Fundamental Properties of the Core Matching Functions for Information Retrieval

D.W. Song [1] K.F. Wong [1] P.D. Bruza [2] C.H. Cheng [1]

[1] Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong

[2] Distributed Systems Technology Center, Building 78, Staff House Road
University of Queensland, Qld 4072 Australia

Abstract

Traditional benchmarking methods for information retrieval (IR) are based on experimental performance evaluation. Although the metrics precision and recall can measure the performance of a system, it cannot assess the functionality of the underlying model. Recently, a theory of "aboutness" has been studied and used for reasoning about functional of IR models. Latest research shows the functionality of an IR model is largely determined by its retrieval mechanism, i.e., the matching function; in particular, containment and overlapping (either with or without a threshold value) are core IR matching functions. The objective of this paper is to model the containment and overlapping matching functions within an aboutness-based framework, reason and analyze the inherent functionality of them from an abstract and theoretical viewpoint. Separate aboutness relations for containment, pure-overlapping (i.e. without threshold) and threshold-overlapping are defined, and the sets of properties supported by them are derived and analyzed respectively. These three relations can be used to explain the functionality supported by an IR system and their effects to its performance; and moreover, they provide the design guidelines for new IR systems.

1. Introduction

The traditional experimental evaluation of information retrieval (IR) focuses on the effectiveness of the system. Measurements in terms of recall and precision are taken as performance indicators, but they are unable to assess the functionality supported by the retrieval models.

To overcome this predicament, logic-based inductive evaluation has been proposed. It provides a uniform representation of information and its semantics, and a framework for reasoning properties of IR independent of any IR models. Most noticeable works in this area are based on "aboutness" (Bruza & Huibers 1994; Bruza & Huibers 1996; Huibers 1996; Hunter 1996; etc.) where matching is modeled by the aboutness relationship between the document and a query. Recent investigations have centered on formalizing the notion of aboutness by axiomatizing its properties (inference rules) in terms of a neutral, theoretical framework. There is yet no consensus regarding to this framework except that it is logic-based (Lalmas 1998; Lalmas & Bruza 1998; Sebastiani 1998).

Song *et al*. (1999) and Wong *et al*. (1999) proposed the *functional benchmarking* of IR models. Song et al. (1999) proposed to use an aboutness-based symbolical and axiomatical method for IR functional benchmarking, which involves a 3-dimensional scale (i.e., representation, matching function and transformation) to model the three classes of essential functionality of IR. Wong et al (1999) defined a functional benchmark suite based on the most fundamental aboutness framework proposed by Bruza (Bruza & Huibers 1994; Bruza & Huibers 1996), and a formal evaluation methodology. The proposed benchmark has been applied to evaluate various classical and logic-based IR models. The results allow us to qualitatively compare their functionality. Wong et al. (1999) led to the following fundamental observations:

- The distinction between classical and logical models is on whether they can capture semantic transformation of information, which can be modeled separately by deep (with transformation) and surface (without transformation) containment. Bruza's original framework doesn't distinguish them.
- The functionality of an IR model is largely determined by the matching function it supports. Two classes of matching function are widely used: *containment* (the document contains the content of the query) and *overlapping* (the document and query share some content). IR models using the same class of matching function show similar functionality yet differ in some expressive power (e.g., some aboutness postulates may not be applicable for some models while they can be expressed in others). These functionalities may be desirable only with respect to the corresponding matching function instead of commonsense. On the other hand, the set of postulates of Bruza's framework is not comprehensive enough to model the functionality supported by containment and overlapping matching functions.

To model the aforesaid observations, we propose a basic framework to analyze the functionality determined by overlapping and containment from an abstract viewpoint. We first define a simple representation of an information carrier. Using it as the basis, the other operators such as information composition, information containment integrating deep containment and surface containment, and preclusion are defined. The information carrier is then used as an abstract entity to study the properties of aboutness.

Instead of giving a general set of aboutness properties, which is studied as commonsense properties of aboutness in another paper (Bruza, Song & Wong 1999b), we define

two aboutness relations, namely *containment aboutness* and *overlapping aboutness,* based on the two widely used matching functions. Their properties are discussed. Each discussion is divided into two parts — definition and derived rules. The former models the corresponding matching function. Based on the definition, a set of rules representing the properties of the aboutness relation (i.e., the functionality of the corresponding matching function) can be derived.

Note that in this paper, we make the assumption that containment and overlapping (either with or without transformation of information involved) are core matching functions for aboutness decision. This can be illustrated by the following figures:

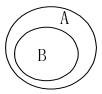

Fig-3 Containment without transformation

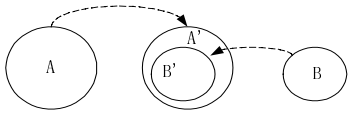

Fig-4 Containment with transformation

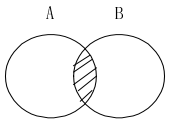

Fig-1 Overlapping without transformation

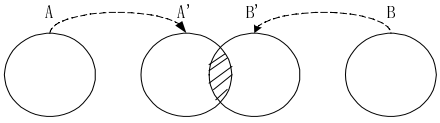

Fig-2 Overlapping with transformation

The other detailed formulas in different IR models can be considered as their variations. Many IR systems assign term weights to document and query, and calculate degree ([0, 1]) of overlapping between them. Even though different systems may use different formulas, their fundamental nature is the same. The difference is manifested in the resultant rankings. Moreover, a threshold value is always adopted to determine the relevance. Thus, in this paper, we divide the overlapping aboutness into two types: *pure-overlapping* (i.e. no threshold, or zero-threshold overlapping) and *threshold-overlapping* (i.e. the threshold value is greater than zero).

Thus, it is crucial to understand what are the properties inherent in containment, pure- and threshold- overlapping notions of aboutness. From a theoretical point of view, they can be used to verify, predict and explain the functionality of an IR system whose matching function is known to be one of them, and moreover, provide the guideline for the design of an IR system.

The rest of the paper is organized as follows. In the next section (Section 2), we define the basic concepts for the study of aboutness, e.g., information carrier, surface containment and deep containment, etc., and propose the methodology. In Section 3 and 4, containment aboutness and overlapping (including pure-overlapping and threshold-overlapping) aboutnesses are respectively defined, and their derived rules modeling the inherent functionality of these matching functions are proposed and proved. Moreover, the relationships between these rules and the effectiveness of IR system are analyzed quantitatively. Conclusion of the paper and discussion for further research are given in Section 5.

2. The Framework

Definition 0: *Basic Information Carrier*
Basic information carrier is the minimal piece of information, which cannot be divided any smaller, e.g., *salmon* and *fish*.

Definition 1: *Information Carrier*

- Let **IC** be the set of information carriers.
- Let A, B be information carriers. They are the sets of basic information carriers, e.g., A={salmon}, B={fish}.
- Information carrier is closed under $\oplus$, i.e., $A \oplus B = A \cup B$ is also an information carrier, where the symbol $\oplus$ denotes a relation namely information composition.

Note that the notions of *information carrier* and *information composition* are broader concepts than set and set union. For example, the representation of an information carrier might be a weighted or an un-weighted set of terms, a sequence (ordered set) of terms, or more complicatedly, some structured or semi-structured representation (Song et al. 1999). We do not consider a sequence of terms in this paper. Moreover, the baseline of matching between two (semi-) structured information carriers involves sets of terms. Thus, we use this definition throughout the rest of the paper. In the future, information composition will be extended to ordered set (sequence).

Definition 2: *Characterization of information carrier*
Let function c(A) be the set of characterizations of an information carrier A, including the intentional descriptions of aspects of A's content. This function models the semantics of an information carrier. For example, c({fish}) is a set of characterizations, or properties, which can identify fish; c({salmon}) is c({fish}) plus the distinct characterizations of salmon.

Definition 3: *Information preclusion ($\perp$)*
Not all information carriers can be meaningfully composed, e.g., $A \oplus B$ may be meaningless because the information carried by A contradicts the information

carried by B. This phenomenon is termed *information preclusion*, denoted by A⊥B. Information preclusion is symmetric, and its negation ($\not\perp$) is decided by close world assumption, i.e., A $\not\perp$ B iff A⊥B does not exist.

Note that to complete the framework, we introduce the preclusion operator. However, we do not really use it in this paper as it is foreign for the containment and overlapping matching functions. It will be used in the investigation of commonsense aboutness.

Definition 4: *Information Containment (→)*

Information containment is a binary relation on IC, i.e., → ⊆ IC×IC. Given two information carriers A and B, A→B iff either A⊇B (surface containment) or for A and B, c(A)⊇c(B) (deep containment).

Information containment models the intuition that the information is explicitly or implicitly nested. Explicit nesting is referred to as surface containment. For example, a document d consisting of two sections A and B (i.e., d=A⊕B, then d→A and d→B). Deep containment occurs when information containment arises at semantic level, e.g., {*salmon*}→{*fish*}. Information containment is assumed to support the following properties:

(1) Reflexivity: A→A
(2) Transitivity: A→B, B→C A→C
(3) Asymmetry: A→B doesn't imply B→A
(4) Containment-Composition: A⊕B→A; A⊕B→B;
(5) Absorption: A→B A⊕B=A;
(6) Non-conflict containment: A→B ⇔A $\not\perp$ B
(7) Containment-Preclusion (CP): A→B, B⊥C ⇔A⊥C

In the next sections, we use the above basic framework to separately define containment aboutness, pure-overlapping and threshold-overlapping aboutness and discuss their properties. Our approach is as follows:

1. Give the formal definitions of containment, pure-overlapping and threshold-overlapping aboutness.
2. Propose and prove the properties of them.
3. Quantitatively discuss the functionality of them, and the effect to performance of IR systems built on them.

3. Containment aboutness

Definition 5 *Containment Aboutness ($\models_C$)*

$\models_C \subseteq$ IC×IC is a binary relation. For two information carriers A and B, A$\models_C$B if and only if A→B.

Note that given information carriers A, B and A→B. Surface containment of A and B implies A⊇B. If it is deep containment, then c(A)⊇c(B). To simplify the representation, we use A⊇B to represent the above two cases generally.

Theorem 3.1 Containment aboutness supports Reflexivity, Containment, Transitivity, Left composition monotonicity, And, Mix and Cut.

(1) Reflexivity (R): $A \models_C A$

Proof: A→A A$\models_C$A

(2) Containment (C): $\dfrac{A \to B}{A \models_C B}$

Proof: A→B A$\models_C$B

(3) Transitivity (T): $\dfrac{A \models_C B \;\; B \models_C C}{A \models_C C}$

Proof: A$\models_C$B, B$\models_C$C A→B, B→C A→C A$\models_C$C

(4) Left composition monotonicity (LM): $\dfrac{A \models_C B}{A \oplus C \models_C B}$

Proof: A$\models_C$B A→B A⊇B A⊕C= A∪C ⊇ B
A⊕C→ B A⊕C$\models_C$B

(5) And (A): $\dfrac{A \models_C B \;\; A \models_C C}{A \models_C B \oplus C}$

Proof: A$\models_C$B, A$\models_C$C A→B, A→C A⊇B, A⊇C
A ⊇ B∪C A ⊇ B⊕C A→B⊕C A $\models_C$B⊕C.

(6) Mix (M) can be derived trivially from LM
$\dfrac{A \models_C C \;\; B \models_C C}{A \oplus B \models_C C}$

(7) Cut (C): $\dfrac{A \oplus B \models_C C \;\; A \models_C B}{A \models_C C}$

Proof: A$\models_C$B A⊇B A⊕B=A∪B=A
Then, A⊕B $\models_C$C A $\models_C$C.

Reflexivity states that an information carrier is about itself. From an IR perspective reflexivity seems to be a reasonable property as we expect a document to be retrieved if it is itself the query. *Left composition monotonicity* (LM) means once an aboutness relation has been established, any expansion (if it is applicable) on the left side (i.e., documents) cannot break it. Note that *right composition monotonicity* (RM) does not hold, which means the expansion of the query may not preserve the aboutness relation. The monotonicity normally improves recall at the cost of precision. However, for containment aboutness, *And*, a cautious form of RM, holds instead of RM. This can help promote precision. Moreover, containment aboutness is *transitive*. This can be explained by the rule *containment,* which states that an information carrier is about the information it contains, and the transitivity property of information containment relation. On the surface this seems reasonable. However, consider the following containment chain: Tweety → penguin → bird → animal → living thing → entity. Although "entity" is nested informationally within "Tweety", it would be unnatural to view "Tweety" being about an "entity". It seems more natural to state that "Tweety" is about a "bird". In other words, when traversing the information containment relation, there may be a point where the aboutness relationship falls apart. Brooks (1995) conducted a phenomenological study that reflects this

observation. It was found that the distance from relevance to non-relevance is approximately three steps when traversing the "broader than" relationship of the thesaurus (i.e., deep containment in our terminology). The result of Brook's study suggests a refinement of Containment (C) property: $\frac{A \xrightarrow{2} B}{A \models_C B}$

This specifies that the aboutness relationship remains preserved within 2 steps along the information containment relationship.

4. Overlapping Aboutness

4.1 Pure-Overlapping Aboutness

Definition 6 *Pure-Overlapping (PO) Aboutness* ($\models_{PO}$)

$\models_{PO} \subseteq IC \times IC$ is a binary relation. For two information carriers A and B, $A \models_{PO} B$ iff $\exists C \in IC \mid A \rightarrow C \wedge B \rightarrow C$.

Overlap between information carriers A and B is modeled by an information carrier C which is contained (i.e., shared) by both A and B.

Note that given information carriers A, B, C, $A \rightarrow C$ and $B \rightarrow C$, we use $A \sqsupseteq C$ and $B \sqsupseteq C$ to generalize the representations of surface and deep containment.

Theorem 4.1 Pure-overlapping aboutness supports Reflexivity, Containment, Symmetry, Left composition monotonicity, Right composition monotonicty, And and Mix.

(1) *Reflexivity (R)*: $A \models_{PO} A$

Proof: $A \rightarrow A \quad A \models_{PO} A$

(2) *Containment (C)*: $\frac{A \rightarrow B}{A \models_{PO} B}$

Proof: $A \rightarrow B, B \rightarrow B \quad A \models_{PO} B$

(3) *Symmetry (S)*: $\frac{A \models_{PO} B}{B \models_{PO} A}$

Proof: $A \models_{PO} B \quad \exists C \in IC \mid A \rightarrow C \wedge B \rightarrow C \quad B \models_{PO} A$

(4) *Left composition monotonicity (LM)*: $\frac{A \models_{PO} B}{A \oplus C \models_{PO} B}$

Proof: $A \models_{PO} B \quad \exists D \in IC \mid A \rightarrow D \wedge B \rightarrow D \quad A \sqsupseteq D \wedge B \rightarrow D$

$A \oplus C = A \cup C \sqsupseteq D \wedge B \rightarrow D \quad A \oplus C \rightarrow D \wedge B \rightarrow D \quad A \oplus C \models_{PO} B.$

(5) *Right composition monotonicity (RM)*: $\frac{A \models_{PO} B}{A \models_{PO} B \oplus C}$

Proof: using symmetry and LM.

(6) *And (A)* can be derived trivially from *RM*.
$\frac{A \models_{PO} B \quad A \models_{PO} C}{A \models_{PO} B \oplus C}$

(7) *Mix (M)* can be derived trivially from *LM*.
$\frac{A \models_{PO} C \quad B \models_{PO} C}{A \oplus B \models_{PO} C}$

Overlap is a common intuition with respect to aboutness. Almost all information retrieval systems function according to this assumption. For example, the vector space model measures the overlap between a query and document vector by computing the cosine of the angle between the respective vectors. In the above discussion, we assume the case that d is about q if *cos(d, q)>0*. Besides R, C, LM, and And, overlapping aboutness also supports Symmetry, RM (And is a special case of RM) and Mix (trivially derived from LM). At first sight, symmetry seems to be an acceptable property. There are, however, cases where it is unsound. For example, the movie "Saving Private Ryan" is about "saving an American soldier during WW2". It seems unnatural to say that the latter carrier is about the movie. An information retrieval system supporting RM and LM cannot "lose" aboutness relationships. This means in practice that *d* can never be removed from the result set, irrespective of any expansions of document and query. For example, consider *surfing⊕Hawaii* $\models_{PO}$ *surfing*. RM permits *surfing⊕Hawaii* $\models_{PO}$ *surfing⊕Australia*, which has the natural language interpretation: "surfing in Hawaii" is about "surfing in Australia"!

Some current IR systems circumvent this by employing threshold values. The vector space model operates according to the following aboutness definition: $d \models q \Leftrightarrow \cos(\vec{d}, \vec{q}) \geq \partial$. Then, in the next sub-section, threshold-overlapping is discussed.

4.2 Threshold-overlapping Aboutness ($\models_{TO}$)

Definition 7 *Threshold-overlapping (TO) aboutness* ($\models_{TO}$) Let *A* and *B* be n-dimensional vectors, where n is the cardinality of vocabulary T. Then,

$A \models_{TO} B$ *iff* $\cos(A, B) > \partial$ where $\partial > 0$.

Theorem 4.2 Threshold-Overlapping aboutness supports Reflexivity and Symmetry.

Proof:

- Reflexivity (R) is implied as cos(A, A)=1.
- Containment (C) is not supported. Consider the case that A contains B but A is much larger than B, then cos (A, B) may be a very small value, even less than ∂.
- Symmetry (S) is supported, as given $\cos(A,B) > \partial$, $\cos(B,A) = \cos(A,B) > \partial$.
- To verify Left compositional Monotonicity (LM), it must be shown that: for an arbitrary x and under the premise $\cos(A,B) > \partial$, $\cos(A \oplus x, B) > \partial$. 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 $\cos(A \oplus x, B) \leq \partial$. Hence, TO does not support LM. (The argument that TO does not support RM follows a similar line).
- And (A) is not supported. Consider the case where $\cos(A,B) > \partial$, $\cos(A,C) > \partial$, $A \cap B = A \cap C = B \cap C = D$. In

this case, cos(A, B⊕C) is not necessary to be still greater than ∂. Similarly to And, Mix (M) is supported.

This definition allows the retraction of aboutness relationships whenever the document d or query *q* is expanded, and the cosine of the respective vectors drops below ∂. The document is then no longer retrieved (i.e. the original aboutness relationship $d \models q$ had been retracted). Although this definition realizes desirable nonmonotonic behavior with respect to aboutness, it is not "clean" from a theoretical point of view because the value ∂ is *not* determined by the retrieval model, but is extraneous to it. (In practice, the value is determined experimentally).

5. Conclusion and Discussion

Summary: We have defined a logical framework for aboutness and used it to model containment, pure and threshold overlappings, which are the most common IR matching functions, and to reason about their functionality. This lays down foundation for theoretical information retrieval research. More specifically,

(a) Given an IR system and its matching function (either containment, pure- or threshold overlapping), the set of functionality of this system can be predicted and verified from a theoretical viewpoint according to the properties of the corresponding aboutness relation;

(b) If an IR system is being designed to support one of the core matching functions, the corresponding set of aboutness rules can provide a guideline for the design, i.e., what functionality the system should support.

Discussion: The sets of properties of containment and overlapping aboutness are sound with respect to their definitions. However, this does not mean that all of the rules supported by containment, pure- and threshold-overlapping aboutness are sound from a commonsense perspective. Some individual rules may have negative impact to information retrieval performance as we have discussed in Section 3 and 4. Thus, it is also necessary to study the commonsense properties independent of any given matching functions. Aboutness is a subjective concept in some sense. However, we believe there is a core set of properties agreeable by different users and this core agreement can be treated formally. Our hypothesis is that a better understanding of aboutness (from a fundamental point of view) will lead to more effective aboutness reasoning system for information retrieval. The objective of our future work is to study commonsense aboutness within a richer logical framework. It could be applied to the following fields:

- *IR functional benchmarking*. The containment, overlapping and commonsense aboutnesses together can be used to benchmark the functionality of IR models.
- *Intelligent agents.* The commonsense aboutness can form the basis of an aboutness reasoning system, which will serve as intelligent agent for relevance judgement in both the retrieval/filtering and query expansion processes.

Remark: The classical probabilistic model is not investigated in this paper. It is difficult to be mapped into aboutness framework, as it does not directly deal with the aboutness relationship between a document and a query, but the probability of relevance. This is still an open question to be worked out.

References

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

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.* pp. 112-121.

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

Bruza, P.D., Song, D.W., & Wong, K.F. (1999a). Fundamental properties of aboutness. In *Proceedings of the Twenty-Second Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, Berkeley, USA, 1999.

Bruza, P.D., Song, D.W., & Wong, K.F. (1999b). Aboutness from a commonsense perspective. Submitted to *Journal of American Society for Information Science*.

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

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

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.

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 Conference.* pp. 389-393.

Wong, K.F., Song, D.W., Bruza, P.D. & Cheng, C.H. (1999) Application of aboutness to functional benchmarking in information retrieval. Submitted to *ACM Transactions on Information Systems*.