Situations, a General Framework for Studying Information Retrieval

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

*P.D. Bruza**
School of Information Systems
Queensland University of Technology
GPO Box 2434
Brisbane Q 4001
Australia
bruza@qut.edu.au

Abstract

This paper presents a framework for the theoretical comparison of information retrieval models based on how the models decide aboutness. The framework is based on concepts emerging from the field of situation theory. So called *infons* and *profons* represent elementary information carriers which can be manipulated by union and fusion operators. These operators allow relationships between information carriers to be established. Sets of infons form so called situations which are used to model the information born by objects such as documents. An arbitrary information retrieval model can be mapped down into the framework. Special functions are defined for this purpose depending on the model at hand. An important aspect is the inference mechanism which is mapped to inference between situations. Two examples are given based on the Boolean retrieval and coordination level matching models. The framework allows the comparison of retrieval models at an abstract level. Starting from an axiomatization of aboutness, retrieval models can be compared according to which axioms they are governed by. This approach is highlighted by the theoretical comparison of Boolean retrieval with coordinate level matching.

*This work was partly performed while employed at the *Utrecht University*.

1 Introduction

In recent years the logical approach to information retrieval has gained quite a deal of attention (see for example [Bru93],[CC92],[LR92],[Nie92], and [Rij89]). Various inference mechanisms have been proposed. Furthermore, the expressive power of the logical framework has been demonstrated by several authors. However, expressiveness has been restricted showing how existing retrieval models can be mapped to logical inference mechanisms.

In addition a number of principles have been launched like for example the *"logical uncertainty principle"* [Rij86]. Notions such as the soundness and completeness of an inference mechanism have been defined within the context of information retrieval, but so far there have not been soundness and completeness results as is typical in logic. This seems to indicate that the logical framework used in information retrieval does not have an arsenal powerful enough to allow the proofs of such results. In particular an axiomatization of information retrieval theory lacks. If such an axiomatization were available, information retrieval systems could be classified according to which axioms they are governed by. This opens the door to an inductive information retrieval theory.

The theme of this paper is an axiomatization of aboutness in information retrieval. Section 2 introduces an information theory based on the situation theory. In section 3 information retrieval concepts are defined in the context of this situation theory. We formulate axioms that describe some properties of aboutness as used in information retrieval. In section 4 two information retrieval mechanisms will be modelled as situation inference, namely the Boolean inference mechanism and the coordination level matching inference mechanism, which will be theoretically compared in section 5. At the end we summarize and discuss options for further research.

2 Modelling Information Retrieval Concepts using Situation Theory

Information retrieval is the problem of retrieving those documents that are likely to be relevant for a certain information need. In 1971, Cooper introduced an objective notion of relevance termed *logical relevance*.

> A stored sentence is *logically relevant* to a representation of an information need if and only if it is a member of some minimal premiss set of stored sentences for some component statement of that need [Coo71].

He added the remark that if you want to look more closely to which part of a document is judged as being relevant, the following auxiliary definition may be adopted:

> A document is relevant to an information need if and only if it contains at least one sentence which is relevant to that need.

This last definition we take as our point of departure. However, in order to avoid the heavily loaded word *relevance*, we will instead use the term *aboutness*. Our goal is to study aboutness within a general theory of information. Such theories are emerging from the field of situation theory (see for example [BE87], [BE90], and [Dev91]). We propose that situation theory is a sufficiently powerful theory of information for studying the aboutness relation between documents or between a document and a query. Lalmas & Van Rijsbergen have recently expounded the merits of situation theory in information retrieval [LR92]. Our approach differs to theirs as we view situation theory not as a tool to drive information retrieval but as a vehicle to analyze theoretical properties of information retrieval mechanisms.

2.1 Infons: the Atomic Information Carriers

In order to go beyond the rather trivial approaches of equating aboutness with some overlap measure between two respective object characterizations, a more sophisticated theory of information must be available. The starting point for such a theory can be found in Farradane's relational indexing [Far80a, Far80b]. In Farradane's work, information was carried by a

fixed set of relationship types over an underlying set of terms. This conception bears a close resemblance to the fundamental information carriers of situation theory, the *infons* [BE87],[BE90],[Dev91]:

Definition 2.1 (Infons) An infon is a structure $\langle\langle \mathrm{R}, a_1, \ldots, a_n; i \rangle\rangle$ that represents the information that the relation R holds (if $i = 1$) or does not hold (if $i = 0$) between the objects $a_1, \ldots, a_n$.

The value i is referred to as the polarity of the infon. As the name suggests, this value is used to denote positive ($i = 1$) or negative ($i = 0$) information. In relational indexing, the objects are terms, which are or are not related to each other. A concrete example of an infon suitable for information retrieval is the following, which is based on the relational indexing approach. Imagine that there is a very small document comprising the text "The author wrote a book". The information carried by this document can be modelled by the infon: $\langle\langle \mathrm{F}, \texttt{author}, \texttt{book}; 1 \rangle\rangle$. The letter F denotes a functional relationship type between the terms `author` and `book`. Such relationships could be used to provide contextual information for driving information retrieval. Due to practical considerations, however, it is still not possible to use this idea in real-life information retrieval systems. Typically the information inherent in documents is partially modelled by a set of terms called *keywords*, simply because there are efficient algorithms to automatically index such keywords from the document. As a result the relationships in which the keywords were involved are no longer available. Information retrieval systems essentially try to infer what the relationships could have been when trying to determine whether a document is about a given information need.

Using keywords to model information results in very primitive infons. In a sense these keyword based infons can be considered "sub-informational" particles just like protons are to atoms. For this reason they will be referred to as *profons*. Profons are a sub-class of infons based on a relation I. This relation I signifies an unspecified unary relation reflecting the fact that the indexing process has deprived us of all knowledge of relations that the keyword was a part of. If a document d is indexed with the keyword `author`, one can consider that the information conveyed by `author` is inherent, or holds in d. The notation $\langle\langle \mathrm{I}, \texttt{author}; 1 \rangle\rangle$ will be used to denote the corresponding profon. It is interesting to note that current information retrieval research focusses on positive infons, particularly with regard to indexing. To the authors' knowledge there are no indexing systems which produce 'negative' keywords, that is, profons with polarity zero.

Up till now we have viewed infons and profons as a mechanism that describes keywords. The question remains as to how documents are to be modelled. As stated earlier, documents are information carriers and as such they can be modelled as a set of infons. This set is termed a *situation* and is an abstract representation of the information born by the document.

2.2 Informational Fusion and Union

A feature of information is that it can be manipulated. For example, two pieces of information can be composed to form a new piece of information. At the level of infons, composition is achieved by special operators depending on the strength of the composition required. Consider the keywords `water` and `pollution`. These can be combined together to form the phrase `water pollution`. This is an example of *informational fusion* as the respective keywords are composed very tightly. Note how `water pollution` bears precisely the combination of the information born by `water` and `pollution`. Fusion is modelled at the level of infons by the operator $\odot$: $\langle\langle$I,`water`; 1$\rangle\rangle\odot\langle\langle$I,`pollution`; 1$\rangle\rangle$. The result is an infon $\langle\langle$R,`water`,`pollution`; 1$\rangle\rangle$ that is based on the two keywords standing in a relation R drawn from a set of predefined relations. Fusion is used by the information retrieval mechanism to try and approximate the original relations lost in the indexing process.

It is also possible to fuse infons whose underlying relations differ in arity. For instance, fusing the infon $\langle\langle$R,`water`, `pollution`; 1$\rangle\rangle$ with the infon $\langle\langle$I,`river`; 1$\rangle\rangle$ results in
$\langle\langle$R',`river`,`water`,`pollution`; 1$\rangle\rangle$. The latter infon expresses the fact that there is some relation R' between the keywords `river`, `water`, and `pollution`.

Sometimes it is not possible, or not desirable, to actually fuse the infons. In such cases *informational union* can be employed, whereby the respective infons are combined to form a situation which models that the infons are unrelated. Informational union is denoted by $\oplus$. For example, $\langle\langle$I,`water`; 1$\rangle\rangle$ $\oplus$ $\langle\langle$I,`air`; 1$\rangle\rangle$ yields the situation $\{\langle\langle$I,`water`; 1$\rangle\rangle$, $\langle\langle$I,`air`; 1$\rangle\rangle\}$.

To summarize, given $x, y \in \mathcal{I}$ the set of infons and R $\in \mathcal{R}$, the set of relations, then:

$$\begin{aligned} x \oplus y &= \{x, y\} \\ x \odot y &= \langle\langle \text{R,x,y; i}\rangle\rangle \quad i \in \{0, 1\} \end{aligned}$$

We assume, for the moment, that $\odot$ is idempotent, commutative and associative. Note that by definition $\oplus$ has all these properties.

Infons constitute the lowest level of information granularity. At a higher level of granularity we find the situations, or in IR-terminology, the documents or queries. There is a natural way to generalize the information fusion operation to work on situations.

Two situations can be composed together to form a new situation, in the same way as the infons. First we handle the fusion of situations, (see this as composing some information together). One way of defining situation fusion is by fusing each infon of the first situation with all the infons from the second situation. Situation union, on the other hand combines all the infons of the first situation together with those of the second one.

$$\begin{aligned} S \odot T &= \{x \odot y \mid x \in S \wedge y \in T\} \\ S \oplus T &= S \cup T \end{aligned}$$

Having generalized the fusion and union to the level of situations, they can now be used as connectives in a language that is designed to reason with situations. The set of situations will be denoted by $\mathcal{S}$.

2.3 The Aboutness Relation between Situations

The relation $\mid\!\!\sim$ between two situations expresses that one situation carries information about another situation. In the literature, one can find a number of informal characterizations of this notion like *"topically related"* [Coo71], *"about"* [Mar77], *"likely to contain information about"* [Rij92], and *"correspondent to"* [Nie92]. As mentioned earlier, in this article the neutral term *aboutness* is preferred. It is formalized as a relation over situations: $\mid\!\!\sim \subseteq \mathcal{S} \times \mathcal{S}$. So, $(S, T) \in \mid\!\!\sim$ or $S \mid\!\!\sim T$ signifies that situation S is about situation T, and $S \not\mid\!\!\sim T$ denotes that S is not about T. Using the notion of situation, Cooper's original definition of logical relevance can be translated into *situation relevance*, or the preferred term *situation aboutness* as follows:

> A situation S is about a situation T if and only if T contains at least one infon i such that situation S is about that infon i.

Note that this definition consists of two implications. Firstly, situation S has not to be about the *complete* situation T. Secondly, the sentence that

"a situation is about one infon" implies the question: "how do we infer that a situation is about a given infon?"

2.4 Inferring Information from Situations

Inference is an integral part of situation theory [BE90]. For example, from the infon
$\langle\langle$R,`water`,`pollution`; 1$\rangle\rangle$ the infon $\langle\langle$I,`pollution`; 1$\rangle\rangle$ can be inferred, as the latter infon is intuitively implied by the former. Inferences of this sort are available to model the strict inference mechanism of a retrieval system. Many proponents of situation theory restrict their attention to strict inference. However, due to the inherent uncertainties of IR we are forced to take plausible inferences of the following sort into account: $\langle\langle$R,`air`,`pollution`; 1$\rangle\rangle$ is inferred from $\langle\langle$I,`pollution`; 1$\rangle\rangle$.

In our model, the symbol $\approx\!\!\!\Rightarrow_{\mathrm{IR}}$ will be used to denote (plausible) inference, that is $\alpha \approx\!\!\!\Rightarrow_{\mathrm{IR}} i$ denotes that infon i can be (plausibly) deduced from the set of infons α within the framework of information retrieval model IR. Just as with the informational union and fusion operators, inference can be generalized to the level of situations in the following way:

$$S \approx\!\!\!\Rightarrow_{\mathrm{IR}} T \quad \text{if and only if } \forall_{i \in T}[S \approx\!\!\!\Rightarrow_{\mathrm{IR}} i]$$

This results in a sufficiently abstract framework in which can be captured the (plausible) inference mechanism of an arbitrary retrieval mechanism, and maps it to inference between situations. In addition, it allows the above informal definition of situation aboutness to be formalized. A situation S is about a situation T if and only if there is an infon i in T which can be inferred from S (see figure 1).

Definition 2.2 (Situation Aboutness)

$$S \;\; |\!\rightsquigarrow \;\; T \text{ if and only if } \exists_{i \in T}[S \approx\!\!\!\Rightarrow_{\mathrm{IR}} i]$$

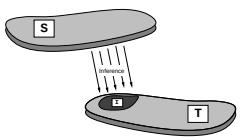

Figure 1: Situation Aboutness

The situation S is termed the *root*, as this situation is the starting point for the inference process. The situation T will be referred to as the *goal* situation.

3 Studying Information Retrieval in the context of Situation Theory

The purpose of an information retrieval system is, with respect to a given information need, to return as many relevant documents as possible while returning a minimum number of irrelevant ones. In this regard a set $\mathcal{D}$ of documents is assumed. In addition to a document, there is the document characterization denoted $\chi(d)$, for $d \in \mathcal{D}$. The document characterization is an approximate representation of the document's content. For example, in typical information retrieval systems a document is characterized by a set of keywords. Information retrieval begins when the user enters a request which is also typically based on a set of keywords. The keywords in the query attempt to characterize the user's information need.

Documents contain information. As such, a document can be modelled as a set of infons, or in other words, as a situation. We will assume a situation S_d which corresponds to the document d. Just as with documents there is a situation $S_{\chi(d)}$ corresponding to the characterization $\chi(d)$. In reality, $S_{\chi(d)}$ is a crude approximation of S_d. A query can be seen as a request for information. It can therefore also be represented as a situation. In summary, all aspects which play a role in information retrieval can be mapped into situation theory, documents as well as the characterization of the documents. What remains now is the question of how to express information retrieval effectiveness in the situation theoretic framework.

3.1 Recall and Precision

As mentioned in the introduction, the theme of this paper is to provide an axiomatization of the aboutness relation whereby the behavior of an information retrieval system can be characterized. The advantage of this approach is that it opens the possibility for the theoretical comparison of systems according to which axioms they are governed by. The axiomatization presented in the next section consists of nine axioms. In order to be able to motivate how a given rule affects recall and precision, it is necessary to bring these concepts within the bounds of our situation theoretic framework. Basically the recall and precision of a retrieval system will be studied within the context of situation inference, which is the situation theoretic counterpart of the retrieval process.

Definition 3.1 (Honest Situation Inference) *The inference mechanism of an information retrieval system is termed* honest *w.r.t.* $\chi(d)$ *if and only if the following holds:*

$$\textit{if } S_{\chi(d)} \ \mid\!\!\sim \ T \ \textit{then } S_d \ \mid\!\!\sim \ T$$

Let us assume that the situation T corresponds to a query. This definition states that if a situation corresponding to a document characterization is about T, then the associated document situation is as well. Honest situation inference corresponds to an information retrieval system with maximal precision.

Definition 3.2 (Total Situation Inference) *The inference mechanism of an information retrieval system is called* total *w.r.t.* $\chi(d)$ *if and only if the following holds:*

$$\textit{if } S_d \ \mid\!\!\sim \ T \ \textit{then } S_{\chi(d)} \ \mid\!\!\sim \ T$$

With the assumption that the situation T corresponds to a query, this definition states that if a situation corresponding to a document is about T, then the situation of the characterization is also about T. Total situation inference corresponds to an information retrieval system with maximal recall.

Some information retrieval systems are designed to be precise, that is an inference mechanism as honest as possible, to estimate a maximal precision. Other systems attempt to approximate a maximal recall, with a total inference mechanism. In the next section an indication is given which axioms characterize honest inference mechanisms and which one characterizes total inference mechanisms.

3.2 Reasoning with Situation Aboutness

In our theory *aboutness* is treated as a relation between situations. Therefore aboutness is treated as a fundamental notion with regard to information. This differs from other approaches, in which aboutness can be expressed in terms of so called information containment. In this section a set of axioms is presented as a series of rules which establish properties of the aboutness

relation between situations. Note that the axioms should not be interpreted in a strict logical sense. We shall see later that some axioms are not universally valid but only hold within the context of a particular retrieval system. This offers the possibility to compare retrieval systems according to which axioms they satisfy.

The interpretation of the following rules is as usual in logical systems, i.e., $\boxed{\begin{array}{c} A \\ \hline B \end{array}}$ means that if A is valid in an information retrieval model, then B is also valid. Note that if B consists of two parts separated by an operator (*or*, *and*), then a validity is supposed to distribute over this operator.

Basic Axioms

The first axiom of aboutness states that a situation is about itself.

Reflexivity:	$S \mid\!\mapsto S$

Reflexivity seems to be an inherent property of many retrieval systems.

Symmetry:	$\dfrac{S \mid\!\mapsto T}{T \mid\!\mapsto S}$

Symmetry expresses the requirement that there is no difference between concluding that a situation S is about a situation T, or concluding that a situation T is about a situation S. In some retrieval systems, for example, Boolean retrieval, symmetry is precluded by the strict inference mechanism. Coordination Level Matching turns out to be symmetric. Symmetry is reflected in several retrieval models wherein the matching function is based on some overlap measure, for example, vector space retrieval. The overlap measure is primarily intended to promote recall. At the level of situation inference, the symmetry property is a derivative of plausible inference. In short, symmetry promotes totality at the expense of honesty.

Transitivity:	$\dfrac{S \mid\!\mapsto T \qquad T \mid\!\mapsto U}{S \mid\!\mapsto U}$

Transitivity states that if you conclude that $S \mid\!\sim T$ and $T \mid\!\sim U$ you are allowed to draw the conclusion that $S \mid\!\sim U$. Transitivity is not reflected in those models wherein the matching function is based on some overlap measure.

Union Axioms

Monotonic Union:

$$\text{(Left)}\quad \frac{S \mid\!\sim T}{S \oplus U \mid\!\sim T}$$

$$\text{(Right)}\quad \frac{S \mid\!\sim T}{S \mid\!\sim T \oplus U}$$

The term "monotonicity" stems from the fact that aboutness is preserved under informational union. Note that there are two versions of the monotonic union property depending on whether the union takes place at the root situation (Left Monotonic Union) or at the goal situation (Right Monotonic Union).

An example of Left Monotonic Union is the following. The example is based on situations constructed from profons. For reasons of notational convenience only the underlying keyword is denoted. Given that `water` $\mid\!\sim$ `river`, a new situation is formed by informationally uniting `water` and `pollution`. Left Monotonic Union allows us to conclude that `water`$\oplus$`pollution` is about `river`.

Union Decomposition:

$$\text{(Left)}\quad \frac{S \oplus T \mid\!\sim U}{(S \mid\!\sim U) \text{ or } (T \mid\!\sim U)}$$

$$\text{(Right)}\quad \frac{S \mid\!\sim T \oplus U}{(S \mid\!\sim T) \text{ or } (S \mid\!\sim U)}$$

The decomposition property expresses that if a union of two situations is about a given situation, it is always possible to decide which part of this union is about that situation (or both). By way of illustration using Left Union Decomposition: Given that `water` $\oplus$ `pollution` $\mid\!\sim$ `river`, then `water` $\mid\!\sim$ `river`, or, `pollution` $\mid\!\sim$ `river`, or both.

Negative Union:

$$\text{(Left)}\quad \frac{S \not\mid\!\sim U \qquad T \not\mid\!\sim U}{S \oplus T \not\mid\!\sim U}$$

$$\text{(Right)}\quad \frac{S \not\mid\!\sim T \qquad S \not\mid\!\sim U}{S \not\mid\!\sim T \oplus U}$$

Negative Union is the converse of Monotonic Union. Left Negative Union states that given two situations are not about a goal situation, then informational union cannot induce aboutness. Right Negative Union takes the root situation as the focus. By way of illustration using Right Negative Union: Assuming that `fire` $\not\mid\!\!\rightarrow$ `river` and `fire` $\not\mid\!\!\rightarrow$ `pollution` then, `fire` $\not\mid\!\!\rightarrow$ `river` $\oplus$ `pollution`.

Fusion Axioms

Monotonic Fusion:

$$\text{(Left)}\quad \frac{S \mid\!\!\rightarrow T}{S \odot U \mid\!\!\rightarrow T}$$

$$\text{(Right)}\quad \frac{S \mid\!\!\rightarrow T}{S \mid\!\!\rightarrow T \odot U}$$

Monotonic Fusion states that, in systems which satisfy this axiom, aboutness is preserved under fusion. For example, given that `river` $\mid\!\!\sim$ `water`, Left Monotonic Fusion permits the conclusion that `river` $\odot$ `pollution` $\mid\!\!\sim$ `water`.

Fusion Decomposition:

$$\text{(Left)}\quad \frac{S \odot T \mid\!\!\rightarrow U}{(S \mid\!\!\rightarrow U) \text{ and } (T \mid\!\!\rightarrow U)}$$

$$\text{(Right)}\quad \frac{S \mid\!\!\rightarrow T \odot U}{(S \mid\!\!\rightarrow T) \text{ and } (S \mid\!\!\rightarrow U)}$$

Fusion Decomposition expresses that aboutness divides over fusion. For example, given that `water` $\mid\!\!\sim$ `river` $\odot$ `pollution`, Right Fusion Decomposition permits the conclusion `water` $\mid\!\!\sim$ `river` *and* `water` $\mid\!\!\sim$ `pollution`. This conclusion may seem strange but it can be defended in the following manner. The initial premise was that `water` $\mid\!\!\sim$ `river` $\odot$ `pollution` meaning that it is possible to infer from `water` the infon `river` $\odot$`pollution`. The latter infon contains precisely the information inherent in the combination of `river` and `pollution`. Therefore it can be argued that a part of the pollution contains river pollution and hence water is referring to that part of the pollution.

Negative Fusion:

$$\text{(Left)}\quad \frac{S \not\mapsto U \qquad T \not\mapsto U}{S \odot T \not\mapsto U}$$

$$\text{(Right)}\quad \frac{S \not\mapsto T \qquad S \not\mapsto U}{S \not\mapsto T \odot U}$$

Left Negative Fusion states that given two situations are not about a goal situation, then informational fusion cannot induce aboutness. Right Negative Fusion takes the root situation as the focus. If `water` $\not\mapsto$ `air` and `pollution` $\not\mapsto$ `air`, then Left Negative Fusion permits the conclusion `water` $\odot$ `pollution` $\not\mapsto$ `air`.

4 Two Information Retrieval Mechanisms modelled as Situation Inference

4.1 The Boolean Inference Mechanism

In Boolean retrieval the characterization consists of a set of terms which originate from a vocabulary $\mathcal{T}$. In particular, $t_1 \in \chi(d)$ is a reflection of the assumption that the document d is about term t_1. Stated in terms of situation theory, the profon corresponding to t_1 holds in the situation S_d. If $t_2 \in \chi(d)$, the Boolean retrieval operates under the assumption that d is about the fusion of t_1 and t_2 whereby the fusion relation R is embodied by the logical conjunction operator $\bigwedge$.

In order to translate Boolean retrieval to situations, the function *sit* is used. To begin with, the primitive propositions, or terms are translated. Let $t \in \mathcal{T}$, then

$$sit(t) = \{\langle\langle \mathrm{I}, t; 1\rangle\rangle\}$$

Situations corresponding to a document characterization ($S_{\chi(d)}$), consist of a fusion of $\mid \chi(d) \mid$ infons:

$$S_{\chi(d)} = \bigodot_{t \in \chi(d)} sit(t)$$

Example 4.1 (Julius Caesar) Let us consider two information carriers: information carrier d_1 carries the information that "Caesar $l$oves Brutus" and information carrier d_2 "Antony $h$ates Brutus". With the characterization language described above the information carriers could be transformed to as follows:

$$\begin{aligned}
\mathcal{T} &= \{C, L, B, A, H\} \\
\chi(d_1) &= \{C, L, B\} \\
\chi(d_2) &= \{A, H, B\} \\
S_{\chi(d_1)} &= sit(C) \odot sit(L) \odot sit(B) \\
S_{\chi(d_1)} &= \{\langle\langle \bigwedge, C, L, B;\ 1\rangle\rangle\} \\
S_{\chi(d_2)} &= \{\langle\langle \bigwedge, B, A, H;\ 1\rangle\rangle\}
\end{aligned}$$

In Boolean retrieval the request is specified as a formula. These formulas are constructed from the vocabulary $\mathcal{T}$ using the logical connectives $\wedge$, $\vee$ and $\neg$. We have seen that $\wedge$ drives a process of information fusion. Informational union, on the other hand, is driven by $\vee$. A formula may contain negation, expressing for example that the user wants documents that are not about a certain keyword t. All documents that do not have t in their characterization satisfy the query $\neg t$. In other words, Boolean retrieval operates under a *Closed World Assumption* [Rij86]. Negation is modelled at the level of profons as follows: $sit(\neg t) = \{\langle\langle \mathrm{I}, \bar{t};\ 1\rangle\rangle\}$.[1]

Without loss of generality we require the formula to be in a disjunctive normal form. The translation of Boolean formulas to situations, proceeds as follows:

$$\begin{aligned}
sit(t) &= \{\langle\langle \mathrm{I}, t;\ 1\rangle\rangle\} \\
sit(\neg t) &= \{\langle\langle \mathrm{I}, \bar{t};\ 1\rangle\rangle\} \\
sit(\psi_1 \vee \psi_2 \vee \ldots \vee \psi_n) &= sit(\psi_1) \oplus sit(\psi_2) \oplus \ldots \oplus sit(\psi_n) \\
sit(\psi_1 \wedge \psi_2 \wedge \ldots \wedge \psi_n) &= sit(\psi_1) \odot sit(\psi_2) \odot \ldots \odot sit(\psi_n)
\end{aligned}$$

The function sit is a bijective function, therefore a function f can be defined that will, given a situation as input, result in a corresponding formula. The function f is the inverse of the sit-function.

[1] The polarity 0 is not used to model negation, because we want to explicitly express that the negation of the keyword *holds*.

Example 4.2 (Caesar) Given the query-formula $(C \wedge B) \vee \neg(A \vee \neg L)$. First we bring this formula, as required, in a disjunctive normal form. $(C \wedge B) \vee (\neg A \wedge L)$. After using the sit-function, the result will be
$\{\ \langle\langle\bigwedge,C,B;1\ \rangle\rangle,\langle\langle\bigwedge,\bar{A},L;\ 1\rangle\rangle\ \}$

In Boolean Retrieval, aboutness is driven by an inference mechanism similar to that of the propositional calculus (see [Bru93] for more details).

Definition 4.1 (Boolean Retrieval Situation Aboutness) Let S be a situation, i an infon, and $\vdash$ the inference mechanism of Boolean retrieval: $S \approx\!\!\!\Rightarrow_{\mathrm{BS}} i$ if and only if $f(S) \vdash f(\{i\})$.

Example 4.3 (Caesar) $\{\langle\langle\bigwedge,C,L,B;1\rangle\rangle\} \approx\!\!\!\Rightarrow_{\mathrm{BS}} \langle\langle \mathrm{I},C;1\rangle\rangle$ as $(C \wedge L \wedge B) \vdash C$

Still using our *Julius Caesar* example, the Boolean retrieval system, and the defined $\approx\!\!\!\Rightarrow_{\mathrm{BS}}$, we can prove that a situation is about another situation. Using definition 4.1:

Example 4.4 (Caesar)

$$\begin{aligned} S_q &= \{\langle\langle\bigwedge,C,\bar{A};\ 1\rangle\rangle, \langle\langle\bigwedge,\bar{A},L;\ 1\rangle\rangle\} \\ S_{\chi(d_1)} &= \{\langle\langle\bigwedge,C,L,B;\ 1\rangle\rangle\} \\ S_{\chi(d_2)} &= \{\langle\langle\bigwedge,B,A,H;\ 1\rangle\rangle\} \end{aligned}$$

It can be shown that $S_{\chi(d_1)} \ \ |\!\!\rightsquigarrow\ \ S_q$ as follows: $S_{\chi(d_1)} \approx\!\!\!\Rightarrow_{\mathrm{BS}} \langle\langle\bigwedge,C,\bar{A};\ 1\rangle\rangle$, because,

$$\begin{array}{lcll} C \wedge L \wedge B & \vdash & C & \\ C \wedge L \wedge B & \nvdash & A & \textit{closed world assumption} \\ C \wedge L \wedge B & \vdash & \neg A & \\ C \wedge L \wedge B & \vdash & C \wedge \neg A & \\ S_{\chi(d_1)} & \approx\!\!\!\Rightarrow_{\mathrm{BS}} & \langle\langle\bigwedge,C,\bar{A};\ 1\rangle\rangle & \end{array}$$

and $\langle\langle\bigwedge,C,\bar{A};\ 1\rangle\rangle \in S_q$.
On the other hand $S_{\chi(d_2)} \ \ |\!\!\not\rightsquigarrow\ \ S_q$ as there is no infon $i \in S_q$, such that $S_{\chi(d_2)} \approx\!\!\!\Rightarrow_{\mathrm{BS}} i$. The reason for this is that for all $i \in S_q \ \ f(S_{\chi(d_2)}) \nvdash f(\{i\})$

4.2 The Coordination Level Matching Inference Mechanism

Just as was the case with Boolean retrieval, documents in coordination level matching are characterized by a set of keywords drawn from the vocabulary $\mathcal{T}$. The information need, however, is not represented a formula but as also as a set of keywords. In coordination level matching situations corresponding to a document characterization consist of the informational union of the terms in the characterization. This is fundamentally different from the approach taken in Boolean retrieval because informational union presupposes no relationships between the terms. Another fundamental difference is that coordination level matching has no notion of negation. The situation $S_{\chi(d)}$ corresponding to characterization $\chi(d)$ is rendered as follows:

$$S_{\chi(d)} = \bigoplus_{\forall_{t \in \chi d}} sit(t)$$

The translation of individual terms from the vocabulary $\mathcal{T}$ proceeds as in Boolean retrieval:

$$sit(t) = \{\langle\langle \mathrm{I}, t;\, 1\rangle\rangle\}$$

Note that in coordination level matching there is no information fusion. As a consequence there is no possibility to step beyond the level of the profons.

Example 4.5 (Caesar) Using our running example:

$$\begin{aligned} S_{\chi(d_1)} &= sit(C) \oplus sit(L) \oplus sit(B) \\ S_{\chi(d_1)} &= \{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle, \langle\langle \mathrm{I}, L;\, 1\rangle\rangle, \langle\langle \mathrm{I}, B;\, 1\rangle\rangle\} \\ S_{\chi(d_2)} &= \{\langle\langle \mathrm{I}, B;\, 1\rangle\rangle, \langle\langle \mathrm{I}, H;\, 1\rangle\rangle, \langle\langle \mathrm{I}, A;\, 1\rangle\rangle\} \end{aligned}$$

Coordination level matching retrieval is driven by a poor inference mechanism, which simply checks whether there is an overlap between two sets of keywords. This inference mechanism is mapped directly to inference between situations ($\approx\!\!\!>_{\mathrm{CLM}}$) as follows:

Definition 4.2 (Coordination Level Matching Situation Aboutness) Let S be a situation and i an infon, then the inference mechanism of CLM is the following: $S \approx\!\!\!>_{\mathrm{CLM}} i$ if and only if
$i \in S$.

Example 4.6 (Caesar) Using this inference mechanism for the situation $S_{\chi(d_1)}$,

$$\{\langle\langle \text{I},C;\, 1\rangle\rangle, \langle\langle \text{I},L;\, 1\rangle\rangle, \langle\langle \text{I},B;\, 1\rangle\rangle\} \quad \approx\!\!\!\Rightarrow_{\text{CLM}} \quad \langle\langle \text{I},C;\, 1\rangle\rangle$$

Example 4.7 (Caesar) In our running example it can be shown that $S_{\chi(d_1)} \;\; |\!\sim\; S_{\chi(d_2)}$ as follows:
$S_{\chi(d_1)} \approx\!\!\!\Rightarrow_{\text{CLM}} \langle\langle \text{I},B;\, 1\rangle\rangle$ and $\langle\langle \text{I},B;\, 1\rangle\rangle \in S_{\chi(d_2)}$.

5 The Theoretical Comparison of Retrieval Systems based on Aboutness Axioms

So far we have demonstrated that the situation theoretic approach is expressive enough to describe different kinds of information retrieval inference mechanisms. Our focus will now shift to the theoretical comparison of information retrieval systems based on what aboutness axioms they are governed by. In this section Boolean retrieval and coordination level matching will be compared based on theorems stating the aboutness properties of their respective situation inference mechanisms.

Theorem 1 (Boolean Retrieval Aboutness Axioms) Let $\mathcal{S}$ be a set of situations such that $S \;\; |\!\sim\; T$ if and only if $\exists_{i \in T}[S \approx\!\!\!\Rightarrow_{\text{BS}} i]$. Then $|\!\sim$ has the following properties:

> Reflexivity, Transitivity, Right monotonic union, Right union decomposition, Right negative union, Left monotonic fusion, Right fusion decomposition, Right negative fusion.

Proof 1 *Sketch: It is sufficient to test the validity of each axiom based on the definition of the situation inference mechanism. For example, the following is the proof that* Right Negative Union *holds:*

> *Assuming* $S \;\; |\!\not\sim\; T$ *and* $S \;\; |\!\not\sim\; U$ *implies that for all infons* i *in* T, $f(S) \nvdash f(\{i\})$, *and for all infons* j *from* U, $f(S) \nvdash f(\{j\})$. *As a consequence,* $f(S) \nvdash f(T) \vee f(U)$, *and hence* $S \;\; |\!\not\sim\; T \oplus U$.

Theorem 2 (Coordination Level Matching Aboutness Axioms) Let $\mathcal{S}$ be a set of situations such that $S \;\; |\!\sim\; T$ if and only if $\exists_{i \in T}[S \approx\!\!\!\Rightarrow_{\text{CLM}} i]$. Then $|\!\sim$ has the following properties:

Reflexivity, Symmetry, Left monotonic union, Right monotonic union, Left union decomposition, Right union decomposition, Left negative union, Right negative union.

Proof 2 *The proof proceeds along the same lines as that for Boolean retrieval, for example, the proof of* Symmetry *is as follows:*

> *Assume that* $S \mid\!\sim T$*, so there is an infon* i *in situation* T *such that* $S \approx\!\!\Rightarrow_{\mathrm{CLM}} i$*. From the definition of coordination level matching this implies that* $i \in S$*. We know that* $i \in T$ *and hence* $T \approx\!\!\Rightarrow_{\mathrm{CLM}} i$*. Therefore,* $T \mid\!\sim S$*.*

Note that as there is no information fusion in coordinate level matching, the axioms involving fusion are not applicable.

Based on the above theorems the following can be concluded (see figure 2):

- CLM is symmetric, hence there is no difference between left and right in the axioms. BS, on the other hand, is not symmetric and left and right distinctions in the axioms are significant.
- CLM has better recall because its aboutness relation is symmetric. As argued earlier, symmetry promotes totality in a situation inference mechanism.

Rule		*BS*	*CLM*
Reflexivity		yes	yes
Symmetry		no	yes
Transitivity		yes	no
Monotonic Union	Left	no	yes
	Right	yes	yes
Union Decomposition	Left	yes	yes
	Right	no	yes
Negative Union	Left	no	yes
	Right	yes	yes
Monotonic Fusion	Left	yes	no
	Right	no	no
Fusion Decomposition	Left	no	no
	Right	yes	no
Negative Fusion	Left	no	no
	Right	yes	no

Figure 2: Boolean retrieval vs. coordination level matching

If the properties for Boolean retrieval and coordination level matching are given, the proofs of aboutness are straight-forward. We will give two examples:

Boolean retrieval Let $\chi(d) = \{C, A\}$ and $q = C \vee \bar{A}$. The following is the proof that $S_{\chi(d)} \;|\!\sim\; S_q$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\{\langle\langle\bigwedge,C;\,1\rangle\rangle\}}{\{\langle\langle\bigwedge,C;\,1\rangle\rangle\} \;|\!\sim\; \{\langle\langle\bigwedge,C;\,1\rangle\rangle\}}\,Ref}{\{\langle\langle\bigwedge,C;\,1\rangle\rangle\} \odot \langle\langle\bigwedge,A;\,1\rangle\rangle\} \;|\!\sim\; \{\langle\langle\bigwedge,C;\,1\rangle\rangle\}}\,LMF}{\{\langle\langle\bigwedge,C;\,1\rangle\rangle\} \odot \{\langle\langle\bigwedge,A;\,1\rangle\rangle\} \;|\!\sim\; \{\langle\langle\bigwedge,C;\,1\rangle\rangle\} \oplus \{\langle\langle\bigwedge,\bar{A};\,1\rangle\rangle\}}\,RMU}{\{\langle\langle\bigwedge,C,A;\,1\rangle\rangle\} \;|\!\sim\; \{\langle\langle\bigwedge,C;\,1\rangle\rangle, \langle\langle\bigwedge,\bar{A};\,1\rangle\rangle\}}\,Def.\ Fusion$$

Coordination level matching Let $\chi(d) = \{C, A\}$ and $q = \{A, B\}$. The following is the proof that $S_{\chi(d)} \ \mid\!\sim \ S_q$:

$$\dfrac{\dfrac{\dfrac{\dfrac{\{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\}}{\{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\} \ \mid\!\sim \ \{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\}}\, Ref}{\{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\} \oplus \langle\langle \mathrm{I}, A;\, 1\rangle\rangle\} \ \mid\!\sim \ \{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\}}\, LMU}{\{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\} \oplus \{\langle\langle \mathrm{I}, A;\, 1\rangle\rangle\} \ \mid\!\sim \ \{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle\} \oplus \{\langle\langle \mathrm{I}, B;\, 1\rangle\rangle\}}\, RMU}{\{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle, \langle\langle \mathrm{I}, A;\, 1\rangle\rangle\} \ \mid\!\sim \ \{\langle\langle \mathrm{I}, C;\, 1\rangle\rangle, \langle\langle \mathrm{I}, B;\, 1\rangle\rangle\}}\, Def.\ Union$$

6 Conclusions

In this paper we have presented a framework for theoretically studying information retrieval mechanisms based on our underlying theory of information. Within this framework, axioms have been outlined, that reflect assumptions made by information retrieval models. The framework presented has been tested. The question naturally arises as to whether this model is sufficiently powerful to cover all kinds of information retrieval models. Although the expressive power of situation theory has been demonstrated, we still have paid no attention to uncertainty aspects. This is an important focus for future research. Another area of continuing interest is to define a complete set of aboutness-axioms and to give a proof of completeness. Even though the theory hinges on the notion of aboutness, it can be applied to investigate characterization-languages in detail.

Acknowledgements

The authors would like to thank Thomas Arts and Bernd van Linder for their helpful (logical) suggestions, and Mounia Lalmas for her helpful comments and criticism on this work.

References

[BE87] J. Barwise and J. Etchemendy. *The Liar, An Essay on Truth and Circularity.* Oxford University Press, 1987.

[BE90] J. Barwise and J. Etchemendy. Information, infons, and inference. In R. Cooper, K. Mukai, and J. Perry, editors, *Situation Theory and its Applications, Volume I*, chapter 2. CSLI, 1990.

[Bru93] P.D. Bruza. *Stratified Information Disclosure, a Synthesis between Hypermedia and Information Retrieval*. PhD thesis, University of Nijmegen, March 1993.

[CC92] Y. Chiaramella and J.P. Chevallet. About retrieval models and logic. *The Computer Journal*, 35(3):233–241, March 1992.

[Coo71] W.S. Cooper. A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7:19–37, 1971.

[Dev91] K. Devlin. *Logic and Information*. Cambridge University Press, 1991.

[Far80a] J. Farradane. Relational indexing, part I. *Journal of Computer Science 1*, pages 267–276, 1980.

[Far80b] J. Farradane. Relational indexing, part II. *Journal of Computer Science 1*, pages 313–324, 1980.

[LR92] M. Lalmas and C.J. van Rijsbergen. A logical model of information retrieval based on situation theory. In *Proceedings of the BCS 14th Information Retrieval Colloquium*. British Computer Society, Springer-Verlag London Ltd, April 1992.

[Mar77] M.E. Maron. On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, pages 38–43, January 1977.

[Nie92] J. Nie. Towards a probabilistic modal logic for semantic-based information retrieval. In N. Belkin, P. Ingwersen, and A.M. Pejtersen, editors, *Proceedings of the Fifteenth Annual Internatiol ACMSIGIR Conference on Research and Development in Information Retrieval*, pages 140–151. ACM, ACM Press, June 1992.

[Rij86] C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, Vol. 29(6):481–485, 1986.

[Rij89] C.J. van Rijsbergen. Towards an information logic. In N.J. Belkin and C.J. van Rijsbergen, editors, *Proceedings of the Twelfth Annual*

Internatiol ACMSIGIR Conference on Research and Development in Information Retrieval, pages 77–86. ACM, ACM Press, June 1989.

[Rij92] C.J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, Vol. 35(3):291–298, 1992.