Systems vs. Methods: an Analysis of the Affordances of Formal Concept Analysis for Information Retrieval*

Francisco J. Valverde-Albacete[1],* and Carmen Peláez-Moreno[2]

[1] Departamento de Lenguajes y Sistemas Informáticos
Univ. Nacional de Educación a Distancia, c/ Juan del Rosal, 16. 28040 Madrid, Spain
fva@lsi.uned.es
[2] Departamento de Teoría de la Señal y de las Comunicaciones
Universidad Carlos III de Madrid, 28911 Leganés, Spain
carmen@tsc.uc3m.es

Abstract. We review previous work using Formal Concept Analysis (FCA) to build Information Retrieval (IR) applications seeking a wider adoption of the FCA paradigm in IR. We conclude that although a number of systems have been built with such paradigm (FCA *in* IR), the most effective contribution would be to help establish IR on firmer grounds (FCA *for* IR). Since such an approach is only incipient, we contribute to the general discussion by discussing affordances and challenges of FCA for IR.

1 Introduction

Modern Information Retrieval (IR) is a wide field with several different concerns pulling in different directions. Under the competitive task evaluation paradigm [29], IR strives to solve *tasks* using any of a variety of *models*, mostly by *Machine Learning* techniques [41]. A glimpse at the main types of IR models can be found in [27], reproduced here as Fig. 1.

Perhaps the simplest and best known task is that of *ad hoc retrieval*, where a corpus of documents is searched with a number of topics (Sec. 2), but certainly the most prevalent task is the familiar *Web retrieval.* They are also typical instances of *batch* and *interactive retrieval tasks*, respectively.

The Formal Concept Analysis (FCA)[3] community has been *implementing* Information Retrieval (IR) systems for well over 25 years, starting with [21]. Yet few mainstream IR practitioner confess to understanding the bases of FCA, a testimony of the scarce impact of the former in mainstream IR.

* FJVA has been partially supported by EU FP7 project LiMoSINe (contract 288024). CPM has been partially supported by the Spanish Government-Comisión Interministerial de Ciencia y Tecnología project TEC2011-26807 for this paper.

[3] This paper is targeted at FCA practitioners, so the reader is expected to be acquainted with the principal results of FCA. For analogue papers targeted at IR practitioners see [35, 42].

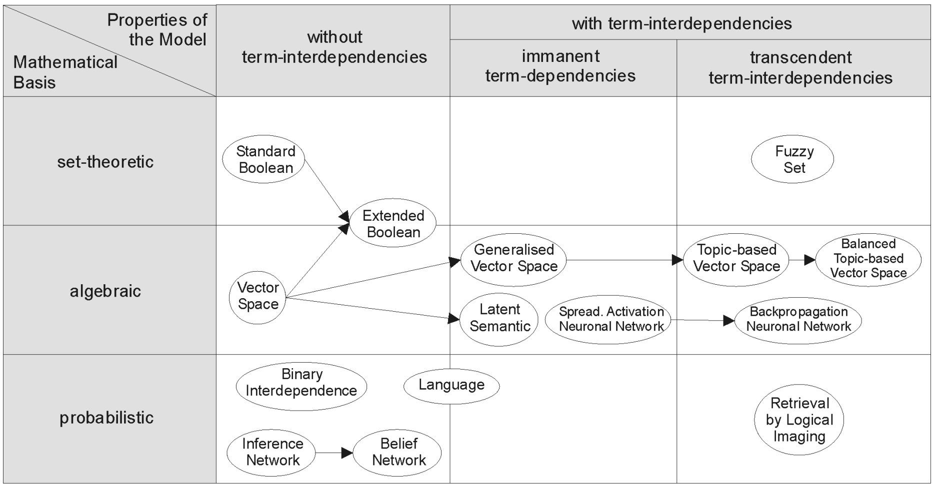

Fig. 1: A schematics showing the main types of Information Retrieval models (after [27])

To the best of our knowledge only two information retrieval-incepted books have realised the potential of FCA for IR: [45] and [11]. On the one hand, Van Rijsbergen briefly notes down that the Boolean Retrieval Model is captured in terms of Galois connections between documents and features (terms) [45, p. 37], although he includes there the inverse index on terms and documents which may best be conceived in terms of a Galois adjunction[42]. On the other hand, Dominich makes a very cursory review of the state-of-the-art up until 2008 [11]. He notes down the work of [37] on faceted information retrieval and that of [6] on browsing Web retrieval results with concept lattices, and the disjunctive approximation to boolean retrieval of [32]. Curiously, the data-driven nature of FCA is downplayed in this work.

In the FCA camp, the broadest review is still [6] but [5, 23, 37, 38] have narrower foci. Notice that both [11, 37] review work in lattice-based IR systems prior to the groundbreaking [21], but pre-FCA emphasis is in *designing* the lattice instead of *obtaining* it from the relevance relation: the *data driven* quality of FCA is missing in this early work, e.g. [33].

We believe that part of the explanation for this divide may be that only the most simple, basic tasks in IR—and using the oldest IR models—have been successfully tackled with FCA techniques. After all, IR in some 60+ years has developed its own set of techniques, methods for research and testing and is practised by, probably, the most thriving community in ICT. It is only natural that FCA can only be considered as a subsidiary discipline to such endeavour. Or not?

In this paper we want to put forward the distinction between FCA *in* IR and FCA *for* IR, that is *implementing IR systems with FCA* vs. *augmenting IR with the methods and ideas of FCA*. We claim that most of the work so far has been in *FCA in IR* and the time is ripe to expound on a *FCA for IR*, that is a theory of the *affordances* and *challenges* of using FCA to solve IR tasks, already started in [6]. Here we use affordances in the sense of [34], to refer to "the actionable properties between the world and an actor", that is, the 'world' of FCA and the 'actor' that is an IR practitioner.

This paper is about raising awareness of these two conceptions of the role of FCA vis-à-vis IR. For this purpose, we introduce in Sec. 2 a prototypical information retrieval task to make explicit what types of problems an IR practitioner comes up with. In Sec. 3 we review to what extent FCA actually solves such problems by supplying a set of *affordances* of FCA *for* IR. Finally, we discuss in Sec. 4 what are further challenges that FCA has to solve for a wider adoption in a number of data-intensive application domains, including IR.

2 A prototypical information retrieval task

To guide our exposition we will discuss the *ad-hoc retrieval task*, that is, the task where the IR system is expected to produce the documents relevant to an arbitrary user need as expressed in a one-off, user-initiated query [29, p. 3]. Although Web retrieval is perhaps the prevalent IR task at present, ad-hoc retrieval is the best studied one and it admits many different models. In the following, we expand the *modelling* of this task propounded in [42] as a script to discuss affordances and challenges in using FCA for IR tasks.

A model for batch ad-hoc tasks To fix notation, we adapt the formal model put forward by Fuhr [17] reproduced in Fig. 2—although we interpret the signs there differently—and we let $\overline{Q}$, $\overline{D}$, and $\overline{R}$ respectively stand for a *set of information needs* for a querying user, a set of *information-bearing percepts* and a *psychological capability* whereby a particular user is going to judge the relevance of the information percepts for her information needs.

$$\begin{array}{ccccc}
\overline{Q} & \xrightarrow{\alpha_Q} & Q & \xrightarrow{\beta_Q} & Q' \\
\overline{R}\Big\| & & R\Big\| & & R'\Big\| \\
\overline{D} & \xrightarrow{\alpha_D} & D & \xrightarrow{\beta_D} & D'
\end{array}$$

Fig. 2: An adaptation of the conceptual model of Fuhr [17] with the concepts dealt with in this paper highlighted.

Figure 2 highlights the data and models we will address in this paper: let Q, D and R be the outcome of as many instantiation processes of the above-mentioned information needs, information supplies and relevance judgments, respectively. We will call them *queries*, *documents* and *relevance judgments* and assume that the relevance judgment representations adopt the form of a relevance relation, $R \subseteq D \times Q$. Finally, let Q', D', R' be the *query representations*, *document representations* and the *relevance judgments in representation space* respectively, so that $R' \subseteq D' \times Q'$.

Whereas Fuhr's model considers queries, documents and judgements to be *inside* the information retrieval system, we consider them *both inside and outside*, since they are more properly conceived as *(multimedia) recordings* of the psychological entities and processes considered above. They have an immanent existence independent of the system yet are related to them by their representations. However, representations arise when we try to approximate the information content of queries and documents inside an IR system, hence they are sometimes called *surrogates or surrogate representations (for their records).*

Although the model posits four maps between the above-introduced domains, for practical reasons it is common to concentrate on only two

A query representation process, $\beta_Q : Q \to Q'$, mapping from queries to query representation suitable for processing in a particular information retrieval system.

A document representation process, $\beta_D : D \to D'$, mapping from documents to document representations.

Therefore, we limit ourselves to the domains, mappings and sets enclosed by the square in Fig. 2, the recording- and representation-related domains.

Assessment The *ideal IR system* $S_{D,Q}(R) =< \varrho_R >$ would consist in a *relevance function* ϱ_R describing relevant documents where $\varrho_R(q_i)$ is the set of documents *relevant to query* q_i as dictated by the ideal relevance relation R. But in the process of building an IR system we may incur modelling errors, approximation, etc., whence we accept that the actual relation implemented will be the approximated relevance $\hat{R} \neq R$ for the *implemented IR system* $S_{D,Q}(\hat{R}) =< \varrho_{\hat{R}} >$. Its *retrieval function* may only return $\varrho_{\hat{R}}(q_i)$ the set of documents *retrieved for the same query as dictated by the approximate relevance* $\hat{R}$,

$$\begin{aligned} \varrho_R &: Q \to 2^D & \varrho_{\hat{R}} &: Q \to 2^D \\ q_i &\mapsto \varrho_R(q_i) = \{d_j \in D | d_j R q_i\} & q_i &\mapsto \varrho_{\hat{R}}(q_i) = \{d_j \in D | d_j \hat{R} q_i\} \ . \end{aligned} \tag{1}$$

The batch retrieval task can be subjected to the so-called "Cranfield model of Information Retrieval system evaluation" [31], where a set of document records, or *collection*, $D_T \subseteq D$, a set of sampled query records, *topics*, $Q_T \subseteq Q$, and a set of relevance judgments involving documents and query records, $R_T \subseteq D_T \times Q_T$ are known. Assessing the quality of $S_{D,Q}(\hat{R})$ means, essentially, comparing R and $\hat{R}$: For a given query q, the system would retrieve documents $\varrho_{\hat{R}}(q)$ whereas the relevant documents are given by the prescribed relevance as $\varrho_R(q)$. Therefore the retrieved relevant documents for each query $q \in Q$ would be

$\varrho_R(q) \cap \varrho_{\hat{R}}(q)$, and we would have *precision* $P_{\hat{R}}$ and *recall* $R_{\hat{R}}$—or any measure derived therefrom—as

$$P_{\hat{R}}(q) = \frac{|\varrho_R(q) \cap \varrho_{\hat{R}}(q)|}{|\varrho_{\hat{R}}(q)|} \qquad R_{\hat{R}}(q) = \frac{|\varrho_R(q) \cap \varrho_{\hat{R}}(q)|}{|\varrho_R(q)|} \ . \tag{2}$$

A decomposition of the problem. We believe it is convenient to conceptually decompose the synthesis of $S_{D,Q}(\hat{R})$ into the following problems[cfr. 6, §. 4]:

Problem 1 (Representation). Given different spaces of queries Q and their representations Q' find a mapping β_Q between them. Do likewise for documents D, their representations D' and a *surjective* mapping β_D between them.

Problem 2 (Generalization). Given local information about the relevance relation R in the form of a training subset $R'_T = D'_T \times Q'_T$, extend/generalise such information to $\hat{R}' \subseteq D' \times Q'$.

Problem 3 (Surrogate implementation). Given domains of documents D and queries Q (whether they be descriptions or representations), a querying hypothesis and an estimated relevance relation $\hat{R}$, build an information retrieval system that faithfully implements the prescribed relevance[4].

Once solved these problems we can build the retrieval set as

$$\varrho_{\hat{R}}(q) = \beta_D^{-1}[\varrho_{\hat{R}'}(\beta_Q[q])] \tag{3}$$

where we have taken the precaution of making all of the functions apply over *sets* rather than singletons.

Problem 4 (Post-retrieval interaction). Given the answer set to a query $\varrho_{\hat{R}}(q)$ present it to the user in an effective manner.

Note that in standard IR engineering practice the steps of retrieving document representations and then finding their original document are often aggregated by means of an inverted index. Also, (3) is often complemented with *retrieval status value* for each result, a number stating the *degree of relevance* of each retrieved document to the query.

3 Affordances of FCA for IR

This list is going to be informally structured as a sort of proof: first we state what we consider the affordances of FCA for IR and then we explain the reasoning behind our assertion.

Affordance 1 (Solving problem 3 in the conjunctive Boolean Model). *FCA implements the (conjunctive) Boolean Keyword model.*

[4] We use here $\hat{R}$ as a variable ranging over possible relation values, not necessarily the optimal one.

Suppose that there exists a set of keywords T [5], queries are represented as keywords $Q' \equiv T$, documents are represented as set of keywords $D' \equiv 2^T$, and estimated relevance $\hat{R}'$ is defined by means of the inclusion relation $d'\hat{R}'q' \Leftrightarrow d' \ni q'$. The retrieval function is easy to write $\varrho_{\hat{R}'}(\{t\}) = \{d \in D \mid q \in d\}$, but what are we to expect when supplying several queries, that is, several keywords?

To implement *conjunctive querying* we produce the intersection of the result sets, that is, for $B = \{q_i\}_{i\in I}$ we have

$$\varrho^1_{\hat{R}'}(\{q_i\}_{i\in I}) = \{d \in D \mid \forall i \in I, q_i \in d\} = \cap_{i\in I}\{d \in D \mid q_i \in d\} \ .$$

In that case, the more keywords a query has the less documents the retrieval function returns, that is, $q'_1 \subseteq q'_2$ implies $\varrho(q'_1) \supseteq \varrho(q'_2)$. Then we realise that this retrieval function is the *query polar* $\varrho^1_{R'}(q)$ of the Galois Connection in Fig. 3.(a)

$\varrho^1_{\hat{R}'} : 2^Q \to 2^D$	$\varrho^2_{R} : 2^Q \to 2^D$
$\varrho^1_{\hat{R}'}(B) = \{d'_i \in D' \vert \forall q' \in B, d'_i\hat{R}'q'\}$	$\varrho^2_{R}(B) = \{d'_i \in D' \vert \exists q' \in B, d'_i\hat{R}q'\}$
$\iota^1_{\hat{R}'} : 2^{D'} \to 2^{Q'}$	$\iota^2_{\hat{R}'} : 2^{D'} \to 2^{Q'}$
$\iota^1_{\hat{R}'}(A) = \{q' \in Q' \mid \forall d' \in A, d'R'q'\}$	$\iota^2_{\hat{R}'}(A) = \{q' \in Q' \mid d'R'q' \Rightarrow d' \in A\}$
(a) Galois connection	(b) Galois adjunction

Fig. 3: Galois connection and adjunction between two powersets of terms that implement the conjunctive and disjunctive models of Boolean retrieval, respectively.

This is one of the contributions of [21], the first paper to use *FCA in IR*, that is, to build a Galois connections that implements an IR system, to the best of our knowledge. Most of the work in FCA in IR uses this model [3, 9, 13, 37], with the notable exception of the work starting with [32], who define relevance in a way that leads to the disjunctive model of Fig. 3.(b). In this case, $\varrho^2_{\hat{R}'}(B) = \{d_i \in D \vert \exists q \in B, d_iRq\}$, but, since there are some tricks to representing this in a concept lattice [44], the authors of [32] develop a browsing model of their own.

Affordance 2. *FCA implements query term expansion*

In fact, the Galois connection has "another half", the *document polar.* Let $A \subseteq D'$ be a set of documents. Then the set of queries for which *all* those documents are relevant is $\iota^1_{\hat{R}'}(A) = \{q' \in Q' \mid \forall d' \in A, d'R'q'\}$. Actually retrieval sets come in pairs called *formal concepts*[6]. In our example, a formal

[5] This is sometimes called the *bag-of-keywords* model of documents.
[6] In [21] they were originally called "complete pairs".

concept (A, B) is a pair of a set of documents A and a set of queries B so that all the documents in A are relevant to all the queries in B, and dually, $\left(A = \varrho^1_{\hat{R}'}(B), B = \iota^1_{\hat{R}'}(A)\right)$. These pairs come from the properties of the polars in the Galois connection, as described in Fig. 4: the composition of the polars are extensive, idempotent operators, that is, *closure operators*.

$$
\begin{tikzcd}
2^D \arrow[r, bend right, "\iota^1_{\hat{R}'}"'] \arrow[d, bend right, "\gamma^1_D"'] \arrow[dr, "\varrho^1_{\hat{R}'}"] & 2^Q \arrow[l, bend right, "\varrho^1_{\hat{R}'}"'] \arrow[d, bend right, "\gamma^1_Q"'] \arrow[dl, "\varrho^1_{\hat{R}'}"'] \\
\underline{\mathfrak{B}}_D(D, Q, \hat{R}) \arrow[u, bend right, "\hookrightarrow_D"'] \arrow[r, bend right, "\iota^1_{\hat{R}}(P)"'] & \underline{\mathfrak{B}}_Q(D, Q, \hat{R}) \arrow[l, bend right, "\varrho^1_{\hat{R}}"'] \arrow[u, bend right, "\hookrightarrow_Q"']
\end{tikzcd}
$$

(a) The concept lattice $\underline{\mathfrak{B}}(D, Q, \hat{R})$ and two closure operators γ^1_D and γ^1_Q on documents and queries.

$$
\begin{tikzcd}
2^D \arrow[r, bend right, "\iota^2_{\hat{R}}"'] \arrow[d, bend right, "\kappa^2_D"'] \arrow[dr, "\varrho^2_{\hat{R}}"] & 2^Q \arrow[l, bend right, "\varrho^2_{\hat{R}}"'] \arrow[d, bend right, "\gamma^2_Q"'] \arrow[dl, "\varrho^2_{\hat{R}}"'] \\
\mathfrak{N}^Q_D(D, Q, \hat{R}) \arrow[u, bend right, "\hookrightarrow_D"'] \arrow[r, bend right, "\iota^2_{\hat{R}}(P)"'] & \mathfrak{N}^Q_Q(D, Q, \hat{R}) \arrow[l, bend right, "\varrho^2_{\hat{R}}"'] \arrow[u, bend right, "\hookrightarrow_Q"']
\end{tikzcd}
$$

(b) The lattice of neighbourhood of queries $\mathfrak{N}^Q(D, Q, \hat{R})$, an interior operator κ^2_D on documents and a closure operator γ^2_Q on queries.

Fig. 4: Galois connections describing conjunctive (left) and disjunctive (right) boolean retrieval.

Note that for a set of queries $B \in Q$, $\gamma^1_{\hat{R}'}(B) \geq B$ hence querying through formal concepts expands the query sets in a data-dependent manner. This was noted cursorily in [21] but is thoroughly explained in [6, Chap. 3] whose authors have contributed the most to this line of work.

Affordance 3. *FCA provides for integrated browsing and querying.*

As previously noted, query submission in a concept lattice-based IR system is just an application of the query polar, which obtains the concept whose extent is the retrieval set, and whose intent is the extended query. This acts as a querying mechanism.

On the other hand, formal concepts have a natural order based in the inclusion order of extents or the dual inclusion order of intents, $(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \Leftrightarrow B_1 \supseteq B_2$. Furthermore, the Fundamental Theorem of Concept Lattices asserts that this order between concepts is a complete lattice [20, p. 20], representable as an *order diagram*.

Godin et al. [21, 22] put forth the idea that lower and upper neighbours as well as parallel concepts define *a topology for browsing in a (concept) lattice* (see Fig. 5). Consider a *concept in focus* C,

– Below it lie its lower covers, those concepts with more stringent (higher cardinality) query sets.
– Above it lie its upper covers, those which have less stringent (lower cardinality) query sets.
– To each side of the concept in focus stand those *sibling concepts* sharing parents (and descendants) with it. They have incompatible query sets (inconsistent with the focus concept intent).

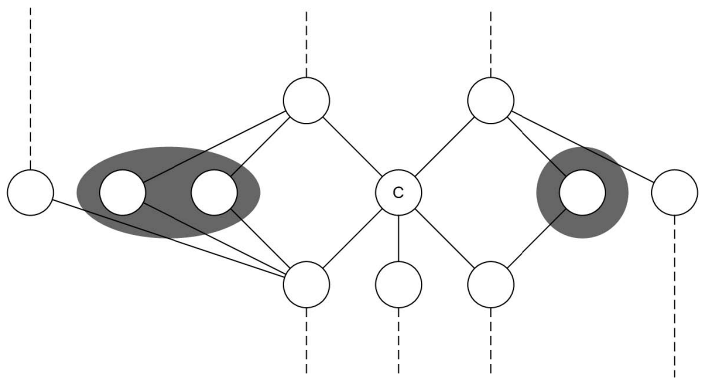

Fig. 5: Schematic representation of a concept C in focus in a concept lattice with upper and lower neighbours, from [12] rather than [21]. Sibling nodes are within shaded areas.

Although previous work had noted the interest of lattices for navigation, to the extent of our knowledge, Godin et al. were the first to tie the modification of queries (and therefore retrieval sets) to navigation in a systematic manner. For in-detail reviews of this affordance in the context of Personal Information Systems see [12, 13].

Affordance 4. *FCA provides visualization schemes for the document-query lattice at different scales.*

The scales we refer to in this affordance are those related to the visual and informational complexity of the lattice. Complexity scales are in other contexts termed the *micro-, meso-*, *macro-* and *mega-scales.*

The local neighbourhood of a formal concept illustrated in Fig. 5 was posited in [21, 22] and developed in a number of works [12]. It is a micro-visualization device depicting the part of the lattice surrounding a particular concept in focus whether incorporating a *fisheye view* [5, 21] or not [12].

On the other hand, the order diagram of the concept lattice acts as a meso-scale visualization technique. Similarly, visualizing only the concepts that lie below—or above—a focus concept produces visualization devices of comparable complexity and can be considered meso-scale visualizations. Here we consider the mapping of the downset of the focus as a tree as in [4]. Furthermore, the use of *attribute and object projections* on the whole lattice, *reduced labelling* and *nested line diagrams* [20, 39] are all tools that help us balance displayed information vs. visual complexity allowing us to display complex lattices at the mesoscale.

For those cases when these complexity-reducing strategies are not sufficient, very little work has been done on observing lattices at the macro-scale—let alone the mega-scale—sacrificing concept and local structure readability for the quick glimpse of emerging features like height, width, overall shape, concept density, etc. For an illustration of such problems, see the lattices in [24]. Recently, [14] have proposed a technique to embed any concept lattice onto a boolean lattice of similar complexity which acts as a representation space disposing of a lot of information: its usability is, as yet, unassessed.

Affordance 5 (Solving problem 4). *FCA provides retrieval-set navigation.*

This niche application of FCA is perhaps the best-know to the IR community [25, § 10.7]. It is a natural consequence of treating the retrieval set as a subcorpus (of snippets, possibly) and using FCA to establish ordering relations between them as induced by their terms. Perhaps the first to propose this use of concept lattices is [3], and it is thoroughly explained in [6], usability studies included. Systems implementing also a post-retrieval visualization of Web retrieval searches or Meta-searches can be found in [8–10, 26].

Affordance 6. *FCA captures naturally occurring (immanent) term dependencies*

If terms were independent, then concept lattices, at least from the perspective of terms, should be boolean: all possible combinations of terms would arise as intents, but this is never the case. Since the inception of the first FCA in IR systems it was noticed that particular groupings of terms occur naturally in documents and this is reflected in the system of intents. Of course, this dovetails into the Automatic Expansion of Queries mentioned in Affordance 2: modelling term dependencies is how automatic query expansion is catered to. In terms of IR models, this means that the model implemented by FCA is actually in the empty square in Fig. 1. Carpineto and Romano [7] have investigated this issue heavily both from the point of view of IR and from that of FCA [see 6, §3.1 for a rather extensive review].

Affordance 7. *FCA scaling implements faceted search & navigation.*

Sometimes certain sets of attributes have different multiple possible values and/or special relationships between those values—such as hierarchies—and it is interesting for navigation purposes to see the collection of documents through the prism of those relations. This is called *faceted information retrieval.*

In FCA, discrete multi-valued attributes or otherwise-related attributes may be rendered in a data-dependent fashion by means of the process of *scaling attributes* [20]. But the effectiveness of this process depends extraordinarily on the experience of the expert user doing the encoding of attributes.

Although faceted navigation is explicitly mentioned in [21], it seems that FaIR was the first actual implementation using FCA [11, 37], albeit for a restricted application, thesaurus exploration. A review of faceted boolean IR can be found in [13]—as applied to Personal Information Systems—with an emphasis on usability, visualisation and navigation.

An alternative to scaling is *logical concept analysis, LCA* where any logical formula may be used to characterize intents [16], and it has been used to build a Personal Information Retrieval system for photos based in metadata [15]. Note that LCA is a *proper generalization* of FCA.

Although a number of other topics suggest themselves for this review—such as Semantic Filesystems [15, 30] or the duality of Information Pull & Push—to put them in context would demand more space than we have at our disposal.

4 Discussion: challenges of IR for FCA

Dealing with redundancy and noise in data. As in other subfields of machine learning and pattern recognition, functions β_Q and β_D of Fig. 2 can be thought of as functions that reduce unnecesary redundancy and noise.

For instance, when dealing with text we should be aware that natural language is widely-acknowledged to be *extremely redundant*: many words, expressions, constructions, etc. convey the same ideas and essentially make the complexity of the system grow. Furthermore, if words are considered terms for IR, every single word encountered when tokenizing a text *invokes* all of the senses conventionally assigned to it in a language. Since it is these senses that are purported to mediate the actual relation between the terms and documents, serendipity may reinforce not just the *originally intended* sense but also some *unrelated senses* due to surrounding context. This is a manifestation of *noise*, e.g. undesired content. And these problems can only be compounded by the ubiquity of synonymy and polysemy in Natural Language.

On top of the excess complexity incurred by redundancy, it is well-known that FCA is very sensitive to the spureous absence or presence of crosses in the incidence relation between documents and terms: the addition or deletion of any such incidences may as much as double or halve the number of concepts in the lattice[20]. If FCA is to succeed in dealing with such problems it has to devise methods to cope with this kind of noise at the incidence relation level.

Big data, supervised operation and training. The main challenge for FCA to be of any help to IR is *scalability.* Perhaps the maximum reported size for FCAinIR systems is some thousands of documents [39], while it is customary for present-day IR systems to have millions of documents. There is no easy way to overcome this inherent limitation for concept lattices: building them is just too costly in time and space[28].

One way to address the complexity of Big Data would be to assume the data-driven paradigm of Machine Learning or Pattern Recognitition [41]. However, FCA is an *unsupervised* machine learning technique: all of the information in the lattice stems from the information in the documents, the terms and the incidence relation between them. But the solution to Problem 2 seems to entail a supervised procedure whereby the training topic judgements can be used to improve unseen topics. At present, relevance in the boolean case is dictated *a priori* and there is no room for such supervision, only for post-retrieval assessment.

Unless this mismatch is addressed, machine learning-inspired techniques will still outperform FCA or address tasks which FCA simply cannot attempt.

Catering to more complex IR models. The history of IR seems to be an account of progressively complex modelling of textual data. From the boolean bag-of-word models, conceived as boolean vectors, it is easy to take a conceptual jump towards softer weighting schemes in the Vector Space Model. From constant-dimension vectors in the Vector Space Model, it is an easy jump to probability-weighted formal series, that is *(generative) language models*. Similarly, from vector description in non-orthogonal systems of generators it is easy to conceive an orthonormal basis wherein to represent vectors, which is the essence of Latent Semantic Indexing, and so on. All such conceptual leaps are steps in a process of continual algebraization of the underlying models that entail better modelling or learning capabilities in IR.

Such a process has barely begun in FCA with the so-called *generalizations* of FCA, [1, 2, 43]. Nevertheless, coincidences can be seen in all such evolutions: it seems that the basis for any possible generalization of FCA is the theory of residuated semirings [36], while many of the models in IR have semiring-based costs (probabilities, log-probabilities, etc.)

In a similar tone, most of the implementations of FCA in IR deal with the conjunctive querying case, with the previously noted exception of [32], which implements a sort of disjunctive model. If FCA wants to embrace all possible "conceptualization modes" for queries, it needs to standardize and use habitually the whole gamut of Galois connections available [44].

A concluding note... On the one hand, the FCA community has an increasing collective expertise in the development of IR applications (FCA in IR) in different domains and tasks, but has achieved only limited impact in IR proper, for the reasons explained above among others.

On the other hand, FCA has strong theoretical foundations that can help IR understand better its own models and basic assumptions (FCA for IR). Yet FCA would very much profit by the assessment-oriented approach to task-solving now prevalent in the field of IR. It would seem FCA only needs to embrace the new generalization efforts outgrowing from the dynamic flourishing of FCA these past 15 years to do so.

At the risk of being too poetical, since IR is highly empirical (and in the quest for firmer theoretical grounding) and FCA highly theoretical yet completely data-driven (but still needs to come to terms with task-realities) there is still hope for a middle ground/sweet spot where someday the twain may meet.

Bibliography

[1] R. Belohlavek. Fuzzy Galois connections. *Math. Logic Quarterly*, Jan. 1999.

[2] A. Burusco and R. Fuentes-González. The study of the L-fuzzy concept lattice. *Mathware & soft computing*, Jan. 1994.

[3] C. Carpineto and G. Romano. ULYSSES: A lattice-based multiple interaction strategy retrieval interface. In *Human-Computer Interaction, 5th International Conference, EWHCI'95*, pages 91–104, Berlin, Heidelberg, 1995. Human-Computer Interaction, 5th International Conference, EWHCI'95.

[4] C. Carpineto and G. Romano. Exploiting the potential of concept lattices for information retrieval with CREDO. *Journal of Universal Computer Science*, 10:8, 2004.

[5] C. Carpineto and G. Romano. *Concept Data Analysis.* Theory and Applications. John Wiley & Sons, Ltd, Chichester, UK, Sept. 2005.

[6] C. Carpineto and G. Romano. Using concept lattices for text retrieval and mining. In *Formal Concept Analysis*, pages 161–179. Springer Verlag, 2005.

[7] C. Carpineto and G. Romano. A Survey of Automatic Query Expansion in Information Retrieval. *Journal of the ACM*, 44(1):1–50, Jan. 2012.

[8] C. Carpineto, A. Della Pietra, S. Mizzaro, and G. Romano. Mobile clustering engine. In M. Lalmas, editor, *Advances in Information Retrieval*, pages 155–166. Springer, 2006.

[9] J. M. Cigarrán, J. Gonzalo, A. Peñas, and F. Verdejo. Browsing search results via Formal Concept Analysis: Automatic selection of attributes. In *Concept Lattices*, volume LNAI 2961, pages 311–331. Springer, 2004.

[10] J. M. Cigarrán, A. Peñas, J. Gonzalo, and F. Verdejo. Automatic selection of noun phrases as document descriptors in an FCA-based information retrieval system. In Ganter and Godin [18], pages 49–63.

[11] S. Dominich. *The Modern Algebra of Information Retrieval.* The Information Retrieval Series. Springer, 2008.

[12] J. Ducrou and P. Eklund. SearchSleuth: The conceptual neighbourhood of an web query. In *Proceedings of the 5th International Conference on Concept Lattices and their Applications, (CLA07)*, pages 253–263, Montpellier, France, Oct. 2007.

[13] J. Ducrou and P. W. Eklund. Faceted document navigation. In P. Hitzler and H. Schärfe, editors, *Conceptual Structures in Practice.* Chapman & Hall / CRC Press, 2009.

[14] J. M. Fernández-Calabozo, C. Peláez-Moreno, and F. Valverde-Albacete. WebGeneKFCA: an On-line Conceptual Analysis Tool for Genomic Expression Data. In L. Szathmary and U. Priss, editors, *Concept Lattices and Applications (CLA 2012)*, pages 345–350, Oct. 2012.

[15] S. Ferré. CAMELIS: Organizing and Browsing a Personal Photo Collection with a Logical Information System. *Proceedings of the 5th International Conference on Concept Lattices and their Applications, (CLA07)*, pages 112–123, Sept. 2007.

[16] S. Ferré and O. Ridoux. Introduction to logical information systems. *Information Processing and Management*, 40:383–419, Jan. 2004.

[17] N. Fuhr. Probabilistic models or information retrieval. *The Computer Journal*, 35(3):243–255, 1992.

[18] B. Ganter and R. Godin, editors. *Proceedings of the 3nd International Conference on Formal Concept Analysis, (ICFCA2005)*, number 3403 in LNAI, Berlin, Heidelberg, February 2005. Springer.

[19] B. Ganter and L. Kwida, editors. *Supplementary Proceedings of the 4th International Conference on Formal Concept Analysis, (ICFCA2006)*, February 2006. Verlag Allgemeine Wissenschaft.

[20] B. Ganter and R. Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin, Heidelberg, 1999.

[21] R. Godin, E. Saunders, and J. Gecsei. Lattice model of browsable data spaces. *Information Sciences*, 40:89–116, 1986.

[22] R. Godin, J. Gecsel, and C. Pichet. Design of a browsing interface for information retrieval. In *Proceedings of the 12th International Conference on Research and Development in Information Retrieval (ACM SIGIR '89)*, pages 32–39, Cambridge, MA, 1989. ACM.

[23] R. Godin, G. Mineau, R. Missaoui, and H. Mili. Méthodes de classification conceptuelle basées sur les treillis de Galois et applications. *Revue d'intelligence artificielle*, 9(2):105–137, 1995.

[24] T. Hannan and A. Pogel. Spring-based lattice drawing highlighting conceptual similarity. *Formal Concept Analysis*, pages 264–279, 2006.

[25] M. Hearst. *Search User Interfaces*. Cambridge University Press, 2009. ISBN 9780521113793.

[26] B. Koester. FooCA: Enhancing Google information research by means of Formal Concept Analysis. In Ganter and Kwida [19], pages 1–17.

[27] D. Kuropka. *Modelle zur Repräsentation natürlichsprachlicher Dokumente. Ontologie-basiertes Information-Filtering und -Retrieval mit relationalen Datenbanken*. Advances in Information Systems and Management Science. logos-verlag.de, 2004.

[28] S. O. Kuznetsov and S. A. Obiedkov. Comparing performance of algorithms for generating concept lattices. *Journal Of Experimental & Theoretical Artificial Intelligence*, 14(2-3):189–216, Apr. 2002.

[29] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[30] B. Martin. Formal concept analysis and semantic file systems. In P. Eklund, editor, *Concept Lattices*, pages 88–95. Springer Berling Heidelberg, 2004.

[31] C. T. Meadow, B. R. Boyce, and D. H. Kraft. *Text Information Retrieval Systems*. Library and Information Sciences. Academic Press, San Diego and San Francisco, second edition, 2000.

[32] N. Messai, M. D. Devignes, A. Napoli, and M. Smail-Tabbone. BR-Explorer: An FCA-based algorithm for Information Retrieval. In *Proceedings of the 4th International Conference on Concept Lattices and their Applications, CLA '06*, 2006.

[33] C. Mooers. A mathematical theory of language symbols in retrieval. In *Proceedings of the International Conference on Scientific Information*, Washington, D.C., 1958.

[34] D. A. Norman. Affordance, conventions, and design. *Journal of the ACM*, 6(3):38–43, May 1999.

[35] R. Pedraza-Jiménez, F. J. Valverde-Albacete, and A. Navia-Vázquez. A generalisation of fuzzy concept lattices for the analysis of web retrieval tasks. In *Proceedings of the International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems, (IPMU'06)*, July 2006.

[36] S. Pollandt. *Fuzzy-Begriffe. Formale Begriffsanalyse unscharfer Daten.* Springer, Heidelberg, 1997.

[37] U. Priss. Lattice-based information retrieval. *Knowledge Organization*, 27 (3):132–142, 2000.

[38] U. Priss. Formal concept analysis in information science. In B. Cronin, editor, *Annual Review of Information Science and Technology (ARIST)*, pages 521–543. Information Today, Inc., Jan. 2006.

[39] T. Rock and R. Wille. Ein TOSCANA-Erkundungssystem zur Literatursuche. In Stumme and Wille [40], pages 239–253.

[40] G. Stumme and R. Wille, editors. *Begriffliche Wissensverarbeitung; Methoden und Anwendungen*, Heidelberg, 2000. Springer.

[41] S. Theodoridis and K. Koutroumbas. *Pattern Recognition.* Academic Press, third edition, 2006.

[42] F. J. Valverde-Albacete. Combining soft and hard techniques for the analysis of batch retrieval tasks. In E. Herrera-Viedma, G. Pasi, and F. Crestani, editors, *Soft Computing for Information Retrieval on the Web. Models and Applications*, volume 197 of *Studies in Fuzziness and Soft Computing*, pages 239–258. Springer, 2006. ISSN 1434-9922 (print edition). ISSN (electronic edition) 1860-0808, ISBN-10 3-540-31588-8, ISBN-13 978-3-540-31588-9.

[43] F. J. Valverde-Albacete and C. Peláez-Moreno. Towards a generalisation of formal concept analysis for data mining purposes. *Concept Lattices. Proceedings of the International Conference on Formal Concept Analysis (ICFCA 06)*, LNAI 3874:161—176, Dec 2006.

[44] F. J. Valverde-Albacete and C. Peláez-Moreno. Extending conceptualisation modes for generalised Formal Concept Analysis. *Information Sciences*, 181: 1888–1909, May 2011.

[45] C. van Rijsbergen. *The Geometry of Information Retrieval.* Cambridge University Press, Cambridge, New York,Melbourne, Madrid and Cape Town, 2004.