A theoretic study of retrieval in the PLBR logical model of IR starting from a characterization of knowledge revision

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

[1] AIlab,
Departamento de Computación. Facultad de Informática.
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@usc.es

Abstract.

1 Introduction

In logical models of Information Retrieval (IR) documents and queries are represented as logical formulas. Relevance of a document relative to a query could be given by classical entaiment, 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 [11]. In the PLBR logical model of IR [6][8], 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 models of formulas d and q. For the purpose of retrieval, the measure of distance can be transformed into a similarity measure and finally the logical model can provide a ranking of documents given a query.

In the field of Belief Revision, distances between logical interpretations have been profoundly studied with the purpose of obtaining knowledge revision with minimal change. In particular, in this paper we retrieve an important result of Katsuno and Mendelzon which characterizes the revision schemes which satisfy the Gärdenfors rationality postulates with minimal change with respect to ordering among interpretations [5]. With this result, in section (2) we aim to indirectely characterize the ranking of documents with respect to a query.

Computation of distances between documents and queries represented as logical formulas is very much demanding from the computational point of view. In PLBR, a syntactic restriction allows to compute the similarity between documents and queries in polynomial time. In section (3) we revisit this computation and again we aim to theoretically characterize the ranking.

In sections (2) and (3) distances are obtained only 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* are introduced with the purpose of testing whether or not the properties presented in the previous sections are preserved. The paper discuss some open question in section 5 and ends with some conclusions.

2 Theoretical basis

2.1 Dalal's distance. Documents having single models

In PLBR documents and queries are represented with propositional formulas where propositional letter 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 [2] can be directly translated in an order of documents relative to a query 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 $\{T, F\}$. 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

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

Let $\mathcal{I}$ be the set of interpretations of 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 single models. The distance *Dist* defines a total preorder $\leq_q$ over the set of models of documents, as

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

where m_{d_i} and m_{d_j} are the single models of documents d_i and d_j.

Since 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 [5]:

1. 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.
2. If $m_{d_i} \in Mod(q)$ and $m_{d_j} \notin Mod(q)$, then $m_{d_i} <_q m_{d_j}$ holds.
3. 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 with q, if we use $\leq_q$ for ranking documents having single models m_d, we guarantee these three properties.

The first and second properties say that documents which completely satisify the information need represented with 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. These properties seem reasonable requirements for a logical formulation for retrieval although one could think that they could be of little practical interest because basically they only differentiate between models and non-models of the query. But, more important is that minimal change is obtained selecting the minimal models with respect to an ordering among interpretations that satisfy these properties as it was proved by Katsuno and Mendelzon in [5]. Actually, this notion of minimal change is established in a particular setting: the AGM postulates for Belief Revision. The AGM postulates [1][4] 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. A functional assignment of a propositional formula ψ to a pre-order $\leq_\psi$ that satisfies the above three properties is a *faithful assignment*. Katsuno and Mendelzon proved that 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 KB 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 satifisfies the AGM postulates for KBs. That is, given a KB ψ, a new information μ and a total pre-order $\leq_\psi$ defined as $\leq_q$ 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.

Of course what we are doing here is an indirect characterization of what is a good retrieval ranking in terms of knowledge revision. All what we say is that the pre-order $\leq_q$ that defines a ranking of documents having single models 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. And although the main interest of this paper is logical retrieval, this result can be applied to other IR tasks. Nie and other researches in [10] 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 [7][8] 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, PLBR 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.

2.2 Document and queries having sets of models

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

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

This distance can be transformed in similarity measure:

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

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 $Dist(Mod(q), m)$, $distance(d, q)$ and $BRsim(d, q)$ would be useless in practical IR systems where alphabets are large. In [8] a syntactic restriction for the document and query formulas, allows 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 if the measures of distance obtained with this polynomial time algorithms preserve the property of faithfulness and the implications for IR applications.

3 Avoiding the computation of an exponential number of models

A Disjunctive Normal Formula (DNF) ψ can be represented by a set of conjunctive clauses, $\psi = \{\psi_1, \psi_2, ..., \psi_n\}$, where each conjunctive clause ψ_i is a set of literals representing their conjunction. 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 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 every model represented in the document

clause has the same truth value for that propositional letter an every model represented in the query clause has 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 $L = \{a, b, c\}$, 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 the two models of the document, to compute the Dalal's distance between these two models and the query model and to average. Observe that in this example the query clause has only one model. But the procedure is the same when a query clause is a compact representation of several models. In this case this procedure avoids to enumerate the models of the document, to compute the Dalal's distance between these models and the closest query model and to average.

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

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

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

We can translate the definition of faithfulness to total pre-orders over the set of document clauses induced by query clauses and test if $\leq_{qcl}$ satisfy it. The functional assignment of propositional conjunctive clauses qcl to total pre-orders $\leq_{qcl}$ is faithful it it satisfies the following properties:

1. 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.
2. If $Mod(dcl_l) \subseteq Mod(qcl)$ and $Mod(dcl_m) \not\subseteq Mod(qcl)$, then $dcl_l <_{qcl} dcl_m$ holds.
3. If $qcl \equiv qcl'$, then $\leq_{qcl} = \leq_{qcl'}$.

That is,

1. 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.
2. 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.
3. For conjunctive clauses qcl, qcl', if $qcl \equiv qcl'$, then $\leq_{qcl} = \leq_{qcl'}$.

The reasons for proposing this definition of faithfulness are the following. The pre-order $\leq_q$ over the set of logical interpretations (documents having a single model) is based on Dalal's distance between logical interpretations. Minimal elements in this pre-order are those that show minimal change with respect to the query according with a well-known and rational notion of change. This is a consequence of the existence of a faithful assignment between queries q and

total pre-orders $\leq_q$ over the set of logical interpretations. Now, we are restricted to work with document and query clauses instead of the enumeration of the logical interpretations represented by these formulas, and what the new definition of faithfulness provides is a partial satisfaction of the original definition of faithfulness in the unrestricted setting. To see this, let us consider the following simple example. Let $L = \{a, b, c\}$ be a propositional alphabet, $qcl = a \wedge b$ a query clause and document clauses $dcl_1 = a \wedge b$ and $dcl_2 = a \wedge c$. The distances of these document clauses to the query clause are $distqcl(dcli_1, qcl) = 0$ and $distqcl(dcli_2, qcl) = 0.5$. Starting from $distqcl$ we can suppose that every model of $dcli_1$ is at distance 0 from the query and that every model of $dcli_2$ is at distance 0.5 from the query without enumerating and computing the distances from these models to the query. We denote interpretations by the set of letters that are mapped into true. The problem is that the interpretation $\{a, b, c\}$ is a model of $dcli_1$ and $dcli_2$. So we can say that $distqcl$ or $\leq_{qcl}$ are providing two pre-orders over logical interpretations. In one of the pre-orders the relations $\{a, b\} \leq \{a, b, c\}$, $\{a, b, c\} \leq \{a, b\}$, $\{a, b\} < \{a, c\}$ and $\{a, b, c\} < \{a, c\}$ hold; in the other $\{a, b\} < \{a, b, c\}$, $\{a, b\} < \{a, c\}$, $\{a, c\} \leq \{a, b, c\}$ and $\{a, b, c\} \leq \{a, c\}$ hold. If we could compute Dalal's distance between logical interpretations we would obtain pre-order where the first subset of relations hold. For this reason we say that the new definition of faithfulness provides a partial satisfaction of the original one. So, we are measuring distances between sets of interpretations as a whole, under a rational notion of closeness.

It has still to be proved that the mapping of query clauses into total pre-orders $\leq_{qcl}$ is faithful:

1. 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 $distqcl(dcl_l, qcl) = 0$, $distqcl(dcl_m, qcl) = 0$.
2. 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, m at least has to map a letter into true where there is a corresponding negative literal for that letter in the query clause, or m has to map a letter into false where 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 $distqcl(dcl_m, qcl) > 0$.
3. If $qcl \equiv qcl'$, since qcl and qcl' are conjunctive clauses, necessarily they have the same set of literals, implying that $\leq_{qcl} = \leq_{qcl'}$.

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:

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

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 $Mod(dm) \nsubseteq Mod(qcl)$, then $dl <_{qcl} dm$ holds.
3. If $qcl \equiv qcl'$, then $\leq_{qcl} = \leq_{qcl'}$.

In this case the properties are satisfied:

1. The models of a DNF document di, dm, are the elements of the result of 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 j and k going over their entire range of values. 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 $Mod(dm) \nsubseteq Mod(qcl)$, there exist at least a conjunctive clause dm_{cl_k} such that $Mod(dm_{cl_k}) \nsubseteq Mod(qcl)$. 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 distance from a document clause dcl to the query $qDNF$ is the distance to the closest query clause.

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

where q_{cl_j} are the conjunctive query clauses of $qDNF$ and $distqc$ 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:

$$dcl_l \leq_{qDNF} dcl_m \quad iff \quad distqDNF(dcl_l, qDNF) \leq distqDNF(dcl_m, qDNF). \tag{9}$$

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 propostional DNF formulas $qDNF$ to total pre-orders $\leq_{qDNF}$ is faithful it it satisfies the following properties:

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

The mapping of query clauses into total pre-orders $\leq_{qDNF}$ 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. $Mod(dcl_2) \subseteq Mod(qDNF)$) 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 $distqDNF(dcl_2, qDNF) > 0$. Since $dcl_3 \models q_{DNF}$ also holds but $distqDNF(dcl_3, qDNF) = 0$, the first property does not hold (i.e. $dcl_3 <_{qDNF} dcl_2$ holds). The second property does not hold because $dcl_2 \models q_{DNF}$ and $dcl_1 \not\models q_{DNF}$ but $distqDNF(dcl_2, qDNF) = 0.5$ and $distqDNF(dcl_1, qDNF) = 0.5$, then $dcl_2 <_{qDNF} dcl_1$ does not hold. The third property is not satisfied because $q_{DNF} \equiv q'_{DNF}$ but $dcl_2 \not<_{qDNF} dcl_1$ and $dcl_2 <'_{qDNF} dcl_1$.

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 $\leq_{qDNF}$ 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.

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 [9]. 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 single models because it is enough to show the important concepts. We first incorporate *idf* information in non-matching terms, then we study how to incorporate *idf* information in 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 PLBR 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 is each propositional letter whose interpretation is different in I and J.

For the case of documents having single models 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 faitfulness 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. If $m_{d_i} \in Mod(q)$ and $m_{d_j} \notin Mod(q)$, $Dist'(Mod(q), m_{d_i}) = 0$ 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 in 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, $dist''(I, J)$,

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

where t is each propositional letter whose interpretation is the same in I and J.

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 faitfulness are preserved with the same proviso concerning the idf factor.

Although it could be counterintuitive, another alternative could be to consider that matching terms also increment the distance. We can consider a new 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 [9]. 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 Further work

In this section we introduce several issues that compose an agenda of tasks to do for a better theoretically characterization of PLBR.

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 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 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 application.

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 [3] 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 [8] in order to implement retrieval situations (see section (2.1)).

Another issue that needs further research is the incorporation of matching terms in PLBR. 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 macthing terms as discriminating elements make differences among models of the theory. Reconciling this two issues is a pending task.

Finally, PLBR, 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 a is formula to represent a set of models. For this reason it is natural to incorporate in PLBR global information, such as *idf*. Incorporating information associated to a particular document, such as term frequency in the document, without endangering the formalism has still to be achieved.

6 Conclusions

References

1. 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.

2. 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.
3. A. del Val. *Belief Revision and Update*. Ph. D. Thesis. Stanford University, Stanford, CA, 1993.
4. P. Gärdenfors. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. Bradford Books/MIT Press, Cambridge, MA, 1988.
5. H. Katsuno and A.O. Mendelzon. Propositional knowledge base revision and minimal change. *Artificial Intelligence*, 52:263–294, 1991.
6. 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.
7. 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.
8. 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.
9. 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, 2003.
10. J.-Y. Nie, M. Brisebois, and F. Lepage. Information retrieval as counterfactual. *The Computer Journal*, 38:643–657, 1995.
11. C.J. van Rijsbergen. A non-classical logic for information retrieval. *The Computer Journal*, 29:481–485, 1986.