART visualizing concept lattice with over 20000 concepts.

formal concept it focuses on it. Then the user can fine-tune the search. The idea originated from the IR field and was first proposed by Godin et al. [26].

Two underlying information seeking models stand behind his approach. First, user tend to start with a short query and then refine their needs. Hearst [27] write while referring to [35, 4]:

> A commonly-observed search strategy is one in which the information seeker issues a quick, imprecise query in the hopes of getting into approximately the right part of the information space, and then doing a series of local navigation operations to get closer to the information of interest

Second, it easier for the user to choose from suggestions than to formulate a query. Aula writes [3]:

> Considered in cognitive terms, searching is a more analytical and demanding method for locating information than browsing, as it involves several phases, such as planning and executing queries, evaluating the results, and refining the queries, whereas browsing only requires the user to recognize promising-looking links.

Third, the incremental navigation prevents the user from getting empty result lists. This is related to the design principle from Shneiderman: Prevent the user from making mistakes. But they can still fail when they then search.

Godin et al. [25] evaluated their work in comparison to boolean retrieval and hierarchical retrieval and proclaim that "[their experiment] suggests that retrieval using a Galois lattice structure may be an attractive alternative since it combines a good performance for subject searching along with browsing potential." Galois lattice is a synonym for concept lattice.

3.3.2 Applications

Carpineto and Romano picked up the idea from Godin et al. and developed a FCA search engines ULYSSES [7, 8]. The user can fine-tune what neighboring vertices are displayed by bounding the information seeking space. They are not only showing directly adjacent vertices but also vertices that do not exceed a given distance. It is also possible to restrict the space to vertices which are above, below, left or right of the focus. In figure 3.4 is a screenshot of the system.

In their following work, CREDO [10], Carpineto and Romano followed the look of ordinary search engines. The presentation of the concept lattice is not oriented at the Hasse Diagram. It looks more like a folder structure. Shown in 3.5. Work that is similar comes from Koester [32], Dau et al.[16], Nauer and Yannik [36] and Cigarran et al. [14]. In all this cases, the concept lattice is built on the fly from the result list of search engines. This come with special interest like performance which is not inserting to us. And it is - again like many FCA application - restricted to relatively small number of formal concepts. Let us move on to some application of FCA to document collections.

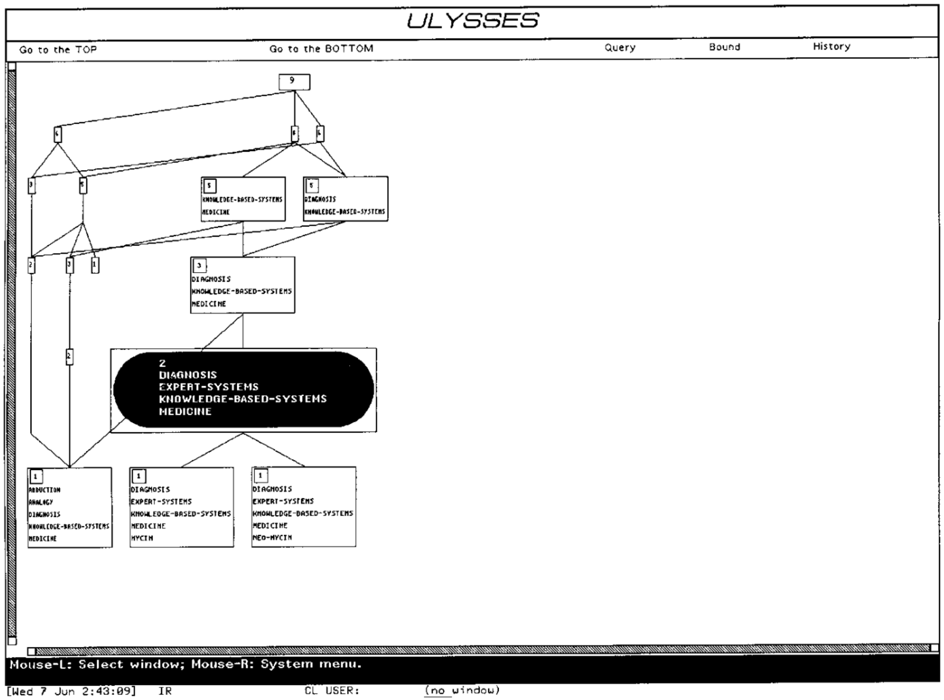

Figure 3.4: ULYSSES with focus on the black node [8]

Eklund et al. applied FCA to email organization[20], image browsing [18, 17] and a later work is the 'Virtual Museum of the Pacific' [19, 21]. Let us review this museum because: it visualizes a static documents collection as we do and it is a rare example of FCA outside academia. It is also web-based like our desired interface. It was created after doing a lot of research into his field. It was built in 2000 - so it is fairly recent. They conducted a usability study with museum experts and non-experts. [21]

3.3.3 Virtual Museum of the Pacific

The museum was created to give user the possibility to browse images of museum objects. It is available on the web [1] and it is advised to take a look at it before continue reading.

After the login it is not clear where to start. If you are completely new

[1] `http://epoc.cs.uow.edu.au/vmp/` - Credentials are required. Use username: filter and password: 45755

Figure 3.5: Screenshot of CREDO, after query 'xml' and browsing after 'editor(5)' and 'windows free(2)' [10]

to the topic, it is a good idea to start with the word cloud. Here comes the Mantra into play: Overview first - details on demand. No real overview is given, this might confuse user.

The interface sets the objects in he focus of the interface. This is, in my opinion, a good decision. The user is probably more interest in the object than on the structure which was built around the objects (the lattice).

Above the objects is the current intent of the formal concept shown. It is possible to remove terms from the current selection to generalize. Below the objects are terms suggested to specify the information needs. When the user clicks on the term the view changes with the newly added item. The sidebar categorizes the attributes/terms and it possible to filter out uninteresting concepts. This categorization contradicts the whole purpose of FCA but might help the users.

To get details on demand, you have to click on an item. From there you

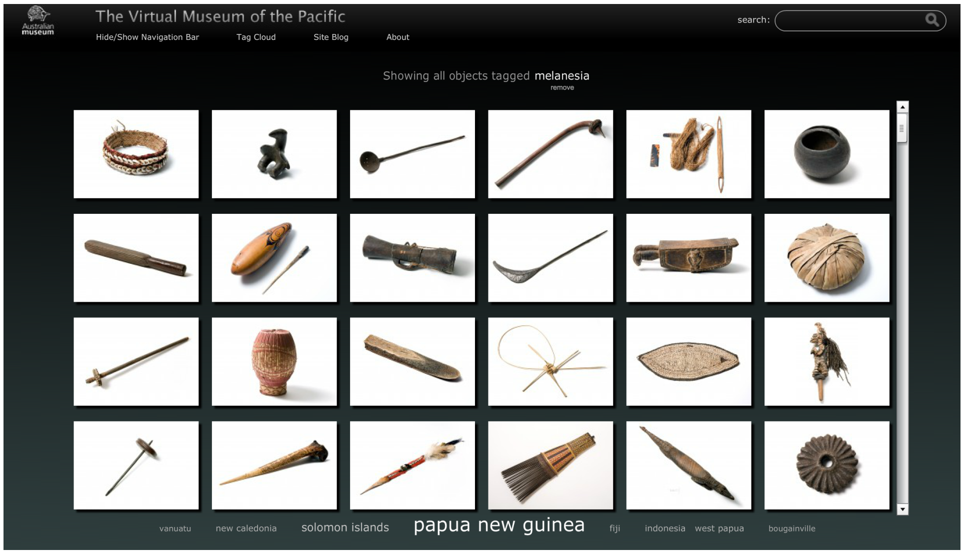

Figure 3.6: The Virtual Museum of the Pacific, focus on concept 'melanesia'

can also find related items.

It is possible to search but this search is very rudimentary. It is very easy to implement a type ahead (autocomplete/suggestion/autofill) for static data set. This is a real problem and there is no excuse not to stick to well-proven user search interface principles as described by Hearst [27].

The biggest problem with the his interface is the missing home/reset button and the absence of any form of history. As Shneiderman [43] pointed out "Information exploration is inherently a process with many steps, so keeping the history of actions and allowing users to retrace their steps is important." FCA is a exploratory technique. Not implementing *any* possible to go back is a utter failure. You can this in the evaluation as they point out [21]: "Users also had difficulty understanding the notion of conceptual navigation as a way of navigating an information space, rather than a fixed hierarchy with a well defined 'home' state." and "[..] many users felt 'lost' within [the FCA-based] style of navigation. A recommendation was put forward so that users could at least back track through the navigation sequences (in the form of a 'Back' button) or that users could easily go back to a 'home' or 'reset' state." They are not going to discuss this problems - even thought the navigation is *the* important problem with FCA. They defend their failure with unfamiliarity of the users [21]: "there were a number of common issues, mostly relating to the unfamiliarity of the interface". This is not a real argument because the users are always unfamiliar with new interfaces.

Combined with meaningless quotes from participants [21] : "There should be more of this kind of thing, but there's room for improvement" or "I think it's really good—it gives specialized information." and a weak summary [21]:

> "Overall, the study concludes that conceptual based browsing can offer significant merit for browsing Pacific collections, and that by resolving the identified issues, the core functionality of searching and browsing via FCA can be considerably easier for users to learn and engage with."

(Easier than what? There was NO comparison with other IR techniques in the whole study) this paper leaves a bitter aftertaste. It started with a good idea but the implementation - in my opinion - is a complete failure. Instead of blaming the stupidity of the users, they should have looked at their own mistakes and learned from them.

3.3.4 Conclusion

While FCA is famos in the academic stuff it does not really help the world at all. There exist only a few useful examples and most of them were conducted on small lattices. Focusing on formal concept on only showing neighboring nodes is the way to built interfaces. Before describing the idea of my interface in detail, let us review some related techniques which is similar to FCA and go into the industry. Feel free to skip this section and go straight to section XX.

3.4 Dynamic Taxonomy and Faceted Search

Sacco and Tzitzikas [42] write: "Although FCA and Dynamic Taxonomy are apparently two distinct approaches to information modeling and access, and they use a different terminology, they are closely related". Completely introducing Dynamic Taxonomy (DT) would exceed this thesis. The interested reader is guided to study the work from Sacco and Tzitzikas [42].

The whole point of this section is to deconstruct a techniques which is called *conceptual scaling*[9]. The technique is applied to many-valued data sets to change them into single-valued data sets. FCA only knows single-valued data sets. Let us take a look at table 3.1.

Table 3.1: many-valued data set

Name	Age	Hometown
Johannes	24	Schwerin
Ana	35	Madrid
Peter	47	Magdeburg

The objects are human beings and they are categorized by the attributes name, age and hometown. With conceptual scaling the attributes for the object Johannes are: "Name_Johannes", "Age_24", "Hometown_Schwerin" and correspondingly for the other people. Now we have only a single-valued data-set which is needed for FCA.

But this method, albeit it is widespread, is, at least for IR, problematic. From a user-centered perspective, complexity reduction is the key point. How do we break massive, complex data set down to make them understandable for the users? So this transformation sounds like a contradiction - and it is. Zobel describes this kind of research as follows:

> Much research — far too much — is just misguided. People investigate problems that are already solved and well understood, or solve problems that technology has made irrelevant, or don't realize that the proposed improvement actually makes the method worse.

So for many-valued data, FCA is not a appropriate technique. The use of facet search is shown in figure 3.7.

You can see that the search is currently narrowed down be restricting location to 'Asia' and Shapes etc. to 'fabrics'. It is possible to continue becoming more specific by choosing attributes and value from the sidebar on the left. Or becoming more general by removing current choices. It is very similar to the local view of FCA described in section 3.3. Sacco and Tzitzikas [42] compare DT to FCA:

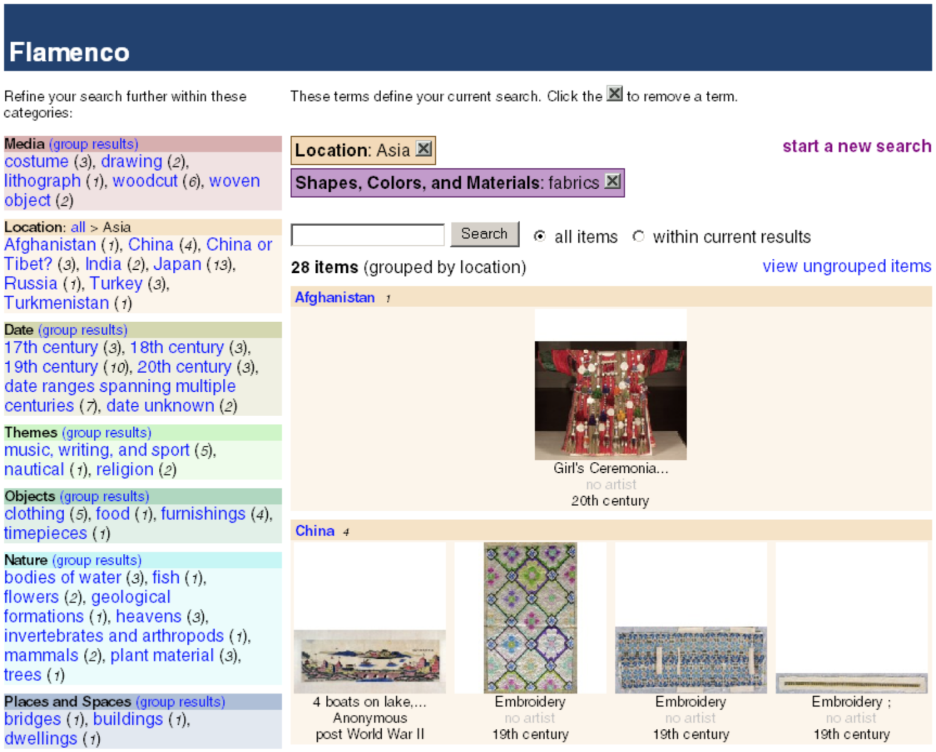

Figure 3.7: Flamenco with ongoing focus

> DT is less informative, but it ensures the results displayed to users is manageable because they are expressed in terms of the original taxonomy. This reduces the cognitive effort of users and is a simpler and more intuitive representation than the line diagrams generally used in FCA.

To summarize the use for IR: Use DT if you have many-value datasets and FCA for single-valued datasets. Let us continues the design of the application of the system.

4. Fancy FCA 1.0

We describe the idea inferred from discussing related work in section 3. Then we describe the implementation of the frontend followed by the backend.

4.1 Idea

The general idea comes from the digital museum but we identified the weaknesses and make it better.

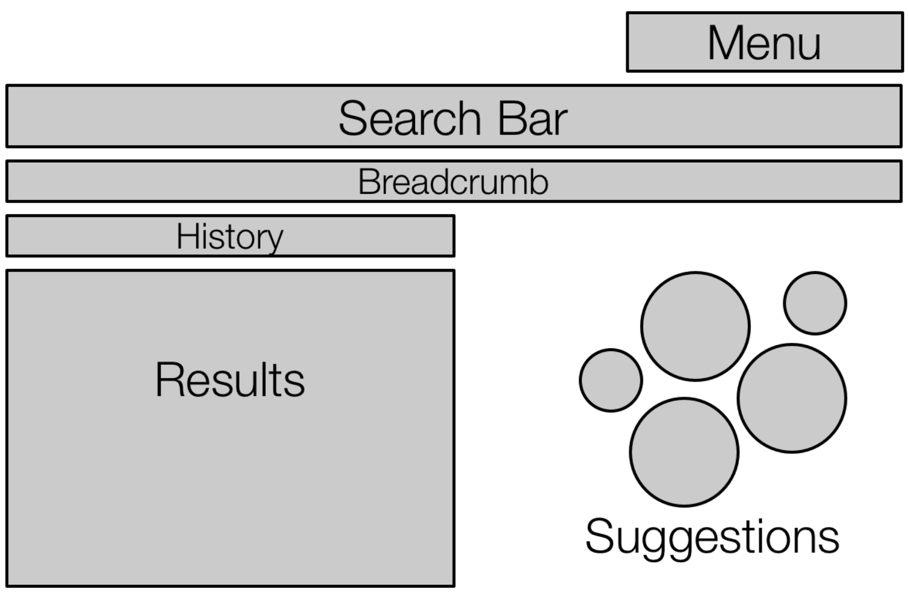

Figure 4.1: The Interface

First, the search interfaces did not followed any design principles. The interface should remind users of modern search engines: Search bar on top vertical result list bottom. Each result has a title withs links to the original page of the map. A snippet is shown with highlighted words from the query. We could not provide thumbnails of the maps because we did not had access to the pictures itself. It was planned to add this feature in later iterations.

Second, there was a lack of history of actions or at least orientation. For this we have two functionalities: Breadcrumbs and Navigation History. Breadcrumbs save the current path of search and allow it to go easily pack. The navigational history allows more possibilities to backtrack. If the user is logged in, it save the history forever. If she is not logged in, it only saves the session history.

Third, the user can save documents which she found interesting.

Fourth, if the user is logged in, all actions are logged and send to the server. This was to evaluate the user behavior.

Fifth, the neighboring concepts (the 'suggestions') are shown right next to the results as a word cloud. The words are presented in a bubble. The font size and bubble radius depend on the documents which are presented in there.

4.2 Frontend

The interface was implemented with web techniques. So HTML, CSS, Javascript. Instead of writing pure HTML, we use the Jade templating engine [1]. Instead of writing Javascript, we use CoffeeScript [2] which compiles to Javascript but offers semantic sugar. For the the word cloud I use D3[3] [5] as background. D3 utilized SVG to draw elements which are also accessible via the DOM API. The word cloud is a customized graph force-directed layout based on the work from Vallandingham [47]. Bootstrap[4] gives the overall good look to the site.

[1] http://jade-lang.com
[2] http://coffeescript.org
[3] http://d3js.org
[4] http://getbootstrap.com

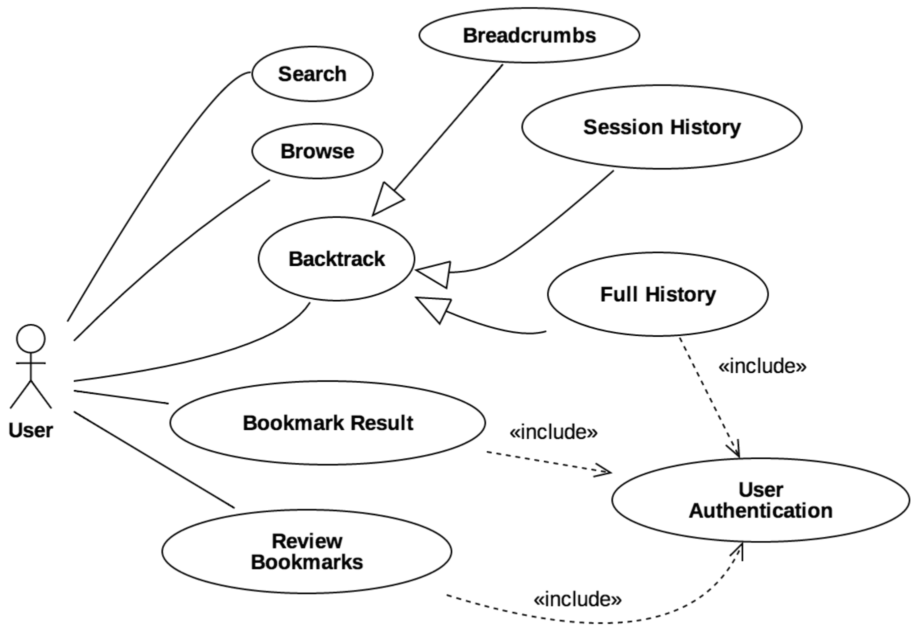

Figure 4.2: The Interface

4.3 Backend

For the backend, Node.JS with the Express Framework [5] was chosen. Reason one, it is good at handling multiple request in parallel. Reason two, I can use the same programming language for frontend and backend: Javascript (CoffeeScript). For data we use MySQl as database because it was already installed on the server. But we use an ORM, Sequelize [6], so you can easily change the database.

4.4 Problems

The whole implementation has one problem: The data is provided as a JSON file. The clients loads the whole file in the beginning. All the navigation happens in the front end. While it a good thing, that during the navigation the client does not have to talk with the server, it is a problem that the client has to deal with over 80 megabytes of data. The server compress the JSON

[5] `http://expressjs.com`
[6] `http://sequelizejs.com`

file to eight megabytes to make the time to download it okay. But it may only run on powerful computers with sufficient RAM because the browser will consume a lot of it. The design decision was taken before it was clear that the JSON file would be so huge. In the beginning it was under ten megabytes.

That the thumbnails of the maps are not presented is a problem. Albeit all the analysis was done on the metadata, a visual representation would be very useful.

5. Evaluation

The design of an interface is highly subjective. User studies can help to evaluate an interface but computer scientists are not experts in human studies. For this, a brief review of different techniques for human studies is given first. Then we describe the setup of the study, describe the results and finally conclude.

5.1 Background

We will only scratch the surface here. A comprehensive introduction into "Methods for Evaluating Interactive Information Retrieval Systems with User" gives Kelly [30]. A shorter introduction gives Hearst in her book on "User Search Interfaces" in Chapter 2. [27].

5.1.1 Introduction

So we want to measure usability, but how is it defined? The ISO 9241-11, 1998 [29] defines three aspects of usability:

- Effectiveness: Accuracy and completeness with which users achieve specified goals.
- Efficiency: Resources expended in relation to the accuracy and completeness with which users achieve goals.
- Satisfaction: Freedom from discomfort, and positive attitudes towards the use of the product.

5.1.2 Experiment vs. Evaluation

It is important to distinguish between the terms 'experiment' and an 'evaluation'. Kelly [30] writes:

> "Evaluations are conducted to assess the goodness of a system, interface or interaction technique and can take many forms [..] Experiments have historically been the main method for interactive system evaluation, but experiments can also be conducted to understand behavior [..] Two important characteristics of experiments are that there are at least two things being compared (e.g., system type) and that some manipulation takes place [..] In some types of [interactive information retrieval] studies only a single system is evaluated. This is a weaker form of evaluation since it is not possible to demonstrate how much better users perform or how different their behaviors and inter actions are since there is no point of comparison. Traditional usability tests are examples of this type of evaluation. Traditional usability tests are usually conducted with a single version of a system, with the goal of identifying potential usability problems."

In this thesis, the system is evaluated to find usability problems. No comparison among other systems are conducted but should be done in further investigations.

5.1.3 Informal Usability Testing

Hearst [27] describes in her book several kinds of studies. We conduct informal usability studies because the research has no experience in studies and we do not have any professional equipment.

Hearst [27] describes the this shortly as "Showing designs to participants and recording their responses". It is often used in short iterative cycles to quickly evaluate a design.

5.1.4 Questionnaires

A questionnaire comprises a set of questions and is cheap and fast way to gather information from people. Kelly et al. [31] describe two types of questions as follows:

> Questionnaires can be comprised of closed questions, open questions or a mixture of both. *Closed questions* are questions that provide a fixed set of responses with which subjects must respond. It is common practice for usability questionnaires to include closed questions in the form of statements such as, the system was easy to learn to use. Subjects are typically provided with 5–7-point Likert-type scales for responding, where one scale end-point represents strong agreement and the other represents strong disagreement. [..] *Open questions*, on the other hand, do not provide a response set and subjects are able to provide any type of response they feel is appropriate.

5.1.5 Thinking Aloud

Kelly [30] writes by referring to Ericsson and Simon [23]: "The think-aloud method asks subjects to articulate their thinking and decision-making as they engage in [interactive information retrieval]". The comments from the participants have to be collected. Either by recording the session or by taking notes. It is hoped that the conductors can learn from the thinking process of the participant. There exist variations of this techniques. Because it can be exhausting, challenging and awkward to talk to yourself all the time, participants are encouraged to report either at some fixed times or when the feel the need.

5.2 Setup

The evaluation was done with five people with a background in humanities. First, an introductory presentation was given in Spanish. Then they were split into two groups. Each group interacted with the interface. The people in the group rotated. There were three tasks they should do. The participants were encouraged to talk out loud. Some times the instructions asked them what they are doing or what they feel. The one group was recored with an iPhone 6 Plus[1]. The session of the other group was not recored and could not be evaluated at all. They had to fill they task what they are feeling. After they session they had to fill in an questionnaire. We use ten closed questions from the USE questionnaire [34] and four common as asked by Kelly [31]. The participants gave the results in Spanish which was translated afterwards.

[1] `https://www.apple.com/iphone-6`

5.3 Results

We group thinking aloud, remarks and open questions together because they all allowed comments on the interface. In praxis the evaluatuon should be done seperatly but because of the bad experiment setup their

5.3.1 Thinking Aloud, Remarks and Open Questions

It turned out, that the evaluation was also a evaluation of the whole research project. All participants mentioned that the interface does not suite their needs. All wanted to refine the search by location and time range and map creators. FCA cannot provide this kind of functionality.

Several participants gave advise how to improve the result list. First, shorten the snippets so that they could read more titles. This could be done by restricting the text length and offering a "show more" button. Second, they mentioned that the thumbnail should be added.

Overall, they give positive comments regarding the interface. They found the navigation with the breadcrumb helpful.

They were not interested in the "Bookmark" feature. Maybe it was because it was an artificial meeting and not a real work environment. They found it a good idea to link to the original site.

There were some problems. They were irritated that they could not search for arbitrary terms. Word for instance like "Valencia" were not in the collection and so also not in the lattice. It was for them not clear that the terms has to be decided before. They suggested that this should get fixed.

They search interface was not used very often. They used the bubble cloud to navigate through the lattice.

One participant mentions that he wanted the see related field to a given field. This would be helpful for his research. FCA can easily provide this method and it is a fail in the interface that it was not included.

5.3.2 Closed Questions

Out of the five participants only three filled out the questionnaire. The results are in the annex B. The results correspond to the comments made by participants. The weakest results got the sentences "It meets my needs" and "It does everything I would expect it to do". The best result got the sentence "It is easy to use".

5.4 Accessibility Analysis

In addition to the the evaluation with users, there was a evaluation regarding accessibility features. It was conducted by people from the UNED: Miguel Angel Marqueta and Covadonga Rodrigo San Juan. It can be found in appendix A. They comments were overall positive buy they highlighted some problems with the interface: It is possible to zoom in the bubble cloud, but there are no buttons on the screen for it. The zooming has to be done with a mouse or a touchpad.

5.5 Conclusions

Zobel [48] proclaims: "Far too many human studies in computer science are amateurish and invalid." and with this study we proved him true. They overall setup was improvised all instructions had no idea about user studies. Even though it cannot stand scientific standards, it is useful as informal evaluation. We can conclude the used method for this collection is in regard to the human experts useless.

For the interface itself, it looks like they liked it. Even though there are some problems like the result list which should be addressed. As a response to this fail, the tool will go away from the fixation onto this one dataset and will visualize arbitrary concept lattice. This leads to the development of Fancy 2.0 which will also address the problems that were found here. The next section described the changes which had be done to the interface.

6. Fancy FCA 2.0

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like "Huardest gefburn"? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

7. Conclusions

Starting from the concept lattices built from metadata from digitized histroical maps, I built and interactive application for exploring the concept lattices. The navigtional

Bibliography

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. *ACM SIGMOD Record*, 22(2):207–216, 1993.

[2] Ch Aswani Kumar and S. Srinivas. Concept lattice reduction using fuzzy K-Means clustering. *Expert Systems with Applications*, 37(3):2696–2704, 2010.

[3] Anne Aula. *Studying user strategies and characteristics for developing web search interfaces.* Tampereen yliopisto, 2005.

[4] Marcia J Bates. Where should the person stop and the information search interface start? *Information Processing & Management*, 26(5):575–591, 1990.

[5] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D 3 : Data-Driven Documents. (March), 2011.

[6] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. Readings in Information Visualization: Using Vision to Think. In *Information Display*, volume 1st, page 686. 1999.

[7] Claudio Carpineto and Giovanni Romano. ULYSSES: a lattice-based multiple interaction strategy retrieval interface. *Human-Computer Interaction*, pages 91–104, 1995.

[8] Claudio Carpineto and Giovanni Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45(5):553–578, 1996.

[9] Claudio Carpineto and Giovanni Romano. *Concept data analysis: Theory and applications.* John Wiley & Sons, 2004.

[10] Claudio Carpineto and Giovanni Romano. Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science*, 10(8):985 – 1013, 2004.

[11] Claudio Carpineto and Giovanni Romano. Using Concept Lattices for Text Retrieval and Mining. *Formal Concept Analysis*, pages 161–179, 2005.

[12] Ángel Castellanos, Ana García-serrano, Etsi Informática Uned, Juan Cigarrán, and Etsi Informática Uned. Concept-based Organization for semi-automatic Knowledge Inference in Digital Humanities : Modelling and Visualization.

[13] J Cigarrán and A Castellanos. A step forward in Topic Detection Algorithms : An Approach based on Formal Concept Analysis . 16.

[14] Juan M. Cigarrán, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo. Browsing Search Results via Formal Concept Analysis: Automatic Selection of Attributes. *Concept Lattices*, 2961:201–202, 2004.

[15] Brock Craft and Paul Cairns. Beyond guidelines: What can we learn from the visual information seeking mantra? *Proceedings of the International Conference on Information Visualisation*, 2005:110–118, 2005.

[16] Frithjof Dau, Jon Ducrou, and Peter Eklund. Concept similarity and related categories in SearchSleuth. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5113 LNAI(October 2007):255–268, 2008.

[17] Jon Ducrou and Peter Eklund. an Intelligent User Interface for Browsing and Searching Mpeg-7 Images Using Concept Lattices. *International Journal of Foundations of Computer Science*, 19(02):359–381, 2008.

[18] Jon Ducrou, Björn Vormbrock, and Peter Eklund. FCA-based browsing and searching of a collection of images. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 4068 LNAI:203–214, 2006.

[19] P. Eklund, P. Goodall, T. Wray, B. Bunt, a. Lawson, L. Christidis, V. Daniel, and M. Van Olffen. Designing the digital ecosystem of the Virtual Museum of the Pacific. *2009 3rd IEEE International Conference on Digital Ecosystems and Technologies, DEST '09*, (June):377–383, 2009.

[20] Peter Eklund, Jon Ducrou, and Peter Brawn. Concept lattices for information visualization: Can novices read line-diagrams? *Concept Lattices. Lecture Notes in Computer Science*, 2961:57–73, 2004.

[21] Peter Eklund, Tim Wray, Peter Goodall, and Amanda Lawson. Design, information organisation and the evaluation of the Virtual Museum of the Pacific digital ecosystem. *Journal of Ambient Intelligence and Humanized Computing*, 3(4):265–280, 2012.

[22] David Eppstein. Hass diagramm, with the integers 1 to 10 as objects and attributes square, prime, composite, even, and odd. `https://commons.wikimedia.org/wiki/File:Concept_lattice.svg`, 2006. Accessed on 16. June 2015.

[23] K A Ericsson and H A Simon. *Protocol analysis: Verbal reports as data*, volume 23. 1993.

[24] Bernhard Ganter and Rudolf Wille. *Formal concept analysis: mathematical foundations*. Springer Science & Business Media, 2012.

[25] Robert Godin, Rokia Missaoui, and Alain April. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. *International Journal of Man-Machine Studies*, 38(5):747–767, 1993.

[26] Robert Godin, C Pichet, and J Gecsei. Design of a browsing interface for information retrieval. In *ACM SIGIR Forum*, volume 23, pages 32–39. ACM, 1989.

[27] Marti a Hearst. Search User Interfaces. *Search User Interfaces*, 54(Ch 1):404, 2009.

[28] Ivan Herman, Guy Melançon, and M Scott Marshall. Graph visualization and navigation in information visualization: A survey. *Visualization and Computer Graphics, IEEE Transactions on*, 6(1):24–43, 2000.

[29] ISO. Ergonomic requirements for office work with visual display terminals (vdts) - part 11 : Guidance on usability. ISO 9241-11, International Organization for Standardization, Geneva, Switzerland, 1998.

[30] Diane Kelly. Methods for Evaluating Interactive Information Retrieval Systems with Users. *Foundations and Trends® in Information Retrieval*, 3(1—2):1–224, 2007.

[31] Diane Kelly, David J. Harper, and Brian Landau. Questionnaire mode effects in interactive information retrieval experiments. *Information Processing and Management*, 44(1):122–141, 2008.

[32] Bjoern Koester. Conceptual Knowledge Retrieval with FooCA : Improving Web Search Engine Results with. *Knowledge Creation Diffusion Utilization*, pages 176–190, 2006.

[33] Sergei Kuznetsov, Sergei Obiedkov, and Camille Roth. Reducing the Representation Complexity of Lattice-Based Taxonomies. pages 241–254, 2007.

[34] Arnold M Lund. Measuring usability with the use questionnaire. *Usability interface*, 8(2):3–6, 2001.

[35] Gary Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4):41–46, 2006.

[36] Emmanuel Nauer and Yannick Toussaint. CreChainDo: an iterative and interactive Web information retrieval system based on lattices. *International Journal of General Systems*, 38(4):363–378, 2009.

[37] A A Neznanov and A A Parinov. About Universality and Flexibility of FCA-based Software Tools. *Proceedings of the International Workshop "What can FCA do for Artificial Intelligence?" (FCA4AI 2014)*, page 59, 2014.

[38] Jakob Nielsen. 10 usability heuristics for user interface design. *Fremont: Nielsen Norman Group.[Consult. 20 maio 2014]. Disponível na Internet*, 1995.

[39] Donald A Norman. *The design of everyday things: Revised and expanded edition.* Basic books, 2013.

[40] J Poelmans and So Kuznetsov. Formal concept analysis in knowledge processing: a survey on models and techniques. *Expert systems with ...*, (2003):1–40, 2013.

[41] J Poelmans and So Kuznetsov. Formal concept analysis in knowledge processing: a survey on models and techniques. *Expert systems with ...*, (2003):1–40, 2013.

[42] Giovanni Maria Sacco and Yannis Tzitzikas. *Dynamic taxonomies and faceted search: theory, practice, and experience*, volume 25. 2009.

[43] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. *Visual Languages, 1996. Proceedings., IEEE . . .*, pages 336–343, 1996.

[44] Ben Shneiderman, Catherine Plaisan, Maxine Cohen, and Steven Jacobs. *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 5th edition, 2010.

[45] Gerd Stumme, Rafik Taouil, Yves Bastide, Nicolas Pasquier, and Lotfi Lakhal. Computing iceberg concept lattices with TITANIC. *Data and Knowledge Engineering*, 42(2):189–222, 2002.

[46] Patrik Svensson. The Landscape of Digital Humanities. *DHQ: Digital Humanities Quarterly*, 4(1):1–34, 2010.

[47] Jim Vallandingham. Building a bubble cloud, 2012.

[48] Justin Zobel. *Writing for Computer Science*. 2004.

List of Figures

List of Tables

A. Appendix: Accessibility Report

Informe de

Descripción/Objeto	Brief Accessibility Analysis for Visualization Tool
Destinatario	Johannes Filter

Editado	Revisado y Aprobado
Miguel Angel Marqueta	**Covadonga Rodrigo**

Índice

1. Destinatario

2. Documentación/herramientas de referencia

3. Contenido

This a brief accessibility analysis for a visualization tool for NLP group of LSI department of UNED. It is only a list of good and bad things pointed out in a first view. It would be necessary, perhaps, a deeper analysis of this tool but in any case it must be considered as a starting point of a better accessibility document.

Good points:

- Good, nice and clear interface.
- Good color contrast. It must be checked in detail for each component with a specific tool (Contrast Color Analyzer).
- Good size and type font.
- Good layout of interface (Disposition Interface in two parts).
- Good decision of edit box. Simple and clear (with suggestion elements).
- Good navigation with tab key. Except for alternating with two display parts.
- Good choice of “heart icon” for “like button”. But it must be checked with Contrast Color Analyzer.

Bad points (They must be checked):

- It is necessary to add tooltips for 4 main links (Home, login…)
- It is necessary to check in accessibility of “right panel” in order to verify whether screen readers can read JavaScript code for “cloud images”. It must be with checked it with tools like “Jaws” or “NVDA” (this last one is free, open source).
- It is necessary to add access key for resizing, (zoom in, and zoom out) now it’s only mouse implemented.
- Although keys in “Navigation History” it’s a good design with breadcrumbs It is necessary to add keyboard navigation by arrows.
- It is necessary change bold font in “No result” advice. It must be checked in with Contrast Color Analyzer to verify it. It is necessary whether screen reader can read this notice when it appears on screen. Probably it will be possible if focus is set to this notice automatically, but it must be checked.

Conclusion:

In general accessibility of this tools is good, but we propose to check in detail this earlier points and to evaluate the full accessibility of these to ensure a better knowledge of the real accessibility state of this tool. At least we propose some changes and to verify some other to ensure a basic accessibility.

B. Appendix: Closed Questions

Results Closed Questions

	I	II	III	IV	V	Average
Age, Gender		60, F	49, F	44, M		
Background	Historian Expert	Historian Expert	Historian Expert	Historian Expert	Historian Expert	
Usefulness						
It is useful.		7	6	7		6 2/3
It meets my needs.		4	5	4		4 1/3
It does everything I would expect it to do.		4	5	4		4 1/3
Ease of Use						
It is easy to use.		7	7	7		7
It requires the fewest steps possible to accomplish what I want to do with it.		5	6	6		5 2/3
I don't notice any inconsistencies as I use it.		4	5	6		5
Ease of Learning						
I learned to use it quickly.		7	7	6		6 2/3
I easily remember how to use it.		7	7	6		6 2/3
Satisfaction						
I am satisfied with it.		7	5	5		5 2/3
It is fun to use.		4	6	7		5 2/3

C. Appendix: Open Questions

- What were the most positive things about using this system and why?
- What were the most negative things about using this system and why?
- How would you improve this system and why?
- Is there anything else that you would like to tell us about this system and your experiences using it?
- In comparison with other systems (Interface of Catálogo Colectivo de las colecciones de Mapas and previous work of this research group). Do you think it is more useful? If yes, which part of the system is more useful.