Towards Functional Benchmarking of Information Retrieval Models

From: Proceedings of the Twelfth International FLAIRS Conference. Copyright © 1999, AAAI (www.aaai.org). All rights reserved.

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

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

[2] School of Information System
Queensland University of Technology, Brisbane, QLD, Australia
bruza@icis.qut.edu.au

Abstract

To evaluate the effectiveness of information retrieval (IR) system, empirical methods (performance benchmarking) are widely used. Although they are useful to evaluate the performance of a system, they are unable to assess its underlying functionality. Recently researchers use logical approach to model IR properties so that inductive evaluation of IR could be performed. This approach is known as functional benchmarking. The *aboutness* framework has been used for this purpose. Aboutness based functional benchmarking is promising but yet ineffective due to the lack of a holistic view of the evaluation process. To overcome the ineffectiveness of the existing aboutness frameworks, we apply the idea of reasoning about function to IR and introduce a new strategy for IR functional benchmarking, which involves the application of a symbolic and axiomatic method to reason about IR functionality. This strategy consists of three parts, namely *definition, modeling* and *evaluation.* To facilitate the unified logical representation of an IR model in definition part and effective reasoning in the modeling part, a three-dimensional scale, which can identify the classes of essential IR functionality (representation, matching function, and transformation) is proposed in this paper. With this scale, the deficiencies of the existing aboutness frameworks could be overcome.

1. Introduction

The evaluation of information retrieval (IR) systems centers on effectiveness. Traditionally, IR systems are evaluated and compared experimentally. The well-known evaluation measurements are precision and recall. Experimental retrieval evaluations are always conducted in laboratory environment, and based on test collections consisting of a corpus, a query set, and sets of relevance judgement (one for each query, a singleton set). Although many important results have been obtained, there are some criticisms concerning the subjectivity in relevance judgement and limitation in corpus construction. Moreover, the experimental methods can not explain why an IR system shows such performance. Thus, an objective evaluation approach is necessary. It should be independent of any given IR model, and be able to predict the underlying functionality of an IR system. In this way, the upper and lower bounds of the systems effectiveness could be approximated.

To fill this gap, logic based inductive evaluation has been discussed by a number of researchers. It is shown by (Lalmas and Bruza 1998) that "*the logic-based approach was launched to provide a richer and more uniform representation of information and its semantics*". Also, It "*provides a framework for studying IR theory independent of the formalisms and idiosyncracies of any given IR model.*" The properties of IR can be modeled by such a logical framework, through which the IR models could be evaluated and compared inductively. Most noticeable works in this area are based on "aboutness" (Bruza and Huibers 1994; Bruza and Huibers 1996; Huibers 1996; Hunter 1996; etc.) where the IR process is assumed to be driven by determining aboutness between two information carriers (i.e., document and query). Recent investigations have centered on formalizing the notion of aboutness by axiomatizing its properties in terms of a neutral, theoretical framework. This framework is important as it allows aboutness to be studied independently. There is as yet no consensus in aboutness except that it is logic-based (Lalmas 1998; Lalmas and Bruza 1998; Sebastiani 1998).

Recently, (Wong *et al.* 1998) proposed "*functional benchmarking*" of IR models. They argued that the traditional experimental approaches could provide useful performance indicators but unable to reflect the functionality of an IR system. This could be overcome by inductive evaluation based on aboutness. The former could serve as *performance benchmark* and the latter as *functional benchmark* for IR. The two would be complementary to each other. Wong et al. first adopted the most fundamental aboutness framework proposed by Bruza[1] (Bruza and Huibers 1994; Bruza and Huibers 1996) as the initial basis, applied it to benchmark the

[1] For two information carriers A and B, information containment ($A \rightarrow B$) means that B is informationally contained in A. The composition of information is denoted by $A \oplus B$, which contains the information born by both A and B. $A|=B$ means A is about B. The properties of aboutness are modeled by a set of postulates.

functionality of various typical IR models and then reassessed the effectiveness of Bruza's framework. Results have shown that the application of Bruza's aboutness framework was feasible to functional benchmarking. However, some deficiencies were identified, including:

(1) The framework could not distinguish between different types of information.

(2) Some concepts, e.g., information composition, were difficult to employ as their rules change according to information carriers.

(3) The framework could not distinguish between surface and deep containment.

(4) The set of aboutness properties was ineffective.

These deficiencies are mainly brought about by the lack of a holistic view of inductive evaluation strategy. The existing frameworks attempt to study the properties of aboutness, but they are either lack of generality, i.e., model-independence, or not enough to cover the essential functionality of IR. Moreover, they are ineffective as they can only reveal the properties of an IR model and may give qualitative interpretations separately on some certain aspects (e.g., precision, recall), but do not provide a formal method, i.e., an evaluation function, to compare the overall effectiveness of different IR models independently.

To overcome the deficiencies mentioned above, we apply the idea of reasoning about function (Chittaro and Kumar 1998; Kumar 1994; Sticklen and Mcdowell 1995; Winsor and Maccallum 1994) to IR, and introduce a holistic strategy for studying functional benchmarking of IR. This strategy is based on the relationships among the purpose, functionality and behavior of an IR model. It involves three parts, namely *definition*, *modeling* and *evaluation*. To achieve the benchmarking strategy, a 3-dimensional scale identifying the essential functionality of IR is proposed. However, the classical theory to define functionality, which is state or flow-based, is developed mainly for applications such as diagnosis and design, etc., and is unsuitable to model IR. The reason is that IR is actually an aboutness decision process between each document and query pair, so that its nature is quite different. In this paper, we introduce a logic-based symbolic and axiomatic method to formalize functional reasoning for IR.

The rest of the paper is organized as follows. In the next section, we describe our strategy for functional benchmarking. A functional model of IR is proposed in Section 3. A three-dimensional scale for classification of essential IR functionality is proposed in Section 4. In Section 5, conclusion of the paper and discussion for further research are given.

2. Our Strategy

Our strategy for functional benchmarking of IR is shown below:

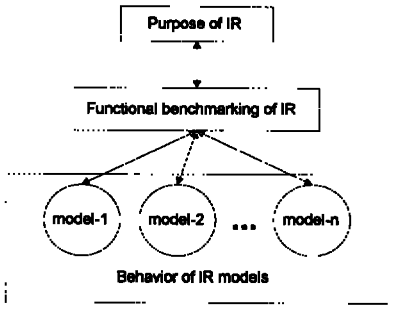

Fig-1 Architecture of functional benchmarking of IR

To apply the idea of reasoning about function to IR, it is essential to clearly identify the relationships between purpose, functionality, and behavior of IR. We consider the purpose of information retrieval as a performance issue. An IR system should retrieve as many as and as precise as possible the relevant documents with respect to a user's request, i.e., the highest effectiveness (ideally, 100% precision-recall). Of course, the judgement of relevance is subjective with respect to different users. This purpose determines what's in the core of IR, i.e., the essential functionality of IR, and functional benchmarking reveals how IR achieves the purpose. The word "*essential functionality*" here means it should reflect the nature of IR and should be the most important factors, which affect the behavior of an IR model, i.e., the effectiveness of an IR system built on this model. Imagine an IR system as a black box. Performance benchmarking evaluates and compares various IR systems through their output behaviors, without understanding their internal mechanisms. Functional benchmarking intends to open up the black box, and examines the essential functionality of an IR model underlying the system. It finds out why the model shows such behavior, and predicts as well as evaluates the model's behavior independently according to the functionality it supports.

Based on the above discussion, we propose a strategy for functional IR benchmarking. This strategy involves three parts: *definition*, *modeling*, and *evaluation*. In the *definition* part, information carriers and the corresponding operators are defined. The essential functionality of IR can be modeled as aboutness properties by a set of axioms and inference rules in *modeling* part. We believe an axiomatic and symbolic method can model IR functionality, because aboutness is an ordered binary relation, despite many IR systems use numerical approaches to produce a list of ranked documents. Actually, the ranking value of 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, it is possible to define the framework symbolically and reason the functionality of IR axiomatically. In the *evaluation* part, an *evaluation function* reflecting the relationship between the functionality and the effectiveness of IR should be defined for comparison of the IR models inductively and independently according to the sets of functionality they separately support. However, none of the existing aboutness frameworks addressed this issue systematically. We believe without this function, the functional benchmark for IR is incomplete.

Thus, in a functional benchmarking exercise, the IR model concerned would be mapped into a unified logical representation in the first part. In the modeling part, the functionality of the IR model will be inductively analyzed using a set of symbolic and axiomatic rules. The analyzed results will then be evaluated in the last part.

3. A Functional Model of IR

The IR process is depicted in Fig-2.

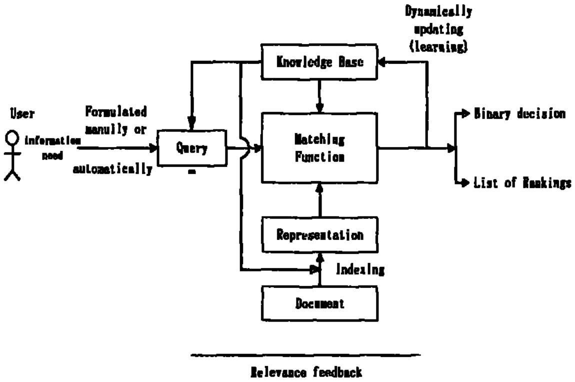

Fig-2 Architecture of IR

Documents in a collection are indexed into internal representations. Independently, user's information need is formulated in form of a query either manually or automatically. Each document (internal representation) is matched against the query based on the matching function. The output of the matching process is either the binary decision of relevance or a list of ranked documents with the degree of relevance. When the user is unsatisfied with the original results, a relevance feedback process may be conducted to reformulate the query or reweigh the index terms of the query. A knowledge base may be constructed either manually or systematically to provide supporting domain knowledge to the indexing, matching and query expansion. Further, the knowledge base could be dynamically updated through learning.

Then, we propose a functional model (see Fig-3) to identify the core, i.e., the essential functionality, of IR.

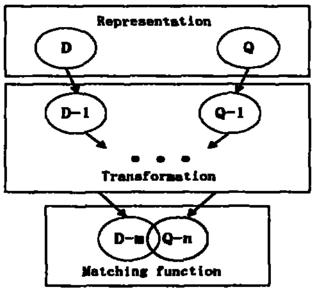

Fig-3 Functional model of IR

The underlying idea of the functional model is that IR could be considered as an uncertain reasoning process. Uncertainty exists in three aspects:

1. Uncertainty of the representations of document and user's information need.
2. Uncertainty of other supporting knowledge.
3. Uncertainty of the inference process.

As shown in the functional model, information representation, information transformation and matching function play the most important roles in coping with the above uncertainty.

Firstly, the representations of the document and query are basis of the other two, and closely related to the first aspect of uncertainty. Secondly, to handle uncertainty, van Rijsbergen proposed the Logical Uncertainty Principle (van Rijsbergen 1986):

> *"Given any two sentences x and y; a measure of the uncertainty of y→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→x."*

Later, some researchers proposed two variations of van Rijsbergen's Principle, which are separately based on the extent of the representations of the document and query. All of these extensions are transformation processes.

After a series of transformation, a matching function is used to match the (transformed) document with the (transformed) query in order to determine the aboutness relation between them. Different choices of the matching function could lead to different performance.

4. Essential Functionality of IR

According to above proposed functional model, the following 3-dimensional scale (Fig-4) identifies these three classes of essential IR functionality.

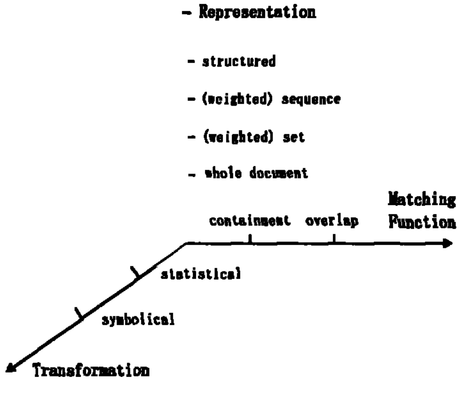

Fig-4 Classes of essential IR functionality

4.1 Representation

This dimension models the representation of both the document and the query within an IR model. The simplest representation is full texts. A most commonly used representation is a set of (weighted) index terms (keywords, key phrases, etc.). Furthermore, some applications may employ an ordered set where the index terms are ordered. For example, user may submit a query requiring "Chinese" and "University" where "Chinese" should occur before "University". Another type of representation is (semi-) structured representation, e.g., the structure of the document representation could be defined according to SGML. A document may be represented as a structure including section, paragraph, index terms, and background information such as author, etc. Experiment has indicated representation is an important role affecting the effectiveness of IR, e.g., weighted set could lead to higher precision-recall than a binary set of index terms.

Based on this dimension, the first two deficiencies of Bruza's framework could be solved by defining an inner structure of information carrier. Different types of information could be identified and the operators of the information carrier could be formally defined through this unified inner structure. At the same time, information carrier is maintained as an abstract entity when studying aboutness relation.

4.2 Matching Function

The containment matching function means the contents of the query are completely contained in the information of the document. An example is the boolean model. *Containment* is represented by following rules:

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

and

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

On the other hand, overlap matching means that a document is retrieved even if it only partially matches the query, e.g., the vector space model. *Overlap* is represented by the following:

$$\frac{A \to C \wedge B \to C}{A \models B}$$

and

$$\frac{A \models B}{A \to C \wedge B \to C}$$

Note that overlap is a superset of containment, i.e., an IR model which supports overlap matching function also supports containment, and the set of documents retrieved by overlap matching encapsulates the set retrieved by containment matching.

4.3 Transformation

There are two kinds of transformation in IR: *statistical* one and *symbolical* one. Statistical transformation is always employed in *relevance feedback* to reweigh index terms based on the results of the previous match. On the other hand, not all of the information items contained in the objects (document and/or query) are explicit. They may be obtained through the transformation of implicit items. For example, in order to match a document containing an index term "fish" and a query containing "salmon", one could transform "salmon" in the query to its superclass "fish". This is symbolical transformation, often referred to as *query expansion*, involving manipulation of "word semantics". It has been proved in practice that relevance feedback and query expansion can significantly improve the effectiveness of IR.

Strictly speaking, statistical transformation is a special case of symbolical transformation. If we consider aboutness as a partially ordered relation, the re-assignment of term weights is equivalent to the changing of the ordering of aboutness. In this paper, we mainly discuss symbolical transformation.

To represent the ability of IR model in handling hidden semantics in transformation, information containment is extended to:

- *surface containment*
- *deep containment*

Information containment involving information carriers literally is surface containment. On the other hand, deep containment involves information transformation. It is possible to compare systems based on the level of containment they support. This overcomes the third deficiency of Bruza's aboutness framework.

Besides information containment, there are some other types of term relation, such as term association relation obtained by some statistical method, e.g., term co-occurrence, which is commonly used in many IR systems. Using these term relations, the information objects can be transformed symbolically.

Within transformation, two important aspects should be captured: *negation handling* and *conflicts resolution*. Assume that α is a formula. It can be applied to IR in two levels: index term and aboutness decision. Negation handling reflects how an IR model handles the relationship between *not-exist*(α) and *negation*(α). In Close World Assumption (CWA), if an index term is not explicitly mentioned to be true, it is assumed to be false; also, if a document is not shown to be relevant to query, it is assumed non-relevant. However, the information in the real world is boundless and non-deterministic. Thus, in Open World Assumption (OWA), α is false only when it is explicitly mentioned to be false.

Conflicts resolution is to handle the conflicts occur in the transformation process. In monotonic reasoning, the set of consequences increases monotonically with the set of antecedents. In IR, monotonicity is represented as follows:

$$\frac{A \models B}{A \oplus C \models B} \quad \textit{(Left Composition Monotonicity)}$$

and

$$\frac{A \models B}{A \models B \oplus C} \quad \textit{(Right Composition Monotonicity)}$$

Non-monotonic logic attempts to formalize common sense reasoning. A conclusion may no longer be valid when new information is collected. For example, the new information may contradict the antecedents and moreover, it is stronger than the latter. The non-monotonicity concerns which and how much information should be extended in order to preserve monotonicity. This is consistent with van Rijsbergen's logical uncertainty principle. In Hunter's work (Hunter 1995), non-monotonicity of IR is represented by the default rule:

$$\frac{\text{Condition: Justification}}{A \models B}$$

Note that some research has suggested that non-monotonicity manifest at fine level of information granularity (index terms). On the other hand, coarse level of granularity (i.e., document and query) exhibits monotonic characteristic. Consider a document D which is about a query Q. If a section S is added to D (yielding D⊕S), then D⊕S is still about Q even though the "strength" of the aboutness relation may be more or less than that between D and Q. Contrast this when the information granularity is fine. It is reasonable to assume that *web⊕ surfing* |= *surfing*. But using right composition monotonicty, the conclusion "*web⊕ surfing* |= *wave⊕ surfing*" could be derived!

Based on the dimensions "matching function" and "transformation", a set of more effective inference rules will be proposed to model the essential functionality of IR, which is closely related to the effectiveness. An evaluation function on the set of rules an IR model supports will be also defined. In this way, the fourth deficiencies of Bruza's framework can be overcome.

5. Conclusion and discussion

We have proposed a strategy for functional benchmarking which involves the application of a symbolic and axiomatic method to reason about IR functionality. The aboutness framework is adopted for this purpose. However, the direct application of aboutness is not flawless. A 3-dimensional scale, which identifies the classes of essential IR functionality, is proposed. It provides a way to improve Bruza's aboutness framework.

In the future, we will establish a more effective logical framework for functional benchmarking of IR. It is based on Bruza's framework and the 3-dimensional scale proposed in this paper. The definitions including aboutness relation, the inner structure of information carrier (representation), surface containment, deep containment, information composition, information preclusion, etc., and their semantics will be formally defined. A set of inference rules representing aboutness properties will be proposed to model transformation and matching function. An evaluation function will be defined to compare the IR models according to the functionality they support. Different IR models can then be mapped into the benchmarking so that their functionality can be assessed and their effectiveness compared inductively.

References

Bruza, P.D. and 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, 112-121.

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

Chittaro, L and Kumar, A.N. 1998. Reasoning about function and its applications to engineering. *Artificial Intelligence in Engineering 12*, 331-336.

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

Hunter, A. 1995. Using default logic in information retrieval. In *Symbolic and Quantitative Approaches to Uncertainty, vol. 946. Lecture Notes in Computer Science*, 235-242.

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.

Kumar, A.N. 1994. Function based reasoning. *The Knowledge Engineering Review 9*, 3, 301-304.

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

Lalmas, M. and Bruza, P.D. 1998. The use of logic in information retrieval modeling. *Knowledge Engineering Review*. In press.

Nie, J. 1989. An information retrieval model based on modal logic. *Information Proceeding & Management 25*, 5, 477-491.

van Rijsbergen, C.J. 1986. A non-classical logic for information retrieval. *The Computer Journal 29*, 6, 481-485.

van Rijsbergen, C.J. and Lalmas, M. 1996. An information calculus for information retrieval. *Journal of America Society for Information Science 47*, 5, 385-398.

Salton, G. 1988. *Automatic Text Processing.* Addison-Wesley.

Salton, G. 1992. The state of retrieval system evaluation. *Information Processing & Management 28*, 4, 441-449.

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

Sticklen, J. and Mcdowell, J. 1995. Future directions in function-based reasoning. *Applied Artificial Intelligence 9*, 1-3.

Winsor, J. and Maccallum, K 1994. A review of functionality modelling in design. *The Knowledge Engineering Review, 9*, 2, 163-199.

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