Using a Belief Revision Operator for Document Ranking in Extended Boolean Models

David E. Losada and Alvaro Barreiro
Dept. Computer Science
University of A Corunna
A Corunna, Spain
{losada,barreiro}@dc.fi.udc.es

Abstract

This paper claims that Belief Revision can be seen as a theoretical framework for document ranking in Extended Boolean Models. For a model of Information Retrieval based on propositional logic, we propose a similarity measure which is equivalent to a P-Norm case. Therefore it shares the P-Norm good properties and behaviour. Besides, it is theoretically ensured that this measure follows the notion of proximity between the documents and the query. The logical model can naturally deal with incomplete descriptions of documents and the similarity values are also obtained for this case.

1 Introduction

Logical approaches have been proposed to model Information Retrieval (IR) in a formal framework. Van Rijsbergen was the pioneer in thinking that logic could help in the retrieval of relevant documents [21]. Moreover, he proposed logic as a new theoretical framework for investigating IR. Given d, a logical representation of a document, and q, a logical representation of a query, retrieval is simply establishing whether $d \models q$. However, the evaluation of $d \models q$ in classical logic is not enough for IR purposes. Basically, classical entailment is too strong and cannot represent partial relevance [13]. In many cases, even though $d \not\models q$ the relevance of d to q can be high. In his seminal work [21], Van Rijsbergen proposed the notion of *minimal revision*: we can augment d in some minimal way until $d+\Delta d \models q$ holds. The measure of Δd can be used to give a measure of $d \models q$ and to rank several documents with respect to the query q. Van Rijsbergen made explicit the requirement of a non-classical logic for IR that should follow the logical uncertainty principle:

> " 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 $y \to x$ "

In that paper, Van Rijsbergen did not provide a method to quantify the *minimal extent*. However, he briefly outlined a probabilistic revision. Also, he considered a measure of nearness between documents which could be defined using Algorithmic Information Theory [4].

Since the original proposal, several logical approaches for IR, based on logical imaging [6], general logical imaging [7], modal logic [18], preferential structures [2], situation theory [22], terminological logics [17] and probabilistic Datalog [10], have been developed. It is worth referring to [5], [13] and [14] for a compendium and surveys about the use of logical models in IR.

In this paper we propose the use of a Belief Revision (BR) process to quantify the document-query similarity. BR can be considered as a quantitative formalism where measures are based on distances between interpretations. The notion of minimal revision is formally captured and models of IR can freely benefit from it. We propose a Propositional Logic for representing documents. Although the representation is restricted to the use of binary weights, a comparison with Extended Boolean Models (EBMs) is possible. The BR operator we will use is the one proposed by Dalal [8]. Katsuno and Mendelzon [12] have proved that Dalal's operator satisfies some important BR postulates. This result ensures a minimal revision following the notion of proximity and we find it of great importance to the field of IR ranking. We show how the proposed measure is equivalent to a P-Norm case and, therefore, it shares the interesting properties of this EBM measure.

It has been said that no consensus has still been reached regarding what is the best logical approach to model IR [14]. Though further analysis of existing logical models using the framework here presented can be done, and specially of those approaches that integrate non-monotonic reasoning and information retrieval [2], we have restricted ourselves to a limited scope in order to present some clear results.

The rest of the paper is organised as follows. Section 2 explains the logical model and the use of BR processes to compute the similarity measure. Section 3 shows the close relationship between BR measures and P-Norm measures.

Section 4 extends the proposal to deal with partial representations of documents. Section 5 discusses an alternative way of making the revision process. In the Conclusion section some advantages and drawbacks of the approach are pointed out.

2 A similarity measure using Belief Revision

We will use a model for representing documents and queries which is a direct translation of the Boolean Model to Logic. Let L be a propositional alphabet where each letter represents a term. Documents are represented as conjunctions of literals so that a positive literal represents the appearance of the term in the document and a negative one represents the absence. Queries are unrestricted propositional formulae. Here we consider the appearance/absence of all the terms (i.e. documents are *complete theories*). This is a common assumption in IR. Afterwards, we will see how the inherent partiality of logic can benefit retrieval models.

Let us consider the evaluation of $d \models q$ in order to decide whether the document represented by d is relevant to the user need represented by q[1]. If we evaluate $d \models q$ in Classical Logic, this criterion to retrieve documents provides only a two-value decision, as in the conventional Boolean Model. We propose a logical EBM where a rank of documents is built via a non-monotonic process. The model for documents and queries remains the same and we will use a BR process to compute the retrieval status values for documents in the line of EBMs. Some notions of Belief Revision are introduced before presenting the similarity measure.

2.1 Belief Revision

Belief Revision deals with the accommodation of a new piece of information into an existing knowledge base. We will restrict ourselves to the propositional case which is the one we will use for document ranking. A BR operator takes a theory and a new formula and decides the way of making an update in the theory, with the policy of producing minimal change. There are two lines of work to revision and update, *formula-based* approaches and *model-based* ones. We will focus on the latter because these approaches fit well with our purposes. In model-based approaches an order is established among interpretations. The minimal change needed to make the revision is provided by the minimal elements in the order.

In the following we denote by L a finitary propositional language, K the knowledge base, A the revising formula and $\circ$ the revision operator. Interpretations are functions from a propositional alphabet to the set $\{t, f\}$. In the following we will denote interpretations as the set of atoms which are mapped into true. A *model* of a propositional formula is an interpretation that makes the formula true. A model of a theory is an interpretation which is a model of every formula of the theory. Given a formula ψ we denote its set of models as $Mod(\psi)$.

Model-based approaches select models of the revising formula, A, on the basis of some notion of proximity to the models of the knowledge base, K. These selected models are the models of the revised theory. Next, we give the model-theoretic characterisation of this process. Let $\mathcal{I}$ be the set of all interpretations of L. A *pre-order* $\leq$ over $\mathcal{I}$ is a reflexive and transitive relation on $\mathcal{I}$. We can also define $<$ as: $I < I^{'}$ iff $I \leq I^{'}$ and not $I^{'} \leq I$. A pre-order is *total* if for each I, $J \in \mathcal{I}$ either $I \leq J$ or $J \leq I$ holds. An assignment is a function which maps a propositional formula ψ to a pre-order $\leq_\psi$ on $\mathcal{I}$. Katsuno and Mendelzon [12] defined the property of *faithfulness* for assignments. This property says that a model of ψ cannot be strictly less (according to $\leq_\psi$) than any other model of ψ and must be strictly less than any non-model of ψ.

An interpretation I is minimal in a subset, $\mathcal{M}$, of $\mathcal{I}$ with respect to $\leq_\psi$ if $I \in \mathcal{M}$ and there is no $I^{'} \in \mathcal{M}$ such that $I^{'} <_\psi I$. The set of these minimal interpretations is denoted $Min(\mathcal{M}, \leq_\psi)$.

If we consider $\leq_\psi$ as a measure representing the closeness between $Mod(\psi)$ and interpretations, i.e. $I^{'} \leq_\psi I$ means that $I^{'}$ is closer to $Mod(\psi)$ than I. Then, $Min(\mathcal{M}, \leq_\psi)$ can be seen as the set of all the closest interpretations in $\mathcal{M}$ to $Mod(\psi)$.

Gärdenfors and other researchers established a set of postulates that revision operators must satisfy [1, 11]. Their work was located in the framework of *knowledge sets* or deductively closed set of formulae. Katsuno and Mendelzon formulated an equivalent set of postulates in the framework of knowledge bases [12].

It has been proved that any proposed revision operator which satisfies the above mentioned postulates, accomplishes an update with minimal change to the set of models of the knowledge base. Specifically, Katsuno and Mendelzon proved a theorem relating the satisfiability of the postulates of Belief Revision with the existence of a faithful assignment. This result has great importance because, given an operator that satisfies the postulates then it is ensured that it can produce an update of the knowledge base with minimal change (and following the notion of proximity).

In the present work we focus only in Dalal's revision operator. As we will see later this operator will produce the desired behaviour for ranking. Moreover, Katsuno and Mendelzon's work proves Dalal's revision produces minimal change revision.

Dalal's revision. This revision uses the cardinality of the symmetric difference between two interpretations I and J, (i.e. the number of propositional letters on which two interpretations differ) as a measure of *distance* between them, $dist(I, J)$.

Then, the distance between $Mod(\psi)$ and I is defined as:

$$dist(Mod(\psi), I) = min_{J \in Mod(\psi)} dist(J, I) \qquad (1)$$

Finally, a faithful assignment of a total pre-order $\leq_\psi$ can be defined:

$$I \leq_\psi J \text{ iff } dist(Mod(\psi), I) \leq dist(Mod(\psi), J) \qquad (2)$$

This assignment joins the conditions of faithfulness. Therefore it can be concluded that Dalal's operator satisfies the postulates of Belief Revision.

Given a new piece of information, A, Dalal's revision operator, $\circ_D$, establishes the models of the revised theory as the models of the new piece of information closest to the theory:

$$Mod(\psi \circ_D A) = Min(Mod(A), \leq_\psi) \qquad (3)$$

2.2 Using Dalal's operator to build a similarity measure in IR

In this subsection we show how to use Belief Revision as a tool to construct a rank in the framework of EBMs.

[1] In an interesting study [20], Sebastiani has compared this criterion with other model-theoretic and proof-theoretic criteria

Let $L = \{t_1, t_2, \ldots, t_n\}$ be a finitary propositional alphabet. Documents are represented as conjunctions of literals, such that every document representation has one occurrence of every t_i, either positive or negative. Queries are unrestricted propositional formulae using the alphabet L.

Belief Revision selects the models of the new information which are the closest to the models of the theory. Let us consider a query, q, as a theory, and a representation of a document, d, as a new piece of information. We could consider the alternative formulation where d is the theory and q is the new piece of information. The study of this formulation is postponed until section 5. There it is shown that this option is not well suited for our purposes. Since the BR process establishes a measure of distance from models of the new information to the set of models of the theory, we can obtain a measure of distance from the model of the document (documents are complete theories and, therefore, they have only one model) to the models of the query, and, hence, from the document to the query itself. Considering each document of the document base as a new piece of information it is possible to compute the document-query similarity via several BR processes. We must remark on the fact that we are not interested in the final result of the revision process (i.e. the revision of the query) but only in the measure of the closeness between documents and the query.

Formally, the query has associated a set of models, $Mod(q)$, and the document has only one model, m_d. Then, we can use Dalal's distance as follows:

$$dist(Mod(q), m_d) = min_{J \in Mod(q)} dist(J, m_d) \qquad (4)$$

This formula uses the distance between each model of the query (J) and the model of the document (m_d), given by the cardinality of the symmetric difference between interpretations. This difference is a measure of the change needed to make both interpretations the same. Then, we establish a measure of distance between the models of the query and the model of the document as the minimum of the cardinalities. As documents have only one model, we can identify a document by its model and define the distance between documents and queries as:

$$distance(d, q) = dist(Mod(q), m_d). \qquad (5)$$

This definition corresponds to the intuitive meaning that distance from a document to a query is given by the number of terms that must be added (or eliminated for negative query literals) to the document to satisfy the query.

We can define a total pre-order $\leq_q$ (established by the query) between the documents:

$$d_1 \leq_q d_2 \text{ iff } distance(d_1, q) \leq distance(d_2, q) \qquad (6)$$

where the distance for the document d_1 is computed in the revision of the theory q with the new information d_1 and the distance for document d_2 is computed in the revision of the theory q with the new information d_2. We can also define $<_q$ as $d_1 <_q d_2$ iff $d_1 \leq_q d_2$ and not $d_2 \leq_q d_1$ and $=_q$ as $d_1 =_q d_2$ iff $d_1 \leq_q d_2$ and $d_2 \leq_q d_1$.

Distance (5) can be used to define a similarity measure (7) normalised in the interval $[0, 1]$. The normalisation uses the fact that $distance(d, q)$ can vary from 0 to k, where k is the number of atoms which appear in the query.

$$BRsim(d, q) = 1 - distance(d, q)/k \qquad (7)$$

The query q is an unrestricted propositional formula over the alphabet L. The above definition can be directly applied to queries consisting of conjunctions of literals, either negative or positive. However some considerations must be done in the processing of queries involving negations and disjunctions. The management of negative literals is naturally integrated in the formalism. In queries we eliminate negations with wide scope using the Morgan's laws and double negation. Therefore, it is not necessary to define the similarity measure for a negation query from its corresponding positive formulation.

One can notice that in an OR-query this measure produces only two values, 0 and 1. Some retrieval processes rank the documents which satisfy an OR-query considering the number of query terms which are satisfied. This is equivalent to take the set of documents that satisfy the OR-query and rank them considering the query as an AND. This process can be accomplished in the logical retrieval model proposed. The previous considerations lead to the definition of the similarity measure for OR-queries:

$$BRsim(d, q_{or}) = BRsim(d, q_{and}) \qquad (8)$$

The processing of general queries involving both conjunctions and disjunctions is defined in the usual way for systems based on EBMs. For instance, consider the query $q = a \wedge b \vee c$, the similarity between a document d and the query is computed as:

$$sim(d, q) = \frac{sim(d, a \wedge b) + sim(d, c)}{2} \qquad (9)$$

In the next section we will prove the equivalence between the proposed measure and P-Norm with binary weights and $p = 1$. In fact, the reader could have observed that definitions provided for OR and complex queries with conjunctions and disjunctions are the same that in the mentioned P-Norm case.

In our formulation a formula representing a document is the new piece of information. From a BR perspective it would be possible to consider the formula whose set of models is the union of the models of all the documents as the new information. As documents are identified by its only model the order in the models induces the order in the documents. However, this option would not be valid if document representations are partial. In this case, each document can have several models and this alternative is inconvenient because BR establishes an order in the models, which is not directly an order in the documents. Moreover, as the measure of distance between the query and each document is absolute (independent of the rest of documents), this choice does not provide additional information and, as it has been shown, can blur the notion of document.

Example 1: Let the propositional alphabet L, documents d_i and a query q be defined as:

$L = \{a, b, c, d\}$

$d_1 = a \wedge \neg b \wedge c \wedge d$
$d_2 = \neg a \wedge \neg b \wedge c \wedge \neg d$
$d_3 = a \wedge b \wedge \neg c \wedge d$
$d_4 = \neg a \wedge b \wedge c \wedge d$
$d_5 = a \wedge b \wedge \neg c \wedge \neg d$
$d_6 = \neg a \wedge \neg b \wedge c \wedge d$

$q = a \wedge b$

Fig. 1 (a) shows the symmetric differences between query and document models. Fig. 1 (b) shows cardinalities of

docs → query ↓	d_1 $\{a,c,d\}$	d_2 $\{c\}$	d_3 $\{a,b,d\}$	d_4 $\{b,c,d\}$	d_5 $\{a,b\}$	d_6 $\{c,d\}$
$\{a,b\}$	$\{b,c,d\}$	$\{a,b,c\}$	$\{d\}$	$\{a,c,d\}$	$\emptyset$	$\{a,b,c,d\}$
$\{a,b,c\}$	$\{b,d\}$	$\{a,b\}$	$\{c,d\}$	$\{a,d\}$	$\{c\}$	$\{a,b,d\}$
$\{a,b,c,d\}$	$\{b\}$	$\{a,b,d\}$	$\{c\}$	$\{a\}$	$\{c,d\}$	$\{a,b\}$
$\{a,b,d\}$	$\{b,c\}$	$\{a,b,c,d\}$	$\emptyset$	$\{a,c\}$	$\{d\}$	$\{a,b,c\}$

(a) Symmetric differences between query and document models

docs → query ↓	d_1 $\{a,c,d\}$	d_2 $\{c\}$	d_3 $\{a,b,d\}$	d_4 $\{b,c,d\}$	d_5 $\{a,b\}$	d_6 $\{c,d\}$
$\{a,b\}$	3	3	1	3	0	4
$\{a,b,c\}$	2	2	2	2	1	3
$\{a,b,c,d\}$	1	3	1	1	2	2
$\{a,b,d\}$	2	4	0	2	1	3
$dist(Mod(q), m_d)$	1	2	0	1	0	2

(b) Cardinalities of symmetric differences

Figure 1: Computation of the distance between documents and a query

symmetric differences. Documents are represented by its only model and each row represents a query model. Last row of Fig. 1 (b) shows the distance between each document and the query.

Then using the total preorder $\leq_q$ established by the query:

$$d_3 =_q d_5 <_q d_1 =_q d_4 <_q d_2 =_q d_6$$

We can observe how the distances and the induced order correspond to what intuitively was expected. Moreover, following Katsuno and Mendelzon's above mentioned theorem, as Dalal's operator verifies the BR postulates, it is theoretically ensured that these distances correspond to the notion of proximity between the documents and the query. The similarity measures are:

$$BRsim(d_1,q) = 0.5$$
$$BRsim(d_2,q) = 0$$
$$BRsim(d_3,q) = 1$$
$$BRsim(d_4,q) = 0.5$$
$$BRsim(d_5,q) = 1$$
$$BRsim(d_6,q) = 0$$

3 Comparison with Extended Boolean Models

IR systems based on the Boolean Model are efficient and can give high performance in terms of recall and precision if the query is well formulated. However, conventional boolean retrieval systems cannot provide ranked retrieval output in order of query-document similarity. In addition to support ranking facilities, Extended Boolean Models (EBMs) [19] preserve the query structure inherent in Boolean Models and incorporate weighted terms into both queries and documents.

Usually, systems based on EBMs only compute the retrieval status values (RSVs) of those documents which satisfy the corresponding boolean query. In the case of P-Norm with binary weights and $p = 1$, all the documents satisfying the query have the same RSV. To motivate this kind of systems, the general strategy to compare queries with the document collection suggested by Salton in the concluding remarks of [19], can be applied: a broadened query is used to obtain a set of potentially relevant documents. Then, the RSVs are computed only for the documents in this set, and using the initial P-Norm query.

We will show now how BR measure is equivalent to P-Norm with p=1 and binary term weights. As it is well known, when p reaches the value 1 the Extended Boolean Model is equivalent to the vector space model with the inner product query-document matching function. Despite that, as complex queries (involving both AND and OR) are allowed, comparison of our proposal with EBM is straightforward. Then, following Lee's work [15], some interesting properties of the measure can be obtained. We use the formal definition of EBMs used by Lee in his work (here restricted to binary weights) and show how the proposed model fits into this framework.

An IR system based on EBM with binary weights is defined by the quadruple $< T, Q, D, F >$ where:

- T is a set of index terms used to represent queries and documents.
- Q is a set of queries that are boolean expressions involving index terms and logical operators AND, OR and NOT.
- D is a set of documents. Every document d has the form $\{(t_1, w_{d1}), \ldots, (t_n, w_{dn})\}$ where $w_{dj} \in \{0, 1\}$
- F is a retrieval function, $F : D \times Q \rightarrow [0, 1]$, which establishes a measure of similarity between documents and queries.

Our logical model can be described in Lee's fashion:

- T is a propositional alphabet where each element represents one index term.
- Q is a set of queries that are well formed propositional formulae on the alphabet T.
- D is a set of documents. Every document d_i is represented as a conjunction of literals on the alphabet T.
- F is a retrieval function, $F : D \times Q \rightarrow [0, 1]$, which is the similarity measure calculated with BR techniques, $BRsim$.

Let a document d and an AND-query q_{and} be represented in the usual way in EBMs:

$$d = \{(t_1, w_{d1}), \ldots, (t_n, w_{dn})\}$$
$$q_{and} = \{(t_1, w_{q1}), \ldots, (t_n, w_{qn})\}$$

Let us consider the P-Norm measure for $p = 1$ and binary weights for both query and document:

$sim(d, q_{and}) = \frac{w_{q1}w_{d1}+\ldots+w_{qn}w_{dn}}{w_{q1}+\ldots+w_{qn}}$

In the proposed logical model, first we analyse the case in which only positive query-terms appear (negative terms are left until the discussion of NOT-queries). Let k be the number of those positive query-terms. In the EBM formulation this corresponds to the fact that k query-terms have weight 1 and $n-k$ query-terms have weight 0. Then, the similarity value is:

$$sim(d, q_{and}) = \frac{\sum_{w_{ql}=1} w_{dl}}{k}$$

Since each w_{dl} is a binary weight, the sum represents the number of query-terms that also appear in the document. Given that $BRsim(d, q) = 1 - distance(d, q)/k$, and $distance(d, q)$ counts the number of terms in which d and q differ, it follows that BR measure is completely equivalent to the P-Norm case.

As it has been shown by Salton and other researchers [19], in OR-queries, P-Norm with $p = 1$ produces the same values as the ones produced for AND-queries. In subsection 2.2 BR similarity measure for OR-queries was defined by means of the measure for AND-queries.

In NOT-queries P-Norm is defined as $sim(D,\text{NOT } Q) = 1 - sim(D, Q)$. Therefore the computation of similarity for queries with negations implies the recurrent use of the previous formula. Within the BR measure the management of negations is completely homogeneous to the management of positive terms. Actually, instead of the recurrent use of the formula, we have to move negations inwards as it has been explained in subsection 2.2. We present now an example with a query involving negative literals.

Example 2: Let the propositional alphabet L, documents d_i and a query q be defined as:

$L = \{a, b\}$

$d_1 = a \wedge \neg b$
$d_2 = \neg a \wedge \neg b$

$q = a \wedge \neg b$

Using P-Norm:

$d_1 = \{(a, 1), (b, 0)\}$
$d_2 = \{(a, 0), (b, 0)\}$

$q = a$ AND NOT b

$sim(d, q) = \frac{sim(d,a)+sim(d,NOTb)}{2} = \frac{sim(d,a)+1-sim(d,b)}{2}$

$sim(d_1, q) = \frac{1+1-0}{2} = 1$
$sim(d_2, q) = \frac{0+1-0}{2} = 0.5$

Fig. 2 shows the tables of symmetric differences and their cardinalities to compute distances.

The results are $BRsim(d_1, q) = 1$ and $BRsim(d_2, q) = 0.5$.

It holds that $BRsim(d, q_{not}) = 1 - BRsim(d, q)$. However it is clear that this kind of transformation is not needed in the computation of the similarity. Similarity measure for complex queries with AND and OR connectives was defined in the way used in P-Norm. From the definition and the equivalences studied for the basic cases, the equivalence for the complex case follows.

Finally, following Lee's work it is proved that BR measure does not have the Single Operand Dependency, Non-associativity and Unequal Importance problems. These problems arise in some EBM measures, even using only binary weights. Associativity is directly inherited from associativity in Propositional Logic and it is clear that the proposed measure gives the same importance to all the terms. Besides, the P-Norm measure provides good experimental results and subsume other measures. Evaluation results for the case of p=1 and binary query and document weights have been provided [19], concluding that substantial improvements in retrieval effectiveness over the Boolean Model are obtained.

docs → query ↓	d_1 $\{a\}$	d_2 $\emptyset$
$\{a\}$	$\emptyset$	$\{a\}$

docs → query ↓	d_1 $\{a\}$	d_2 $\emptyset$
$\{a\}$	0	1
$dist(Mod(q), m_d)$	0	1

Figure 2: An example with a query with negative literals

4 Partial Information

One of the major breakthroughs of logical retrieval models lies in the representation of incomplete knowledge. Classical IR systems are based on the assumption that document representation is complete. In fact, this is equivalent to consider that IR systems have adopted the policy of closed world assumption, i.e.: if no information is available about the presence/absence of a certain term in a document, it is assumed that it does not appear in the document. Actually logic is hardly needed when knowledge is total and partiality is an important feature captured by logic [3, 20]. Indeed, the closed world assumption is too strong in cases where an automatic content analysis cannot be fully accomplished. For instance, think of multimedia documents. The proposed logical model can naturally deal with partial knowledge.

Let $L = \{t_1, t_2, \ldots, t_n\}$ be the propositional alphabet. Documents are represented as conjunctions of literals such that every t_i can appear or not in the document representation. Now document representations can be *partial theories*. Queries are unrestricted propositional formulae.

Then, we apply the BR formulation to the case of incomplete document representation. The roles of queries as theories and documents as new pieces of information remain the same. A query, q, has associated a set of models, $Mod(q)$, and each document, d, has also a set of models, $Mod(d)$.

Dalal's distance is used to measure the closeness between the query and each model of the document:

Firstly, for each $m_d \in Mod(d)$:

$$dist(Mod(q), m_d) = min_{J \in Mod(q)} dist(J, m_d) \quad (10)$$

Using the faithful assignment suggested by Dalal, we can define a total pre-order $\leq_q$ (which is established by the query) between the models:

$$m_1 \leq_q m_2 \text{ iff } dist(Mod(q), m_1) \leq dist(Mod(q), m_2) \quad (11)$$

We suggest to use an average measure to obtain a measure for each document:

$$distance(d, q) = \frac{\sum_{m \in Mod(d)} dist(Mod(q), m)}{|Mod(d)|} \quad (12)$$

This corresponds to the intuitive meaning that distance from a document to a query is given by the average of the number

docs → / query ↓	d_1 m_{11} $\{a,b,c\}$	m_{12} $\{a,c\}$	d_2 m_{21} $\{c\}$	m_{22} $\{a,b,c\}$	m_{23} $\{a,c\}$	m_{24} $\{b,c\}$	d_3 m_{31} $\{a,b\}$	m_{32} $\{a,b,c\}$
$\{a,b\}$	$\{c\}$	$\{b,c\}$	$\{a,b,c\}$	$\{c\}$	$\{b,c\}$	$\{a,c\}$	$\emptyset$	$\{c\}$
$\{a,b,c\}$	$\emptyset$	$\{b\}$	$\{a,b\}$	$\emptyset$	$\{b\}$	$\{a\}$	$\{c\}$	$\emptyset$

(a)Symmetric differences between query and document models

docs → / query ↓	d_1 m_{11} $\{a,b,c\}$	m_{12} $\{a,c\}$	d_2 m_{21} $\{c\}$	m_{22} $\{a,b,c\}$	m_{23} $\{a,c\}$	m_{24} $\{b,c\}$	d_3 m_{31} $\{a,b\}$	m_{32} $\{a,b,c\}$
$\{a,b\}$	1	2	3	1	2	2	0	1
$\{a,b,c\}$	0	1	2	0	1	1	1	0
$dist(Mod(q), m_{ij})$	0	1	2	0	1	1	0	0
$dist(d_i, q)$	0.5		1				0	

(b)Cardinalities of symmetric differences and proposed measure of distance

Figure 3: Computation of distance with partial information

of terms that must be changed in the models of the document in order to satisfy the query. Other measures have been taken into account. For instance, when we take the *worst* model of each document (i.e. the one whose distance to the query is the biggest) some problems arise. Given the alphabet $L = \{a, b, c\}$, the query $q = a \wedge b$, and the documents $d_1 = a \wedge c$ and $d_2 = a \wedge \neg b \wedge c$, the *worst* approach produces the same RSV for d_1 and d_2, against the expected and intuitive behaviour. We think a measure that takes into account all the models is useful and fair. Another option is to do the closed world assumption. Notice that if we had considered this option we would have taken into account only one model, the one that denies all the non-appearing terms.
Finally, BR similarity measure is:

$BRsim(d, q) = 1 - distance(d, q)/k$, where k is the number of atoms which appear in the query.

Example 3: Let the propositional alphabet L, documents d_i and a query q be defined as:

$L = \{a, b, c\}$

$d_1 = a \wedge c$
$d_2 = c$
$d_3 = a \wedge b$

$q = a \wedge b$

The symmetric differences between respective models, their cardinalities and the distances are shown in Fig. 3. The similarity values are:

$BRsim(d_1, q) = 0.75$
$BRsim(d_2, q) = 0.5$
$BRsim(d_3, q) = 1$

5 The revision $d \circ_D q$

The BR process could be conceived as the revision of the theory d for the new information q. In this section we show how this alternative does not fit well with the purposes of document ranking. In the case of complete information a revision in the form $d \circ_D q$ could be used to produce the same ranking as our proposal. However, when incomplete representations of documents appear, the option $d \circ_D q$ does not produce the desired results.

Let $L = \{a, b, c\}$ be a propositional alphabet, $d = a \wedge b \wedge \neg c$ a document representation and $q = a \wedge c$ a query. Fig. 4 shows how the revision $d \circ_D q$ can be used to build the

$q \circ_D d$ / document → / query ↓	m_d $\{a,b\}$
$\{a,c\}$	2
$\{a,b,c\}$	1
$dist(Mod(q), m_d) = min_{m \in Mod(q)} dist(m, m_d)$	1

(a) $q \circ_D d$ revision

$d \circ_D q$ / query → / document ↓	mq_1 $\{a,c\}$	mq_2 $\{a,b,c\}$
$\{a,b\}$	2	1
$dist(Mod(d), mq_i) = min_{m \in Mod(d)} dist(m, mq_i)$	2	1

(b) $d \circ_D q$ revision

Figure 4: $q \circ_D d$ and $d \circ_D q$ revisions with complete descriptions of documents

similarity measure. The previous proposed process, $q \circ_D d$, is also shown for comparison. Only cardinalities of symmetric differences appear in the figure.

The BR process $d \circ_D q$ provides a measure of distance from each model of the query to the model of the document. An order $\leq_d$ over the models of q is obtained by the comparison of their respective distances to the model of d ($mq_2 \leq_d mq_1$). A measure of the distance from the document to the query can be obtained by taking the minimum of the distances from the models of the query to the model of the document. Obviously the revisions are different: $Mod(d \circ_D q) = \{mq_2\}$ and $Mod(q \circ_D d) = \{m_d\}$. The model mq_2 is the model of the query closest to the document and m_d is the model of the document closest to the query. However, both distances are the same in the case of complete description of documents given that there is only one document model.

When the document representation is partial, the desired result cannot be obtained via $d \circ_D q$. Let $L = \{a, b, c\}$ be a propositional alphabet, $d = c$ and $d' = a \wedge b \wedge c$ document representations and $q = a \wedge b$ a query. Fig. 5 presents both approaches. With the revision $d \circ_D q$ measures of distance from each query model to the set of models of the document are obtained. Actually, the distance from a query model to a document is the distance from the query model to the closest document model. The distances from models of q, mq_1 and mq_2, to the document d are 1 and 0 respectively. Consider the document d' also shown in Fig. 5. The distances of the models of q to d' are also 1 and 0. Obviously, any measure combining both distances would produce the same retrieval

$q \circ_D d$ document $\rightarrow$ query $\downarrow$	d m_1 $\{c\}$	m_2 $\{a,b,c\}$	m_3 $\{a,c\}$	m_4 $\{b,c\}$
$\{a,b\}$	3	1	2	2
$\{a,b,c\}$	2	0	1	1
$dist(Mod(q), m_i) = min_{m \in Mod(q)} dist(m, m_i)$	2	0	1	1
$dist(d,q)$	1			

(a) $q \circ_D d$ revision

$d \circ_D q$ query $\rightarrow$ document $\downarrow$	q mq_1 $\{a,b\}$	mq_2 $\{a,b,c\}$
$\{c\}$	3	2
$\{a,b,c\}$	1	0
$\{a,c\}$	2	1
$\{b,c\}$	2	1
$dist(Mod(d), mq_i) = min_{m \in Mod(d)} dist(m, mq_i)$	1	0

(b) $d \circ_D q$ revision

$d' \circ_D q$ query $\rightarrow$ document $\downarrow$	q mq_1 $\{a,b\}$	mq_2 $\{a,b,c\}$
$\{a,b,c\}$	1	0
$dist(Mod(d'), mq_i) = min_{m \in Mod(d')} dist(m, mq_i)$	1	0

(c) $d' \circ_D q$ revision

Figure 5: $q \circ_D d$ and $d \circ_D q$ revisions with partial descriptions of documents

status values for both d and d' given the query q. Clearly, this fact is counterintuitive. However these distances could be useful if we want to rank queries given a document.

6 Conclusion

Our proposal implies a theoretical formalisation of the document ranking process while other proposed measures are based on empirical experimentations. It has been guaranteed that the BR measure follows the notion of proximity between documents and queries, and provides an homogeneous management of positive and negative terms. We have shown the equivalence between the BR measure and a P-Norm case, achieving the interesting properties and good behaviour of P-Norm.

The main drawback lies in the fact that the expressiveness of Propositional Logic is limited. The representation of weights in the interval $[0,1]$ needs a more expressive language. We think BR is a promising framework to analyse established models for IR. Also, it is a particularly attractive formalism to formulate document ranking in the scope of logical models of IR based on sublanguages of First Order Logic [9, 17]. An interesting practical result along this line of research is the possibility of directly using algorithms developed for Circumscription also for BR, as it has been shown by Liberatore [16].

Acknowledgements: This work was supported in part by project PB97-0228 from "Ministerio de Educacion y Cultura", Spain. We thank Ramón P. Otero and Pedro Cabalar for their valuable comments and suggestions.

References

[1] C. E. Alchourron, P. Gärdenfors, and D. Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50:510–530, 1985.

[2] P. Bruza. Preferential models of query by navigation. In F. Crestani, M. Lalmas, and C. J. Van Rijsbergen, editors, *Information Retrieval, Uncertainty and Logics:advanced models for the representation and retrieval of information*, Norwell, MA, 1998. Kluwer Academic Publishers.

[3] P. Bruza and M. Lalmas. Logic-based information retrieval: Is it really worth it? In *Proc. of EUFIT 96, Fourth European Congress on Intelligent Techniques and Soft Computing*, Aachen, Germany, 1996.

[4] G.J. Chaitin. Algorithmic information theory. *IBM Journal of Research and Development*, 21:350–359, 1977.

[5] F. Crestani, M. Lalmas, and C.J.Van Rijsbergen, editors. *Information Retrieval, Uncertainty and Logics: advanced models for the representation and retrieval of information.* Kluwer Academic Publishers, Norwell, MA, 1998.

[6] F. Crestani and C. J. Van Rijsbergen. Information retrieval by logical imaging. *Journal of Documentation*, 51(1):3–17, 1995.

[7] F. Crestani and C. J. Van Rijsbergen. Probability kinematics in information retrieval. In *Proc. of SIGIR-95, the 18th ACM Conference on Research and Development in Information Retrieval*, pages 291–299, Seattle, Washington, 1995.

[8] M. Dalal. Investigations into a theory of knowledge base revision: Preliminary report. In *Proc. of AAAI-88, the 7th National Conference on Artificial Intelligence*, pages 475–479, 1988.

[9] N. Fuhr. Modelling information retrieval in Datalog. In R. Kuhlen and M. Rittberger, editors, *Hypertext - Information Retrieval - Multimedia, Synergieeffekte elektronischer Informationssysteme*, pages 163–174. Universitätsverlag Konstanz, Konstanz, 1995.

[10] N. Fuhr. Probabilistic Datalog - a logic for powerful retrieval methods. In *Proc. of SIGIR-95, the 18th ACM Conference on Research and Development in Information Retrieval*, pages 282–290, Seattle, Washington, 1995.

[11] P. Gärdenfors and D. Makinson. Revisions of knowledge systems using epistemic entrenchment. In *Proceedings of the Second Conference on Theoretical Aspects of Reasoning about Knowledge*, pages 83–95, 1988.

[12] H. Katsuno and A. O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.

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

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

[15] J. H. Lee. Properties of extended boolean models in information retrieval. In *Proc. of SIGIR-94, the 17th ACM Conference on Research and Development in Information Retrieval*, Dublin, Ireland, 1994.

[16] P. Liberatore and M. Schaerf. Reducing belief revision to circumscription (and vice versa). *Artificial Intelligence*, 93(1-2):261–296, 1997.

[17] C. Meghini, F. Sebastiani, U. Straccia, and C. Thanos. A model of information retrieval based on a terminological logic. In *Proc. of SIGIR-93, the 16th ACM Conference on Research and Development in Information Retrieval*, pages 298–307, Pittsburgh, PA, 1993.

[18] J. Y. Nie. An information retrieval based on modal logic. *Information Processing & Management*, 25(5):477–491, 1989.

[19] G. Salton, E. A. Fox, and H. Wu. Extended boolean information retrieval. *Communications of the ACM*, 26(12):1022–1036, 1983.

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

[21] C.J. Van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.

[22] C.J. Van Rijsbergen and M. Lalmas. An information calculus for information retrieval. *Journal of the American Society for Information Science*, 47(5):385–398, 1996.