A theoretical study of logical retrieval in terms of a characterization of knowledge revision

Alvaro Barreiro[1] and David E. Losada[2]

[1] AIlab,
Department of Computer Science
University of A Coruña, SPAIN
barreiro@udc.es
[2] Intelligent Systems Group
Department of Electronics and Computer Science
University of Santiago de Compostela, SPAIN
dlosada@dec.usc.es

Abstract. In this paper we investigate the application for Information Retrieval of the relationship between knowledge revision with minimal change and orders among logical interpretations. The study is presented in the framework of a logical model of information retrieval. Following this approach, documents and queries are represented by propositional formulas and document ranking is based on measures of distances between logical interpretations. The discussion is shifted from a pure model theoretic scenario to one of syntactic restrictions where efficient ranking is possible. The result is a certain epistemic justification of logical retrieval which is independent of considerations about relevance.

1 Introduction

Models of Information Retrieval (IR) are "theoretical descriptions of the IR task that could serve both as specifications for building running systems, and as theoretical tools for abstractly investigating the relative effectiveness of systems built along their guidelines" (Sebastiani, 1998). As a first step towards the second of these goals, in this paper we address the issue of giving a pure epistemic justification of logical ranking. We tackled this problem in the particular framework of the PLBR (Propositional Logic & Belief Revision) logical model of IR (Losada and Barreiro, 1999, 2001; Losada, 2001).

In logical models of Information Retrieval documents and queries are represented as logical formulas. Given a query, relevance of a document could be given by classical entailment, i.e. document d is relevant to query q iff $d \models q$. But this criterion is too strict because it does not consider partial matching (van Rijsbergen, 1986). In the PLBR logical model of IR, documents and queries are represented as Propositional Logic formulas, and a measure of distance between documents and queries is obtained from measures of distances between the sets of logical interpretations of formulas d and q. For the purpose of retrieval, the measure of distance can be transformed into a similarity measure and, finally, the PLBR model can provide a ranking of documents given a query. Since distances between logical interpretations, and consequently distances between formulas, can be obtained with different procedures, the main motivation of this paper is give a theoretical justification of the logical ranking that the PLBR model provides.

In the field of Belief Revision (BR), distances between logical interpretations have been profoundly studied with the purpose of obtaining knowledge revision with minimal change. A number of rationality postulates (AGM postulates) establish formally the principles that any revision operator should fulfil (Alchourrón et al., 1985; Gärdenfors, 1988). Katsuno and Mendelzon characterized the revision schemes which satisfy the AGM postulates with respect to an ordering among logical interpretations (Katsuno and Mendelzon, 1991). Specifically, the fulfilment of the AGM postulates for knowledge revision with minimal change is equivalent to the existence of an ordering among logical interpretations that exhibits some particular properties. In section (2) of this paper we will revisit this result to characterize the ranking of documents given a query which is constructed within the PLBR model. We aim to give a pure theoretical support of the ranking produced in the PLBR model.

Computation of distances between logical interpretations associated with the logical formulas representing documents and queries, is very much demanding from the computational point of view. In the PLBR model, a syntactic restriction in the formulas allows to compute the similarity between documents and queries in polynomial time. This was necessary to implement and evaluate the model in large collections (Losada and Barreiro, 1999, 2003b). In section (3), the validity of the results presented in section (2) is studied under this restricted formulation.

In sections (2) and (3) distances are obtained from propositional logical formulas where propositional letters represent index terms. In section (4) more information, such as for example specifity of terms, is incorporated in the measure. Different heuristics to incorporate *idf* (*inverse document frequency*) are introduced with the purpose of testing whether or not the properties presented in the previous sections are preserved. In section (5), the paper discusses some open questions and introduces future work. The paper ends with some conclusions.

2 Theoretical basis

2.1 Dalal's distance. Documents having a single model

In the PLBR model, documents and queries are represented by propositional formulas where propositional letters represent index terms. Let us consider the case of documents represented by propositional formulas which have a single model, i.e. documents with complete knowledge about the presence or absence of every term of the alphabet. Dalal's distance (Dalal, 1988) can be directly translated into an order of documents which can be used for ranking.

Let L be a propositional language. A logical interpretation is a function from the set consisting of all the propositional letters in L to $\{true, false\}$. Given two propositional logical interpretations I, J, Dalal's distance between them, $dist(I, J)$, is the number of propositional letters on which they differ. i.e. whose interpretation is different in I and J. A *model* of a propositional formula ψ is an interpretation that makes ψ true; $Mod(\psi)$ denotes the set of all the models of ψ. A measure of distance between the set of models of a query q and a document d having a single model m_d can be defined as

$$Dist(Mod(q), m_d) = min_{m_q \in Mod(q)} dist(m_q, m_d). \tag{1}$$

Let $\mathcal{I}$ be the set of all interpretations built on L. A *pre-order* $\leq$ is a reflexive and transitive relation on $\mathcal{I}$. A pre-order is *total* if for every $I, J \in \mathcal{I}$, either $I \leq J$ or $J \leq I$. Let us consider documents having a single model. Without imposing further restrictions, a document can be seen as its associated model and the set of possible documents (or, equivalently, the set of models of these possible documents) is $\mathcal{I}$. The distance *Dist* defines a total preorder $\leq_q$ over the set of models of documents:

$$m_{d_i} \leq_q m_{d_j} \quad iff \quad Dist(Mod(q), m_{d_i}) \leq Dist(Mod(q), m_{d_j}), \tag{2}$$

where m_{d_i} (m_{d_j}) is the model of document d_i (d_j).

Because the measure of distance defined in (1) is based on Dalal's distance, for the total pre-order $\leq_q$ the following three properties hold (Katsuno and Mendelzon, 1991):

- (PI1) If $m_{d_i}, m_{d_j} \in Mod(q)$, then $m_{d_i} <_q m_{d_j}$ does not hold, where $<_q$ is defined from $\leq_q$ in the usual way.
- (PI2) If $m_{d_i} \in Mod(q)$ and $m_{d_j} \notin Mod(q)$, then $m_{d_i} <_q m_{d_j}$ holds.
- (PI3) If $q \equiv q'$, then $\leq_q = \leq'_q$.

That is, (1) a model of q cannot be strictly less than any other model of q; (2) it must be strictly less than any non-model of q; and (3) logically equivalent query formulas produce equal pre-orders. Therefore, given an information need represented by q, if we use $\leq_q$ for ranking documents having a single model m_d, we guarantee these three properties.

The first and second properties say that documents which completely satisfy the information need represented by the query must be minimal in the order and strictly less that documents that do not completely satisfy that information need. The third property establishes a principle of irrelevance with respect to different syntactic but logically equivalent queries. Although these properties seem reasonable requirements for a logical formulation for retrieval, one could think that they could be of little practical interest because basically they only differentiate between models and non-models of the query. More importantly is that knowledge revision with minimal change is obtained selecting the minimal models with respect to an ordering among interpretations that satisfies these properties as it was proved by Katsuno and Mendelzon in (Katsuno and Mendelzon, 1991). Actually, this notion of minimal change is established in a particular setting: the AGM postulates for Belief Revision. The AGM postulates (Alchourrón et al., 1985; Gärdenfors, 1988) are a proposal of rational principles that every operator of knowledge revision must satisfy. Originally formulated in a very general setting and on philosophical grounds, AGM postulates can be instantiated for the propositional logic case, where the work of Katsuno and Mendelzon is restricted. Next we remind the result of Katsuno and Mendelzon in a formal manner.

A functional assignment of a propositional formula ψ to a pre-order $\leq_\psi$ that satisfies the above three properties is a *faithful assignment*. With $\psi \circ \mu$ we denote the result of the revision of ψ with a new information μ. Katsuno and Mendelzon proved the following theorem. A revision operator $\circ$ satisfies the AGM postulates formulated for propositional KBs (Knowledge Bases) if and only if there exists a faithful assignment that maps each ψ to a total pre-order $\leq_\psi$ such that $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.

It follows that Dalal's revision operator $\circ_D$ assures that knowledge revision satisfies the AGM postulates for KBs. That is, given a KB ψ, a new information μ and a total pre-order $\leq_\psi$ over $\mathcal{I}$ defined as in (2), the selection of the models of μ which are minimal w.r.t. $\leq_\psi$ produces a revision of ψ with the new information μ with minimal change.

What we are doing here is a characterization of a retrieval ranking in terms of knowledge revision. The pre-order $\leq_q$ that defines a ranking of documents having a single model with respect to q, has the property that its minimal models produce minimal change with respect to the query. That is, if we measure distances from documents to the query that lead to $\leq_q$, we know that we are doing it according with a rational well-defined notion of closeness. We are not saying that finding a faithful assignment is sufficient to have a good retrieval ranking. One can imagine a faithful assignment of propositional formulas ψ to total pre-orders $\leq_\psi^{\models}$ which mirrors classical entailment. The pre-order $\leq_\psi^{\models}$ is easily constructed from a measure of distance $Distce$, defined as: if $\mathcal{I} \in Mod(\psi)$, then $Distce(Mod(\psi), \mathcal{I}) = 0$; otherwise $Distce(Mod(\psi), \mathcal{I}) = 1$. This faithful assignment is not useful for IR purposes because it does not deal with partial matching. Dalal's distance was used in the PLBR model because it translates to the logical formulation of retrieval the behaviour of coordination level matching (Losada and Barreiro, 1999). The result of Katsuno and Mendelzon gives an additional justification to that decision. Moreover, we want to apply the result of Katsuno and Mendelzon in the case of documents and queries having several models and particular syntactic restrictions -section (3)-, or in the case of modifications of Dalal's distance with practical IR purposes -section (4)-.

Also we could consider the PLBR model under the paradigm of "probability of inference" for probabilistic IR (van Rijsbergen, 1986). This paradigm aims at computing the probability that a document represented by a formula d, logically implies a query represented by q. In order to compute this probability, several approaches were studied in (van Rijsbergen, 1992) taking into account their adequacy for IR. Although in the PLBR model we do not estimate the probability of the logical implication, the distance from d to q is computed. This distance is a measure of how far is d from logically implying q. Another important characteristic of the "probability of inference" approach is that the models under this paradigm do not explicitly cope with the concept of relevance. On the contrary, the classical probabilistic IR model is explicitly relevance-oriented or under the paradigm of "probability of relevance". This model assumes that some pairs query-document are judged relevant by the user. The Probability Ranking Principle (PRP) gives a theoretical justification of the classical probabilistic IR model: optimum retrieval is achieved when documents are ranked according to decreasing values of the probability of relevance with respect to a given query. A monotonically increasing relationship between probability of inference and probability of relevance is assumed. Therefore, in order to do ranking, it is sufficient to rank documents under the "probability of inference" paradigm. The application of the result of Katsuno and Mendelzon to logical modeling of IR is an epistemological, non-relevance oriented, theoretical justification of the ranking based on orders among logical interpretations. The justification provided by the PRP is stronger because it precisely characterizes the whole rank while in the logical setting the assessment is in terms of the connection between the measure of distance that produces the rank and knowledge revision with minimal change.

2.2 Document and queries having several models

When documents have several models we want to define the distance from the document to the query averaging the distances from individual models. If we use the average mean:

$$distance(d, q) = \frac{\sum_{m \in Mod(d)} Dist(Mod(q), m)}{|Mod(d)|}. \tag{3}$$

This distance can be transformed into a similarity measure:

$$BRsim(d, q) = 1 - \frac{distance(d, q)}{k}, \tag{4}$$

where k is the number of letters appearing in q.

The number of possible models of a propositional formula grows exponentially with the size of the alphabet. Therefore a direct implementation of the measures $Dist(Mod(q), m)$, $distance(d, q)$ and $BRsim(d, q)$ would be useless in practical IR systems where alphabets are large. In (Losada and Barreiro, 2001) a syntactic restriction for the document and query formulas, drives the design of algorithms that avoid the computation of all the models. Details of the algorithms can be found in the referenced work. In this paper, we study whether the measures of distance obtained with this polynomial time algorithms preserve the property of faithfulness or not.

3 Avoiding the computation of an exponential number of models

We impose the restriction that formulas d and q must be DNF formulas. A Disjunctive Normal Formula (DNF) has the form: $c_1 \vee c_2 \vee ...$ where each c_j is a conjunction of literals (also called *conjunctive clause*). A conjunctive clause has the form: $l_1 \wedge l_2 \wedge ...$ where each l_j is a literal (a propositional letter or its negation). A DNF formula is usually represented by the set of its conjunctive clauses. A conjunctive clause is usually represented by the set of its literals. So we first study the basic case of conjunctive clauses.

3.1 Conjunctive clauses for queries

Conjunctive clauses are compact representations of sets of models. This fact allows the computation of the distance $distqcl(dcl, qcl)$ from a document clause dcl to a query clause qcl skipping the enumeration of all the models. According to Dalal's distance, contradicting literals in query and document clauses should produce an increment of 1 to the distance, because all models of the document clause have the same truth value for that propositional letter, whereas all models of the query clause have the opposite truth value. A query literal not mentioned by the document clause should increase 0.5 the value of distance, because half of the models of the document will map the corresponding propositional letter into the same truth value than the query and half of the models will map it into the opposite truth value. Consider an alphabet $L = \{a, b, c, d, e\}$, a query $q = a \wedge b \wedge c$ and a document $d = \neg a \wedge b$. The computation of distance between the document and the query clause is: 1 (for the contradicting literal for letter a) plus 0.5 (for the query literal c that does not appear in the document clause). This procedure avoids to enumerate every model of the document. Otherwise, the distances from each individual model of the document to the query would have to be computed.

This distance, $distqcl$, induces a total pre-order $\leq_{qcl}$ over the set of document clauses:

$$dcl_i \leq_{qcl} dcl_k \quad iff \quad distqcl(dcl_i, qcl) \leq distqcl(dcl_k, qcl), \tag{5}$$

where dcl_i and dcl_k are document clauses and qcl is the query clause.

We propose to translate the definition of faithfulness to total pre-orders over the set of document clauses induced by query clauses and test if it is satisfied by $\leq_{qcl}$.

The functional assignment of propositional conjunctive clauses qcl to total pre-orders $\leq_{qcl}$ over the set of document clauses is faithful it it satisfies the following properties:

- (PC1) If $Mod(dcl_l) \subseteq Mod(qcl)$ and $Mod(dcl_m) \subseteq Mod(qcl)$, then $dcl_l <_{qcl} dcl_m$ does not hold, where $<_{qcl}$ is defined from $\leq_{qcl}$ in the usual way.
- (PC2) If $Mod(dcl_l) \subseteq Mod(qcl)$ and $Mod(dcl_m) \not\subseteq Mod(qcl)$, then $dcl_l <_{qcl} dcl_m$ holds.
- (PC3) If $qcl \equiv qcl'$, then $\leq_{qcl}=\leq_{qcl'}$.

That is,

- (PC1) For conjunctive clauses dcl_l, dcl_m and qcl, if $dcl_l \models qcl$ and $dcl_m \models qcl$, then $dcl_l <_{qcl} dcl_m$ does not hold.
- (PC2) For conjunctive clauses dcl_l, dcl_m and qcl, if $dcl_l \models qcl$ and $dcl_m \not\models qcl$, then $dcl_l <_{qcl} dcl_m$ holds.
- (PC3) For conjunctive clauses qcl, qcl', if $qcl \equiv qcl'$, then $\leq_{qcl}=\leq_{qcl'}$.

The purpose of the definition of these properties is to grasp the meaning of an hypothetical pre-order over the set of logical interpretations. The new definition of faithfulness is weaker than the original because the new setting is restricted. In the general case, individual interpretations are not accessible now; but for the particular case of documents having a single model, $\leq_{qcl}$ preserves the original properties:

- Fulfilment of PC1 implies that for the particular case of dcl_l and dcl_m having a single model, $\leq_{qcl}$ behaves like an pre-order over logical interpretations that satisfies PI1.
- The same argument applies to the fulfilment of PC2.
- The pre-order $\leq_{qcl}$ over the set of document clauses is independent of different logically equivalent formulations of qcl. Again, we can say that for the particular case of documents having a single model, $\leq_{qcl}$ behaves like an pre-order over logical interpretations that satisfies PI3.

It has still to be proved that the mapping of query clauses into total pre-orders $\leq_{qcl}$ is in compliance with the new definition of faithfulness:

- (PC1) If $Mod(dcl_l) \subseteq Mod(qcl)$ and $Mod(dcl_m) \subseteq Mod(qcl)$, since dcl_l, dcl_m and qcl are conjunctive clauses, literals of qcl are a subset of literals of dcl_l and a subset of literals of dcl_m, and $distqcl(dcl_l, qcl) = 0$, $distqcl(dcl_m, qcl) = 0$.
- (PC2) If $Mod(dcl_l) \subseteq Mod(qcl)$ then $distqcl(dcl_l, qcl) = 0$. If $Mod(dcl_m) \not\subseteq Mod(qcl)$, then there exists at least a model m of dcl_m that is not a model of qcl. All the models of qcl have the same truth value for the letters corresponding to the literals appearing in qcl. Because m is not a model of qcl, either m has to map at least one letter into true and there is a corresponding negative literal for that letter in the query clause, or m has to map at least one letter into false and there is a corresponding positive literal for that letter in the query clause. Let a be one of these propositional letters which make that m is not a model of qcl. If all the models of dcl_m map a into the same truth value than m, then there exists a contradicting literal, i.e. the positive literal a appears in qcl and the negative $\neg a$ in dcl_m or vice versa. If dcl_m has models with different truth values for the letter a, then a is not a literal appearing in dcl_m and, hence, it is the case of a query literal not mentioned by the document clause. Both cases imply that $distqcl(dcl_m, qcl) > 0$.
- (PC3) If $qcl \equiv qcl'$, since qcl and qcl' are conjunctive clauses, necessarily they have the same set of literals, implying that $\leq_{qcl}=\leq_{qcl'}$.

If documents are DNF formulas, their models are the result of the union of the set of models of the component conjunctive clauses. Given a document di and a query clause qcl, we can define a new measure of distance which averages the distances between the document clauses and the query clause, avoiding the computation of every model of di as it would be necessary in the definition of (3).

$$distanceqcl(di, qcl) = \frac{\sum_{di_{cl_j} \in di} distqcl(di_{cl_j}, qcl)}{|di|}, \tag{6}$$

where di_{cl_j} are the conjunctive document clauses of di and $|di|$ is the number of these conjunctive document clauses.

Now we can define a mapping of query clauses qcl to total pre-orders $\leq_{qcl}$ over the set of DNF documents, where $\leq_{qcl}$ is defined as:

$$di \leq_{qcl} dk \quad iff \quad distanceqcl(di, qcl) \leq distanceqcl(dk, qcl), \tag{7}$$

Since the pre-order is induced by the query clause, we write $\leq_{qcl}$ to denote it. It must not be confused with the pre-order of definition (5), which is a pre-order over the set of document clauses. For similar reasons to those presented before, we can give a new definition of faithfulness that has to take into account that now we have to work with DNF documents (sets of conjunctive clauses) instead of only a conjunctive clause per document. The mapping of query clauses qcl to total pre-orders $\leq_{qcl}$ over the set of DNF documents is a faithful assignment if it satisfies:

1. If $Mod(di) \subseteq Mod(qcl)$ and $Mod(dm) \subseteq Mod(qcl)$, then $di <_{qcl} dm$ does not hold, where $<_{qcl}$ is defined from $\leq_{qcl}$ in the usual way.
2. If $Mod(dl) \subseteq Mod(qcl)$ and $Mod(dm) \not\subseteq Mod(qcl)$, then $dl <_{qcl} dm$ holds.
3. If $qcl \equiv qcl'$, then $\leq_{qcl} = \leq_{qcl'}$.

In this case the properties are satisfied:

1. The set of models of a DNF document is the union of the sets of models of its constituent conjunctive query clauses. If $Mod(di) \subseteq Mod(qcl)$ and $Mod(dm) \subseteq Mod(qcl)$, then $Mod(di_{cl_j}) \subseteq Mod(qcl)$ and $Mod(dm_{cl_k}) \subseteq Mod(qcl)$ where j and k respectively range over the constituent conjunctive clauses of d_i and d_m. It follows that $distqcl(di_{cl_j}, qcl) = 0$ and $distqcl(dm_{cl_k}, qcl) = 0$, for all j and k. Finally, after averaging, $distanceqcl(di, qcl) = 0$ and $distanceqcl(dm, qcl) = 0$.
2. If $Mod(dl) \subseteq Mod(qcl)$, then $distanceqcl(dl, qcl) = 0$. If $Mod(dm) \not\subseteq Mod(qcl)$, then there exists at least a conjunctive clause dm_{cl_k} such that $Mod(dm_{cl_k}) \not\subseteq Mod(qcl)$. Therefore $distqcl(dm_{cl_k}, qcl) > 0$ and $distanceqcl(dm, qcl) > 0$
3. If $qcl \equiv qcl'$, since qcl and qcl' are conjunctive clauses, necessarily they have the same set of literals, implying that $\leq_{qcl} = \leq_{qcl'}$.

3.2 DNF queries

In the case of a DNF query, the PLBR model defines the distance from a document clause dcl to the query q_{dnf} as the distance to the closest query clause.

$$distqdnf(dcl, q_{dnf}) = min_{q_{cl_j} \in q_{dnf}} distqcl(dcl, q_{cl_j}), \tag{8}$$

where q_{cl_j} are the conjunctive query clauses of q_{dnf} and $distqcl$ was defined in section (3.1).

This measure of distance allows us to define $\leq_{q_{dnf}}$ that is a total pre-order over the set of document clauses:

$$dcl_l \leq_{q_{dnf}} dcl_m \quad iff \quad distqdnf(dcl_l, q_{dnf}) \leq distqdnf(dcl_m, q_{dnf}). \tag{9}$$

We can translate the definition of faithfulness to total pre-orders over the set of document clauses induced by DNF queries. A functional assignment of propositional DNF formulas q_{dnf} to total pre-orders $\leq_{q_{dnf}}$ is faithful it it satisfies the following properties:

1. If $Mod(dcl_l) \subseteq Mod(q_{dnf})$ and $Mod(dcl_m) \subseteq Mod(q_{dnf})$, then $dcl_l <_{q_{dnf}} dcl_m$ does not hold, where $<_{q_{dnf}}$ is defined from $\leq_{q_{dnf}}$ in the usual way.
2. If $Mod(dcl_l) \subseteq Mod(q_{dnf})$ and $Mod(dcl_m) \not\subseteq Mod(q_{dnf})$, then $dcl_l <_{q_{dnf}} dcl_m$ holds.
3. If $q_{dnf} \equiv q'_{dnf}$, then $\leq_{q_{dnf}} = \leq_{q'_{dnf}}$.

The mapping of DNF queries into total pre-orders $\leq_{q_{dnf}}$ over the set of document clauses is not faithful. It can be seen with a simple example. Let us consider the propositional alphabet $L = \{a, b\}$, the document clauses $dcl_1 = a$, $dcl_2 = b$ and $dcl_3 = \neg a \wedge b$, and the queries $q_{dnf} = (\neg a \wedge b) \vee (a \wedge b)$ and $q'_{dnf} = b$. Observe that $dcl_2 \models q_{dnf}$ (i.e. $Mod(dcl_2) \subseteq Mod(q_{dnf})$) but the distance of dcl_2 to each of the query clauses is 0.5 because there is a query literal in each clause (a, $\neg a$) not appearing in the document clause. Therefore $distqdnf(dcl_2, q_{dnf}) > 0$. Since $dcl_3 \models q_{dnf}$ also holds but $distqdnf(dcl_3, q_{dnf}) = 0$, the first property does not hold (i.e. $dcl_3 <_{q_{dnf}} dcl_2$ holds). The second property does not hold because $dcl_2 \models q_{dnf}$ and $dcl_1 \not\models q_{dnf}$ but $distqdnf(dcl_2, q_{dnf}) = 0.5$ and $distqdnf(dcl_1, q_{dnf}) = 0.5$, then $dcl_2 <_{q_{dnf}} dcl_1$ does not hold. The third property is not satisfied because $q_{dnf} \equiv q'_{dnf}$ but $dcl_2 \not<_{q_{dnf}} dcl_1$ and $dcl_2 <_{q'_{dnf}} dcl_1$.

If documents are DNF formulas, a new measure of distance, $distanceqdnf$, which averages the distances between document clauses and the DNF query can be defined as in (6). Since the mapping of query clauses into total pre-orders $\leq_{q_{dnf}}$ is not faithful, we can only conclude that $distqdnf$ and $distanceqdnf$ are good measures of distance just because they are based on $distqcl$ and Dalal's distance. In section (5) we reconsider this point and study some alternatives to skip the problems caused by DNF representation of queries.

4 Incorporating information about terms in the measure of distance

It is possible to incorporate information about the index terms in the PLBR logical model but keeping the propositional formalism. To do that, the measures of distance have to take into account that information, while the representations of documents and queries remain the same. There exist many alternative ways to modify the measures of distance revealing different heuristics or intuitions. Just one of them was implemented an evaluated in a small collection in (Losada and Barreiro, 2003a). In this section we check whether different heuristics preserve or not the properties studied in the previous sections. In order to illustrate the study, we suppose that we want to incorporate inverse document frequency (*idf*) into the basic PLBR model. We restrict the study to the case of documents having a single model because it is enough to show the important concepts. We first incorporate *idf* information for non-matching terms, then we study how to incorporate *idf* information for matching and non-matching terms.

4.1 Incorporating *idf* information in non-matching terms

Dalal's distance is a measure of distance between logical interpretations. Since, in the PLBR model propositional letters represent index terms, we can modify Dalal's distance to reflect the intuition that index terms do no contribute equally to that distance. We can define a new distance between logical interpretations where differing propositional letters contribute to the distance according to their *idf*. That is,

$$dist'(I, J) = \sum_{t} idf(t), \tag{10}$$

where t ranges over propositional letters whose interpretation is different in I and J.

For the case of documents having a single model we can define a new measure of distance $Dist'$ as in (1) but based on $dist'$ instead of $dist$, and a new preorder based in $Dist'$ as in (2). For this new mapping of queries to total pre-orders $\leq_q$ over the set of logical interpretations, the properties that constitute faithfulness are preserved for the following reasons:

1. If $m_{d_i}, m_{d_j} \in Mod(q)$, $Dist'(Mod(q), m_{d_i}) = 0$ and $Dist'(Mod(q), m_{d_j}) = 0$; then $m_{d_i} <_q m_{d_j}$ does not hold.

2. Consider now that $m_{d_i} \in Mod(q)$ and $m_{d_j} \notin Mod(q)$. Then $Dist'(Mod(q), m_{d_i}) = 0$ and $Dist'(Mod(q), m_{d_j})$ can be forced to be strictly greater than zero just using appropriate *idf* (for example, logarithmic) factors. With this proviso, $m_{d_i} <_q m_{d_j}$ and the second property holds.
3. $Dist'$ is also based on distances between logical interpretations. Therefore the total pre-orders are independent of different syntactic but logically equivalent formulations of the query.

4.2 Incorporating *idf* information for matching and non-matching terms

We can make use of different heuristics to incorporate *idf* information in matching and non-matching terms. In this section, we do not aim to exhaustively compare different heuristics with the purpose of experimentation and evaluation at a later time. We just want to show that different heuristics can be theoretically compared studying whether or not faithfulness is preserved.

Dalal's distance, $Dist$ and $Dist'$ are measures of distance or disagreement. For this reason only non-matching (differing) terms contribute to the measures. We can consider the importance of the set of matching terms as a factor modifying the measure of distance. We can define another modification of Dalal's distance, $dist''(I, J)$,

$$dist''(I, J) = \frac{dist'(I, J)}{1 + \sum_t idf(t)}, \tag{11}$$

where now t is each propositional letter whose interpretation is the same in I and J. Remember that $dist'(I, J)$ contains the contribution of differing terms.

We define a new measure of distance $Dist''$ as in (1) but based on $dist''$ instead of $dist$, and a new preorder based in $Dist''$ as in (2). For this new mapping of queries to total pre-orders $\leq_q$ over the set of logical interpretations, the properties that constitute faithfulness are preserved with the same proviso concerning the *idf* factor.

Although it could appear not natural, another alternative could be to consider that matching terms also increment the distance. We can consider a new measure of distance, $dist'''(I, J)$,

$$dist'''(I, J) = dist'(I, J) + \alpha \times \sum_t (1 - idf(t)), \tag{12}$$

where t is each propositional letter whose interpretation is the same in I and J and α is a tuning value in $[0, 1]$ measuring the importance of the contribution of matching terms. This is a slightly different heuristic of the approach implemented and evaluated in (Losada and Barreiro, 2003a). There, it showed good performance results as it was expected because considering the contribution of matching terms to distance is a discriminating mechanism. However, if we define $Dist'''$ as in (1) but based on $dist'''$ instead of $dist$, and a new preorder based in $Dist'''$ as in (2), the properties that constitute faithfulness are not preserved. It is easy to see that the contribution of matching terms causes that there can be models of the query at a distance strictly greater that zero, or equivalently, only a subset of models of the query are minimal in the order.

5 Discussion and further work

An important remark is that in knowledge revision only minimal models have to be incorporated in the revised theory. In retrieval we need the ranking, being a more complex problem. In (del Val, 1993) del Val presented several algorithms that implement Dalal's revision in polynomial time for several syntactic restrictions in the theory and new information. Following that techniques, an algorithm to do polynomial time Dalal's revision for a theory and new information in DNF was presented in (Losada and Barreiro, 2001) in order to implement retrieval situations.

Nie and other researches in (Nie et al., 1995) proposed the name of retrieval situations to comprise all the factors (with the exception of those included in the representation of documents and information needs) affecting relevance: semantic relations between terms, knowledge and intentions of the user, etc. They also proposed counterfactual conditional logic to model retrieval situations. Following this line of research, in (Losada and Barreiro, 2000, 2001) Dalal's BR operator was proposed to model the revision process of a retrieval situation S with the new knowledge represented by a document d. Therefore $(S \circ_D d)$ is a revision of S with the new information d that produces minimal change. Consequently, the relevance test for retrieval taking into account a retrieval situation S could be $(S \circ_D d) \models q$. Again, since this criterion is too strict, the PLBR model provides a measure of distance between the sets of models of the result of the revision $(S \circ_D d)$ and q, that can be transformed into a similarity measure.

Next we introduce several issues that compose an agenda of tasks to do for a better theoretically characterization of the PLBR model.

The first issue is the use of minimal DNF representations for queries (minimal w.r.t. the number of conjunctive clauses). An *implicant* D of a formula f is a conjunction of literals such that $D \models f$ and D does not contain two complementary literals, i.e. D has at least one model. Models of f are implicants in which each possible letter appears exactly once, as a positive literal it it is assigned true in the model, as a negative literal otherwise. A *prime implicant* of f is an implicant D such that for every other implicant D' of f, $D' \not\subset D$, where $D' \not\subset D$ is a relation between the sets of literals associated to the implicants. It is well known that a minimal DNF representation of a formula f is a disjunction of some of its prime implicants. In the example of section (3.2), $q_{dnf} = (\neg a \wedge b) \vee (a \wedge b)$ could be reduced to a minimal form $(q'_{dnf} = b)$ avoiding the problems caused by the non minimal form. However other results are of interest for IR. For monotone formulas (formulas that have only positive literals or, equivalently, the corresponding Boolean function is monotone) there exist an unique minimal DNF representation. Therefore the reduction of monotone queries to its minimal DNF equivalent can solve the above mentioned problems. But arbitrary formulas can generally have multiple minimal DNF representations, so it is necessary to study if these minimal representations preserve faithfulness. Independently of the theoretical study, the use of minimal DNF queries would be of interest also for allowing a more efficient computation of the distance to the query. Finally we must remember that obtaining a minimal DNF representation is a NP-complete problem and the real limitation of this result has to established in the specific context of the IR applications.

Another issue that needs further research is the incorporation of matching terms in the PLBR model. In section (4.2) we presented two possible ways. In $dist''$ matching terms act only as a modifying factor. In $dist'''$ the use of matching terms is counterintuitive and contradicts basic assumptions of the underlying formalism. Following the rationality of Belief Revision, models of a theory to be revised must be minimal in the pre-orders defined from the measures of distance to the theory. The use of matching terms as discriminating elements make differences among models of the theory. Reconciling these two issues is a pending task.

The PLBR model, since it is based on Dalal's distance, is a model based on distances among interpretations. Actually, the notion of document is a derived one. A document is formula to represent a set of models. For this reason it is natural to incorporate in the PLBR model global information, such as *idf*. Incorporating information associated to a particular document, such as term frequency in the document, without endangering the basic properties of the formalism, has still to be achieved.

Finally, we have to investigate if an epistemic characterization of the whole ranking is possible. Also the relationship between the justification based on knowledge revision with minimal change and other epistemic justifications, and the connection with the "probability of relevance" approach, have to be studied.

6 Conclusions

In this paper we have applied a result of Belief Revision to the logical modeling of Information Retrieval. We have seen how the relationship between knowledge revision with minimal change and orders between logical interpretations can be seen as a certain justification of the ranking produced in the PLBR model of IR. The study has been done taking account the syntactic restriction which allows efficient ranking in the PLBR model and the extension of the basic formalism which can deal with global term information such as inverse document frequency.

Acknowledgements

The work reported here was co-funded by "Ministerio de Ciencia y Tecnología" and FEDER funds under research project TIC2002-00947 (R&D program: "Tecnologías de la Información y las Comunicaciones"). The second author is supported in part by "Ministerio de Ciencia y Tecnología" and in part by FEDER funds through the "Ramón y Cajal" R&D program.

Bibliography

C.E. Alchourrón, P. Gärdenfors, and D. Makinson. On the logic of theory change: partial meet contraction and revision functions. *J. Symbolic Logic*, 50:510–530, 1985.

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

A. del Val. *Belief Revision and Update*. Ph. D. Thesis. Stanford University, Stanford, CA, 1993.

P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Bradford Books/MIT Press, Cambridge, MA, 1988.

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

D. E. Losada and A. Barreiro. Using a belief revision operator for document ranking in extended boolean models. In *Proc. SIGIR-99, the 22nd ACM Conference on Research and Development in Information Retrieval*, pages 66–73, Berkeley, USA, August 1999.

D. E. Losada and A. Barreiro. Retrieval situations and belief change. In *Proc. LUMIS'2000 DEXA'2000 Int. Workshop on Logical and Uncertainty Models for Information Systems*, pages 531–537, Greenwich, UK, September 2000. IEEE Computer Society Press.

D. E. Losada and A. Barreiro. A logical model for information retrieval based on propositional logic and belief revision. *The Computer Journal*, 44(5):410–424, 2001.

D. E. Losada and A. Barreiro. Embedding term similarity and inverse document frequency into a logical model of information retrieval. *Journal of the American Society for Information Science and Technology*, 54(4):285–301, 2003a.

D. E. Losada and A. Barreiro. Propositional logic representations for documents and queries: a large-scale evaluation. In *Proc. ECIR-2003, the 25th European Conference on Information Retrieval Research, Lecture Notes in Computer Science*, volume 2633, pages 219–234, Pisa, Italy, Abril 2003b.

D.E. Losada. *A logical model of information retrieval based on propositional logic and belief revision*. PhD Thesis. University of A Corunna, A Corunna, 2001.

J.-Y. Nie, M. Brisebois, and F. Lepage. Information retrieval as counterfactual. *The Computer Journal*, 38:643–657, 1995.

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

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

C.J. van Rijsbergen. Probabilistic retrieval revisited. *The Computer Journal*, 35(3):291–298, 1992.