Towards A Commonsense Aboutness Theory for Information Retrieval Modeling

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

Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong, Shatin, N.T. Hong Kong
{dwsong, kfwong, chcheng, wjli}@se.cuhk.edu.hk

[*] Distributed Systems Technology Center
University of Queensland, Qld 4072 Australia
bruza@dstc.edu.au

ABSTRACT

Information retrieval (IR) can be viewed as a process to determine the "aboutness", or sometimes "relevance", relationship between information carriers (e.g. document and query). Thus, the concept of aboutness lies at the heart of IR. A better understanding of aboutness would lead to more effective IR systems. In this paper, we give a review of the status of current research on aboutness. It is shown that important outcomes have been obtained. However, there are still several problems to be further worked out, including *soundness of aboutness properties, ordering of aboutness inferences* and *conservative monotonicity of aboutness relation.* We discuss these problems in detail and propose a strategy for further investigation of aboutness, i.e. the fundamental and sound properties of aboutness from commonsense perspective. The goal of our work is to define a generic commonsense aboutness theory for IR modeling. The applications of aboutness theory are also highlighted.

Keywords: *Information Systems Development; Intelligent Information retrieval.*

1. INTRODUCTION

Today information can be globally shared via Internet and can be accessible from any part of the world. However, the increasing complexity and growing information volume of the WWW render information retrieval (IR) increasingly difficult. This urges the need of more effective information retrieval systems. In the early years IR was conducted experimentally and by trial and error. Since the operational semantics of IR is not clearly defined, traditional approach was not so effective in understanding the issues of effectiveness and relevance. To overcome this, researchers endeavor to model IR process. Such models formally define the essential semantics of IR in terms of mathematical or logical accounts respectively. IR is very subtle and it involves some forms of information transformation. The classical mathematical models, e.g. vector space model, cannot capture these subtleties precisely. Recently some logical IR models, e.g. situation theory based and possible world based models, etc., have been proposed to improve this situation. These mathematical or logical models form the basis of a number of practical IR systems. Nevertheless, existing IR models cannot clearly explain the performance of an IR system, which is commonly measured by precision and recall. The properties of an IR system are determined by its underlying model. But it is still unclear how these properties could be explicitly expressed? Are they sound? Are they complete? How do they affect precision and recall? From a more general and abstract (i.e. model-independent) perspective, what are the sound (see Section 3 for a detailed discussion on soundness) properties of IR? These questions are crucial to further improve the effectiveness of an IR system. Thus, being a complement of current experimental IR research, there is a pressing need of a theoretical framework to study the fundamental IR properties.

IR can be viewed as a process determining the "*aboutness*", or sometimes "relevance", relationship between information carriers (e.g. document-query, document-document, or query-query). Aboutness plays a prominent role in 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 cited experiments wherein the agreement between subjects judging documents with respect to a query was around 60% [8]. This suggests that aboutness have an inter-subjective core of agreement, which in our opinion is amenable to formal treatment. Thus, it is possible to define a theory of aboutness independent of any given IR models to formally model this core of agreement. We have already conducted some research in this area, and shown that it is promising. However, there are still some important issues, i.e., *soundness of aboutness properties, ordering aboutness inferences* and *conservative monotonicity of aboutness relation,* to be further solved. We believe that a further investigation of a general aboutness theory for IR will lay down a significant theoretical foundation for IR and in turn lead to more effective IR systems. The primary objective of this paper is to give a review of existing research on aboutness (Section 2), discuss the aforesaid three issues in detail (Section 3) and propose a strategy for further studying aboutness theory for IR (Section 4). Finally, Section 5 concludes the paper.

2. A REVIEW OF EXISTING RESEARCH ON ABOUTNESS

Researches on aboutness have appeared sporadically in the literature for more than two decades. Cooper proposed a definition of "relevance" (i.e., aboutness) between a piece of stored information and the information need of an IR system user [9]. The notion of relevance is divided into "*logical relevance*", alias "topic-appropriateness", and "*utility*", which has to do with the ultimate usefulness of the piece of information to the user. It is only *logical relevance* which is explicated in Cooper's work. The suggested definition explicates relevance as "*logical consequence*", or "*logical implication*", which is founded on the assumption that the data stored and the user's request are expressed in the form of well-formed sentences. Even though Lalmas showed that the classical logical consequence is too restrictive to model the IR process [17, 18], Cooper's work sheds the light to use logic to formally define the notion relevance couched in IR theoretical terms. Van Rijsbergen proposed a non-classical logical approach, namely *logical uncertainty principle* [23] – "*Given any two sentences x and y; a measure of the uncertainty of $y \to x$ relative to a given data set is determined by the minimal extent to which we have to add information to the data set, to establish the truth of $y \to x$*". It forms the basis of many logic-based IR models today.

Hutchins provided a thoughtful early study of the topic [15]. This account attempts to define a notion of aboutness in terms of a combination of linguistic and discourse analyses of a text. He introduces the thematic and sementic progressions of a text. 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. It is argued that for the purpose of indexing, the "aboutness" of document is to be found among the presuppositions of authors concerning the knowledge of their potential readers.

Maron tackled aboutness by relating it to a probability of satisfaction [19]. 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. A document D is O-about (i.e. objective 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). 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. developed a qualitative framework for relevance using possibility theory [10]).

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 [3, 4, 13, 14, 20] culminating in an axiomatic theory of information retrieval developed by Huibers [13]. The use of aboutness postulates as the basis of functional benchmarking (i.e., an inductive, rather than experimental, evaluation) of IR models [25, 26] and information discovery [22] have been shown promising. Broadly speaking, all these works on aboutness view IR as a reasoning process, determining aboutness between two information carriers. 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. Existing aboutness frameworks, however, suffer from model-dependent as well as from the lack of expressive power [26]. The main reason is the lack of holistic and independent view of aboutness and its properties. Up to now, there is yet no consensus regarding this framework except that it should be logic-based [17, 18]. Although a number of aboutness properties are commonly discussed in the literature, e.g. reflexivity, transitivity, symmetry, and left (right) monotonicity, etc., there is thus far no agreement on a core set of aboutness postulates. The disagreement stems partially from the framework chosen to formalize aboutness. Once the framework has been fixed, certain aboutness properties are implied by it. Moreover, some authors tried to propose aboutness properties as completely as possible. This is necessary to model the functionality of different IR models. However, some properties, e.g., transitivity and symmetry, etc., may be sound only within certain IR models, and some of them may lead to negative effects to the effectiveness of IR system [6].

In order to overcome the aforesaid problems, our recent researches [6, 7] proposed the concept of commonsense aboutness. Viewing aboutness (and its dual, non-aboutness) from a fundamental and commonsense perspective leads to a set of reasonable (i.e. sound) properties of aboutness and non-aboutness, which is generic, i.e., independent of any IR model, and is examined within an information based, abstract framework. Nevertheless, the following important issues are still left untackled: *soundness of aboutness properties, ordering aboutness inferences* and *conservative monotonicity of aboutness relation.*

3. CONSERVATIVE MONOTONICITY, ORDERING OF ABOUTNESS INFERENCES AND SOUNDNESS

3.1 Conservative monotonicity

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 added (composed) to either A or B. On the other hand, *non-monotonicity* allows for the aboutness relation to be broken under certain circumstances, thus realizing a more sensitive retrieval system. Current systems mimic non-monotonic behavior via the use of threshold values for inclusion in the document ranking. This is in some sense undesirable as the threshold value is *not* determined by the retrieval model in question, but is external to the model.

Our recent case study [7] based on monotonicity shows that many current IR systems are either monotonic (e.g. boolean and non-thresholded vector space models) or non-monotonic (e.g. thresholded vector space and probabilistic models). An interesting class of IR models, namely those that are conservatively monotonic, does not exist. *Conservative monotonicity* can be considered as a special form of nonmonotonicity. It allows aboutness relation to be preserved under certain guarded conditions. The conservatively monotonic models are interesting because there are indications that IR is conservatively monotonic, rather than purely non-monotonic. Take query expansion as an example. 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. Our previous work [26] 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 behavior can be characterized by symbolic rules making them more transparent than the latter. We believe that the right conservative form of monotonicity means the optimal trade-off between precision and recall.

3.2 Ordering of aboutness inferences

We studied the consistency of the proposed aboutness and non-aboutness properties, and found that some forms of inconsistency were unresolvable at this stage. However they seem resolvable. Aboutness relationships could be ordered to relfect the intuition that some aboutness relationships are "stronger" than others. 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 A |≠ B is stronger than A |= B. This provides a basis for prefering the former. Thus inconsistency can be resolved. Only in the case where both inferences are strong more information would be 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.

3.3 Soundness

We have claimed that the proposed aboutness properties should be sound (reasonable). When investigating the issue of soundness with respect to (non-) aboutness rules, a frame of reference must be defined within which soundness can be verified. In classical logic, the frame of reference is a model. However, due to the lack of an underlying model theory for IR, this approach cannot be taken. It is also arguable whether such a theory exists. One of the complications here is that the "retrieval is inference" view [9, 23] cannot be considered independent of the user. One existing approach for soundness evaluation with respect to aboutness is that the soundness of an aboutness proof system is investigated in the context of an underlying aboutness definition [13]. Although valid from a formal point of view, it does not consider the user. We argue that aboutness inference is "psychologistic" in nature. As a consequence, aboutness reasoning cannot be studied independently from the agents involved. With respect to IR, sound aboutness rules can only be established via cognitive studies. It is interesting to draw an analogy 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 [11, 12, 21]. This not only demonstrates that nonmonotonic reasoning is psychologistic in nature, but also that soundness must ultimately be grounded from a cognitive perspective. We advocate the same to aboutness inference. More studies similar to Brooks' work [2] are needed to gain a clearer understanding of how users perceive aboutness, and how they reason about it.

4. OUR PROPOSAL FOR FUTURE RESEARCH

In our opinion, aboutness theory is most suitable to model the nature of IR (see Section 1). However, as introduced in last section (Section 2), even though many researchers, including ourselves, have done a lot of work in this area, the existing aboutness frameworks insufficient. In particular, the aforesaid three problems (i.e., *soundness*, *ordering inference rules*, and *conservative monotonicity*) need to be further studied.

We propose to better understand the nature of IR (specifically, fundamental and sound properties of aboutness). Based on the understanding, we intend to define a theory of aboutness. Our research shall focus on the issues: *soundness*, *inference rules ordering* and *conservative monotonicity.*

4.1 A Theory of Commonsense Aboutness

4.1.1 Overall Strategy

A symbolic and axiomatic method will be used to define the aboutness theory in an abstract and information-based manner. We believe an axiomatic and symbolic method can model aboutness effectively. In our opinion aboutness is an ordered binary relation. Its properties can be modeled by a set of inference rules. Then the aboutness decisions could be reasoned by those inference rules. Moreover, many IR systems use numerical approaches to produce a list of ranked documents. The important thing is not the ranking values, but the ordering of the documents in the list. Similarly, the weight of index term itself is not important. The importance lies on its ranking. Thus, by formally defining the orderings of aboutness decisions and inference rules, it is possible to define the framework symbolically and reason aboutness decision axiomatically.

4.1.2 The Core of Commonsense Aboutness Theory

(1) Conservative monotonicity, ordering inference rules and soundness

It has been shown that conservative monotonicity can be viewed as the compromise between monotonicity and non-monotonicity, and it allows aboutness relation to be preserved only under certain guarded conditions. We argue that IR is a conservatively monotonic, rather than purely monotonic or non-monotonic process. Therefore, a set of inference rules to model the conservative monotonicity lies in the core of our aboutness theory. An example of conservative monotonicity rules is *Qualified Left Compositional Monotonicity (QLM),* which is proposed in our previous work: for information carriers A, B, and C, if A is about (|=) B and B does not preclude ($\perp$) C, then the composition ($\oplus$) of A and C is also about B. This rule[1] is formulated as below (For detailed semantics of these operators, refer to [7]):

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

To illustrate it, let's look at the famous Tweety-bird example in logic: Tweety (t) is a bird (b); Tweety is a penguin (p); penguins are birds; birds are about flying (f) (b|=f); penguins do not fly (p⊥f); Tweety cannot fly (t⊥f). The monotonicity allows b⊕t|=f (Tweety, which is a bird, is about flying) to be inferred from b|=f (A bird is about flying). This is an unsound inference as Tweety is a penguin which cannot fly. QLM prevents this via the qualifying preclusion t⊥f.

There are several other forms of inference rules expressing conservative monotonicity discussed in the literatures, e.g., *Cautious Left (Right) Monotonicity, And, Mix, Rational Left (Right) Monotonicity,* etc. There is yet no agreement, but it is evident that conservative monotonicity is at the core of aboutness theory. We will further investigate this issue to find the most appropriate set of inference rules to model it.

Aboutness should be an ordered relation. The ordering of aboutness decisions is derived from the ordering of inference rules involved. The ordering of aboutness should be explicitly defined. It is necessary to investigate the relative importance of different inference rules according to their degrees of contribution to aboutness decision, and define a mathematical model to represent the inference rules ordering. This model should also formulate how to calculate the ordering of final aboutness decisions involving multiple rules.

Soundness of aboutness inference rules is "psychologistic" in nature. With respect to IR, sound aboutness rules can only be established via cognitive studies. Therefore, we will conduct series of experiments (in forms of survey and case study) to investigate the soundness of the proposed aboutness rules and their orderings. We will restrict the subjects of the investigation to several specific areas, e.g., computer science, financial news, etc. Also, we shall classify the users into three levels of populations: *1. Experts on the subject; 2. People who have medium knowledge on the subject; 3. People who know little about the subject.*

Here, we agree with Nie in that there are strong connection and interaction between cognitive understanding and formal modeling of aboutness. A better understanding leads to a more accurate formal modeling. On the other hand, formal modeling enables us to express and study our cognitive understanding on a formal basis; thus it helps us to see inside the still-empirical notion of aboutness, to detect and handle our misunderstandings and the inconsistencies in our observations [20].

(2) Other relations

Relations for modeling other IR aspects, e.g. "similarity" which can model document clustering, will also be defined, and the relationship between these relations and aboutness will be studied.

(3) Structure of aboutness inferences

Two levels of aboutness inference - micro and macro structures will be defined. The micro structure involves fine granularity (i.e., index terms); and the macro structure involves the coarse level, e.g. paragraphs. We will also study the close interrelationship between these two structures.

(4) Theorems and propositions

We will propose and formally prove several theorems and propositions which reflect the properties of the proposed aboutness theory. Furthermore, we will evaluate the functionality of several classical and logical IR models using the aboutness theory.

4.2 Development and Prototyping

The commonsense aboutness theory is founded upon notions such as information containment, 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. Aboutness and preclusion relationships can be derived via relevance feedback [1, 5]. 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 and more directly. More sensitive IR systems would then result; in particular those which are conservatively monotonic with respect to composition. The lack of such systems currently is attributed in part to the difficulty to effectively "operationalize" those notions.

We will implement a prototype system to operationalize the aforesaid concepts. This will be the foundation for applying aboutness theory directly large-scale real IR systems. Based on aboutness theory, such IR systems would be more effective.

5. CONCLUSION

A better understanding of aboutness will lead to more effective and more intelligent IR systems. The proposed aboutness theory could be directly applied to the following areas:

- *IR functional benchmarking*. The traditional experimental methods are good at evaluating the performance of a system, but they are 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.

[1] We represent aboutness rules in form of condition/ consequence. This representation is widely used in logic, e.g. the preferential reasoning system [16].

- *Intelligent information agents*, including *(1) Information retrieval and filtering; (2) Query expansion; (3) Document clustering;* etc. These are widely used techniques in IR, and could be described in terms of an aboutness reasoning system. The sound properties of aboutness would form the basis of the inference rules guiding the above process.

Remark: *Completeness of the aboutness theory***.** In our recent research we have found out that the aboutness is a fundamentally incomplete. The completeness can be established by introducing certain unsound rules. This is consistent with the interaction between precision and recall. We will mainly focus on the soundness of aboutness rules in the future.

References:

[1] Amati, G. & Georgatos, K. (1996). "Relevance as deduction: a logical view of information retrieval." In *Proceedings of the Second Workshop on Information Retrieval, Uncertainty and Logic (WIRUL'96)*, 21-26.

[2] 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.

[3] 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.

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

[5] Bruza, P.D. & van Linder, B. (1998). Preferential models of query by navigation. In *Information Retrieval: Uncertainty and Logics.* The Kluwer international series on Information Retrieval.

[6] 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)*, 1999.

[7] Bruza, P.D., Song, D.W., & Wong, K.F. (1999b). Aboutness in commonsense perspective. Accepted by *Journal of American Society for Information Science*.

[8] 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.

[9] Cooper, W.S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*, 7, pp. 19-37, 1971.

[10] 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.

[11] 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.

[12] 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.

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

[14] Hunter, A. (1996). 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.

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

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

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

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

[19] Maron, M.E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, 28 (1): 38-43.

[20] Nie, J., Brisebois, M., & Lepage, F. (1995). Information retrieval as counterfactual. *The Computer Journal 38*, 8, pp.643-657.

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

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

[23] van Rijsbergen, C.J. (1986) A non-classical logic for Information Retrieval. *The Computer Journal*, *29, 6, 1986*.

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

[25] 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*, pp. 389-393..

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