Investigating Aboutness Axioms using Information Fields

P.D. Bruza*
School of Information Systems
Queensland University of Technology
GPO Box 2134
Brisbane Q 4001
Australia
bruza@qut.edu.au

T.W.C. Huibers
Department of Computer Science
Utrecht University
P.O. Box 80.089
3508 TB Utrecht
The Netherlands
theo@cs.ruu.nl

Abstract

This article proposes a framework, a so called information field, which allows information retrieval mechanisms to be compared inductively instead of experimentally. Such a comparison occurs as follows: Both retrieval mechanisms are first mapped to an associated information field. Within the field, the axioms that drive the retrieval process can be filtered out. In this way, the implicit assumptions governing an information retrieval mechanism can be brought to light. The retrieval mechanisms can then be compared according to which axioms they are governed by. Using this method it is shown that Boolean retrieval is more powerful than a strict form of coordinate retrieval. The salient point is *not* this result in itself, but *how* the result was achieved.

1 Introduction

The logic based approach to information retrieval has been around for some time now. So far, a number of inference mechanisms, both strict and plausible, have been proposed for driving the retrieval process [15, 6, 4, 12]. Furthermore, the expressive power of the approach has been demonstrated by several authors [14, 13, 7, 2]. We feel, however, that its real potential has not yet been tapped. Research has not yet produced a powerful enough framework whereby information retrieval systems can be compared *inductively* instead of *experimentally*. A breakthrough in this area would mean that a theorem could be proven stating, for example, that vector space retrieval is more effective that Boolean retrieval. Such a result would not only spare us the efforts of experimentation, but more importantly, it would allow us to side step the controversies surrounding the experimental process. This paper is an attempt at laying some ground work for an inductive theory of information retrieval.

2 Informational Fundamentals

In information retrieval, the question of aboutness appears in a number of guises. For example, in order to drive the retrieval mechanism aboutness is studied as a relation between a document and a query. Sometimes, the question whether one document is about another document arises, for example, in document clustering. Aboutness between descriptors occurs within the framework of a characterization language, for example, a term independence assumption is just another way of expressing that two terms aren't about each other.

We begin by abstracting from notions such as descriptors, documents and queries and introduce the notion of an *information carrier*. The central theme of this paper is to axiomatize the notion of when one information carrier is about another one. A rather strict notion of aboutness will be employed:

> an information carrier i will be said to be *about* information carrier j if the information born by j holds in i

This definition can be found explicitly or implicitly in several papers on logic-based information disclosure [8, 15, 2]. The intuition behind it closely approximates the notion of a *model* in logic, for this reason the notation $i \models j$ will be used to denote that information carrier i is about carrier j. Note that this conception of aboutness is not applicable for document clustering where aboutness is determined by the overlap between respective document characterizations. The definition is however useful for studying aboutness between a document and a query.

*This work was performed while at Utrecht University

Information Carriers

Information carriers are descriptions of situations, real or possible worlds etc. These descriptions may be long and detailed, for example, a document, or may be short, for example, a query. Some information carriers convey more information about a situation(s) than others. According to Landman [11] and Barwise [1] information can be partially ordered with respect to information containment (denoted by $\rightarrow$):

$i \rightarrow j$ iff the information which i carries already *contains* the information which j carries

In other words, carrier i bears more information that carrier j. Bear in mind that carrier i "is less than" carrier j in the ordering. Information containment is related to specificity. For example, the information carrier little green martians is more specific than green martians which is more specific than martians. Note that little green martians $\rightarrow$ green martians $\rightarrow$ martians. The information containment relation $\rightarrow$ is reflexive, antisymmetric and transitive. Transitivity in the context of information containment is also known as the Xerox Principle [1].

Information Composition

A document can be seen as a composition of various pieces of information. It can therefore be viewed as a complicated description of a diverse range of situations, for example, an encyclopedia. The way information carriers can be composed to form complexer information carriers functions according to rules. At lower levels of information granularity, composition is typically governed by linguistic considerations, for example, how subelements of a sentence may be combined to form a sentence. Above the level of sentences, composition rules become structural in nature. For example, how sentences are composed to form paragraphs and paragraphs to sections etc. Document specification languages such as SGML are specifically designed to provide the rules for governing such structural composition.

As an illustration of information composition, consider the information carriers green martians and little martians. These can be composed to form the carrier little green martians. Note how the latter carrier bears precisely the information furnished by the combination of the two previous carriers. This is the fundamental property of information composition. More formally,

$i \oplus j$ is the largest (as defined by $\rightarrow$) information carrier that precisely contains the information born by i and by j.

Recent work with index expressions also shows how connectors can be used to combine characterizations [4, 2]. For example, river $\circ$ pollution and pollution in australia can be composed to render the carrier river $\circ$ pollution in australia. Within Boolean retrieval $\oplus$ is embodied by $\wedge$.

If one carries information composition to the extreme one attains total information, which according to Landman is *too much information* [11]. Imagine if all the information carriers of mankind were to be composed into a single carrier! The total information carrier will be denoted by ∞ which constitutes the bottom element of the partial order of information carriers. As a consequence, all information carriers are informationally contained in ∞. The carrier ∞ is intuitively similar to *falsum* in the context of the propositional calculus in the sense that all formulae can de derived from it.

Information Preclusion

Not all information carriers can be meaningfully composed. The reason for this is that they are incompatible; the information they share clashes, or is contradictory. In other words, carriers i and j are said to preclude each other, denoted $i \perp j$. It is natural to assume that facts (viewed as information carriers) preclude their negation, for example martians are green $\perp$ martians are not green. Information preclusion, however, is not restricted to being a relation over facts. It can be argued that green martians precludes blue martians because martians are either blue or green, but not both. This intuition behind this phenomenon can be explained in terms of possible worlds: After characterizing a world as being a "green martian" world, it cannot be re-characterized as a "blue martian" world. Note that little martians does not preclude green martians. Several authors regard information preclusion as being fundamental to a theory of information [11, 1].

Even though green martians precludes blue martians this does not necessarily mean that their composition isn't an information carrier. One can imagine an information carrier formed from the composition of the previous two carriers but the information it bears is discordant:

green martians	$\oplus$	blue martians	$\rightarrow$	green martians	and
green martians	$\oplus$	blue martians	$\rightarrow$	blue martians	*but*
green martians	$\perp$	blue martians			

An information carrier k is termed *harmonious* if and only if it doesn't bear discordant information. That is, there are no information carriers i, j such that $k \rightarrow i$ and $k \rightarrow j$ where $i \perp j$.

Information preclusion is sometimes defined in terms of composition [11]: Two information carriers i and j are deemed to preclude each other if and only if their composition results in total information: $i \perp j$ iff $i \oplus j = \infty$. In other words, non-harmonious information carriers are synonymous with total information. To illustrate this definition index expression based information carriers will be used [2].

Example 2.1

Figure 1 depicts a partial ordering of index expressions under information containment. In the context of this ordering and using the above definition of preclusion air ∘ poll ⊥ riv ∘ poll *and* aus ⊥ holl, *but* poll $\not\perp$ aus. □

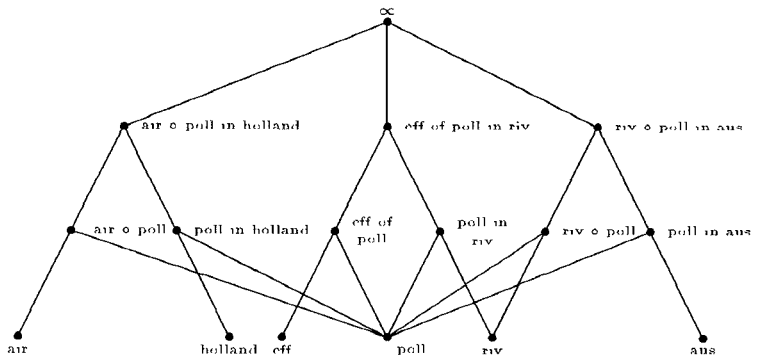

Figure 1. Example poset index expression carriers

Preclusion is interesting for information disclosure because if it is known that two information carriers preclude each other, then this may be used to determine aboutness. For example, if we know that a document is about green martians and that green martians precludes blue martians, we may then be able to infer that the document is *not* about blue martians.

Information Degradation

The converse of information composition is information degradation. The interesting facet of this notion is that it is related to uncertainty. For example, when one wants to retrieve all documents about i or j in Boolean retrieval systems, the expression $i \vee j$ is used. The uncertainty comes in as follows: if a given document is about $i \vee j$ it need not be about i. The expression $i \vee j$ can be considered to be a carrier which bears vague information, hence the term *information degradation*. We will not further deal with information degradation, although it does play a role in information retrieval. Instead, attention will be focussed on how aboutness behaves in the light of information containment, composition and preclusion.

Information Fields

To summarize, a framework has been proposed in which the notion of information carrier is fundamental. Such a framework is termed an *information field*. An information fields draws its underlying concepts from theories of information being currently developed in situation theory (specifically infon algebras [1]) and data semantics [11]. The intention is that an information field offers the necessary building blocks to axiomatize the the notion of aboutness employed in a given information retrieval mechanism. More about this in the next section.

Definition 2.1 (Information Field)

An information field *is a structure* $(\Im, \rightarrow, \oplus, \perp, \infty)$ *such that*

1. $\Im$ *is a non-empty set of information carriers*
2. $(\Im, \rightarrow)$ *is a poset*
3. $\infty \in \Im$ *and for all* $i \in \Im$, $\infty \rightarrow i$
4. *if* $i, j \in \Im$ *then* $i \oplus j \in \Im$, *where* $i \oplus j$ *is the largest information carrier such that* $i \oplus j \rightarrow i$ *and* $i \oplus j \rightarrow j$
5. $\perp \subseteq \Im \times \Im$

□

3 Axiomatizing Aboutness

We sometimes speak about information retrieval theory, but what are the axioms that govern this theory? Certainly they exist because the researchers who propose a retrieval mechanism entertain some assumptions as to what produces "good" retrieval. The problem is that these assumptions are often not expressed, and when they are, they tend not to be expressed in a general enough way to determine whether two different retrieval mechanisms are governed by the same or similar sets of assumptions. At the moment, the question as to whether the assumptions behind coordinate retrieval are the same as those that govern vector space retrieval cannot be satisfactorily answered. If it could, the two systems could be formally compared, for example, by using representation theorems. Such an approach has been used with success in comparing non-monotonic reasoning systems.

This section proposes a system of axioms expressed in terms of concepts from the information field. The term "axiom" should not be interpreted in a strict logical sense, but rather in a more intuitive fashion: The axioms are intended to characterize the *assumptions* inherent within a given retrieval mechanism with regard to aboutness. Retrieval mechanisms can then be compared according to which axioms they are governed by. The notion of aboutness is embodied by a binary relation $\models$ over the set of information carriers $\Im$. The first axiom posits that an information carrier is about itself.

Axiom 1 (Reflexivity (R))

$$i \models i$$

An information carrier is about the information it contains. This is the premise behind the Containment Axiom.

Axiom 2 (Containment (C))

$$\frac{i \rightarrow j}{i \models j}$$

Say we have a document d which is about green martians. It is natural to assume green martians $\rightarrow$ martians. Therefore, d is about martians. This is an example of Right Containment Monotonicity.

Axiom 3 (Right Containment Monotonicity (RCM))

$$\frac{k \models i \quad i \rightarrow j}{k \models j}$$

Right Containment Monotonicity is fundamental to many systems proposed in situation theory [1]. Note that Left Containment Monotonicity is not a sensible axiom for governing aboutness.

$$\frac{k \models j \quad i \rightarrow j}{k \models i}$$

For example, let d be a document which is about martians and green martians $\rightarrow$ martians. Left Containment Monotonicity permits the conclusion that d is about green martians, which does not necessarily have to be the case. Left Containment Monotonicity lurks behind the lack of precision in a recently proposed context free plausible inference mechanism [3].

Axiom 4 (Context-Free And (C-FA))

$$\frac{k \models i \quad k \models j}{k \models \ i \oplus j}$$

Boolean retrieval, for one, is founded on this axiom. For example, if a document d is about river and the same document is about pollution it is assumed that d is about river $\wedge$ pollution. Recent research has shown that this can be a dubious assumption, particularly at lower levels of information granularity [5, 16]. The problem lies in the fact that the carrier river $\wedge$ pollution bears implicitly the assumption that river and pollution are related which doesn't have to be the case. By way of illustration, river valleys $\oplus$ air pollution is about both river and pollution, but the carrier is not about river pollution. In order to alleviate this problem, a context sensitive approach can be adopted.

Axiom 5 (Context-Sensitive And (C-SA))

$$\frac{i \rightarrow c \quad c \models j \quad c \models k}{i \models \ j \oplus k}$$

The idea is that information composition may only occur if the component carriers j, k are a part of the same informational context c. As a matter of passing, the carrier c may be of a different type than i, as is demonstrated by the following example based on Farradane's relational indexing [9]. Farradane introduced nine relationship types between terms, for example, air **A** pollution denotes an *action* relationship type between the terms air and pollution. This carries the information that pollution is something which affects, or acts on the air. The relationship type can be considered as a context which relates two terms, so with d = river valleys $\oplus$ air pollution, the following would be a valid application of CS-A within Boolean retrieval:

$$\frac{d \rightarrow \text{air A pollution} \quad \text{air A pollution} \models \text{air} \quad \text{air A pollution} \models \text{pollution}}{d \models \text{air} \wedge \text{pollution}}$$

Note that it is not possible to affirm river $\wedge$ pollution as these terms are unrelated in the carrier d. Current retrieval systems tend only to support syntactic contexts, for example, the terms must appear in the same paragraph. Finally, the Context-Free And axiom is in fact derivable from the Context-Sensitive And by choosing the broadest possible context ($i = c$).

If a document is not about pollution, then it can't be about river pollution. This is the intuition behind the so called Negation Rationale.

Axiom 6 (Negation Rationale (NR))

$$\frac{k \not\models i}{k \not\models i \oplus j}$$

It has been recently proven that inference network models implicitly embody this property if the topology of the network is determined by the information containment relation [4].

If it can established that a component part of an information carrier i is about another carrier k, then composing i with another carrier j will not violate this, provided no preclusions are apparent. In other words, aboutness is preserved under composition.

Axiom 7 (Compositional Monotonicity (CM))

$$\frac{i \models k \quad j \not\perp k}{i \oplus j \models k}$$

Note that the carrier j must not preclude k in order to avoid paradoxes of aboutness. Assume that green martians $\models$ space creatures and earth creatures $\perp$ space creatures. Application of Compositional Monotonicity without looking at preclusion relations would allow the following:

$$\frac{\text{green martians} \models \text{space creatures}}{\text{green martians} \oplus \text{earth creatures} \models \text{space creatures}}$$

The paradox arises as follows: It follows from the definition of an information field that

$$\text{green martians} \oplus \text{earth creatures} \quad \rightarrow \quad \text{earth creatures}$$

and therefore, via the Containment Axiom:

$$\text{green martians} \oplus \text{earth creatures} \quad \models \quad \text{earth creatures}$$

but Rational Negation (see later) allows this result and the preclusion to yield:

$$\text{green martians} \oplus \text{earth creatures} \quad \not\models \quad \text{space creatures}$$

which is a contradiction of the unprecluded compositional monotonicity axiom. This example also demonstrates how the axioms can interact with each other in quite complex ways.

If information carriers preclude each other, then it doesn't seem unreasonable to assume that they are not about each other. Applications of this assumption can be readily found in information retrieval. For example, in Boolean retrieval the formula α precludes $\neg\alpha$ and certainly $\alpha \not\models \neg\alpha$ and vice versa. Furthermore, term vectors in vector space retrieval are orthogonal to each other. This is a geometric expression of information preclusion. These examples are concrete cases of the so called Preclusion Axiom:

Axiom 8 (Preclusion (P))

$$\frac{i \perp j}{i \not\models j \text{ and } j \not\models i}$$

For *harmonious* information carriers, an axiom governing when a carrier is *not* about another carrier can be stated. This axiom is illustrated as follows. Given that d = little green martians is harmonious. Furthermore, d is (via Containment) about green martians. Assuming, as before, that green martians $\perp$ blue martians, then Rational Negation permits the conclusion that d is *not* about blue martians.

Axiom 9 (Rational Negation (RN))

$$\frac{i \models j \quad j \perp k}{i \not\models k}$$

Last in the list of axioms pertinent to determining aboutness is the rather infamous *Closed World Assumption* which bedevils a number of retrieval models, most notably Boolean retrieval [14, 2].

Axiom 10 (Closed World Assumption (CWA))

$$\frac{i \not\models j \quad j \perp k \quad i \not\perp k}{i \models k}$$

There are a number of issues that deserve attention. Do the axioms capture all the properties of aboutness desirable for information retrieval? Are they complete? Are they sound? In the rest of this section, these important questions will be addressed.

One method to examine the soundness and completeness of the axioms is to examine them within the context of an existing logic. For this purpose, the propositional calculus is chosen. The basic idea is to treat propositional formulae as information carriers. Within this realm the notion of aboutness between formulae can neatly be mapped onto entailment. Information containment is captured by the inference mechanism of the propositional calculus. This corresponds to the intuition that the information contained by a given formula is all the theorems provable from it. Information composition is embodied by conjunction. Furthermore, a formula is deemed to preclude its negation. More specifically, let $\mathcal{B}(P)$ be the set of formulae based of the atomic propositions P and the connectives $\neg, \wedge, \Rightarrow$ and where $f \in P$ denotes *falsum*. Let $\vdash_{PC}$ and $\models_{PC}$ respectively denote the inference and entailment relations of propositional calculus. Furthermore, let $(\mathcal{B}(P), \vdash_{PC}, \wedge, \perp, f)$ be an information field such that $i \perp \neg i$ for $i \in \mathcal{B}(P)$.

To investigate soundness is it sufficient to prove that the axioms preserve the truth of $\models_{PC}$. Well known theorems for the propositional calculus can be applied. For example, the following is the proof for Right Containment Monotonicity:

$$\begin{array}{lll} & k \models_{PC} i \quad i \rightarrow j & \langle \texttt{completeness} \rangle \\ \Longrightarrow & k \vdash_{PC} i \quad i \rightarrow j & \langle \texttt{definition} \rangle \\ \Longrightarrow & k \vdash_{PC} i \quad i \vdash_{PC} j & \langle \texttt{cut} \rangle \\ \Longrightarrow & k \vdash_{PC} j & \langle \texttt{soundness} \rangle \\ \Longrightarrow & k \models_{PC} j & \end{array}$$

It turns out that, barring the Closed World Assumption, all other axioms examined within an information field based on the propositional calculus preserve truth. That is, the axiomatization is not sound in the context of this logic. As a consequence, the notion of aboutness cannot be mapped onto truth. The fact that the first nine axioms are sound is disappointing in the sense that it reveals that the axiomatization does not extend much further than the propositional calculus. In particular, non-monotonic behaviour, which is surely inherent in information retrieval, needs to be axiomatized.

The axioms are complete, that is: if $i \models_{PC} j$ then $i \models j$. This means that within the propositional calculus, truth can be mapped directly onto aboutness. The proof is as follows:

$$\begin{array}{lll} & i \models_{PC} j & \langle \texttt{completeness} \rangle \\ \Longrightarrow & i \vdash_{PC} j & \langle \texttt{definition} \rangle \\ \Longrightarrow & i \rightarrow j & \langle \texttt{containment} \rangle \\ \Longrightarrow & i \models j & \end{array}$$

4 Boolean Information Fields

In order to illustrate the axioms of the previous section in the context of information retrieval, we will first consider Boolean retrieval due to its simplicity and close relation with logic. To distill the axioms which govern Boolean retrieval requires a definition which establishes aboutness within this model. This definition functions as a foothold which can be used to examine which axioms are valid within the associated information field.

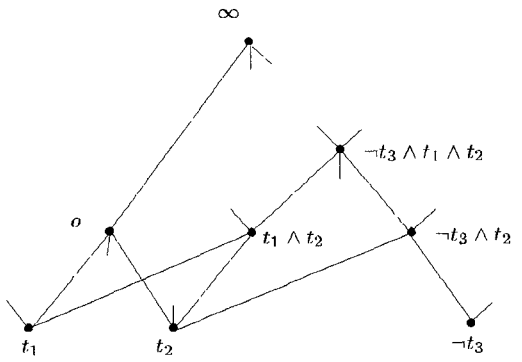

Figure 2. Part of a Boolean Information Field

Definition 4.1 (Boolean Aboutness)
Let T be a vocabulary (set of terms) and $\mathcal{B}(T)$ the Boolean language defined on T using the logical connectives $\neg, \vee, \wedge$. Furthermore, let $_{BR}$ be a set of strict inference rules defined on $\mathcal{B}(T)$. Let $\mathcal{O}$ be a set of objects (documents) whereby for $o \in \mathcal{O}, \chi(o)$ $(\chi(o) \subseteq T)$ denotes the characterization of o. Let $\alpha \in \mathcal{B}(T)$, then

$$o \models \alpha \quad \textit{iff} \quad \chi(o) \vdash_{BR} \alpha$$

□

In other words, a query α holds in a document o if and only if the query can be deduced from o's characterization using the strict inference rules which are defined as follows. For simplicity, only rules involving $\neg$ and $\wedge$ are specified:

1. if $\alpha \in \chi(o)$ then $\chi(o) \vdash_{BR} \alpha$
2. if $\chi(o) \vdash_{BR} \alpha$, $\chi(o) \vdash_{BR} \beta$ then $\chi(o) \vdash_{BR} \alpha \wedge \beta$
3. if $\chi(o) \nvdash_{BR} \alpha$ then $\chi(o) \vdash_{BR} \neg\alpha$

The basis of this inference mechanism (clause 1) can be explained in terms of information containment: $\alpha \in \chi(o)$ is nothing more than an affirmation that the index term α is *informationally contained* in o, hence $o \rightarrow \alpha$. Therefore, both objects and terms are considered information carriers in the Boolean Information Field. This demonstrates that information fields can consist of information carriers of different types. Furthermore, clause 1 and definition 4.1 imply $o \models \alpha$, hence a Boolean Information Field supports the Containment Axiom.

Complex Boolean formulae are also considered to be information carriers and are ordered informationally in a natural way whereby information composition $\oplus$ is modelled by $\wedge$: $\alpha \wedge \beta \rightarrow \alpha$ and $\alpha \wedge \beta \rightarrow \beta$. Note that in this set up information composition is only defined for formulae. There is no composition operator for objects. This is consistent with the view held in the Boolean retrieval model that objects are disjoint, amorphous things with no operators defined on them. As a consequence, the Compositional Monotonicity axiom is not applicable. To complete the Boolean Information field, all formulae are deemed to informationally preclude their negation: $\alpha \perp \neg\alpha$. An illustration of a part of the Boolean Information Field is depicted in figure 2.

Thus far we have established that Boolean Information Fields embody the Containment Axiom. Furthermore, it can directly shown that clause 2 functions according to the Context-Free And axiom as follows:

$$\dfrac{\dfrac{\dfrac{o \models \alpha}{\chi(o) \vdash_{BR} \alpha}\,\text{Def 4.1} \quad \dfrac{o \models \beta}{\chi(o) \vdash_{BR} \beta}\,\text{Def 4.1}}{\chi(o) \vdash_{BR} \alpha \wedge \beta}\,\text{Def 4.1}}{o \models \alpha \wedge \beta}\,\text{Def 4.1}$$

In a similar way, it can be shown that clause 3 cloaks the Closed World Assumption and that the Right Containment Monotonicity (RCM) axiom is also supported. Using RCM it can quickly be demonstrated that the Negation Rationale is also supported by proceeding *reductio ad absurdum*: Given there are

carriers i, j, k and $k \not\models i$, $k \models i \wedge j$ is assumed. By definition, $i \wedge j \rightarrow i$, and applying RCM

$$\frac{k \models i \wedge j \quad i \wedge j \rightarrow i}{k \models i}$$

leads to a contradiction. Therefore, $k \not\models i \wedge j$ must have been the case.

In order to investigate the Rational Negation axiom, we must first insure that all objects are *harmonious* information carriers. This turns out to be the case because $o \rightarrow t$ and $o \rightarrow \neg t$ cannot occur ($t \in \chi(o)$ and $\neg t \in \chi(o)$ is not possible in Boolean retrieval). It can then be shown that the Rational Negation property is supported by a Boolean Information Field. Analysis of the remaining aboutness axioms leads to the following theorem which states which axioms are supported by a Boolean Information Field.

Theorem 4.1 Let BF be a Boolean Information Field with $\models \subseteq \mathtt{BF}_\Im \times \mathtt{BF}_\Im$ and $\models$ defined as in Definition 4.1, then BF satisfies the axioms *C,CF-A,CWA,RCM,NR,RN.*

5 Strict Coordinate Information Fields

When providing a foothold definition for studying coordinate retrieval care must be taken not to confuse two different notions of aboutness. The match function which drives coordinate retrieval measures overlap, for example, between the set of terms characterizing a document d and the set of terms comprising the query q. One way of interpreting aboutness is to deem that d is about q if and only if the overlap between their respective characterizations is non-null. This interpretation of aboutness is adopted to promote recall, however, this is *not* the aboutness being investigated here. Remember q must hold in d, and this is the case when $q \subseteq \chi(d)$. Adopting this interpretation of aboutness allows the investigation of coordinate retrieval, but in a narrower sense. For this reason it will be referred to as *strict* coordinate retrieval.

Definition 5.1 (Strict Coordinate Aboutness)
Let T be a vocabulary, with $\alpha \subseteq T$. Furthermore, let $\mathcal{O}$ be a set of objects whereby for $o \in \mathcal{O}, \chi(o)$ ($\chi(o) \subseteq T$) denotes the characterization of o. Then,

$$o \models \alpha \quad \textit{iff} \quad \alpha \subseteq \chi(o)$$

□

The mapping of strict coordinate retrieval to an information field proceeds in a similar way to Boolean retrieval. Both object characterizations and queries are modelled as information carriers consisting of a set of terms. Information containment within this framework is modelled by the subset relation over $\wp(T)$. Furthermore, the indexing relation once again determines an information containment relation between objects and terms: if $t \in \chi(o)$ then $o \rightarrow \{t\}$. The information composition operator $\oplus$ is realized by set union. The notion of information preclusion is foreign to coordinate retrieval, hence $\perp = \emptyset$. It can be shown that the Strict Coordinate Information Field embodies the following axioms.

Theorem 5.1 Let CF be a Strict Coordinate Information Field with $\models \subseteq \mathtt{CF}_\Im \times \mathtt{CF}_\Im$ and $\models$ defined as in Definition 5.1, then CF satisfies the axioms *C,CF-A,RCM,NR*

Comparing the representation theorems from Boolean retrieval and strict coordinate retrieval, we see that the axioms supported by a Strict Coordinate Information Field is a subset of the axioms supported by a Boolean Information Field. Stated otherwise; a Boolean Information Field is more powerful than a Strict Coordinate Information Field. In terms of information retrieval this means that whatever object a strict coordinate retrieval mechanism can deliver can also be delivered by a Boolean retrieval mechanism. It seems that this result is also reflected in an algebraic setting as there exists a homomorphism from a Strict Coordinate field to a Boolean Field. The homomorphism h is defined as follows assuming the same underlying set of objects O and vocabulary T. First, the information carriers of the Strict Coordinate Information Field CF are mapped to the information carriers of the Boolean field BF.

$$\begin{aligned} h(\infty^{\mathtt{CF}}) &= \infty^{\mathtt{BF}} \\ h(o^{\mathtt{CF}}) &= o^{\mathtt{BF}} \\ h(t^{\mathtt{CF}}) &= t^{\mathtt{BF}} \end{aligned}$$

Information carriers consisting of a set of terms are mapped to a conjunction of those terms:

$$h(\{t_1, \ldots, t_n\}^{\mathtt{CF}}) = \bigwedge_{1 \leq i \leq n} h(t_i^{\mathtt{CF}})$$

Furthermore, if α, β are information carriers in the Strict Coordinate Field, information composition can be mapped as follows:

$$h(\alpha \cup \beta) \quad = \quad h(\alpha) \wedge h(\beta)$$

Algebraic properties are a potentially powerful tool for the comparison of information fields, and thus the associated information retrieval mechanisms.

6 Conclusions and Further Research

The theme of this article is the theoretical comparison of information retrieval mechanisms. Such a comparison occurs as follows: Both retrieval mechanisms are first mapped to an associated information field. Within the field, the axioms that drive the retrieval process can be filtered out. In this way, the implicit assumptions governing an information retrieval mechanism can be brought to light. The retrieval mechanisms can then be compared according to which axioms they are governed by. Using this method it is shown formally that Boolean retrieval is more powerful than a strict form of coordinate retrieval. The salient point is *not* this result in itself, but *how* the result was achieved.

As the theoretical comparison of retrieval mechanisms is in its infancy, there are many avenues for further research. To begin with, the investigation of more existing retrieval mechanisms is needed. We envisage that an ordering of retrieval mechanisms could result. The effectiveness of a new retrieval mechanism could then be examined, not by running experiments, but by mapping it appropriately into the ordering. Detailed investigation into aboutness axioms is also needed. In particular, the axioms presented here should be extended to model non-monotonic aspects of aboutness. It should be possible to investigate under which conditions axioms preserve aboutness within the context of a given information field. This could open the door to soundness results similar to those found in logic. Furthermore, representation theorems (like those used to compare non-monotonic reasoning systems [10]) backed by algebraic comparisons are an interesting avenue for further exploration. The ultimate goal is an inductive theory of information retrieval.

Acknowledgements

The authors thank Bernd van Linder for the interesting discussions on (modal) logic and for his keen and critical eye.

References

1. J. Barwise and J. Etchemendy. Information, Infons, and Inference. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications*, volume 1 of *CSLI Lecture Note Series*, pages 33–78. CSLI, 1990.
2. P.D. Bruza. *Stratified Information Disclosure: A Synthesis between Information Retrieval and Hypermedia.* PhD thesis, University of Nijmegen, 1993.
3. P.D. Bruza and L.C. van der Gaag. Efficient Context-Sensitive Plausible Inference for Information Disclosure. In *Proceedings of the Sixteenth ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 12–21, 1993.
4. P.D. Bruza and L.C. van der Gaag. Index Expression Belief Networks for Information Disclosure. *International Journal of Expert Systems*, 1994. (To appear).
5. J.P. Callan and W. Bruce Croft. An Evaluation of Query Processing Strategies using the TIPSTER collection. In *Proceedings of the Sixteenth ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 347–355, 1993.
6. Y. Chiaramella and J. Nie. A Retrieval Model based on an Extended Modal Logic and its Application to the RIME Experimental Approach. In *Proceedings of the 13th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 25–43, 1990.
7. Y. Chiarmarella and J.P. Chevallet. About Retrieval Models and Logic. *The Computer Journal*, 35(3):233–242, 1992.
8. W.S. Cooper. A Definition of Relevance for Information Retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.
9. J. Farradane. Relational Indexing Part I. *Journal of Information Science*, 1(5):267–276, 1980.
10. S. Krauss, D. Lehmann, and M. Magidor. Nonmonotonic Reasoning, Preferential Models and Cumulative Logics. *Artificial Intelligence*, 44:167–207, 1990.

11. F. Landman. *Towards a Theory of Information.* Foris, 1986.

12. M. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A Model of information Retrieval based on Terminological Logic. In *Proceedings of the Sixteenth ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 298–307, 1993.

13. J. Nie. An Outline of a General Model for Information Retrieval Systems. In *Proceedings of the Ninth ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 495–506, 1986.

14. C.J. van Rijsbergen. A New Theoretical Framework for Information Retrieval. In *Proceedings of the Ninth International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 194–200, 1986.

15. C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29(6):481–485, 1986.

16. G. Salton, J. Allan, and C. Buckley. Approaches to Passage Retrieval in Full Text Information Systems. In *Proceedings of the Sixteenth ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 49–58, 1993.