OpenAIR@RGU

The Open Access Institutional Repository at The Robert Gordon University

http://openair.rgu.ac.uk

This is an author produced version of a paper published in


This version may not include final proof corrections and does not include published layout or pagination.

Citation Details

Citation for the version of the work held in ‘OpenAIR@RGU’:


Citation for the publisher’s version:


Copyright

Items in ‘OpenAIR@RGU’, The Robert Gordon University Open Access Institutional Repository, are protected by copyright and intellectual property law. If you believe that any material held in ‘OpenAIR@RGU’ infringes copyright, please contact openair-help@rgu.ac.uk with details. The item will be removed from the repository while the claim is investigated.
Commonsense Aboutness for Information Retrieval

P.D. Bruza\textsuperscript{1} D.W. Song\textsuperscript{2} K.F. Wong\textsuperscript{2} C.H. Cheng\textsuperscript{2}
\textsuperscript{1}Distributed Systems Technology Center, University of Queensland, Qld 4072 Australia
E-mail: bruza@dstc.edu.au
\textsuperscript{2}Department of Systems Engineering \& Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
E-mail: \{dwsong, kfwong, chcheng\}@se.cuhk.edu.hk

Abstract. Information retrieval (IR) is driven by a process which decides whether a document is about a query. Recent attempts have been made to formalize properties of “aboutness”, but no consensus has been reached. The properties being proposed are largely being determined by the underlying model. Once the framework has been fixed, certain aboutness properties could be implied from it. Moreover, some properties may be sound only within a certain IR model, and some may lead to negative effects to the performance of IR systems. In this paper, by adopting an information-based and abstract framework, model independent, commonsense properties of aboutness are investigated. A set of properties characterizing commonsense aboutness and its dual — non-aboutness is introduced. Issues such as the soundness and completeness of aboutness inference are discussed.

1. Introduction

Aboutness plays a prominent role in information retrieval (IR) systems. If a system determines that a document \(d\) is about a query \(q\), then the document is returned to the user. Thus, the study of aboutness is important to the theoretical foundations of IR.

Aboutness has been studied for several years \[4, 5, 10, 11, 12, 16\]. These studies usually model aboutness as a binary relation over a set of information carriers, and view the relationship as a form of entailment. Recent attempts have been made to formalize properties of aboutness which can be expressed as postulates (rules) in terms of information containment, composition and preclusion. There is as yet no consensus to this framework except that it should be logic-based \[14, 15, 19\]. Although a number of aboutness properties are commonly discussed in the literature, e.g., reflexivity, transitivity, symmetry and left (right) monotonicity, etc., there is no agreement on a core set of aboutness postulates. We consider aboutness from a fundamental, commonsense perspective and define a set of reasonable (hopefully sound) properties of aboutness, which is independent of any given IR model. The theoretical goal of aboutness research is to gain a clearer understanding of IR models from a symbolic, rather than numeric perspective. In addition, research into aboutness has a number of practical aspects, for example, aboutness inference systems can form the basis if information agents that reason about the relevance/nonrelevance of information fragments.

2. Preliminaries

A basic information carrier is the minimal piece of information that cannot be divided further. In IR, basic information carriers correspond to keywords. Let \(B\) denote the set of basic information carriers. Even at the level of basic information carriers, aboutness manifests itself. For example, football has the property of being a sport, and it seems
natural to say that football is about sport. Not all property relationships imply aboutness, e.g. “apple is about (being) round” does not seem as natural as “apple is about fruit”.

More complex information carriers can be composed from basic ones. Information composition is a complex issue [15]. Consider the composition of information carrier A with carrier B, denoted A⊕B. Viewed from a situation-theoretic perspective [14], the latter carrier represents the intersection between the situations supporting A and the situations supporting B. For example, flying⊕tweety represents the intersection of “flying” situations and “Tweety” situations, which is the "Tweety is flying" situations. For ease of exposition, information composition is assumed to be idempotent, commutative and associative.

An information carrier A may clash, or contradict, with another one B. This phenomenon is termed information preclusion, denoted by A⊥B. Information preclusion is symmetric, and its negation (⊥/) is decided by the closed world assumption. Information preclusion is a subtler notion than contradiction in logic. Information carriers may clash due to underlying natural language semantics, or convention. Information preclusion arises in IR as a consequence of an information need, for example, if the information need deals with “wave surfing”, this clashes with documents about “web surfing”. IC represents a set of information carriers constructed from the basic carriers B by information composition. IC is assumed to be closed with respect to the information composition operator ⊕.

Information carriers can not only be composed, but also ordered. In the literature, several authors have proposed that information can be ordered with respect to containment [2]. Information containment models the intuition that information is explicitly and implicitly nested. Explicit nesting is referred to as surface containment, e.g. A⊕B⊇B denotes that the information carried by B is also carried by A⊕B (as B is a syntactic element of A⊕B). Deep containment is when information containment arises at the semantic level, e.g. salmon→fish. In general, information containment (either surface or deep) will be denoted by the symbol →, whereby → is the union of the relations ⊇ (surface) and →(deep) containment. In addition, the information structure (IC, ⊨, ⊆, →, ⊕, ⊥) has the following additional properties:

- Reflexivity (R): A→A
- Transitivity (T): A→B and B→C imply A→C
- Anti-symmetry (AS): A≠B and A→B imply B→A
- Containment-Composition (CC): A⊕B→A; A⊕B→B
- Absorption (AB): if A→B then A⊕B=A
- Containment-Preclusion (CP): if A→B, B≤C then A→C

Aboutness is modeled as a binary relation |= over the information carriers IC.

Huibers introduced the notion of an aboutness proof system, which is founded on an aboutness language [10]:

**Definition 2.1.** Let IC be a set of information carriers. The aboutness language Λ(IC) is the smallest set such that If A, B ∈ IC then A→B, A↔B, A≥B, A≥B, A→B, A→B, A⊥B, A⊥B, A|B, A|B, A=B, A≠B ∈ Λ(IC).

Observe carefully that in this context the symbol |= generally expresses an aboutness relation between A and B. The specific type of aboutness relation will be signified by a subscript, e.g. the next section will deal with overlapping aboutness (|=). The symbol / denotes negation, e.g. A→B means “not A→B”.

**Definition 2.2** An aboutness proof system is a triple ⟨Λ(IC), A, R⟩ where
- A is a decidable subset of Λ(IC), whose elements are called axioms;

---

1 The formalization of aboutness operators (e.g. composition, preclusion, containment, etc.) is dependent on the language of the information carriers.
\( R = \{R_1, \ldots, R_k\} \) is a finite set of rules.

Axioms are elements of the aboutness language that are assumed to be true, e.g., \( A \rightarrow A \).

Rules have premises and a conclusion. For example, the premises of the And rule are \( A \models B \) and \( A \models C \), which yield the conclusion \( A \models B \oplus C \). As in [10], we assume that each rule is decidable as a relation. Axioms and rules can be used to drive inference. The concern of this paper are inferences of the form \( A \models B \) and later \( A \models \neg B \).

### 3. An Intuitive Form of Aboutness: Overlap

A commonly occurring intuition equates aboutness with overlap, i.e., if two information carriers overlap, then they are deemed to be about each other. Almost all information retrieval systems function according to this intuition. For example, the vector space model measures the overlap between a query and a document vector by computing the cosine of the angle between the two vectors. In this section, we investigate the consequences of defining aboutness in this fashion.

**Definition 3.1** Overlapping Aboutness \((\models_o)\)

Let \( A, B \in IC \) and \( \models_o \subseteq IC \times IC \) such that \( (A, B) \in \models_o \Leftrightarrow \exists c,w : [A \rightarrow C \land B \rightarrow C] \).

The more readable convention \( A \models_o B \), instead of \( (A, B) \in \models_o \), will be employed to signify “\( A \) is about \( B \)”.

**Proposition 3.1** \( \models_o \) supports Reflexivity, Containment, Symmetry, Left Compositional Monotonicity, Right Compositional Monotonicity, And, Simplification, Loop and Mix where these properties are defined as follows: (See [7] for proofs)

- Reflexivity states that an information carrier is about itself. From an IR perspective reflexivity seems a reasonable property as we expect a document to be retrieved if it was itself the query.
- Containment states that an information carrier is about the information it contains. On the surface this seems reasonable. However, consider the basic information carrier \( ghiardia \), a water-bound microbe. Observe that deep containment involves semantic transformation, e.g. \( ghiardia \mapsto microbe \mapsto \ldots \mapsto animal \). The Containment postulate permits both \( ghiardia \models_o microbe \) and \( ghiardia \models_o animal \). The former is intuitively acceptable, but the latter much less so. Our contention is that at some point along the information containment chain the aboutness relation can be severely weakened. Brooks documented a user study that supports our contention. Brooks found that the distance to non-relevance (non-aboutness) is approximately three steps [3].
- At first sight, symmetry seems to be an acceptable property. There is evidence to dispute this. For example, in hypertext an information fragment \( A \) is linked to a fragment \( B \), but in many cases it does not make sense to have the link organized the other way round.
- Monotonicity ensures that once an aboutness relationship between two information carriers \( A \) and \( B \) has been established, it cannot be broken irrespective of the other information that is composed to either \( A \) or \( B \). For example, consider the phrase “surfing in
Hawaii”. This phrase deals with surfing, so \( \text{surfing} \oplus \text{Hawaii} \models \text{surfing} \). RM permits \( \text{surfing} \oplus \text{Hawaii} \models \text{surfing} \oplus \text{australia} \), which has the natural language interpretation “surfing in Hawaii” is about “surfing in Australia”. Thus, in response to the query “surfing in Australia”, the document “surfing in Hawaii” is returned. An IR system supporting RM or LM cannot “lose” aboutness relationships. This should not be the case because the terms used to expand the query may invalidate the original aboutness relationship. Some current IR systems circumvent this behaviour by employing threshold values.

Simplification states that the aboutness relationship between a carrier \( A \) and a complex information carrier \( B \oplus C \) implies that \( A \) is about \( B \), or \( A \) is about \( C \). This is debatable. It may well be that \( A \) is about \( B \oplus C \), but \( A \) is not about \( B \) and \( A \) is not about \( C \), since it is precisely their combination which establishes the aboutness relationship.

Loop appears in AI [13] and logic-based IR literatures [1, 5, 7]. These IR accounts view aboutness in terms of a nonmonotonic consequence relation. Loop in the context of overlapping aboutness is implied trivially by symmetry.

In summary, the overlapping view of aboutness is intuitive, but it implies some unforseen properties, namely left, right monotonicity, containment, symmetry and simplification. These properties are unsound from a commonsense perspective and can negatively impact information retrieval precision (i.e., the ratio of retrieved relevant documents to retrieved documents).

4. Commonsense Aboutness

In this section we characterize aboutness more broadly rather than just overlap. We adopt a commonsense point of view in an attempt to establish properties of aboutness acceptable from a human reasoning perspective. This serves not only to illustrate the aboutness postulates, but also to highlight the similarities and differences between the aboutness and nonmonotonic consequence relations (e.g., preferential entailment [6, 13]).

**Example:** Tweety

Let \( t \) (Tweety), \( b \) (bird), \( p \) (penguin) and \( f \) (fly) be basic information carriers. The example is then described as follows: Tweety is a bird \((t \rightarrow b)\); Tweety is a penguin \((t \rightarrow p)\); penguins are birds \((p \rightarrow b)\); birds are about flying \((b \models f)\); penguins do not fly \((p \perp f)\). Applying the properties of information containment results in: \( t \rightarrow t \); \( b \rightarrow b \); \( f \rightarrow f \); \( p \rightarrow p \). Further, applying CP yields: \( t \perp f \).

4.1 Aboutness Postulates

To distinguish the following properties from the aboutness properties associated with overlap (in the previous section), the symbol \( \models \) will be used to denote commonsense aboutness postulates. These postulates build on, but go beyond, the notion of overlapping aboutness. In particular, the problems surrounding the rules dealing with monotonicity and information containment are addressed.

**(R) Reflexivity**

It seems reasonable to assume that an information carrier is about itself.

**(AS) Asymmetry**

Given “Tweety is about a bird”. It is unnatural to immediately conclude that “A bird is about Tweety”. This rule states that aboutness is fundamentally asymmetric.

**(AC) Aboutness Consistency:** \[ A \models B \] \[ \frac{A \perp B}{A \models B} \]

An information carrier should be compatible with what it is about.
(B1) **Semantic Containment:**

\[
\frac{A \rightarrow_t B}{A \models B}
\]

Brook’s study revealed that relevance perceptions are inversely proportional to semantic distance [3]. When broadening, the demarcation point where relevant perceptions degraded to non-relevance was two semantic steps. Following this principle, the aboutness relationship is also severed after traversing three steps along the deep containment relation. The “2” in the above formula (i.e. B1) signifies that aboutness is preserved within two steps along the deep containment relation.

(CT) **Cut:**

\[
\frac{A \uplus B \models C \quad A \models B}{A \models C}
\]

If the composition of two pieces of related information is about another one, then cutting one does not affect the aboutness relation. For example, \(p \uplus t \models b\), \(t \models p \Rightarrow t \models b\). That is, from “Tweety the penguin is about a bird” and “Tweety is about a penguin”, “Tweety is about a bird” can be derived.

In the previous section, monotonicity was shown to be unsound. The following three rules express conservative forms of compositional monotonicity (both left and right).

(CLM) **Cautious Left Compositional Monotonicity:**

\[
\frac{A \models B \quad A \models C}{A \uplus C \models B}
\]

If \(A\) is about \(B\) and \(A\) is about \(C\), then composing the information in \(C\) to \(A\) means adding “compatible” or “related” information to \(A\). Thus, \(A \uplus C \models B\) should hold. For example, from \(t \models p\) (Tweety is about a penguin) and \(t \models b\) (Tweety is about a bird), then \(t \uplus p \models b\) (Tweety the penguin is about a bird) can be inferred.

The following two rules (Mix and And) are also variations on constraining monotonicity.

(M) **Mix:**

\[
\frac{A \models C \quad B \models C}{A \uplus B \models C}
\]

For example, \(p \models b\), \(t \models b \Rightarrow p \uplus t \models b\). Unlike preferential entailment [13], we argue that Mix produces acceptable aboutness inferences even when information clashes.

(A) **And:**

\[
\frac{A \models B \quad A \models C}{A \models B \uplus C}
\]

The previous three rules featured how monotonicity can be constrained solely based on aboutness relationships. The following rules constrain monotonicity by ensuring that information will not clash.

(QLM) **Qualified Left Compositional Monotonicity:**

\[
\frac{A \models B \quad B \cup C}{A \uplus C \models B}
\]

Traditional Left Compositional Monotonicity (LCM) is \(A \models C \Rightarrow A \uplus B \models C\). This allows \(b \uplus t \models f\) (Tweety, which is a bird, is about flying) to be inferred from \(b \models f\) (A bird is about flying). Absorption \((b \uplus t \models t)\) then renders \(t \models f\). (Tweety is about flying), which is undesirable as Tweety is a penguin which cannot fly. QLM prevents this via the qualifying preclusion \(t \perp f\). Observe, however, that \(t \uplus p \models p\) (swimming Tweety is about a penguin) is derivable from \(t \models p\) (Tweety is about a penguin) and \(p \uplus s\) (“penguin” does not clash with “swimming”, as penguins swim).

QLM deviates from several authors who have advocated a variant of Rational Monotonicity:

\[
\frac{A \models B \quad A \cup C}{A \uplus C \models B}
\]

Observe that QLM permits the inference \(p \uplus f \models b\) (Flying penguins are about birds) from \(p \models b\) and \(b \perp f\). We argue that this inference is acceptable, even though penguins preclude flying \((p \perp f)\). Rational Monotonicity [1, 5, 6, 26] prevents such an inference.

(QRM) **Qualified Right Compositional Monotonicity:**

\[
\frac{A \models B \quad A \cup C}{A \models B \uplus C}
\]
RCM is $A|=B \Rightarrow A|=B \oplus C$. For example, $p|=b \Rightarrow p|=f \oplus b$ (penguins are about flying birds) and $t|=b \Rightarrow t|=f \oplus b$ (Tweety is about a flying bird) are unsound aboutness inferences. The qualifying preclusions $p \perp f$ and $t \perp f$ separately prevent $p|=f \oplus b$ and $t|=f \oplus b$ from being inferred. Thus, QLM and QRM can prevent undesired conclusions permitted by LM and RM, thus describing conservative monotonicity of aboutness with respect to information composition.

4.2 Non-aboutness

Bruza, Huibers and Hunter have investigated non-aboutness [4, 10, 11]. In some situations non-aboutness may be easier to determine than aboutness. Information filtering is a good example where reasoning about the non-aboutness of incoming documents with respect to the user profile may be easier than reasoning about their aboutness with respect to the profile.

**Definition 4.1** Non-aboutness ($|\neq$) Let $A, B \in IC$. Then $A|\neq B \subseteq IC \times IC$ denotes $A$ is not about $B$.

Non-aboutness seems mainly to be influenced by information preclusion. It should be noted that in this paper, the initial preclusion relations are assumed, such as $p \perp f$, etc. In IR, the preclusion relations may not always be given explicitly. For example, in an IR system, a sentence “penguin doesn’t fly” may only be indexed to $\{p, f\}$. If the query is “flying bird” ($\{f, b\}$), then the sentence could be judged to be about the query because the preclusion relation “p $\perp f$” is not considered. In the following we describe the commonsense properties of non-aboutness:

(P) **Preclusion:**

\[
\frac{A \perp B}{A \neq B}
\]

Two fragments of clashing information are not about each other, e.g. $p \perp f \Rightarrow p|\neq f$ (Penguins are not about flying).

(B2) **Semantic Containment Non-aboutness:**

\[
\frac{A \rightarrow_2 B}{A \neq B}
\]

It is the complement of the Semantic Containment postulate (B1). If information carrier $B$ is more than 2 semantic (deep containment) steps away from $A$, then the aboutness relation is severed yielding non-aboutness.

(B3) **Inverse Semantic Non-aboutness:**

\[
\frac{A \rightarrow B \ A \neq B}{B \neq A}
\]

Brooks also studied how relevance degrades when traversing against the flow of the deep containment relation. In terms of a thesaurus this means traversing the “narrower than” relationship. The study suggested that aboutness does not flow backwards at all: “It may be best to conclude that one step down in a generic tree produces a neutral perception of relevance verging on non-relevance” [3]. We argue that B3 can be generalized to information containment (both surface and deep containment):

(I-CN) **Inverse Containment Non-aboutness:**

\[
\frac{A \rightarrow B \ A \neq B}{B \neq A}
\]

(P-NA) **Preclusion Non-aboutness:**

\[
\frac{A|=B \ B \perp C}{B \neq A}
\]

This is an expression of the intuition that aboutness involves compatibility between the respective carriers: $A$ cannot be about anything which clashes with $B$. For example, $t|=p$, $p \perp f \Rightarrow t|\neq f$.

5. **Completeness and Soundness**
5.1 Completeness

For any two arbitrary information carriers A and B, the aboutness inference system is deemed complete when it is able to conclude either $A|=B$ or $A|\neq B$. Note that a closed world assumption regarding $|=\$ has been convincingly opposed [10, 18]. We agree with this view and have characterized non-aboutness via constructive means (see Section 4.2). The following proposition asserts that the commonsense aboutness and non-aboutness postulates are incomplete (see [7] for proof).

**Proposition 5.1** The aboutness proof system $\Pi=\langle \Lambda(\text{IC}), \{R\}, \{C, B1, CT, CLM, M, A, QLM, QRM\} \rangle$ and the non-aboutness system $\Omega=\langle \Lambda(\text{IC}), \phi, \{P, B2, I-CN, P-NA\} \rangle$ are incomplete.

The above incompleteness result parallels a similar result using default logic-based theory of aboutness [11]. Bruza et al. show that by introducing unsound aboutness properties such as Right Containment Monotonicity, completeness can be achieved [7]. The price is an increase in the number of unsound aboutness inferences (i.e., loss of precision in IR terms).

5.2 Soundness

When investigating the issue of soundness with respect to (non-) aboutness rules, a frame of reference must be defined within which soundness can be verified. In classical logic, the frame of reference is a model. However, due to the lack of an underlying model theory for IR, this approach cannot be taken. It is also arguable whether such a theory exists. One of the complications here is that the “retrieval is inference” view [8, 18] cannot be considered independent of the user. One existing approach for soundness evaluation with respect to aboutness is that the soundness of an aboutness proof system is investigated in the context of an underlying aboutness definition [10]. Although valid from a formal point of view, it does not consider the user. We argue that aboutness inference is “psychologistic” in nature. With respect to IR, sound aboutness rules can only be established via cognitive studies, as has been done in non-monotonic reasoning [9, 17]. More studies similar to Brooks’s [3] are needed to gain a clearer understanding of how users perceive aboutness, and how they reason about it.

6. Discussion and Conclusions

The aboutness theory presented here has been founded upon notions such as deep and surface containments, information preclusion etc. The question beckons – for practical IR, how will these concepts be embedded into a working system? It is true that current IR systems are not defined in terms of these concepts mainly because they do not view retrieval as an aboutness reasoning process. Aboutness and preclusion relationships can be derived via relevance feedback [1, 6]. For restricted domains, information containment relationships can be derived from ontologies, and the like.

It is interesting to note that aboutness inferences can be ordered on their potential for soundness [7]. For example, there would be more confidence placed in an aboutness inference derived by Cautious Monotoniticy than one derived by Qualified Left Monotonicity, as the former rule is more conservative. In this way, orderings on aboutness inferences can be generated. An area of further investigation would be to extend the work presented here to cater for such orderings resulting in a symbolic aboutness system into
which weighted IR systems can be mapped. It is our feeling that much of IR theory can be 
re-created in a symbolic framework which could possibly extend the explanatory power of 
IR theory, for example, in the area of functional benchmarking [20].

Acknowledgments

The work reported in this paper has been funded in part by the Cooperative Research Centres Program 
through the Department of the Prime Minister and Cabinet of Australia and by the Hong Kong Research 
Grant Committee under the Strategic Research Scheme (Project #: 44M5007).

References

and Lalmas, M. (Eds.), *Proceedings of the Second Workshop on Information Retrieval, Uncertainty and 
ACM SIGIR Conference on Research and Development in Information Retrieval*. Dublin, Ireland, pp.112- 
121, 1994.
Uncertainty and Logics* (Crestani, F., Lalmas, M. and Van Rijsbergen, C.J. eds.) The Kluwer international 
[8] W.S. Cooper, A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7, pp. 19- 
37, 1971.
Netherlands. 1996.
International Conference on Tools with Artificial Intelligence (TAI’96)*, 34-40, IEEE Computer Society Press. 
1996.
[17] F.J. Pelletier and R. Elio, What should default reasoning be, by default? *Computational Intelligence 
on Research and Development in Information Retrieval*, pp.77-86. 1989.
[19] F. Sebastiani, On the role of logic in information retrieval. *Information Processing & Management*, 34, 1, 1- 
retrieval models. In *Proceedings of 12th International Florida Artificial Intelligence Society Conference 