Aboutness from a Commonsense Perspective

P.D. Bruza[1] **D.W. Song**[2] **K.F. Wong**[2]

[1] Distributed Systems Technology Center, Building 78, Staff House Road
University of Queensland, St. Lucia. Qld 4072 Australia
bruza@dstc.edu.au

[2] Department of Systems Engineering & Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{dwsong, kfwong}@se.cuhk.edu.hk

Information retrieval (IR) is driven by a process which decides whether a document is about a query. Recent attempts spawned from logic-based information retrieval theory have formalized properties characterizing “aboutness”, but no consensus has yet been reached. The proposed properties are largely determined by the underlying framework within which aboutness is defined. In addition, some properties are only sound within the context of a given IR model, but are not sound from the perspective of the user. For example, a common form of aboutness, namely overlapping aboutness, implies precision degrading properties such as compositional monotonicity. Therefore, the motivating question for this paper is: Independent of any given IR model, and examined within an information-based, abstract framework, what are commonsense properties of aboutness (and its dual, non-aboutness)? We propose a set of properties characterizing aboutness and non-aboutness from a commonsense perspective. Special attention is paid to the rules prescribing conservative behaviour of aboutness with respect to information composition. The interaction between aboutness and non-aboutness is modeled via normative rules. The completeness, soundness and consistency of the aboutness proof systems are analyzed and discussed. A case study based on monotonicity shows that many current IR systems are either monotonic or non-monotonic. An interesting class of IR models, namely those that are conservatively monotonic, is identified.

1. Introduction

You are sitting in a bus and two people in front of you are talking. The first says to the second, "I went to see so-and-so film last night", to which the second replies, "Oh really, what was it about?" The first then proceeds to describe it. Thus, the notion of "aboutness" is present in our everyday communications, particularly when one agent wishes to inform, or be informed, by another agent.

Aboutness plays a prominent role in information retrieval (IR) systems: If the system determines that a document *d* is topically related (i.e. about) query *q*, then the document is returned to the user. Cleverden (1991) cites experiments wherein the agreement between subjects judging documents with respect to a query was around 60%. This suggests that aboutness has a subjective component. However, there also seems to be an inter-subjective core of agreement, which in our opinion is amenable to formal treatment.

Articles on aboutness have appeared sporadically in the literature for more than two decades. Hutchins (1977) provides a thoughtful early study of the topic. This account attempts to define a notion of aboutness in terms of a combination of linguistic and discourse analyses of a text. At a high level of information granularity, e.g. a sentence, Hutchins introduces *themes* and *rhemes* as the carriers of the thematic progression of a text. Roughly speaking, the theme states what the writer intends to express in the sentence (i.e. what it is about), and the rheme is the “new” information. Thematic elements of a sentence are typically bound textually to the preceding text, or assumed as given within the current context. Hutchins also considers how sequences of sentences combine to form textual elements of lower information granularity such as an episode. In other words, sentences are considered to be a part of the micro structure of the text, whereas an episode is considered to be an element of its macro-structure. Themes and rhemes can be generalized to the macro level. Hutchins asserts “The thematic part of the text expresses what the text is ‘about’, while the rheme expresses what the author has to say about it” (Hutchins, 1977, p31).

Maron (1977) tackled aboutness by relating it to a probability of satisfaction. Three types of aboutness were characterized: S-about, O-about and R-about. S-about (i.e. subjective about) is a relationship between a document and the resulting inner experience of the user. O-about (i.e. objective about) is a relationship between a document and a set of index terms.

More specifically, a document D is about a term set T if user X employs T to search for D. R-about purports to be a generalization of O-about to a specific user community (i.e., a class of users). Let I be an index term and D be a document, then D is R-about I is the ratio between the number of users satisfied with D when using I and the number of users satisfied by D. Using this as a point of departure, Maron further constructs a probabilistic model of R-aboutness. The advantage of this is that it leads to an operational definition of aboutness which can then be tested experimentally. However, once the step has been made into the probabilistic framework, it becomes difficult to study properties of aboutness, e.g. how does R-about behave under conjunction? The underlying problem relates to the fact that probabilistic independence lacks properties with respect to conjunction and disjunction. In other words, one's hands are largely tied when trying to express qualitative properties of aboutness within a probabilistic setting. (For this reason Dubois et al. (1997) developed a qualitative framework for relevance using possibility theory).

During the eighties and early nineties, the issue of aboutness remained hidden in the operational definitions of various retrieval models and their variations. The emergence of logic-based information retrieval in the late eighties planted the seed for fundamental investigations of the nature of aboutness (Bruza & Huibers, 1994; Bruza & Huibers, 1996; Hunter, 1996; Nie *et al.*, 1995) culminating in an axiomatic theory of information retrieval developed by Huibers (1996). Aboutness theory has more recently appeared in context of information discovery (Proper & Bruza, 1999). Broadly speaking, these works view information retrieval (IR) as a reasoning process, determining aboutness between two information carriers (e.g. document and query, or document and document). The properties of aboutness are described by a set of postulates, which can be used to compare IR models depending on which aboutness postulates they support. Up to now, there is as yet no consensus regarding this framework except that it should be logic-based (Lalmas, 1998; Lalmas &Bruza, 1998; Sebastiani, 1998). Although a number of aboutness properties are commonly discussed in the literature, e.g. reflexivity, transitivity, symmetry, simplification, and, right weakening and left (right) monotonicity, etc., there is thus far no agreement on a core set of aboutness postulates. Nevertheless, the use of aboutness postulates as the basis of an inductive, rather than experimental, evaluation of IR models is promising. Existing aboutness frameworks, however, suffer from incompleteness as well as from the lack of expressive power (Wong *et al.*, 1998). The main reason is the lack of holistic and independent view of aboutness and its properties. The purpose of this article is to consider aboutness from a fundamental, neutral perspective, to shed light on the nature of aboutness by formalizing properties describing it, and to define a set of reasonable (hopefully sound) properties of aboutness, which is independent of any IR model.

The remaining of the paper is organized as follows. In the next section, the basic notions of aboutness are introduced. Section 3 outlines a common intuitive form of aboutness - overlapping aboutness. However, it implies some unsound properties; as such it is inadequate to model the general properties of aboutness. Thus, commonsense aboutness and its properties are proposed in Section 4. At the same time, its negation, i.e. non-aboutness is described, and the relationship between them is also discussed. Section 5 investigates the soundness, completeness and consistency of the commonsense aboutness inference system. In the next section (Section 6), an investigation of the relationship between similarity and aboutness is given. Section 7 presents a case study on the relation between the monotonicity property and the more prominent IR models. In Section 8, possible extensions and applications of the theoretical results are discussed. Finally, Section 9 concludes the paper.

2. Preliminaries

A *basic information carrier* is the minimal piece of information that cannot be divided further. In IR, basic information carriers often correspond to keywords. Let **B** denote the set of basic information carriers.

Even at the level of basic information carriers, aboutness manifests itself. This has to do with the relationships between basic information carriers. For example, *football* has the property of being a *sport*, and it seems natural to say that *football* is about *sport*. Note that being a property does not guarantee aboutness, for example, "*apple* is about (being) *round*" does not seem natural. Primitive aboutness relationships need not be property based, for example, "*dancing* is about *having fun*", "*marriage* is about *fidelity*" and "*marriage* is about *commitment*". From the latter two, it seems reasonable to draw the conclusion that "*marriage* is about *fidelity* and *commitment*". This example demonstrates that aboutness may also be preserved under the composition, denoted by ⊕. The previous statement can thus be rendered "*marriage* is about *fidelity*⊕*commitment*". This example does not imply, though, that aboutness relationships involving more complex information carriers are all derived. In fact, many primitive aboutness relationships involve information composition. (by "primitive" we mean aboutness relationships that are assumed to be true, i.e., axioms). By way of illustration, "*surfing* is about *riding*⊕*waves*" and "*politics* is about *greed*⊕*power*". A major concern of this paper is studying how aboutness relates to information composition, and as a consequence, how aboutness relationships can be derived between complex information carriers using more primitive aboutness relationships.

More complex information carriers can be composed from basic ones. Information composition is a complex issue (Lalmas & Bruza, 1998). It can be conceived of as a form of informational "meet". Consider the composition of information carrier A with carrier B, denoted A⊕B. Viewed from a situation-theoretic perspective (Lalmas, 1996), the latter carrier represents the intersection between the situations supporting A and the situations supporting B. For example, *flying*$\oplus$*tweety* represents the intersection of “flying” situations and “Tweety” situations, that is the situations which support the information "Tweety is flying". For ease of exposition, information composition is assumed to be idempotent ($A\oplus A=A$), commutative ($A\oplus B=B\oplus A$) and associative ($((A\oplus B)\oplus C=A\oplus (B\oplus C))$)[1]. Note that the theory presented in this paper can be generalized to cater for information composition that does not possess these properties.

Observe that not all information carriers can be meaningfully composed. For example, $A\oplus B$ is meaningless when the information carried by A clashes, or contradicts, with the information carried by B. This phenomenon is termed information preclusion, denoted by $A\perp B$. Information preclusion is symmetric, and its negation ($\not\perp$) 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. For example, *swimming*$\oplus$*crocodile* is acceptable, but *flying*$\oplus$*crocodile* is meaningless in most contexts. It has also been suggested that information preclusion arises in IR as a consequence of information needs (Bruza & Van Linder, 1998). For example, when searching for documents about *wind surfing*, terms such as *internet*, *web*, *net* etc. may be precluded as the user is not interested in *web surfing*. In some accounts, (e.g. (Landman, 1986; Bruza & Huibers, 1994)), the composition of clashing information is formalized as the “meaningless” information carrier, denoted by 0. It is attributed with properties similar to *falsum* in propositional logic, e.g. $A\perp B\Leftrightarrow A\oplus B = 0$. We do not introduce this notion for two reasons. Firstly, one is quite often confronted with documents containing conflicting information in real life. Although it is convenient, it would seem unrealistic to sweep such documents under the mathematical abstraction 0. Secondly, one can sometimes state what such documents are about. For example, consider “flying crocodiles” and “reptiles”. Under the assumption *flying*$\perp$*crocodiles*, we do not subscribe to the view that there are no aboutness inferences that can be drawn. It may turn out that the inference “*flying*$\oplus$*crocodiles* is about *reptiles*” may be inferred. It depends on the context- more will be said about this issue shortly. **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 $\oplus$.

Information carriers cannot only be composed, but also ordered. For example, we can say "A *contains at least the same information that* B does". In the literature, several authors have proposed that information can be ordered with respect to containment (Barwise & Etchemendy, 1990; Landman, 1986) These accounts take the position that information does not depend on the user (i.e., not subjective) but is analog, like radiation. User’s digitize the information according to their ability. We do not take a position on this issue. It is assumed that the information carriers are nested irrespective of whether the nesting fundamentally exists, or is due to some user’s view of the information carriers. Explicit nesting is referred to as *surface containment*, e.g. $A\oplus B\supseteq B$ denotes that the information carried by B is also carried by $A\oplus B$ (as B is a syntactic element of $A\oplus B$). For example, if a document *d* consists of sections A and B (i.e. $d=A\oplus B$), then $d\supseteq A$ and $d\supseteq B$. *Deep containment* is when information containment arises at the semantic level, e.g. *salmon*$\mapsto$*fish*. In general, information containment (either *surface* or *deep*) will be denoted by the symbol $\rightarrow$, whereby $\rightarrow$ is the union of the relations $\supseteq$ (surface) and $\mapsto$ (deep) containment. In addition, the information structure ($\mathbf{IC}, \mapsto, \supseteq, \rightarrow, \oplus, \perp$) has the following additional properties:

- Reflexivity (R): $A\rightarrow A$
- Transitivity (T): $A\rightarrow B$ and $B\rightarrow C$ imply $A\rightarrow C$
- Anti-symmetry (AS): $A\neq B$ and $A\rightarrow B$ imply $B\not\rightarrow A$
- Containment-Composition (CC): $A\oplus B\rightarrow A$; $A\oplus B\rightarrow B$
- Absorption (AB): if $A\rightarrow B$ then $A\oplus B=A$
- Non-conflict containment (NCC): if $A\mapsto B$ then $A\not\perp B$
- Containment-Preclusion (CP): if $A\mapsto B$, $B\perp C$ then $A\perp C$

CP describes how information preclusion relationships behave in relation to information containment. For example, if you are a vegetarian, then *fruit*$\perp$*meat*. Assuming *apple*$\rightarrow$*fruit*, then CP yields *apple*$\perp$*meat*. In a sense it is a normative property because it expresses what we believe to be the desirable behavior. Barwise & Etchemendy (1990)'s infon algebra also embodies CP whereby information containment is realized via $\Rightarrow$ (involves) over infons (information particles), and $\sigma\perp\overline{\sigma}$ means infon σ precludes its pseudo-complement. Observe that the information containment relation $\rightarrow$ is reflexive, anti-symmetric and transitive. As a consequence, information containment is more general than the hierarchical information structures such as thesauri; it permits a network of containment relationships to be expressed.

[1] The formalization of aboutness operators (e.g. composition, preclusion, containment, etc.) is dependent on the language of the information carriers.

Aboutness is modeled as a binary relation |= over the information carriers **IC**. This symbol should not be confused with logical entailment. We use this symbol for historical reasons as early studies viewed aboutness as a form of entailment (Bruza & Huibers, 1994; Bruza & Huibers, 1996). Bear in mind that aboutness as a broader notion than logical entailment, the details of which will be presented in ensuing sections. Aboutness properties can be expressed in terms of information containment, composition and preclusion. Several authors have studied aboutness and proposed various properties (see (Lalmas & Bruza, 1998) for a recent survey). There are some disagreements about its properties, e.g. Hunter (1996) deems aboutness to be irreflexive whereas Huibers (1996) deems it reflexive. The disagreements stem, partially, from the framework chosen to formalize aboutness. Hunter uses default logic whereas Huibers uses situation theory. Once the framework is fixed, certain aboutness properties are implied by it. In a sense, this is putting the cart before the horse. In this article we attempt to turn things around. By adopting a simple, information-based, abstract framework, we hope to gain enough freedom to propose and discuss a wide range of aboutness postulates without being bound too much by the consequences of the underlying theoretical model.

Huibers (1996) introduced the notion of an *aboutness proof system*. Such a proof system is founded on an aboutness language:

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 $\in$ **IC** then $A \mapsto B$, $A \not\mapsto B$, A⊒B, A⋣B, A→B, A↛B, A⊥B, A$\not\perp$B, A|=B, A$|\neq$B, A=B, A≠B $\in$ 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 ($|=_O$). The symbol / denotes negation, e.g. $A \not\mapsto B$ means "not $A \mapsto B$", or in other words, B is not semantically contained in A.

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
- **R** = $\{R_1,..,R_k\}$ is a finite set of rules

Axioms are elements of the aboutness language that are assumed to be true, e.g., A→A. Rules have premises and a conclusion. For example, the premises of the And rule are A|=B and A|=C, which yield the conclusion A|=B⊕C. As in Huibers (1996), we assume that each rule is decidable as a relation. Axioms and rules can be used to drive inference. The concerns of this paper are inferences of the form A|=B and later A$|\neq$B. The aboutness closure of a proof system is the set of aboutness inferences derivable from a given proof system:

Definition 2.3. Let Π=⟨Λ(**IC**), **A**, **R**⟩ be an aboutness proof system. The aboutness closure, denoted ACl (Π) is defined by:

$$ACl(\Pi) = \{\mathrm{A} \models \mathrm{B} \,\big|\, \mathbf{A} \vdash_{\mathbf{R}} A \models B\}$$

For ease of exposition, we will pretend that the set **A** consists only of axioms involving the aboutness and non-aboutness relation, e.g., A|=A (Reflexivity Axiom). However, in reality there are other axioms, which reflect the properties of the information structure, for example, if *salmon→fish* is an axiom, then so is *salmon* $\not\perp$ *fish* due to the property NCC.

Aboutness inference and context

The issue of aboutness inference and context was alluded to above. Observe that aboutness inferences are drawn within the context of the axioms. More specifically, in this framework here, context is modelled by a set of axioms. These are primitive aboutness relationships, information containment relationships, preclusion relationships that are assumed to hold. For example, assuming a keen wave surfer who wishes to configure an aboutness inference system to infer relevant documents, the following axioms set could be used to establish the context {*surfing* |= *waves*, *surfing* |= *weather*, *surfing*⊥*web*, *surfing*⊥*net*, *wave*⊕*surfing*⊥*wind*⊕*surfing*}. The primitive aboutness relationships establish the interest area; the preclusion relationships are aimed at preventing documents about web surfing, net surfing and wind surfing being filtered. Note that *surfing*⊥*wind* was not specified, as the user does not want to reject documents about wave surfing and wind conditions. Obviously, if the context changes, so to does the set of inferred aboutness relationships.

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* ($|=_O$)

Let $A, B \in$ **IC** and $|=_O \subseteq \mathbf{IC}\times\mathbf{IC}$ such that $(A, B) \in |=_O \Leftrightarrow \exists_{C\in \mathrm{IC}}[A \to C \wedge B \to C]$.

Overlap between information carriers A and B is modeled by an information carrier C which is contained (shared) by both A and B. The more readable convention $\mathrm{A}|=_O\mathrm{B}$, instead of $(A, B) \in\ |=_O$, will be employed to signify that A "is about" B.

Proposition 3.1 $|=_O$ supports *Reflexivity*, *Containment*, *Symmetry*, *Left Compositional Monotonicity*, *Right Compositional Monotonicity*, *And*, *Simplification*, *Loop* and *Mix* where these properties are defined as follows:

$$A\,|=_O A \text{ (Reflexivi ty)}$$

$$\frac{\mathrm{A} \to \mathrm{B}}{\mathrm{A}\,|=_O \mathrm{B}} \text{ (Containment)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B}}{\mathrm{B}\,|=_O \mathrm{A}} \text{ (Symmetry)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B}}{\mathrm{A}\oplus\mathrm{C}\,|=_O \mathrm{B}} \text{ (Left Compositional Monotonicity)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B}}{\mathrm{A}\,|=_O \mathrm{B}\oplus\mathrm{C}} \text{ (Right Compostional Monotonicity)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B} \quad \mathrm{A}\,|=_O \mathrm{C}}{\mathrm{A}\,|=_O \mathrm{B}\oplus\mathrm{C}} \text{ (And)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B}\oplus\mathrm{C}}{\mathrm{A}\,|=_O \mathrm{B} \text{ or } \mathrm{A}\,|=_O \mathrm{C}} \text{ (Simplification)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{B} \quad \mathrm{B}\,|=_O \mathrm{C} \quad \mathrm{C}\,|=_O A}{\mathrm{A}\,|=_O \mathrm{C}} \text{ (Loop)}$$

$$\frac{\mathrm{A}\,|=_O \mathrm{C} \quad \mathrm{B}\,|=_O \mathrm{C}}{\mathrm{A}\oplus\mathrm{B}\,|=_O \mathrm{C}} \text{ (Mix)}$$

Proof:

(1) *Reflexivity (R)*:

$\mathrm{A}\to\mathrm{A} \Rightarrow \mathrm{A}|=_O\mathrm{A}$

(2) *Containment (C)*:

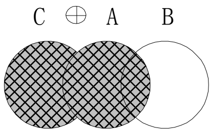
$\mathrm{A}\to\mathrm{B},\ \mathrm{B}\to\mathrm{B} \Rightarrow \mathrm{A}|=_O\mathrm{B}$

(3) *Symmetry (S)*:

$\mathrm{A}|=_O\mathrm{B} \Rightarrow \exists\mathrm{C}|\ \mathrm{A}\to\mathrm{C} \wedge \mathrm{B}\to\mathrm{C} \Rightarrow \mathrm{B}|=_O\mathrm{A}$

(4) *Left compositional monotonicity (LM)*:

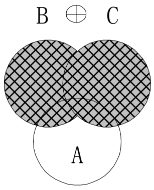
$A \models_O B \Rightarrow \exists D[A \to D \wedge B \to D]$. Observe that $A \oplus C \to A \to D$. Finally, $A \oplus C \to D$ and $B \to D$ yield $A \oplus C \models_O B$. In other words, if A overlaps with B, then adding information C to A will not affect this overlap:



(5) *Right compositional monotonicity(RM)*: Derivable from S and LM.

(6) *And (A)* can be derived from RM.

(7) *Simplification (SM)* can be readily seen when one considers the following two objects. Irrespective of how A overlaps with $B \oplus C$, there must be some overlap between A and B, or A and C, or A and both B and C.



(8) Loop (L) is implied by Symmetry.

(9) *Mix (M)* can be derived trivially from *LM.*

The following lemma establishes that overlap aboutness can be characterized by Reflexivity, Right and Left Compositional Monotonicity. All other properties mentioned above are derivable from these.

Lemma 3.1 Let $\models_O$ be the relation as specified in Definition 3.1 and $\Pi = \langle \Lambda(\mathbf{IC})$, {Reflexivity}, {Right Compositional Monotonicity, Left Compositional Monotonicity}$\rangle$ be an aboutness proof system, Then

$$(A, B) \in \ \models_O \Leftrightarrow A \models_O B \in ACl(\Pi)$$

Proof:

Let A and B be arbitrary information carriers ($A, B \in \mathbf{IC}$). Definition 3.1 states that A is about B, i.e. $(A, B) \in \ \models_O$, if and only if they both contain an X representing the information they both share. Reflexivity states $X \models_O X$. As **IC** is closed under information composition, there must be an information carrier Y such that $X \oplus Y = A$. Similarly, there must be an information carrier X such that $X \oplus Z = B$. Using left monotonicity on $X \models_O X$ yields $X \oplus Y \models_O X$. Application of right monotonicity yields $X \oplus Y \models_O X \oplus Z$. Hence $A \models_O B$.

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* ↦ *microbe* ↦ … ↦ *animal*. The Containment postulate permits both *ghiardia* |= *microbe* (ghiardia is about a microbe) and *ghiardia* |= *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 (1995) documented a user study that supports our contention. Brooks used a hierarchical thesaurus to test whether relevance (aboutness) is inversely proportional to semantic distance. The hierarchical thesaurus contained "broader than" and "narrower than" relationships between terms. The "broader than" relation is equivalent to deep containment, e.g. *Documentation* ↦ *Information Processing* ↦ *Information Services* ↦ *Services*. Semantic distance was measured by the number of steps along the chain, so the semantic distance between *Documentation* and *Services* in the above example is three. Brooks found that the distance to non-relevance (non-aboutness) is approximately three steps.

At first sight, symmetry seems to be an acceptable property. In IR it is a common view that a query q being about a document is the same as stating that d is about q. 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 *surfing⊕Hawaii* |= *surfing*. RM permits *surfing⊕Hawaii* |= *surfing⊕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. In other words, once the system has determined that a document *d* is about a query *q*, the aboutness relation can never be retracted. This means in practice that *d* can never be removed from the result set irrespective of any expansions of query *q*. 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. The vector space model operates according to the following aboutness definition: $d \mid=_O q \Leftrightarrow \cos(\vec{d}, \vec{q}) \geq \partial$. In other words, *d* is deemed to be about *q* iff the cosine of their respective vector representations is greater than or equal to the threshold value δ. This definition allows the retraction of aboutness relationships whenever the query *q* is expanded into a query *q*⊕ *r*, and the cosine of the respective vectors drops below ∂. Document *d* is then no longer retrieved (i.e. the original aboutness relationship *d* |= *q* had been retracted). Although this definition realizes desirable nonmonotonic behavior with respect to aboutness, it is unsatisfactory from a theoretical point of view as the value ∂ is *not* determined by the retrieval model, but is extraneous to it (In practice, the value is determined experimentally).

Simplification states that the aboutness relationship between a carrier A and a complex information carrier B⊕C implies that A is about B, or A is about C. In other words, aboutness can be split into smaller parts and the splitting can go all the way down to the basic information carriers. This is debatable. It may well be that the aboutness relationship between A and B⊕C is dependent on the information granularity of B⊕C. In other words, A is about B⊕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 the AI literature (Kraus et al., 1990} and logic-based IR literature (Amati & Georgatos, 1996; Bruza & Huibers, 1996; Bruza & Van Linder, 1998). These IR accounts all view aboutness in terms of a nonmonotonic consequence relation. Loop when viewed in the context of overlapping aboutness as it is implied trivially by symmetry.

In summary, the overlapping view of aboutness is intuitive, but it implies some unforseen properties, namely left, right montonicty, containment symmetry and simplification. These properties are unsound from a commonsense perspective and can negatively impact information retrieval precision.

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. For the purpose of illustration we will use a variation of the Tweety example. 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 (Bruza & Van Linder 1998, Kraus et al. 1990).

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 (axiom: t→b); Tweety is a penguin (axiom: t→p); penguins are birds (axiom: p→b); birds are about flying (axiom: b|=f); penguins do not fly (axiom: p⊥f). Applying the properties of information containment results in the additional axioms: t→t; b→b; f→f; p→p. Further, applying CP yields the axiom: t⊥f.

4.1 Aboutness Postulates

To distinguish the following properties from the aboutness properties associated with overlap (in the previous section), the symbol |= 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. Dubious rules such as Simplification are dropped.

(R) *Reflexivity*
It seems reasonable to assume that an information carrier is about itself. In terms of information retrieval, this postulate ensures that a document is returned in response to itself being the query.

(AS) *Asymmetry*
Given “football is about sport”, it seems unwarranted to conclude that “sport is about football”. This rule states that aboutness is fundamentally asymmetric.

(AC) *Aboutness Consistency*:

$$\frac{A \models B}{A \perp B}$$

An information carrier should be compatible with what it is about.

(B1) *Semantic Containment*:

$$\frac{A \mapsto_2 B}{A \models B}$$

This property is due to Brook’s (1995). This study revealed that relevance perceptions are inversely proportional to semantic distance. 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. In the absence of similar studies involving surface containment, we generalize Brook’s conclusions to I nformation containment (both surface and deep).

(C) *Containment*:

$$\frac{A \rightarrow_2 B}{A \models B}$$

The containment postulate states that within a restricted context, information is about the information it contains.

(CT) *Cut*:

$$\frac{A \oplus B \models C \quad A \models B}{A \models C}$$

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⊕t|=b, t|=p ⇒ t|=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*:

$$\frac{A \models B \;\; A \models C}{A \oplus C \models B}$$

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⊕C |=B should hold. For example, from t|=p (Tweety is about a penguin) and t|=b (Tweety is about a bird), then t⊕p|=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*:

$$\frac{A \models C \quad \mathrm{B} \models \mathrm{C}}{A \oplus B \models C}$$

For example, p|=b, t|=b ⇒ p⊕t|=b. Unlike preferential entailment (Kraus et al 1990), we argue that Mix produces acceptable aboutness inferences even when information clashes.

(A) *And:*

$$\frac{A \models B \quad \mathrm{A} \models \mathrm{C}}{\mathrm{A} \models \mathrm{B} \oplus \mathrm{C}}$$

The explanation of And is similar to Mix.

The previous three rules featured how compositional 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*

$$\frac{A \models B \quad B \not\perp C}{A \oplus C \models B}$$

Traditional Left Compositional Monotonicity (LCM) is $A|=C \Rightarrow A\oplus B|=C$. This allows $b\oplus t|=f$ (Tweety, which is a bird, is about flying) to be inferred from $b|=f$ (A bird is about flying). Absorption ($b\oplus t=t$) then renders $t|= f$. (Tweety is about flying), which is undesirable as Tweety is a penguin which cannot fly. QLM prevents this via the qualifying preclusion $t\perp f$

QLM deviates from several authors who have advocated a variant of Rational Monotonicity (Bruza & Huibers, 1996; Bruza & van Linder, 1998; Amati & Georgatos, 1996; Wondergem, 1996):

$$\frac{A \models B \quad A \not\perp C}{A \oplus C \models B}$$

Observe that QLM permits the inference $p\oplus f|=b$ (Flying penguins are about birds) from $p|=b$ and $b \not\perp f$. We argue that this inference is acceptable, even though penguins preclude flying ($p\perp f$). Rational Monotonicity prevents such an inference.

(QRM) *Qualified Right Compositional Monotonicity:*

$$\frac{A \models B \quad A \not\perp C}{A \models B \oplus C}$$

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 a bout 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 retract the undesired conclusions from LM and RM, thus describing conservative monotonicity of aboutness with respect to information composition.

4.2 Non-aboutness

Bruza & Huibers (1994), Huibers (1996) and Hunter (1996) have investigated non-aboutness[2]. The practical relevance of these studies is that 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 counterpart 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:*

$$\frac{A \perp B}{A \mid\neq B}$$

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:*

$$\frac{A \mapsto_{>2} B}{A \mid\neq B}$$

[2] Huibers (1996) proposed an additional concept "anti-aboutness" as being distinct from non-aboutness.

It is the complement of the Semantic Containment postulate (B1) and a finding from Brook’s study (Brooks, 1995). If information carrier B is more than 2 semantic (deep containment) steps away from A, then the aboutness relation is severed yielding non-aboutness. Once again, we generalize Brook’s finding to hold for surface containment as well:

(N-C) *Containment non-aboutness:*

$$\frac{A \rightarrow_{>2} B}{A \mathrel{|}\!\neq B}$$

(B3) *Inverse Semantic Non-aboutness:*

$$\frac{A \mapsto B \quad A \neq B}{B \mathrel{|}\!\neq A}$$

Brooks (1995) 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” (Brooks, 1995, p111). For example, imagine an information carrier describing bibliometrics. Subjects in the study tended to give low relevance scores to descriptors more specific than “bibliometrics”, e.g. “citation analysis”. As was the case with the B1 rule, we assume that B3 generalizes to information containment:

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

$$\frac{A \rightarrow B \quad A \neq B}{B \mathrel{|}\!\neq A}$$

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

$$\frac{A \models B \quad B \perp C}{A \mathrel{|}\!\neq C}$$

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⊥f ⇒ t|≠f.

4.3 Interactions Between Aboutness and Non-aboutness

Assume dancing is about “having fun” and dancing is not about “sitting still”, is dancing about “having fun⊕sitting still”? This example demonstrates the interaction between aboutness and non-aboutness. The following properties attempt to characterize the interaction. They are normative meaning the rules are motivated form a particular standard, or perspective. In the following “optimism” connotes a standard whereby aboutness premises are favoured over non-aboutness premises, whereas “pessimism” connotes the converse. It is assumed that those two stances are mutually exclusive, i.e. either an optimistic or a pessimistic stance is adopted, but not mixed.

(OL) *Optimistic Left:*

$$\frac{A \mathrel{|}\!\neq B \quad C \models B}{A \oplus C \models B}$$

For example, t|≠f, b|=f ⇒ t⊕b|=f (Tweety bird is about flying) demonstrates that optimism can lead to dubious aboutness inferences because it implies left monotonicity. The optimism stems from C|=B being favoured to draw the conclusion A⊕C |= B, irrespective of A’s aboutness with B.

$$\frac{A \mathrel{|}\!\neq B \quad A \models C}{A \models B \oplus C}$$

For example, t|≠f, t|=p ⇒ t|=f⊕p. That is, from “Tweety is not about flying” and “Tweety is about a penguin”, an optimist can conclude that “Tweety is about a flying penguin”.

(PL) *Pessimistic Left:*

$$\frac{A \mathrel{|}\!\neq B}{\mathrm{A} \oplus \mathrm{C} \mathrel{|}\!\neq \mathrm{B}}$$

(PM) *Pessimistic Middle:*

$$\frac{A \mid\neq C \quad B \mid= C}{A \mid\neq B}$$

(PR) *Pessimistic Right:*

$$\frac{A \mid\neq B}{A \mid\neq B \oplus C}$$

PL and PR state that non-aboutness is monotonic with respect to information composition, or more informally, “once a pessimist, always a pessimist”. PR is also known as the Negation Rationale (Bruza & Huibers, 1994; Hunter, 1996). PM adopts a selective approach in the vain of “Peter likes skiing, but John doesn’t, so Peter won’t like John”. In terms of aboutness, if penguins are not about flying, and hawks are, then penguins aren’t about hawks. PM enforces a strict standard in favour of non-aboutness.

IR models supporting aboutness defined in terms of overlap are often optimistic. For example, the vector space model is optimistic when the threshold value ∂ is greater than zero: For document *d* and query *q*, ($d \mid\neq q \Leftrightarrow \cos(\vec{d}, \vec{q}) > 0$. Globally speaking, optimism promotes the recall of an IR system, i.e. by tolerating certain possibly unsound aboutness inferences an optimistic IR system would retrieve more relevant documents. Pessimism, on the other hand, attempts to preserve precision (at the expense of recall).

5. Completeness, Consistency and Soundness

5.1 Completeness

Given a description of a user’s preferences for information as a profile, an information filtering system must determine whether an incoming document is about this description, or not. As a consequence, the completeness of an aboutness inference system is an important issue. 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 by van Rijsbergen (1989). We agree with this view and have characterized non-aboutness via constructive means (see Section 4.2). Huibers (1996) also follows this line. The following proposition asserts that the commonsense aboutness and non-aboutness postulates are incomplete. Before the details of this result can be given, a non-aboutness closure is defined. This is simply the set of all non-aboutness inferences derivable from a set of rules prescribing non-aboutness together with non-aboutness axioms.

Definition 5.1. Let Π=⟨Λ(**IC**), **A**, **R**⟩ be a non-aboutness proof system. The non-aboutness closure, denoted *NCl* (Π) is defined by

$$NCl(\Pi) = \{\mathrm{A} \mid\neq \mathrm{B} \mid \mathbf{A} \vdash_{\mathbf{R}} A \mid\neq B\}$$

Proposition 5.1 *The aboutness proof system* Π=⟨Λ(**IC**), {R}, {C, CT, CLM, M, A, QLM, QRM}⟩ and the non-aboutness system Ω=⟨Λ(**IC**), ϕ, {P, N-C, I-CN, P-NA}⟩ *are incomplete.*

Proof:

Assume that **IC** = {x, y, x⊕y} and ⊥=ϕ (there are no preclusion relationships). Assume also that the only non-reflexive information containment relationships are x⊕y→x and x⊕y→y. We will show that $x \mid= y \notin ACl(\Pi)$ and $x \mid\neq y \notin NCl(\Omega)$.

- In order to show $x \mid\neq y \notin NCl(\Omega)$, it must be shown that there is no rule which can derive x|≠y.
  - (P) and (P-NA) cannot derive x|≠y due to the absence of preclusion relationships.
  - (N-C) and (I-CN) cannot derive x|≠y as $x \not\to y$ and $y \not\to x$.

  Therefore, $x \mid\neq y \notin NCl(\Omega)$.

- In order to show $x \mid= y \notin ACl(\Pi)$, it must be shown that there is no rule which can derive x |= y.
  - (C) cannot derive x |= y as $x \not\to y$ and $y \not\to x$.

- In order for the Cut (CT) rule to derive x |= y, there must be an information carrier B such that x⊕z |= y and x |= z. The choices for carrier z are the three carriers in **IC.** None of these suffice to initiate the Cut rule in the desired fashion.
- The relationship x |= y could be derived from conservative left monotonic rules if we can derive x⊕z |= y with the requirement that x→z, x≠z, which renders x |= y because of the absorption principle. Due to our assumptions, z does not exist, so x |= y cannot be derived from the rules (CLM), (M) and (QLM). For similar reasons, x |= y cannot be derived from the conservative right monotonic rules (A) and (QRM).

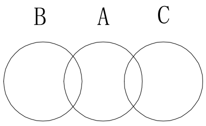
Therefore, $x \models y \notin ACl(\Pi)$.

The above incompleteness result parallels a similar result using default logic-based theory of aboutness (see Proposition 4.3 of (Hunter, 1996)). Both results suggest that aboutness is inherently an incomplete notion meaning that there will always be information carriers A and B for which we cannot determine with reasonable confidence whether A is about B, or A is not about B. As mentioned above, incompleteness is an undesirable property from a practical perspective. It is worth noting that by introducing unsound aboutness properties such as Right Containment Monotonicity, or Equivalence, completeness can be achieved. The price is an increase in the number of unsound aboutness inferences (i.e., loss of precision in IR terms). The following result demonstrates how completeness can be achieved via the inclusion of the Equivalence (E) property.

(E) *Equivalence:*

$$\frac{A \models B \quad B \models A \quad A \models C}{B \models C}$$

Equivalence has been shown to be a sound aboutness property when aboutness is interpreted as a relation of preferential entailment[3] (Amati & Georgatos, 1996; Bruza & Huibers, 1996; Bruza & van Linder, 1998). Note that when aboutness is equated with overlap, it leads to unsound aboutness inferences. This is because the overlapping between A and B and between A and C do not imply the overlapping between C and B. This is demonstrated by the following diagram:



Note, however, that when aboutness is defined in terms of a high degree of overlap, equivalence is a sound property.

Proposition 5.2 *The aboutness proof system* Π=⟨Λ(**IC**), {R}, {C, QLM, QRM, E}⟩ *and the non-aboutness proof system* Ω=⟨Λ(**IC**), ϕ, {P, N-C, I-CN}⟩ *are complete.*

Proof:

Let x and y be arbitrary information carriers:

- If x = y, then x |= y (R)
- If x ≠ y,
  - If x→y then either,
    x |= y from $x \rightarrow_2 y$ (C), or
    x |≠ y from $x \rightarrow_{>2} y$ (N-C)
  - If y→x, then x |≠ y (I-CN)
  - If neither x→y nor y→x
    If x⊥y, then x |≠ y (P)
    If $x \not\perp y$, then: x|=y⊕x (QRM), x⊕y|=x (QLM) and y⊕x|=y (QLM). Applying E (Equivalence) yields x|=y.

[3] A preferentially entails B if and only if the preferred models (e.g., documents) where A holds, B also holds.

Hence, for arbitrary information carriers x and y, either x |=y or x |≠ y.

A similar result is possible by using the unsound Right Containment Monotonicty (RCM) property. For example, when A|=B is derived from the containment (C) rule and assuming C is within 2 semantic steps from B, then RCM permits the conclusion A|=C, even though C is nested more than 2 steps away from A.

(RCM) *Right Containment Monotonicty*:

$$\frac{A \mid= B \quad B \rightarrow_2 C}{A \mid= C}$$

5.2 Consistency

Consider the information carrier A = "Tweety (t) is a penguin (p)" and B = "Jack (j) is a hawk (h)". Assume the additional axioms "Jack flies (j |= f) and "Jack is a bird" (j |= b) to the ones already stated for the Tweety example. Conflicting aboutness inferences can be seen via the following derivations.

$$\dfrac{\dfrac{\dfrac{\dfrac{\mathrm{t} \rightarrow \mathrm{p} \quad \mathrm{p} \perp \mathrm{f}}{\mathrm{t} \perp \mathrm{f}}(CP) \quad \mathrm{t} \mid= \mathrm{b}}{\mathrm{t} \mid\neq \mathrm{f}}(P)}{\mathrm{t} \mid\neq \mathrm{f} \oplus \mathrm{b}}(PR) \quad \dfrac{\mathrm{j} \mid= \mathrm{f} \quad \mathrm{j} \mid= \mathrm{b}}{\mathrm{j} \mid= \mathrm{f} \oplus \mathrm{b}}(And)}{\mathrm{t} \mid\neq \mathrm{j}}(PM)$$

Hence, driven by a pessimistic stance, $\mathrm{t} \mid\neq \mathrm{j}$.

On the other hand, assume that Jack and Tweety are both owned by Bill, so Tweety and Jack don't preclude each other ($t \not\perp j$) Using Qualified Right Monotonicity (QRM) does permit the conclusion that Tweety is about Jack, under the assumption that Tweety does not clash with Jack:

$$\dfrac{\dfrac{\mathrm{t} \mid= \mathrm{b} \quad \mathrm{t} \not\perp \mathrm{j}}{\mathrm{t} \mid= \mathrm{b} \oplus \mathrm{j}}(QRM)}{\mathrm{t} \mid= \mathrm{j}}(Absorption)$$

In the framework presented in this paper, such inconsistencies cannot be resolved. However they are resolvable. Aboutness relationships can be ordered relfecting the intuition that some aboutness relationships are "stronger" than others. IR systems implement this intuition via terms weights. For example if term t has a higher weight than term u in the context of document d, then the relationship d |= t is considered stronger than d |= u. A similar statement can be made about document rankings: If d_1 is ranked higher than d_2 in repsonse to q, then the relationship d_1 |= q is deemed stronger than d_2 |= q. By ordering aboutness inferences it may be the case that t |≠ j is deemed stronger than t |= j. This provides a basis for chosing the former inference over the latter, thus resolving the inconsistency. Only in the case that both inferences are strong would imply that more information is required to break the deadlock. The details on how to extend the aboutness proof theory to produce orderings on aboutness inferences are yet to be worked out. Indications about how it could be done are presented later in this paper.

5.3 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. The soundness can be verified by proving that in all models where the premises of the rule hold, the conclusion also holds. In logic, models are formally defined frameworks representing some slice of reality. It is not possible to verify aboutness rules using the same approach, because IR lacks an underlying and integrating model theory. We believe that the inference processes involved in determining aboutness are "psycho-logistic" in nature- that is the inference processes involved cannot be studied independently of the user doing the reasoning. The only recourse is to perform studies which investigate how users reason about aboutness, and thereby attempt to identify those rules which produce agreeable inferences across a majority of users. Brook's (1995) is an example of such an investigation. It is interesting to draw a parallel with non-monotonic reasoning. The soundness of non-monotonic reasoning systems is investigated in the context of non-monotonic reasoning benchmark problems, which are based on an "agreed" correct solution. The agreement has been established via researchers in the AI community. Studies on first year psychology students have shown significant variations from the agreed solutions of some benchmark problems (Elio & Pelletier, 1994; Elio & Pelletier, 1996; Pelletier & Elio, 1997). This not only demonstrates that non-monotonic reasoning is psycho-logistic in nature, but also that soundness must ultimately be grounded from a inter-subjective user perspective. The inter-subjective agreement can then be formalised as a set of rules for driving the inference process. Inference engines can then be constructed whereby the inferences drawn would be acceptable to a majority of users.

6. Similarity versus Aboutness

The deep containment relation can be used to model the "broader than" thesaurus relation. What about synonyms, which is a similarity relationship between terms? In IR, similarity not only plays a role at the level of terms, but at the level of documents as well via document clustering. The imaging-based logical models rely on a similarity relationship spanning the space of worlds (Crestani & Van Rijsbergen, 1998). It is important to investigate the relationship between similarity and aboutness. This poses a major problem in a pure symbolic framework as similarity is an inherently fuzzy notion. As a consequence, the potential of unsound aboutness inferences increases without the machinery to deal with it effectively.

Similarity is modeled as a reflexive, symmetric relation ~ over the information carriers **IC**. *A~B* denotes that information carrier A is similar to carrier B. We assume that similarity is preserved under context dependent compositional monotonicity:

(MS) *Modus Substituens:*

$$\frac{B \sim C}{A \oplus B \sim A \oplus C}$$

This property states that B can be substituted by C (and vice versa) in the context of A. For example, assuming that sparrows are similar to pigeons (sparrow~pigeon), we may infer that migrating sparrows are similar to migrating pigeons (migrate⊕sparrow~migrate⊕pigeon). The Modus Substituens rule originates from a plausible inference rule of the same name documented in (Bruza, 1993; Chen, 1994).

Sun (1995) prescribed the interaction between similarity and rule-based reasoning. The following rule is taken from this work and placed within the context of aboutness:

(S) *Aboutness Similarity:*

$$\frac{A \sim B \quad B \models C}{A \models C}$$

Reflexivity of aboutness, together with (S) yields the conclusion A|=B from A~B. In other words, similarity is a form of aboutness. Note that the soundness of this inference is proportional to the integrity of the similarity relation.

The above combination yields a very common form of aboutness reasoning pattern. For example, using similarity as the starting point, one can reason that a plumbing device is about a plumbing apparatus:

$$\cfrac{\cfrac{device \sim apparatus}{plumbing \oplus device \sim plumbing \oplus apparatus}}{plumbing \oplus device \models plumbing \oplus apparatus}$$

Sun (1995) argues that information containment is a special case of similarity. This is a contentious argument that contradicts Brook's study (1995). If we do accept Sun's position, inconsistent aboutness inferences may occur due to the well-known problems in inheritance reasoning. By way of illustration, from $penguin \mapsto bird$, one can conclude *penguin~bird*. Under the assumption *bird*|=*fly*, *penguin*|= *fly* could be inferred.

One way to avoid this problem is to introduce a normative rule that states similarity is only permitted between information carriers with the same level of information granularity:

$$\frac{A \sim B}{A \nrightarrow B \text{ and } B \nrightarrow A}$$

If this is too strong a restriction, then some mechanism must be introduced to order aboutness inferences, e.g. by using weights. In short, similarity cannot be comprehensively studied with within a purely symbolic framework.

7. Theoretical Case Study: Monotonicity

Monotonicity is a property that has engendered considerable attention because it is a property that can adversely affect the precision of an IR system (as was argued in Section 3). In this section, several prominent IR models are placed within the perspective of this property. First, the aboutness relation is defined in terms of the IR model in question.

Definition 7.1 (*Boolean retrieval aboutness)* Let document representation *d* be a conjunction of terms (atomic propositions) drawn from vocabulary T. Let query *q* be a Boolean formula whose atomic propositions are drawn from T and combined via the connectives $\wedge$ (conjunction), $\vee$(disjunction) and $\neg$(negation). Then,

$$d \models_B q \Leftrightarrow d \vdash q \cdot$$

In other words, *d* is about *q* if and only if *q* can be derived from document representation *d* (see Bruza & Huibers (1994) for more details). Note that in the Boolean model, information composition is realized via logical conjunction.

Two versions of the vector space model will be studied depending on the threshold value: (In vector based systems, information composition is realized by vector addition.)

Definition 7.2 *(Un-thresholded vector retrieval aboutness)* Let *d* and *q* be n-dimensional vectors, where n is the cardinality of vocabulary T. Then,

$$d \models_{UV} q \Leftrightarrow \cos(d,q) > 0 \cdot$$

Definition 7.3 (*Thresholded vector retrieval aboutness*) Let *d* and *q* be n-dimensional vectors, where n is the cardinality of vocabulary T. Then,

$$d \models_{TV} q \Leftrightarrow \cos(d,q) > \partial \text{ where } \partial>0.$$

In probabilistic retrieval, aboutness between a document representation *d* and a query *q* depends on the event “relevance” with respect to a probabilistic decision rule.

Definition 7.4 (*Probabilistic retrieval aboutness*) Let *d* and *q* be n-dimensional vectors, where n is the cardinality of vocabulary T, and R denote the relevance event. Then,

$$d \models_{PR} q \Leftrightarrow \Pr(R \mid d,q) > \Pr(\neg R \mid d,q)$$

Proposition 7.1 In the context of the aboutness properties {R, LM, RM}

- Un-thresholded vector retrieval (UV) as defined in Definition 7.2 supports {R, LM, RM}
- Boolean retrieval (BR) as defined in Definition 7.1 supports {R, LM}
- Thresholded vector retrieval (TV) as defined in Definition 7.3 supports {R}
- Probabilistic retrieval (PR) as defined in Definition 7.4 does not support any of {R, LM, RM}

Proof:

UV:

cos(d,q) will be greater than zero iff there exists a non-empty subset s of terms (s$\subseteq$T) whereby s$\subseteq$q and s$\subseteq$d. In other words, $\models_{UV}$ is a particular implementation of overlapping aboutness $\models_O$. Therefore, from Proposition 3.1, $\models_{UV}$ supports R, LM and RM.

BR:

- Reflexivity (R) is supported as d |- d.

- Left monotonicity (LM) is supported as, given d |- q, then for an arbitrary x, d∧x |- q.
- Right monotonicity (RM) is not supported. Given d |- q, for an arbitrary x, d |- q∧x cannot be concluded as the fact d |- x is unknown.
- Therefore, $\models_B$ supports R and LM.

TV:

- R is implied by TV as cos(d,d)=1.
- In order to show LM, it must be shown that: for an arbitrary x and under the premise cos(d,q)>∂, cos(d⊕x,q)>∂. Consider the case where x contains many terms that do not exist in d. In such a case it may well turn out that the cosine is diluted to the point where cos(d⊕x,q)≤∂. Hence, TV does not support LM. (The argument that TV does not support right monotonicity (RM) follows a similar line).
- Therefore TV supports R, but not LM and RM.

PR:

- In order to show PR supports R, one must show that Pr(R|d,d)>Pr(¬R|d,d), which simplfies to showing Pr(R|d)>Pr(¬R|d). This relationship need not hold for an arbitary document d. So refexivity is not implied by PR.
- In order to show LM, it must be shown that for an arbitrary x and under the premise Pr(R|d,q)>Pr(¬R|d,q), Pr(R|d⊕x,q)>Pr(¬R|d⊕x,q). Consider the case where x contains many terms that appear in irrelevant documents. In such a case it may well turn out that Pr(¬R|d⊕x,q)>Pr(R|d⊕x,q). Thus, PR does not support LM. (The argument that PR does not support RM follows a similar line).
- Therefore PR does not imply R, LM and RM.

It is interesting to note that each of the above retrieval models are examples of a particular class of model. For example, un-thresholded vector retrieval is an example of a naïve overlap model in which aboutness is determined by a non-zero level of overlap. Boolean retrieval is an example of a containment model (see Wong et al., (1999)). In other words a query q must be derivable[4] from d, for d to be deemed about q. Containment retrieval models are also known as “exact-match” models. Thresholded vector retrieval is an example of a more sophisticated overlap model wherein aboutness is determined by a threshold determining “sufficient” overlap. Finally, probabilistic retrieval represents the class of models employing the probabilistic decision rule. Figure 1 attempts to depict the above models in a spectrum defined by left and right monotonicity. The black hole in the centre represents no aboutness properties at all. Probabilistic models are placed in the middle of the spectrum as they do not embody any monotonic properties. They do not even embody reflexivity, which, even though is not a monotonic property, has been added to the spectrum as a point of reference. In other words, probabilistic models are fully non-monotonic. The thresholded overlap models are also non-monotonic with respect to aboutness, but they distinguish themselves from the probabilistic models, as they are reflexive. Naïve overlap models exist at two extremes. They are fully monotonic (both left and right). On the other hand, the containment models are left monotonic. As a general rule, the precision of a model is inversely proportional to the degree of monotonicity. This suggests that the naïve overlap models are the least precise followed by the containment models followed by the non-monotonic models. This is reflected by the circles in the graph. The size of the circle depicts the number of aboutness inferences present in the closure of an aboutness proof system involving on this rule. We can see, for example, that left monotonicity (LM) implies cautious monotonicity (CM) as the closure of the latter rule is a subset of the former. The larger the closure, the greater the probability of unsound aboutness inferences.

An interesting class of retrieval models is not present in the spectrum, namely the models which are conservatively monotonic. The conservatively monotonic models are interesting because there are indications that IR is a conservatively monotonic, rather than non-monotonic process. Take, for example, query expansion. When expanding a query with additional terms, the terms added are not arbitrary. They must be chosen carefully, i.e., conservative monotonicity is at work here. We suspect that relevance feedback is also fundamentally conservatively mono-

[4] Strict inference is a form of information containment (Bruza & Huibers, 1994; Van Rijsbergen 1989)

tonic processes. In terms of aboutness, such models embody properties such as QLM, CM, QRM, without also supporting LM and RM. The authors are unaware of any IR models, which fall into this class, however it is possible that certain types of weighting schemes employed in conjunction with non-monotonic models may be able to simulate a form of conservative monotonicity. Wong et al. (1998) obtained a similar observation. It is important that this class of model be studied and developed. They have the advantage over non-monotonic models as their behaviour can be characterized by symbolic rules making them more transparent than the "black hole" non-monotonic models.

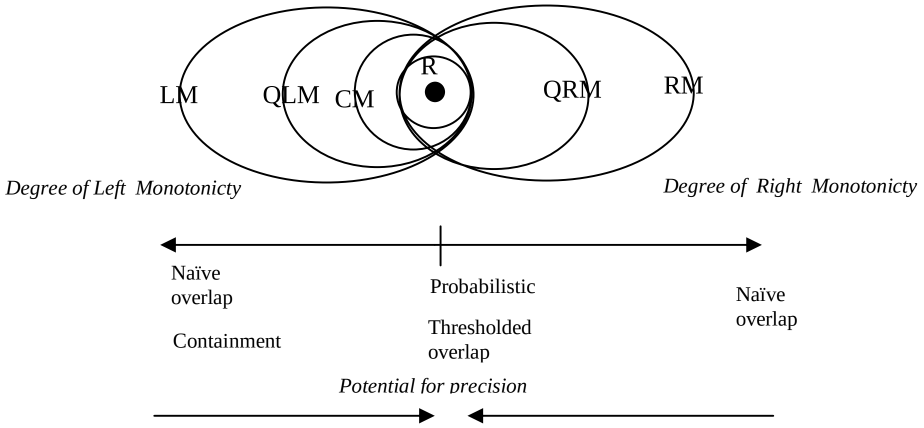


Figure 1: The spectrum of monotonicity

8. Discussion

We begin this section by relating the work presented in previous sections with previous major studies on aboutness. Much of the work presented here follows the same philosophy as Huibers (1996). A major difference is the framework chosen for formalization. Huibers uses situation theory, so concepts such as information composition can be expressed in terms of set union. In addition, situation theory allows the definition of positive and negative information particles known as infons. This allows a more detailed analysis of non-aboutness. Huibers distinguishes carefully between non-aboutness and anti-aboutness, the latter being expressed in terms of negative infons. Our framework lacks the expressiveness to model anti-aboutness, because we do not have the ability to express negation. We feel that our framework can be extended to include negation by defining it extensionally using information preclusion. For example, the negation of an information carrier could be defined as the set of carriers which it precludes. We also feel that for most applications of aboutness theory, the division between aboutness and non-aboutness, as presented in this account, is sufficient.

Huibers introduces three forms of information composition – union and intersection of sets of infons, and a third form which operates directly upon the infons. This permits a finer grain of analysis, but is less general than ours as the formalization is dependent on situation theory.

The set of aboutness and non-aboutness properties proposed in this account is also distinct from those of Huibers. Huibers tried to propose aboutness properties as completely as possible. This is necessary to model the functionality of different IR models. However, in this paper, we study the problem from another point of view (i.e. aboutness itself) and investigate various "reasonable" aboutness properties. We argue that they are sound from a commonsense perspective. For example, Symmetry is included in Huibers's framework because those IR models supporting zero-threshold overlap would satisfy this property. But it is dropped in our proposal as we consider aboutness fundamentally asymmetric. Even though some IR models support Symmetry, this property is in our opinion unsound.

Hutchins (1977, p30) relates aboutness of a text to the semantic networks drawn from it. He distinguishes between the macro and micro structures of the text; both are important with respect to aboutness. Under our framework, we see the semantic network of the macro-structure being based on the surface containment relation, and the micro-structure being based on the deep (semantic) containment relation. In the Figure 2, the upper layer represents the macro-level of the text, in which elements of the macro-structure are ordered via the surface containment relation. In this way, the thematic progression of the text could be modeled. The lower level constitutes the micro-structure of the text. The elements in this level would be at a low level of information granularity (e.g. topics), and their inherent informational ordering would be expressed by the deep containment relation. Additional expressivity at both levels is fostered by the information preclusion relation (⊥) and the similarity relation (~).

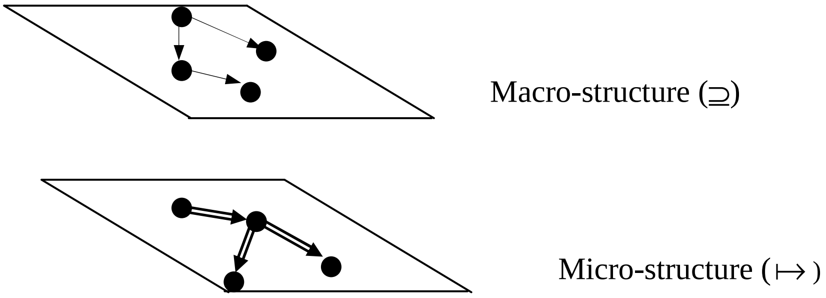


Figure 2 Micro- and macro- structures of text

The aboutness can be formalized using our framework as follows. Extentional aboutness (Hutchins, 1977, p26) has to do with the topics of component parts of a text. This can be modelled as the closure of an aboutness proof system constructed from the micro-structure. Intentional aboutness (Hutchins, 1977, p26) has to do with the topic of the text at a more global level. This could be modelled as the closure of an aboutness proof system constructed form the macro-structure of the text.

Cooper (1971) subdivides the notion of relevance into *utility* and *logical relevance*. Utility has to do with the ultimate usefulness of a piece of information to the user, whereas logic relevance has to do with the topical relatedness. Placed within these definitions, this work can be seen as an attempt to formalize logical relevance by formalizing commonsense properties describing the aboutness relation. Non-aboutness and the interaction between aboutness and non-aboutness are also characterized, the latter via normative rules. The properties are consolidated from the representative work in the area. Our characterization of the aboutness relation shows some similarities to the characterization of preferential entailment (Kraus, Lehmann & Magidor, 1990), with the primary difference in how conservative monotonicity is described. These differences suggest that aboutness relation is similar, but not identical to nonmonotonic consequence.

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. However informational concepts are in the background. We have seen that naïve vector space retrieval can be driven by an aboutness proof system involving {R, LM, RM} (Lemma 3.1). Aboutness and preclusion relationships can be derived via relevance feedback (Amati & Georgatos 1996, Bruza & van Linder 1998). For restricted domains, information containment relationships can be derived from ontologies, and the like. When language processing tools have advanced further, the concepts under the aboutness theory could be applied to IR more easily. More sensitive IR systems would then result; in particular those which are conservatively monotonic with respect to composition. The lack of such systems currently can be attributed in part to the inability to effectively "operationalize" information preclusion.

There has been a good deal of discussion in this article about compositional monotonicity simply because this property has a direct bearing on the precision of an IR system. The spectrum presented in Section 7 could be extended to include other rules prescribing guarded forms of monotonicity, e.g. Mix, And, etc. The position of a rule in the spectrum could be used to order the aboutness inferences with respect to soundness. 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 (as seen by the size of its associated closure). In this way, orderings on aboutness inferences can be generated. An interesting 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 some IR theory can be re-created in a symbolic framework which could possible extend the explanatory power of IR theory. For example, such a framework may lend itself to explaining why certain weighing schemes are superior to others. In addition, we feel that aboutness theory could also be applied to the following areas:

- *IR functional benchmarking*. Song, et al. (1999) and Wong, et al. (1998) proposed the methodology and the overall strategy for aboutness-based functional benchmarking of IR as a complement of the traditional empirical approach (performance benchmarking). Although the latter is good at evaluating the performance of an IR system, it is unable to assess its underlying functionality and explain why the system shows such performance. This problem can be overcome by functional benchmarking, where aboutness plays a central role.
- *Query expansion/relevance feedback*. These are widely used techniques in IR, and could be described in terms of a conservatively monotonic aboutness reasoning system. Commonsense properties of aboutness would form the basis of the inference rules guiding the query expansion process. By way of illustration, based on some primitive aboutness axioms like surfing |= waves and associated preclusions, conservative monotonicity would produce inferences like surfing |= wave⊕conditions, wave⊕surfing |= surfing⊕hawaii, whereby wave⊕conditions and surfing⊕hawaii would be considered as possible expansions of "surfing" [Bruza & Van Linder, 1998].
- *Intelligent information agents*. Aboutness and non-aboutness proof systems could be used to make relevance decisions.

9. Conclusions

A contribution of this paper is a set of aboutness and non-aboutness properties which have been expressed symbolically. As a consequence, they are in the open for further discussion and elaboration. Our hope is that this work promotes a debate on which properties of aboutness are useful and relevant for IR, thus expanding the boundaries of IR theory. We cannot claim the set of commonsense properties to be "the set" which totally unravels aboutness, and non-aboutness, but put them forward as a potentially useful set from which further work can proceed. More specifically, the following can be concluded from this account:

- An analysis of a common form of aboutness used in IR, namely aboutness defined in terms of overlap, implies precision degrading properties, in particular, monotonicity.
- We contribute to the evidence that aboutness is an inherently incomplete concept. Completeness can be achieved, however, via the introduction of unsound rules.
- Aboutness can produce conflicting inferences (inconsistency). Most conflicts could be resolved via orderings on aboutness and non-aboutness inferences.
- The soundness of aboutness inference should be motivated from cognitive studies which examine aboutness reasoning patterns.
- Most common IR models are either monotonic or non-monotonic - another class of IR models, namely those that are conservatively monotonic is missing. Such models are interesting for purposes for producing symbolic inference foundation to query expansion and perhaps even relevance feedback.
- Further work in aboutness should focus on incorporating orderings on aboutness inferences, whereby the ordering reflects the confidence in the inferences.

In our opinion, a good understanding of aboutness will lay down significant theoretical groundwork in IR research. This in turn will lead to more effective IR systems.

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). The first author would like to thank Gianni Amati, Theo Huibers, Bernd van Linder and Keith van Rijsbergen for the interesting discussions in the topic area of this paper.

References

Amati, G. & Georgatos, K. (1996). "Relevance as deduction: a Logical View of Information Retrieval." In Crestani, F. and Lalmas, M. (Eds.), *Proceedings of the Second Workshop on Information Retrieval, Uncertainty and Logic (WIRUL'96)* (Technical Report TR-1996-29 Department of Computer Science, University of Glasgow, 21-26.

Barwise, J. & Etchemendy, J. (1990). Information, Infons and Inference. In *Situation Theory and its Applications* (Cooper, R. *et al.* Eds). CLSI Lecture Notes, 1, 33-78.

Brooks, T. A. (1995) People, Words, and Perceptions: A phenomenological Investigation of Textuality. *Journal of the American Society for Information Science*. 46(2): pp.103-115, 1995.

Bruza, P.D. (1993). Stratified Information Disclosure. Ph.D. Thesis. University of Nijmegen. The Netherlands.

Bruza, P.D. & Huibers, T.W.C. (1994). Investigating aboutness axioms using information fields. In *Proceedings of ACM SIGIR Conference on Research and Development in Information Retrieval.* Dublin, Ireland, pp.112-121.

Bruza, P.D. & Huibers, T.W.C. (1996). A study of aboutness in information retrieval. *Artificial Intelligence Review 10*, pp.1-27.

Bruza, P.D. & van Linder, B. (1998). Preferential Models of Query by Navigation. In *Information Retrieval: Uncertainty and Logics* (Crestani, F., Lalmas, M. and Van Rijsbergen, C.J. eds.) The Kluwer international series on Information Retrieval. Kluwer Academic Publishers.

Chen (1994) On Inference Rules of Logic-based Information Retrieval systems. *Information Processing & Management* 30(1): 43-59, 1994.

Cleverden, C.W. (1991). The Significance of the Cranfield Tests on Index Languages. In *Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and Development in Information Retrieval*, pp 3-12, 1991.

Cooper , W.S. (1971). A Definition of relevance for Information Retrieval. Information Storage and Retrieval, 7, pp. 19-37, 1971.

Crestani, F.& van Rijisbergen, C.J. (1998). A Study of Probability Kinematics in Information Retrieval. *ACM Transactions on Information Systems* 16(3): 225-255.

Dubois, D., Farinas del Cerro, L., Herzig, A., & Prade, H. (1997). Qualitative Relevance and Independence: A Roadmap. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence (IJCAI-97),* pp. 62-67, 1997.

Elio, R. & Pelletier, F.J. (1994) On Relevance in Nonmonotonic Reasoning: Some Empirical Studies. In R. Greiner & D. Subramanian (Eds) Relevance. American Association for Artificial Intelligence Fall Symposium Series, Nov 4-6, 1994 pp 64-67. AAAI Press.

Elio, R. & Pelletier, F.J., (1996) On Reasoning with Default Rules and Exceptions. In *Proceedings of the 18th Conference of the Cognitive Science Society*, pp 131-136, Lawrence Erlbaum. 1996.

Huibers, T.W.C. (1996). *An Axiomatic Theory for Information Retrieval*. Ph.D. Thesis, Utrecht University, The Netherlands. 1996.

Hunter, A. (1996). A. Hunter. Intelligent text handling using default logic, In *Proceedings of the Eighth IEEE International Conference on Tools with Artificial Intelligence (TAI'96),* 34-40, IEEE Computer Society Press.

Hutchins, W.J. (1977). On the problem of 'aboutness' in document analysis. *Journal of Informatics*, 1(1):17-35, 1977.

Kraus, S., Lehmann, D., & Magidor, M. (1990). Nonmonotonic Reasoning, preferential models and cumulative logics. *Artificial Intelligence* 44, pp.167-207.

Lalmas, M. (1996). *Theories of Information and Uncertainty for the modelling of Information Retrieval: An application of Situation Theory and Dempster-Shafer's Theory of Evidence.* Ph.D. Thesis. University of Glasgow. 1996

Lalmas, M. (1998). Logical models in information retrieval: Introduction and overview. *Information Processing & Management* 34(1): 19-33.

Lalmas, M. & Bruza, P.D. (1998). The use of logic in information retrieval modeling. *Knowledge Engineering Review* 13(3): 263-295.

Landman, F.W. (1986). *Towards a theory of information. The status of partial objects in semantics.* Foris, Dordrecht, 1986.

Maron, M.E. (1977). On Indexing, Retrieval and the Meaning of About. *Journal of the American Society for Information Science*, 28 (1): 38-43.

Nie, J., Brisebois, M., & Lepage, F. (1996). Information retrieval as counterfactual. *The Computer Journal 38,* 8, pp.643-657.

Pelletier, F.J. & Elio, R. (1997) What should default reasoning be, by default? *Computational Intelligence* 13(2):165-187.

Proper, H.A. & Bruza, P.D. (1999) What is information discovery about? *Journal of the American Society for Information Science*, 50 (9): 737-750.

Sebastiani, F. (1998). On the role of logic in information retrieval. *Information Processing & Management, 34,* 1, 1-18.

Song, D.W., Wong, K.F., Bruza, P.D., & Cheng, C.H. (1999). Towards functional benchmarking of information retrieval models. In *Proceedings of 12th International Florida Artificial Intelligence Society Conference (FLAIRS '99),* pp. 389-393, Orlando, USA.

Sun, R. (1995). Robust reasoning: integrating rule-based and similarity-based reasoning. *Artificial Intelligence* 75 (1995) pp.241- 295.

Wondergem, B. (1996). *Preferential Structures for Information Retrieval*. MSc thesis. Technical Report INF-SCR-96-21. Department of Computer Science. Utrecht University. The Netherlands.

Wong, K.F., Song, D.W., Bruza, P.D., & Cheng, C.H. (1998). Application of aboutness to functional benchmarking in information retrieval. In revision for *ACM Transactions on Information Systems.*