Formal Concept Analysis and Information Retrieval – A Survey

Victor Codocedo, Amedeo Napoli

Formal Concept Analysis and Information Retrieval - A survey

Victor Codocedo and Amedeo Napoli

LORIA - CNRS - INRIA - Université de Lorraine, France.
victor.codocedo@loria.fr, amedeo.napoli@loria.fr

Abstract. One of the first models to be proposed as a document index for retrieval purposes was a lattice structure, decades before the introduction of Formal Concept Analysis. Nevertheless, the main notions that we consider so familiar within the community ("extension", "intension", "closure operators", "order") were already an important part of it. In the '90s, as FCA was starting to settle as an epistemic community, lattice-based Information Retrieval (IR) systems smoothly transitioned towards FCA-based IR systems. Currently, FCA theory supports dozens of different retrieval applications, ranging from traditional document indices to file systems, recommendation, multi-media and more recently, semantic linked data. In this paper we present a comprehensive study on how FCA has been used to support IR systems. We try to be as exhaustive as possible by reviewing the last 25 years of research as chronicles of the domain, yet we are also concise in relating works by its theoretical foundations. We think that this survey can help future endeavours of establishing FCA as a valuable alternative for modern IR systems.

1 Introduction

Surveying the intersection of Formal Concept Analysis (FCA) [33] and Information Retrieval [3] is not an easy task. The main complexity is that both domains have a an application range so wide that just getting a relevant set of articles to report about is a knowledge discovery process in itself. This is clearly exemplified by the survey presented by Poelmans et al. in 2012 [60] where FCA is used to report on 103 articles related to topics of FCA and IR in a period of only six years (2003-2009) crawled from the Web. In this paper we intend to approach the surveying in a more general and integral manner. We try to answer a very simple question. How have FCA and concept lattices been used in the context of IR applications? We answer this in a chronological narration, trying to cover the last 25 years of research since the first inception of the use of lattice structures to model the space of possible queries (or prescriptions, as they were called) to the last approaches, supporting file systems and semantic technologies.

As we can observe, most of the approaches presented here rest over a limited pool of *ideas and techniques* associated with FCA/IR but applied to a myriad of domains and applications. These ideas are:

1. Using a concept lattice as a model of the description and document spaces
2. Enriching the description space through external knowledge sources

3. Enabling Relevance Feedback
   - Mixing querying and browsing
   - Query-by-navigation
   - Query-by-example
4. Using a concept lattice as a support for automatic retrieval

Our goal in this survey is to catalogue these ideas so future endeavours may have an easier way reaching further domains while developing new different and more interesting techniques. The remainder of this article is as follows: Section 2 introduces some context w.r.t. the use of lattice-based structures in the field of information retrieval. It also introduces the underlying model that generalizes the use of FCA for retrieval purposes. Section 3 describes the first approaches of FCA in the IR domain. Section 4 reviews works using background information to improve retrieval results. Section 5 reviews works based on the paradigm of relevance feedback and automatic document ranking. Section 6 lists the main applications and systems encompassing the ideas and notions described in the previous sections. Finally, Section 7 concludes the paper by introducing some concepts left out of the scope of this paper.

1.1 Related Work

Along with the work of Poelmans [60], there have been other important reviews of the literature regarding FCA and IR [13, 64, 68]. In 2005, Carpineto and Romano [13] described the main possible tasks that FCA could perform regarding querying and indexing by summarizing some of their work in the field. In 2007, Uta Priss [64] dedicated a full chapter to describe the state-of-the-art up to 2004 on FCA-based IR in her paper on *FCA and Information Sciences*. The last of these reviews was presented by Valverde and Peláez-Moreno in 2013 in the first (and sadly, the last) workshop on *Formal Concept Analysis meets Information Retrieval* in the context of the European Conference on Information Retrieval (ECIR 2013)[1]. This work differentiates between what is FCA *in* IR and what is FCA *for* IR, the latter of which refers to the possibility of *"augmenting IR with the methods and ideas of FCA"*. The authors describe these ideas in seven "affordances" of FCA for IR, classifying with them the body-of-work of FCA-based IR approaches.

1.2 Notation and Definitions

Formal Concept Analysis. For the sake of brevity, in this paper we assume a certain degree of familiarity with FCA. In what follows, we use the notation of [33]. A formal context is defined as $\mathcal{K} = (G, M, I)$ where G is a set of documents, M is a set of attributes or descriptors and I an incidence relation set indicating by gIm that document $g \in G$ *has* descriptor $m \in M$. Descriptors denote any kind of metadata associated with documents, being *terms, phrases, symbols, authors, image features, etc.* For the sake of generality in this paper we will refer to M as *the set of descriptors*, unless indicated otherwise.

[1] `http://fcair.hse.ru`

Boolean IR model. The Boolean IR model is considered as the first and one of the simplest techniques to index and retrieve documents [3, 47]. Given a collection of documents G, we can consider each document g as represented by a *conjunction* of Boolean descriptors $g' \subseteq M$, where M is the set of all descriptors (sometimes called "repertory" or "dictionary"). A query (or "request", or "prescription") is defined as a set of descriptors connected by a logical operator AND, OR, NOT. The simplest query is given by a set of descriptors connected by AND and is called a "conjunctive query". Given a conjunctive query Q_{and}, the set of relevant documents to be retrieved (Q'_{and}) are those that contain *at least* all the descriptors in the query. A disjunctive query (using OR) can always be split into its conjunctive parts and the set of relevant documents can be computed by the union of each separate set of relevant documents. A similar approach can be applied for NOT. In this work we will consider every query Q as being conjunctive, unless indicated otherwise.

A query $Q \subseteq M$ is a subset of descriptors usually provided by a given user. In this review we respect the original denominations given by different works to queries (requests, prescriptions, questions, etc.), however we indicate in parenthesis what denominations refer to. Finally, the *"space of documents"* is denoted as $(\wp(G), \subseteq)$ while *"the space of descriptors"* or *"the query space"* is indistinctly denoted as $(\wp(M), \subseteq)$.

2 Pre-FCA history - A lattice to model the description and document spaces

Lattice structures were early adopted by information scientists as a model of document indexing [29, 54]. As early as 1956, Robert A. Fairthorne [29] discussed how to model a library classification system by producing all possible requests (queries) as combinations of categories (descriptors) and logical connectors (AND, OR, NOT) and how this model could be compared to a "free distributive lattice". Some years later, Calvin Mooers [54] would consider two spaces for this model, namely the space of prescriptions P (descriptors) and the space of all possible documents subsets as $L = \wp(G)$.

He realised that L with the set inclusion operator $\subseteq$ was naturally a partially-ordered set (or poset) and that, under certain circumstances (actually when $P = \wp(M)$), P could also be modelled as such. With this, a retrieval system consists in a transformation $T : P \rightarrow L$ that is able to take a prescription (query) into the largest subset of documents that satisfies it (see Figure 1).

It is important to note that Mooers did not describe an actual IR system, but a "model" for retrieval systems that would enable the comparison of different approaches. We can observe that FCA is an instance of this model, where the transformation T is naturally represented by a Galois connection defined between $\wp(G)$ and $\wp(M)$ and where the concept lattice is an elegant solution for the spaces P and L as it represents them in an integrated manner. Particularly, when this Galois connection is defined in terms of the derivation operator ($(\cdot)'$), FCA becomes an implementation of the Boolean IR model.

	patient	laparoscopy	scan	user	medicine	response	time	MRI	practice	complication	arthroscopy	infection
d_1	×	×	×							×		×
d_2			×	×	×	×	×		×	×	×	
d_3		×		×	×			×				
d_4	×				×			×				
d_5				×		×	×					
d_6									×		×	
d_7										×	×	
d_8										×	×	×
d_9											×	×

Table 1: A document-term formal context.

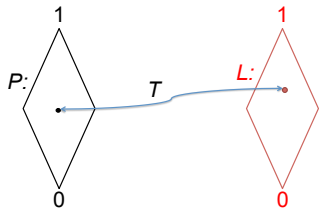


Fig. 1: Mooers' model: *"The space P of all possible retrieval prescriptions (queries), the space L of all possible document subsets, and the retrieval transformation T associating points of P with points of L."*

2.1 The underlying model of FCA-IR

Let us introduce a general model of Boolean retrieval using the FCA framework with an example. In the following sections, we will re-use this model to explain how the tasks of browsing and querying can be performed using a concept lattice. Consider a formal context of documents and descriptors as the one shown in Table 1. Documents for a query $Q \subseteq M$ are retrieved through the derivation $Q' \subseteq G$ which works as the "transformation" T shown in Figure 1. For example, the query $Q = \{arthroscopy, complication\}$ has as an answer documents in $Q' = \{d_2, d_7, d_8\}$.

Key aspects: The query Q can be naturally *extended* to Q'', which of course, contains the same set of answers Q'. In the example, the query $Q = \{MRI\}$ extends to $Q'' = \{MRI, medicine\}$ and they both have the same answer $Q' = \{d_3, d_4\}$. This fact was already discussed by Mooers [54] and has been exhaustively exploited by FCA-IR approaches to provide context to user queries, in this case showing the user that his answer for MRI is within a *medical* context instead of several other possible interpretations [2]. The formal concept formed by (Q', Q'') has been called virtual node, virtual concept or *query concept*, and represents both, the extended query (intent), and the set of retrieved documents (extent). Notice that the latter can be an empty set if there are no documents satisfying the query (hence the name *virtual*). Finally, in this article we will make the distinction between "query extension" and "query expansion". The first of which refers to the *closure* of the query w.r.t. $(\cdot)'$. The second refers to an actual *modification* of the query by taking a set Q_1 where $Q_1'' \neq Q''$ and in general $Q_1 \cap Q \neq \emptyset$ (i.e. finding a query Q_1 related to Q which yields different results).

Throughout all the approaches discussed in this survey, the underlying model described above has not varied much (notice that the book of Barbut and Monjardet which included what will be FCA later was published in 1970! [4, 69]). This fact is in no way a negative point for FCA-based IR approaches, but actually a statement about the adequacy of the model to fit in different tasks and domains. On the other hand, this advantage of FCA is also one of its main drawbacks when dealing with modern IR systems.

[2] `http://en.wikipedia.org/wiki/MRI\_(disambiguation)`

The Boolean IR model was quickly regarded as too limited for the complex tasks involved in the retrieval of documents considering the size of modern document collections or the nature of their descriptions (e.g. numeric instead of Boolean). The IR community would shift to more complex models such as the vector space model (for ranking documents by "relevance" w.r.t. a query) or the probabilistic model (for predicting which are those more "relevant" for a user) . Current introductory books on IR [47, 3] do not mention lattice structures (not to say concept lattices) as valid IR models[3]. In [47], the chapter on the Boolean IR model finishes with the following quote attributed to Calvin Mooers in a book of Fairthorne (1961):

"It is a common fallacy (...) that the algebra of George Boole (1847) is the appropriate formalism for retrieval system design. This view is widely and uncritically accepted as it is wrong."

3 FCA meets IR

The bad scenario for the Boolean retrieval model and its drawbacks did not stop many researchers from developing several different applications using this paradigm. In the '90s, the first FCA-based IR systems were developed, while several other systems based on the use of lattice structures became popular.

3.1 Non-FCA lattice-based IR systems

Pedersen in [58] introduced BRAQUE (BRowse And QUery Environment) as a system that allowed the navigation of a document collection modelled as a *relationship lattice* [59], strongly resembling the features of a concept lattice. At the AT&T Bell labs, Ginsberg [34] introduced WorldViews, consisting *"of a system for automatic document indexing, an information retrieval system and a user interface"* using a taxonomy modelled as a lattice structure. In the work of Bosman et al. [5], a similar approach to Ginsberg's WorldLattice was presented for creating a "Hyperindex" of a *faceted hierarchical thesaurus* using a lattice structure. The lattice supported a "query-by-navigation" approach where the user could "refine" or "enlarge" a query. In the domain of software engineering, Mili et al. [53] proposed a lattice-based index of software descriptions for retrieval purposes based on software reuse needs. The authors describe two types of retrieval namely, *"exact"* which resembles the Boolean retrieval model and "approximate" measuring "proximity" w.r.t a given query.

3.2 FCA-based IR systems

The proposition of Godin et al. [36] revealed the capabilities of concept lattices for indexing and retrieval as an alternative to Boolean querying and hierarchical classifications. This work was built over the initial user interaction design proposed by the same

[3] Actually, in [3] there is an entry of two paragraphs - in a 500 pages book - about lattices in chapter 10 about user interfaces and visualization, referencing [9, 58] as systems for query reformulation (expansion).

authors years before [37, 35] . A major highlight in this work is the efficient browsing capabilities generated from a document collection by the construction of a concept lattice which actually represents a query space. In this manner, the user can pose different queries without explicitly indicating a set of terms to be sought within documents. An important advantage of this model is that users do not have to be completely familiarised with the lexicon used for indexing.

In the same year, Carpineto and Romano presented their system GALOIS [7] for conceptual clustering[4] which would be later implemented for information retrieval purposes through a query browsing interface called ULYSSES [14, 9]. ULYSSES develops further in the model for the unification of querying and browsing plus a third procedure called "bounding". The latter allows the user to restrict the search space within the concept lattice (deriving a sub-lattice) by including into the query sentences such as "all documents indexed by a given term m" (i.e. contained in formal concepts $(A, B) s.t. (A, B) \leq (m', m'')$) and "all documents not indexed by a given term m" (i.e. contained in formal concepts (A, B) such that $B \cap m'' = \emptyset$). Experimentation showed similar results to a plain Boolean retrieval system.

As Fairthorne proposed [29], in an ideal world we could take the descriptions of all the documents in a library and create a map of all the possible requests that could be made (this map would be the P space in Figure 1). However, this is a rather an unlikely scenario as the size of such map grows *"faster than exponentially"* w.r.t. the number of categories [63]. Instead we would prefer to generate a smaller P space that represents the "most meaningful" queries[5]. For this reason, two main strategies were embraced. Firstly, the use of an authoritative source such as the thesaurus-based WordLattice in [34] which would model in a more concise manner the space P. Secondly, the elicitation of this space from document features (lexical properties [5], metadata [58] or terms [36, 14]).

An anecdote. Mooers described the size of the search space of a document collection (L in Figure 1) of one million documents as the number of subsets we can construct from it, being the staggering figure of $10^{310,000}$ [54]. This remembered one of the authors the description of a googol (10^{100}), a number proposed in 1938 by mathematician Edward Kasner to exemplify the difference between "an unimaginably large number and infinity" [6]. While a googol is much larger than the number of particles in the observable universe[7], we can see that the L space is much larger than a googol. Apparently, we were not the first ones to step on this interesting fact. In 1997, a couple of entrepreneurs looking for a name for their search engine, in an attempt to represent the "indexing of an immense amount of data" [8], registered the misspelled version "Google".

[4] Actually, GALOIS is an incremental algorithm for building a concept lattice.

[5] It is worth mentioning that the "meaningfulness" of a request is a matter of perspective. What is meaningful in a domain may not be in another. Meaning also changes with time.

[6] Wikipedia article - `http://en.wikipedia.org/wiki/Googol`

[7] Video about googol from the University of Nottingham - `https://www.youtube.com/watch?v=8GEebx72-qs`

[8] David Koller on the origin of the name "Google" `http://graphics.stanford.edu/~dk/google_name_origin.html`

4 Enriching the description space through external knowledge sources

In the FCA-IR model explained in Section 2.1, attributes are descriptors obtained from the set of documents. As previously explained, this space (P) can be very large but other than that it can suffer from other problems. For example, it can be non-representative of the document set by different reasons (poor document description, poor vocabulary, incompleteness, etc.). Regarding these issues, it may be useful to use an external knowledge source to complement document descriptions. For example, if we are interested in considering synonymia for indexing (e.g. relating documents referring to *"concept lattices"* and *"Galois lattices"*) we may use a thesaurus. If we are interested in considering hierarchical relations (e.g. relating documents referring to *"monkeys"* with those referring to *"primates"*) we may use a taxonomy. If we are interested in considering logical implications (e.g. relating documents written by a French author to those written by a German author using the label *"European literature"*) we may use an ontology.

With these concerns, in 1996 Carpineto and Romano proposed a modified version of the GALOIS system to include "background information" in the form of a thesaurus for document indexing using FCA [8]. The modification was made in the order relation between formal concepts ($\leq_{\mathcal{K}}$) using the order between document descriptors ($\leq_{\mathcal{T}}$) induced by a thesaurus as follows:

$$(A_1, B_1) \leq_{\mathcal{K}} (A_2, B_2) \iff \forall m_2 \in B_2, \exists m_1 \in B_1 \; s.t. \; m_1 \leq_{\mathcal{T}} m_2$$

Furthermore, they redefined the intersection between two descriptor sets as:

$$B_1 \cap^* B_2 = \{m_i \mid m_i \geq_{\mathcal{T}} m_1, m_2, m_1 \in B_1, m_2 \in B_2, m_i \in \mathcal{T}, \nexists m_j \in \mathcal{T}, s.t. \; m_i \geq_{\mathcal{T}} m_j \geq_{\mathcal{T}} m_1, m_2\}$$

From the example in Table 1, consider a thesaurus $\mathcal{T}$ with the relations *arthroscopy, laparoscopy* $\leq_{\mathcal{T}}$ *endoscopy*[9]. Then, $\{laparoscopy\} \cap^* \{arthroscopy\} = \{endoscopy\}$ and we can build the formal concept $(\{d_1, d_2, d_3, d_6, d_7, d_8, d_9\}, \{endoscopy\})$. Consider this analogous to including in the formal context the attribute $endoscopy$ and the relation where each document related either to *laparoscopy* or *arthroscopy* is also related to *endoscopy*.

The authors argue that this approach would lower the complexity associated to computing the concept lattice compared to the more simple approach of adding the thesaurus terms to the initial formal context. In 1997, Uta Priss presented several propositions for a FCA-based IR system in which three main components were discussed [62]. Firstly, a combined formal context comprising document descriptors and other metadata components (e.g. publisher, author, etc.). These kind of fields were coded by many-valued formal contexts which were later scaled (see attribute scaling in [33]). The second component described the inclusion of a thesaurus within the formal context by two approaches, namely by mapping document-descriptor pairs to thesaurus elements, and

[9] Wikipedia categories `http://en.wikipedia.org/wiki/Category:Endoscopy`

by constructing a combined formal context considering documents, descriptors and thesaurus elements in a relational concept analysis (RCA) manner (this RCA proposition is formally different from the one presented by Huchard et al. [40]). The third component referred to the use of "nested line diagrams" to represent in a better manner the combination of different concept lattices in an integrated view offering different description levels within a document collection.

Some of these ideas were later revisited by Cole and Eklund in 1999 [21] where the authors proposed an interactive e-mail retrieval system based on FCA. The formal context was built using "classifier outputs" as attributes which the user was asked to order in a hierarchy (G is a set of emails). *Conceptual scaling* was applied to many-valued attributes deriving views (sub-lattices) that were more manageable for the user to browse than the concept lattice of the entire email collection. In 2003, the authors (plus Gerd Stumme) would propose an extension of their work into a fully integrated system called "HIERMAIL" [22] in which nested-line-diagrams were used to represent conceptual scales (instead of sub-lattices) for knowledge discovery over an e-mail collection. Incidentally, Cole and Eklund had proposed a "folding" and "unfolding" mechanism (using the same notion of conceptual scales) for the concept lattice in a previous work oriented to model a document retrieval system in which documents were indexed by a medical thesaurus called SNOMED [20], although these procedures were not clearly defined.

A similar approach for domain-specific interactive FCA-based IR systems was presented by Mihye and Compton in 2001 [43] and later extended in [44]. An interesting point of this work is that it addresses the fact that taxonomies used to index documents are not static and should evolve through time. For this reason, the concept lattice is used not only to retrieve documents but also to aid users in the annotation of documents and in the evolution of the taxonomy.

5 Relevance Feedback and Automatic Retrieval

5.1 Relevance Feedback

Other than choosing and modelling the kind of data to be used as attributes in a formal context, an important factor in the efficiency of a retrieval system is to help the user closing what is usually called the knowledge or "the cognitive gap" [42]. The cognitive gap describes the distance between the space occupied by the actual information needs of a user and the space occupied by its ability to describe its information needs. For example, consider a user searching for "the book which they made a film about and a wizard appears on it". Somebody could answer "Is it about a girl, a lion, a tin man and a scarecrow?" to which the user may answer "No, there are some kids in a school". Then, the answer could be narrowed down to the 7 books of the "Harry Potter" saga. Here we can see that the cognitive gap can be represented as the distance between the initial query, possibly with the keywords ``book film wizard'', which the user is able to provide, and the query that he actually needs to provide which is ``Harry Potter book''.

In 1971, Rocchio proposed his famous *relevance feedback* model to overcome this issue [65]. In a nutshell, we can see relevance feedback as a "query calibration" system

using extra user inputs. In the previous example, the initial user query was very abstract. Somebody (possibly the librarian), with knowledge about fantasy books asked the user a question based on the assumption that the answer may be "The wizard of Oz". The negative answer provides a feedback of relevance (i.e. "The wizard of oz" is not relevant) which is used to generate the query: `book film wizard school -``the wizard of oz''`[10].

In FCA terms, we can represent this scenario as the join of two object concepts as depicted in Figure 2(b). The initial query yields concept 0 for which the system may propose concept 1 or concept 2.

This approach was proposed by Carpineto and Romano in 1998 through their system REFINER [10]. The user pose a query to which the system generates a "virtual concept in the lattice". By the use of the upper and the lower cover of the virtual concept, REFINER is able to propose minimal query refinements/enlargements (resp.) to the user. Experimental results showed significant better results in the search time employed by a user w.r.t. the Boolean IR model. In 2002, Grootjen et al. [38] proposed a similar rougher approach called "conceptual relevance feedback" further developed as a query expansion method [39] in the lines of pseudo-relevance feedback [47].

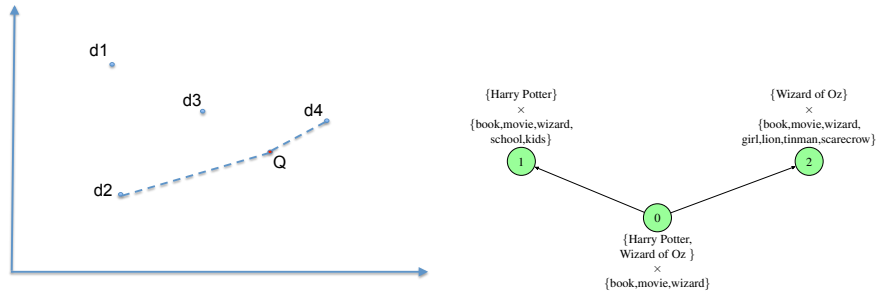
In 2007, Nauer and Toussaint [55] presented a model for "explicit relevance feedback"[11] over a standard Web search engine (such as Google) supported over a concept lattice. This model consisted of a constant iteration of the formal context by "extension" and "reduction" procedures. Extensions were made whenever the user submitted a new query or gave a positive assessment. Reductions were performed whenever the user gave a negative assessment. Explicit relevance feedback was also supported in a previous work by Martines and Loisant [49] for concept lattice-based image retrieval in a similar manner. Users were asked to evaluate images as "good" or "bad". An example of "implicit relevance feedback" can be found in the work of Ducrou et al. [26], supported by a procedure called "query-by-example". Instead of asking the user to give explicit relevance assessments, the query is modified by a sample set manually created by the user.

5.2 Ranking documents

So far we have reviewed approaches that assist the user in navigating the query space and deciding what *is* or *is not* relevant. This is usually achieved by providing an interface that helps them retrieve parts of the concept lattice by the use of "query-by-navigation", "query-by-browsing", "relevance feedback" or "query-by-example". This however is not what we are used to when dealing with search engines. The "file search program" of any operating system, or the mechanics of traditional Web search engines follows a very simple scenario. The user inputs a query and the system outputs a list of documents already ordered by the "relevance" w.r.t. that query. Thus, the system is provided with the notion of *what is relevant* and *what is not*. For instance, in the

[10] In Google query syntax, '-' is used for excluding terms - `http://goo.gl/7RZrQl`

[11] i.e. the user is explicitly asked to make relevance assessments in the system. Opposed to "implicit relevance feedback", where relevance assessments are "inferred" by the interaction of the user with the system.



(a) Vector-space model example. Documents ($d1, d2, d3, d4$) and the query Q are represented as points. Axis are arbitrary description spaces. Dashed lines represent the distance between the query and documents $d2$ and $d4$. The latter is closer to the query than the former meaning that it is more "relevant" for retrieval purposes.

(b) Example of relevance feedback.

Fig. 2: Retrieval paradigm examples.

vector-space model, documents and queries are represented by points in an arbitrary Euclidean-space. "Relevance" in this case may be represented by the distance between a query and a document (the closer the document, the more relevant it is w.r.t. the query) as shown in the example of Figure 2(a) (explaining the meaning of the axis or why the documents and the query are located in the space as they are is out of the scope of this paper. For more information see [47]).

A similar notion was adopted by Carpineto and Romano in 2000 in what they called concept lattice-based ranking (CLR) [11] for a fully automated retrieval system. Using the REFINER model, the virtual concept representing the query is placed in the concept lattice and a series of "concentric rings" around the virtual concept yields a distance that allows to rank documents (e.g. in Figure 2(b), documents in concepts 1 and 2 - and not in concept 0 - are at distance 1, while documents in their super-concepts would be at distance 2). Different measures are also introduced in the work of Ducrou et al. [26] where instead of using the concept lattice structure, differences in extent and intent sets are taken into account. In 2014, Codocedo et al. [18] presented a system for lattice-based ranking using notions of case-based reasoning. This approach inspired in CLR uses the concept lattice to find suitable "query modifications" through pivotal elements called "cousin concepts". Query modifications are evaluated w.r.t. a semantic distance to the original query yielding automatic document ranking. Experimental evidence suggests that such an approach leads to better precision w.r.t. the Boolean querying model and CLR.

6 Applications and Systems

6.1 Applications

Semantic retrieval. How to mix semantic technologies (what is known as semantic web) with IR techniques is still an open question. It is fair to say that modern IR systems are more focused on how to retrieve documents from very large collections than to provide reasoning or inferring capabilities to their engines. Nevertheless, this has not hindered the adoption of some of the semantic web notions such as the knowledge graph in the Google Web engine[12]. Regarding FCA-based IR approaches we can highlight the work of Messai et al. [50] presented in 2005 adapting the ideas of query refinement to support the use of ontologies for generalization purposes. In 2011, Codocedo et al. [17] presented an application of FCA to index songs using semantic similarity among keywords in a concept lattice. In 2012, Ferré et al [30] introduced LISQL, a query language for logical information systems supporting complex relational properties among objects. These ideas were materialised in a geographical information system. Finally, in 2014 the work of Alam et al [2] presents the concept lattice as a classification of SPARQL answers to provide views on linked open data retrieval system.

Recommender systems (RSs). RSs have become increasingly popular at the point that currently, it is an independent research community. Nevertheless, RSs have their roots in IR sharing many notions such as indexing, retrieval and ranking. To phrase it in the terms of [68], an important *affordance* of FCA *for* RSs is the *characterization* it can provide to recommendations, i.e. it can explain why a certain item is being recommended, so the user can have a better experience with the system. This fact was addressed by [41] in 2008 which proposed a system for "well-interpretable recommendations based on FCA" for advertisement keywords using association rules. Previously, in 2006 [23] FCA was used as a method to pre-calculate groups of users that agree in certain groups of items. The notion of *query concept* is in here replaced by the "entry-level concept" of a user or an item. Experimental results suggest that FCA alleviates the otherwise hard task of finding the neighbourhood of a given item or user in the dataset. In 2013, Senatore et al. [66] proposed a recommender system based on an extension of FCA (namely, "Fuzzy FCA" or more precisely, FCA with fuzzy attributes) allowing to include *degrees of similarity* between users (i.e. not just Boolean relations for rating the same item) providing *ranked* recommended items. Finally, in 2014 Castellanos et al. [15] presented an approach based on [23] to extract preferences from a user activity log and derive semantically-enhanced item recommendations from them.

Others: For the sake of brevity, in here we give a summarised overview of some other applications of FCA-based IR systems. **File Systems (FS)** are an interesting application in low-level information retrieval (operating system level). FCA provides a more dynamic interaction with the file system structure where the FS can be represented as a lattice instead of a tree [31, 48, 67] **Source code location** is an important task in software engineering as it enables code refactoring, among other applications [53, 61, 1]. Other interesting applications are **mathematical expression search** [57] and **multimedia indexing** [49, 26].

[12] `http://www.google.com/insidesearch/features/search/knowledge.html`

6.2 FCA-based IR Systems

FaIR (2000) by Uta Priss [63]: A faceted IR system in which formal concepts of documents and descriptors are mapped to thesauri entries. It features a query language built on top of the set of formal concepts with the logical operators AND, OR, NOT.
CREDO (2004) by Carpineto and Romano [12]: CREDO works as the front-end of a Web search engine (such as Google or Yahoo). It implements some of their ideas in query expansion presented in REFINER providing context to an otherwise plain-list of ranked documents. Extensions of CREDO included its port to mobile devices (CREDINO and SmartCREDO [6]).
JBrainDead (2004) by Cigarrán et al. [16]: A FCA-based system that combines standard IR techniques such as term weighting and ranking for automated attribute selection. We highlight in this work the novel evaluation metrics considering the effort needed to find documents within a concept lattice derived from the number of concepts to visit and the percentage of those that represent relevant results.
Mail-Sleuth and the Sleuth Family (2004 - 2009), Ducrou, Eklund et al.: Building on previous work, the authors present a commercial tool called Mail-Sleuth [28], a system for searching and browsing personal email collections under the assumption that novice users are able to manage a line diagram of a lattice structure. The authors extended these ideas to different application domains: ImageSleuth [26] for image browsing and retrieval (discussed in the previous sections), DVDSleuth [24] for browsing Web catalogues , SearchSleuth [25] for browsing results from a standard Web search engine and AnnotationSleuth [27] a system designed for browsing a virtual-museum collection. In 2014, Wray and Eklund presented the application *"A place for art"* [70] which followed in the steps of AnnotationSleuth with a much more elaborated user interface.
FooCA (2005) by Koester [45]: In the steps of CREDO, it also relied in the assumption that users can manage line diagrams of concept lattices, as well as interacting directly with the formal context.
BR-Explorer (2006) by Messai et al [51]: An algorithm for document retrieval the notions of "query concept", "pivoting" and "ranking" for bioinformatic datasets.
Camelis (2007) by Sebastian Ferré: Based on "a generalization of FCA", named Logical Concept Analysis (LCA), where attributes are replaced by logical formulas. Designed to cover four main aspects: mixing query and navigation, expressive query language, genericity in data types, and efficiency for large collections. Camelis integrates several taxonomies different in nature, e.g. geographical (*Paris* $\sqsubseteq$ *France*), numeric ($1999 \leq 2000$) and conceptual (ICFCA $\sqsubseteq$ Conference), allowing complex querying and other tasks previously discussed, such as "query-by-navigation" and "query by example". An extension of Camelis called Sewelis (or Camelis 2) was introduced in [30] for "Query-based Faceted Search" on linked data, introducing an expressive query language called LISQL (Logical Information System Query Language).
CreChainDo (2007) by Nauer and Toussaint [56]: A FCA-based IR system supporting explicit relevance feedback (details in Section 5).

7 Conclusions

Two related topics have been left out of the scope of this review while they remain of extreme importance for FCA-based IR approaches. Firstly, the *use of complex data* for

document indexing. Several approaches have proposed more sophisticated models than the standard Boolean retrieval model defined at the beginning of this article. Mainly, they rely in three FCA extensions for dealing with complex data, Logical Concept Analysis such as in [32], Fuzzy FCA such as the case of [66] and Pattern Structures, such as the case of [19] or [52] (the latter does not explicitly apply pattern structures, but the notions are very similar). Secondly, the application of FCA to large collections of documents or big data (an interesting discussion is provided in [46]). Both of these matters deserve a more extensive treatment than the one we could give them here.

Finally, this paper has presented an exhaustive review of FCA-based IR approaches focusing in the shared ideas and notions they share. We have shown how these ideas can be applied in a variety of domains and applications ranging from standard Boolean retrieval to semantic retrieval or file systems.

References

1. R. Al-Msie'Deen, A. Seriai, M. Huchard, C. Urtado, S. Vauttier, and H. Eyal Salman. Mining Features from the Object-Oriented Source Code of a Collection of Software Variants Using Formal Concept Analysis and Latent Semantic Indexing. In *The 25th International Conference on Software Engineering and Knowledge Engineering*, page 8, États-Unis, 2013. Knowledge Systems Institute Graduate School.
2. M. Alam and A. Napoli. Defining Views with Formal Concept Analysis for Understanding SPARQL Query Results. In *Proceedings of the 2014 International Conference on Concept Lattices and their Applications*, 2014.
3. R. A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Boston, MA, USA, 1999.
4. M. Barbut and B. Monjardet. *Ordre et classification : algebre et combinatoire*. 1970.
5. F. Bosman, R. Bouwman, and P. Bruza. The Effectiveness of Navigable Information Disclosure Systems. In *Proceedings of the Informatiewetenschap*, 1991.
6. C. Carpineto, S. Mizzaro, G. Romano, and M. Snidero. Mobile information retrieval with search results clustering: Prototypes and evaluations. *Journal of the American Society for Information Science and Technology*, 60(5), 2009.
7. C. Carpineto and G. Romano. Galois : An order-theoretic approach to conceptual clustering. *Proceedings of the 10th International Conference on Machine Learning (ICML'90)*, 1993.
8. C. Carpineto and G. Romano. A lattice conceptual clustering system and its application to browsing retrieval. *Machine Learning*, 24(2), Aug. 1996.
9. C. Carpineto and G. Romano. Information retrieval through hybrid navigation of lattice representations. *International Journal of Human-Computer Studies*, 45(5), 1996.
10. C. Carpineto and G. Romano. Effective reformulation of Boolean queries with concept lattices. In *Flexible Query Answering Systems*, volume 1495 of *Lecture Notes in Computer Science*. 1998.
11. C. Carpineto and G. Romano. Order theoretical ranking. *Journal of the American Society for Information Science*, 51(7), 2000.
12. C. Carpineto and G. Romano. Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science*, 10, 2004.
13. C. Carpineto and G. Romano. Using concept lattices for text retrieval and mining. *Formal Concept Analysis*, 2005.
14. C. Carpineto, G. Romano, and F. U. Bordoni. ULYSSES: A Lattice-based Multiple Interaction Strategy Retrieval Interface. In *In Blumenthal et al., Human-Computer Interaction*, 1995.

15. A. Castellanos, A. García-Serrano, and J. Cigarrán. Linked Data-based Conceptual Modelling for Recommendation: A FCA-Based Approach. In *E-Commerce and Web Technologies*, volume 188 of *Lecture Notes in Business Information Processing*. 2014.
16. J. M. Cigarrán, J. Gonzalo, A. Peas, and F. Verdejo. Browsing search results via formal concept analysis: Automatic selection of attributes. In *Concept Lattices*, volume 2961 of *Lecture Notes in Computer Science*. 2004.
17. V. Codocedo, I. Lykourentzou, and A. Napoli. A Contribution to Semantic Indexing and Retrieval Based on FCA - An Application to Song Datasets. In *Proceedings of the 2012 International Conference on Concept Lattices and their Applications*, 2012.
18. V. Codocedo, I. Lykourentzou, and A. Napoli. A semantic approach to concept lattice-based information retrieval. *Annals of Mathematics and Artificial Intelligence*, 2014.
19. V. Codocedo and A. Napoli. A Proposition for Combining Pattern Structures and Relational Concept Analysis. In *Formal Concept Analysis*, volume 8478 of *Lecture Notes in Computer Science*. Springer International Publishing, 2014.
20. R. Cole and P. Eklund. Application of Formal Concept Analysis to Information Retrieval using a Hierarchically Structured Thesaurus. In *International Conference on Conceptual Graphs, ICCS 96, University of New South*, 1996.
21. R. Cole and P. Eklund. Analyzing an Email Collection Using Formal Concept Analysis. In *Principles of Data Mining and Knowledge Discovery*, volume 1704 of *Lecture Notes in Computer Science*. 1999.
22. R. J. Cole, P. W. Eklund, and G. Stumme. Document Retrieval for Email Search and Discovery using Formal Concept Analysis. *Journal of Applied Artificial Intelligence (AAI)*, 17(3), 2003.
23. P. Duboucherryan and D. Bridge. Collaborative Recommending using Formal Concept Analysis. *Knowledge-Based Systems*, 19(5), Sept. 2006.
24. J. Ducrou. Dvdsleuth: A case study in applied formal concept analysis for navigating web catalogs. In *Conceptual Structures: Knowledge Architectures for Smart Applications*, volume 4604 of *Lecture Notes in Computer Science*. 2007.
25. J. Ducrou and P. Eklund. SearchSleuth: The Conceptual Neighbourhood of a Web Query. In *Proceedings of the 2007 International Conference on Concept Lattices and their Applications*, CLA '07, 2007.
26. J. Ducrou, B. Vormbrock, and P. Eklund. FCA-Based Browsing and Searching of a Collection of Images. In *Conceptual Structures: Inspiration and Application*, volume 4068 of *Lecture Notes in Computer Science*. 2006.
27. P. Eklund and J. Ducrou. Navigation and annotation with formal concept analysis. In *Knowledge Acquisition: Approaches, Algorithms and Applications*, volume 5465 of *Lecture Notes in Computer Science*. 2009.
28. P. Eklund, J. Ducrou, and P. Brawn. Concept Lattices for Information Visualization: Can Novices Read Line-Diagrams? In *Concept Lattices*, volume 2961 of *Lecture Notes in Computer Science*. 2004.
29. R. A. Fairthorne. The patterns of retrieval. *American Documentation*, 7(2), 1956.
30. S. Ferré and A. Hermann. Reconciling faceted search and query languages for the semantic web. *IJMSO*, 7(1), 2012.
31. S. Ferré and O. Ridoux. A File System Based on Concept Analysis. In *Computational Logic - CL 2000*, volume 1861 of *Lecture Notes in Computer Science*. 2000.
32. S. Ferré and O. Ridoux. A Logical Generalization of Formal Concept Analysis. In *ICCS*, volume 1867 of *LNCS*, 2000.
33. B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Dec. 1999.
34. A. Ginsberg. A unified approach to automatic indexing and information retrieval. *IEEE Expert*, 8(5), Oct. 1993.

35. R. Godin, J. Gecsei, and C. Pichet. Design of a Browsing Interface for Information Retrieval. In *Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '89, New York, NY, USA, 1989.
36. R. Godin, R. Missaoui, and A. April. Experimental comparison of navigation in a Galois lattice with conventional information retrieval methods. *International Journal of Man-Machine Studies*, 38(5), May 1993.
37. R. Godin, E. Saunders, and J. Gecsei. Lattice model of browsable data spaces. *Information Sciences: an International Journal*, 40(2), Dec. 1986.
38. F. A. Grootjen and T. van der Weide. Conceptual relevance feedback. In *Systems, Man and Cybernetics, 2002 IEEE International Conference on*, volume 2, Oct. 2002.
39. F. A. Grootjen and T. P. van der Weide. Conceptual query expansion. *Data Knowl. Eng.*, 56(2), Feb. 2006.
40. M. Huchard, M. R. Hacene, C. Roume, and P. Valtchev. Relational concept discovery in structured datasets. *Annals of Mathematics and Artificial Intelligence*, 49(1-4), June 2007.
41. D. Ignatov I. and S. O. Kuznetsov. Concept-based Recommendations for Internet Advertisement. In *Proceedings of the 2008 International Conference on Concept Lattices and their Applications*, 2008.
42. P. Ingwersen. Cognitive perspectives of information retrieval interaction: Elements of a cognitive IR theory. *Journal of Documentation*, 52(1), 1996.
43. M. Kim and P. Compton. Formal concept analysis for domain-specific document retrieval systems. In *AI 2001: Advances in Artificial Intelligence*, 2001.
44. M. Kim and P. Compton. Evolutionary document management and retrieval for specialized domains on the web. *International Journal of Human-Computer Studies*, 60(2), 2004.
45. B. Koester. Conceptual Knowledge Retrieval with FooCA: Improving Web Search Engine Results with Contexts and Concept Hierarchies. In *Industrial Conference on Data Mining*, volume 4065 of *Lecture Notes in Computer Science*, 2006.
46. S. O. Kuznetsov. Fitting Pattern Structures to Knowledge Discovery in Big Data. In *Formal Concept Analysis*, volume 7880 of *Lecture Notes in Computer Science*. 2013.
47. C. D. Manning, P. Raghavan, and H. Schtze. *Introduction to Information Retrieval*. July 2008.
48. B. Martin. Formal concept analysis and semantic file systems. In *Concept Lattices*, volume 2961 of *Lecture Notes in Computer Science*. 2004.
49. J. Martinez and E. Loisant. Browsing image databases with galois' lattices. In *Proceedings of the 2002 ACM Symposium on Applied Computing*, SAC '02, New York, NY, USA, 2002.
50. N. Messai, M.-D. Devignes, A. Napoli, and M. Smaïl-Tabbone. Querying a bioinformatic data sources registry with concept lattices. In *Proceedings of the 13th international conference on Conceptual Structures: common Semantics for Sharing Knowledge*, volume 3596 of *Lecture Notes in Computer Science*, July 2005.
51. N. Messai, M.-D. Devignes, A. Napoli, and M. Smaïl-Tabbone. BR-Explorer: An FCA-based algorithm for Information Retrieval. In *Fourth International Conference On Concept Lattices and Their Applications - CLA 2006*, Hammamet/Tunisia, 2006.
52. N. Messai, M.-D. Devignes, A. Napoli, and M. Smaïl-Tabbone. Many-Valued Concept Lattices for Conceptual Clustering and Information Retrieval. In *Proceedings of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 127–131, June 2008.
53. A. Mili, R. Mili, and R. Mittermeir. Storing and retrieving software components: A refinement based system. In *Proceedings of the 16th International Conference on Software Engineering*, ICSE '94, Los Alamitos, CA, USA, 1994.
54. C. N. Mooers. *A Mathematical Theory of Language Symbols in Retrieval*. 1958.

55. E. Nauer and Y. Toussaint. Dynamical modification of context fo an iterative and interactive information retrieval proess on the Web. In *Proceedings of the 2007 International Conference on Concept Lattices and their Applications*, CLA '07, 2007.
56. E. Nauer and Y. Toussaint. CreChainDo: an iterative and interactive Web information retrieval system based on lattices. *International Journal of General Systems*, 38(4), 2009.
57. T. T. Nguyen, S. C. Hui, and K. Chang. A lattice-based approach for mathematical search using Formal Concept Analysis. *Expert Systems with Applications*, 39(5), 2012.
58. G. S. Pedersen. A Browser for Bibliographic Information Retrieval, Based on an Application of Lattice Theory. In *Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '93, New York, NY, USA, 1993.
59. G. S. Pedersen. Relationship lattices for information modelling. *Information Modelling and Knowledge Bases*, 1994.
60. J. Poelmans, D. I. Ignatov, S. Viaene, G. Dedene, and S. O. Kuznetsov. Text mining scientific papers: a survey on FCA-Based information retrieval research. In *Proceedings of the 12th Industrial conference on Advances in Data Mining: applications and theoretical aspects*, volume 7377 of *Lecture Notes in Computer Science*, Berlin, Heidelberg, July 2012.
61. D. Poshyvanyk and A. Marcus. Combining Formal Concept Analysis with Information Retrieval for Concept Location in Source Code. In *15th IEEE International Conference on Program Comprehension (ICPC '07)*, June 2007.
62. U. Priss. A graphical interface for document retrieval based on formal concept analysis. In *Proceedings of the 8th Midwest Artificial Intelligence and Cognitive Science Conference*, 1997.
63. U. Priss. Lattice-based Information Retrieval. *Knowledge Organization*, 27, 2000.
64. U. Priss. Formal concept analysis in information science. *Annual Review of Information Science and Technology*, 40(1), Sept. 2007.
65. J. J. Rocchio. Relevance feedback in information retrieval. In *The Smart retrieval system - experiments in automatic document processing*. 1971.
66. S. Senatore and G. Pasi. Lattice Navigation for Collaborative Filtering by Means of (Fuzzy) Formal Concept Analysis. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, SAC '13, New York, NY, USA, 2013.
67. A. Shah and L. Caves. ConceptOntoFs: A Semantic File System for Interno. In *First International Workshop on Plan 9 and Inferno*. 2006.
68. F. J. Valverde and C. Pelaez-Moreno. System vs. Methods: an Analysis of the Affordances of Formal Concept Analysis for Information Retrieval. In *Proceedings of the Workshop Formal Concept Analysis Meets Information Retrieval (FCAIR 2013)*, 2013.
69. R. Wille. Restructuring Lattice Theory: An approach based on hierarchies of concepts. In *Formal Concept Analysis*, volume 5548 of *Lecture Notes in Computer Science*. 2009.
70. T. Wray and P. Eklund. Using formal concept analysis to create pathways through museum collections. In *Proceedings of the 3rd International Workshop "What can FCA do for Artificial Intelligence"?* 2014.