Análisis de la medida de distancia entre documentos y consultas en el modelo lógico de Recuperación de Información PLBR

Alvaro Barreiro
AIlab. Dept. Computación. Universidad de A Coruña
barreiro@udc.es
David E. Losada
GSI. Dept. Electrónica y Computación. Universidad de Santiago de Compostela
dlosada@dec.usc.es

Resumen

En este trabajo analizamos algunas propiedades de la medida de distancia entre documentos y consulta del modelo lógico de Recuperación de Información PLBR, proporcionando una justificación epistémica del *ranking* de documentos que es independiente de la relevancia de usuario explícita. Esta justificación téorica lo es en cuanto a la capacidad del *ranking* para producir revisión de conocimiento con cambio mínimo.

Palabras claves: *ranking*; similaridad; modelos; lógica; revisión de creencias.

1. Introducción

PLBR (*Propositional Logic* & *Belief Revision*) es un modelo lógico de Recuperación de Información (RI) basado en Lógica Clásica Proposicional y Revisión de Creencias (RC) [8][9][11]. En los modelos lógicos de RI, los documentos y las necesidades de información se representan con fórmulas lógicas. La relevancia de un documento a una necesidad de información pudiera obtenerse por medio de la consecuencia lógica clásica entre las fórmulas que los representan: el documento d es relevante a la necesidad de información expresada con la consulta q si y sólo si $d \models q$. Para su aplicabilidad a RI, el test $d \models q$ es demasiado estricto ya que sólo proporciona un criterio binario de relevancia [13]. Diferentes modelos lógicos han propuesto diferentes aproximaciones para considerar relevancia parcial. Una recopilación de tales trabajos puede encontrarse en [2]. En el modelo lógico PLBR, los documentos y necesidades de información se representan como fórmulas de lógica proposicional d, q. A partir de medidas de distancias entre el conjunto de interpretaciones que satisfacen d y q, se obtiene una medida de distancia entre el documento y la consulta. La medida de distancia se transforma en una medida de similaridad y finalmente se proporciona un *ranking* de documentos dada una consulta.

En el campo de RC se han propuesto diferentes medidas de distancia entre interpretaciones lógicas con el propósito de definir operadores de revisión de conocimiento con cambio mínimo. Los postulados AGM son principios de racionalidad que deben satisfacer los operadores de revisión [1][5]. Katsuno y Mendelzon caracterizaron el cumplimiento de los postulados AGM a partir de propiedades de órdenes entre interpretaciones lógicas [6]. En concreto, el cumplimiento de los postulados AGM de cambio mínimo es equivalente a la existencia de un orden entre interpretaciones lógicas que cumple ciertas propiedades. En la sección (2) aplicamos este resultado al ranking de documentos producido en el modelo PLBR. El objetivo es dar una justificación teórica al *ranking* independiente de los juicios de relevancia de usuario.

Recientemente, diferentes trabajos han resaltado la potencialidad de los modelos lógicos de RI [7][12]: mayor expresividad y generalidad, capacidad para tratar con conocimiento parcial y motivación conjunta de representación y razonamiento. Sin embargo, muy pocos de los modelos recopilados en [2] han sido implementados y evaluados en colecciones grandes (colecciones TREC). En particular, el cómputo de distancias entre interpretaciones lógicas asociadas a las fórmulas que representan documentos y consultas, tiene una demanda computacional que supone una limitación real para su aplicación en RI. Por esta razón, en el modelo PLBR se propone una restricción sintáctica de las fórmulas que representan documentos y consultas, que permite el cómputo de distancias en tiempo polinomial. Este tipo de restricciones sintácticas se habían aplicado en el campo de Revisión de Creencias con el objeto de obtener revisión en tiempo polinomial [4]. En nuestro caso el objetivo es obtener un *ranking* de documentos dada una consulta. De esta forma fue posible implementar y evaluar el modelo en colecciones de gran tamaño [8][10]. En la sección (2) consideramos documentos que tiene un solo modelo (interpretación lógica que satisface la fórmula que representa el documento). En la sección (3) estudiaremos la validez de los resultados presentados en la sección (2) bajo la restricción sintáctica que permite la operatividad del modelo PLBR. En la sección (4) presentamos diferentes heurísticas para incorporar *idf* (*frecuencia inversa de documento*), con el objeto de comprobar si las propiedades presentadas en las secciones anteriores se preservan. El trabajo continúa con las secciones de discusión y trabajo futuro, y finaliza con la presentación de conclusiones.

2. Fundamentos teóricos

2.1. Distancia de Dalal. Documentos con un modelo

En el modelo PLBR, los documentos y consultas son fórmulas proposicionales donde las letras representan términos de indexación. Para el caso en que los documentos sólo tienen un modelo, la distancia de Dalal [3] proporciona de forma directa un orden de los documentos que puede usarse como *ranking*.

Sea L un lenguaje proposicional. Una interpretación lógica es una función del conjunto de letras proposicionales de L en $\{T, F\}$. Dadas dos interpretaciones lógicas I, J, la distancia de Dalal, $dist$, es el número de letras proposicionales en las que difieren, esto es, el número de letras proposicionales cuya interpretación es diferente en I y J. Un *modelo* de una fórmula proposicional ψ es una interpretación que hace ψ cierto (T); con $Mod(\psi)$ denotamos el conjunto de modelos de ψ. Una medida de distancia entre el conjunto de modelos de una consulta q y un documento d que sólo tiene un modelo m_d puede definirse como $Dist(Mod(q), m_d) = min_{m_q \in Mod(q)} dist(m_q, m_d)$.

Sea $\mathcal{I}$ el conjunto de todas las interpretaciones lógicas que se pueden construir sobre L. Un *pre-orden* $\leq$ es una relación reflexiva y transitiva en $\mathcal{I}$. Un pre-orden es *total* si para todo $I, J \in \mathcal{I}$, $I \leq J$ ó $J \leq I$. Consideremos el caso de documentos que sólo tienen un modelo. Sin imponer más restricciones, un documento puede verse como su modelo asociado, y el conjunto de documentos posibles (o, equivalentemente, el conjunto de sus modelos asociados) es $\mathcal{I}$. La medida de distancia $Dist$ define un pre-orden total $\leq_q$ sobre el conjunto de modelos posibles:

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

donde m_{d_i} (m_{d_j}) es el modelo del documento d_i (d_j).

Ya que la medida de distancia $Dist$ está basada en la distancia de Dalal, el pre-orden total $\leq_q$ cumple las tres propiedades siguientes [6]:

- (PI1) Si $m_{d_i}, m_{d_j} \in Mod(q)$, entonces no se cumple $m_{d_i} <_q m_{d_j}$, donde $<_q$ se define a partir de $\leq_q$ de la forma habitual.
- (PI2) Si $m_{d_i} \in Mod(q)$ y $m_{d_j} \notin Mod(q)$, entonces se cumple $m_{d_i} <_q m_{d_j}$.
- (PI3) Si $q \equiv q'$, entonces $\leq_q = \leq'_q$.

La primera y segunda propiedades dicen que los documentos que satisfacen por completo la necesidad de información representada por la consulta deben ser minimales en el orden y estrictamente menores que los documentos que no satisfacen completamente la necesidad de información. La tercera propiedad establece un principio de irrelevancia con respecto a diferentes formulaciones sintácticas lógicamente equivalentes de la consulta. Aunque estas tres propiedades parecen requerimientos razonables para una formulación lógica de la tarea de recuperación, pudiera pensarse que son de poco intéres ya que básicamente sólo diferencian entre modelos y no modelos de la consulta. Sin embargo, debe considerarse que puede obtenerse revisión de conocimiento con cambio mínimo, seleccionando los modelos minimales de un orden entre interpretaciones lógicas que satisface estas tres propiedades, tal y como fue demostrado por Katsuno y Mendelzon en [6]. Esta noción de cambio mínimo lo es en un marco particular: los postulados AGM para revisión de creencias. Los postulados AGM [1][5] son unos principios de racionalidad que cualquier operador de revisión de conocimiento debe cumplir. Originalmente formulados en un marco general y filosófico (epistemológico), los postulados pueden formalizarse para el caso de lógica proposicional, en el que se sitúa el trabajo de Katsuno y Mendelzon. A continuación recordamos de manera formal el resultado de Katsuno y Mendelzon.

Un *mapping* funcional de fórmulas proposicionales ψ en pre-órdenes $\leq_\psi$ que satisface las tres propiedades mencionadas es una *asignación fiel*. Denotamos con $\psi \circ \mu$ el resultado de la revisión de ψ con la nueva información μ. Katsuno y Mendelzon demostraron el siguiente teorema:

Un operador de revisión $\circ$ satisface los postulados AGM formulados para BCs (Bases de Conocimiento) si y sólo si exite una asignación fiel que asigna a cada ψ un pre-orden total $\leq_\psi$ de forma tal que $Mod(\psi \circ \mu) = Min(Mod(\mu), \leq_\psi)$.

Se concluye a partir del teorema que el operador de revisión de Dalal, $\circ_D$, garantiza una revisión de conocimiento que satisface los postulados AGM. Esto es, dada una BC ψ, la nueva información μ y un pre-orden total $\leq_\psi$ sobre $\mathcal{I}$ definido de forma análoga a como hemos definido $\leq_q$, la selección de los modelos de μ que son minimales con respecto a $\leq_\psi$, produce una revisión de ψ con la nueva información μ con cambio mínimo.

El estudio de las propiedades que constituyen una asignación fiel es también una caracterización del *ranking* en términos de revisión de conocimiento. El pre-orden $\leq_q$ que define un *ranking* de documentos con respecto a q, tiene la propiedad de que sus modelos minimales producen cambio mínimo con respecto a la consulta. Por tanto, si tenemos medidas de distancia de los documentos a la consulta que llevan a $\leq_q$, sabemos que responden a una noción de proximidad racional y bien definida. Esto no quiere decir que la obtención de una asignación fiel sea suficiente para tener un buen *ranking*. Podemos imaginar una asignación fiel de fórmulas proposicionales ψ en pre-órdenes totales $\leq_\psi^{\models}$ que refleja la consecuencia lógica clásica. Este preorden se construye fácilmente a partir de una medida de distancia $Distce$, definida como: si $\mathcal{I} \in Mod(\psi)$, entonces $Distce(Mod(\psi), \mathcal{I})$ =0; en otro caso $Distce(Mod(\psi), \mathcal{I})$ =1. Esta asignación fiel no es útil para los propósitos de RI ya que no trata con relevancia parcial. La distancia de Dalal se usó para construir las medidas de distancia del modelo PLBR porque traslada a la formulación lógica de la recuperación de documentos la estrategia de *coordination level matching* [8]. La aplicación del resultado de Katsuno y Mendelzon da una justificación adicional a esta decisión. Además, en este trabajo pretendemos aplicar este resultado al caso de documentos y consultas con varios modelos y restricciones sintácticas particulares -sección (3)-, o al caso de modificaciones de la distancia de Dalal para tratar con información sobre términos -sección (4)-.

También es ilustrativo considerar el modelo PLBR bajo el paradigma de "probabilidad de inferencia" para RI probabilístico [13]. En este paradigma se pretende computar la probabilidad $P(d \models q)$. En [14] se estudiaron varias formas de obtener $P(d \models q)$. Aunque en el modelo PLBR no estimamos la probabilidad de la implicación lógica, obtenemos una medida de lo lejos que está d de implicar q. Otra característica importante de los modelos bajo el paradigma de "probabilidad de inferencia" es que no tratan explícitamente con el concepto de relevancia. Al contrario, el modelo probabilístico clásico de RI está desarrollado explícitamente sobre el concepto de relevancia, esto es, bajo el paradigma de "probabilidad de relevancia". En este modelo se supone que algunos pares documento-consulta se han juzgado como relevantes. El Principio de *Ranking* por Probabilidad (PRP) da una justificación teórica al modelo clásico probabilístico: la recuperación óptima se obtiene en un *ranking* donde los documentos se ordenan por valores decrecientes de probabilidad de relevancia con respecto a una consulta. Si además se supone que la relación entre la probabilidad de inferencia y la probabilidad de relevancia es monótona creciente, es suficiente con obtener el *ranking* bajo el paradigma de “probabilidad de inferencia”. La aplicación del resultado de Katsuno y Mendelzon al modelado lógico de la tarea de recuperación es una justificación epistemológica y no basada en relevancia del *ranking* producido a partir de órdenes entre interpretaciones lógicas. La justificación que proporciona el PRP es más fuerte porque caracteriza con precisión el *ranking*. En nuestro caso la justificación es sólo en cuanto que la medida de distancia que produce el *ranking* puede producir revisión de conocimiento con cambio mínimo.

2.2. Documentos y consultas con varios modelos

Cuando los documentos tienen varios modelos queremos definir la distancia de un documento a la consulta promediando las distancias de los modelos individuales. Si usamos la media:

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

Esta distancia pude transformase en una medida de similaridad:

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

donde k es el número de letras que aparecen en q.

El número de modelos posibles de una fórmula proposicional crece exponencialmente con el tamaño del alfabeto. Por tanto, una implementación directa de $Dist(Mod(q),\ m)$, $distance(d, q)$ y $BRsim(d, q)$ no es factible en sistemas de RI que tienen que tratar con alfabetos muy grandes. En [9] se diseñaron algoritmos que evitan el cómputo de todos los modelos, a partir de una restricción sintáctica en las fórmulas que representan los documentos y consultas. Los detalles de los algoritmos pueden encontrarse en el trabajo referenciado. En este trabajo estamos interesados en estudiar si las medidas de distancia obtenidas con los algoritmos de tiempo polinomial preservan la propiedad de fidelidad.

3. Evitando el cómputo de un número exponencial de modelos

Imponemos la restricción sintáctica de que d y q sean fórmulas DNF. Una Fórmula Normal Disyuntiva (*Disjunctive Normal Formula* - DNF) tienen la forma: $c_1 \vee c_2 \vee ...$ donde cada c_j es una conjunción de literales (o *cláusula conjuntiva*). Una cláusula conjuntiva tiene la forma: $l_1 \wedge l_2 \wedge ...$ donde cada l_j es un literal (una letra proposicional o su negación). Una fórmula DNF suele representarse por el conjunto de sus cláusulas conjuntivas. Una cláusula conjuntiva suele representarse por el conjunto de sus literales. Comenzaremos con el caso básico de cláusulas conjuntivas.

3.1. Consultas representadas con una clásula conjuntiva

Las cláusulas conjuntivas son representaciones compactas de conjuntos de modelos. Este hecho permite el cómputo de la distancia, $distqcl(dcl, qcl)$, de una clásula de un documento dcl a una consulta representada por una cláusula qcl, evitando la enumeración de todos los modelos. De acuerdo con la distancia de Dalal, los literales contradictorios en las cláusulas de documento y consulta debe producir un incremento de 1 en la distancia, porque todos los modelos de la clásula del documento tienen el mismo valor de verdad para esa letra proposicional, mientras que todos los modelos de la cláusula de la consulta tienen el valor de verdad opuesto. Un literal de la consulta que no aparece en la cláusula del documento debe producir un incremento de 0.5 en la medida de distancia, porque la mitad de los modelos del documento asignarán el mismo valor de verdad que la consulta a la letra proposicional correspondiente, y la otra mitad de los modelos del documento le asignarán el valor de verdad opuesto. Considere un alfabeto $L = \{a, b, c, d, e\}$, una consulta $q = a \wedge b \wedge c$ y un documento $d = \neg a \wedge b$. El cómputo de distancia entre el documento y la consulta es: 1 (por el literal contradictorio para la letra a) más 0.5 (por el literal c de la consulta que no aparece en el documento). Este procedimiento evita enumerar todos los modelos del documento. De otra manera, sería necesario computar las distancia de cada modelo individual del documento a la consulta.

La distancia $distqcl$ induce un pre-orden total $\leq_{qcl}$ sobre el conjunto de cláusulas de documentos:

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

donde dcl_i y dcl_k son clásulas de documentos y qcl es la cláusula de la consulta.

Para estudiar las propiedades que constituyen una asignación fiel en este marco de restricciones sintácticas, proponemos trasladar la definición de fidelidad a pre-órdenes totales sobre el conjunto de cláslulas de documentos, inducido por las cláusulas de consulta, y comprobar si esta nueva definición se satisface en $\leq_{qcl}$.

La asignación funcional de cláusulas proposicionales conjuntivas qcl a pre-ordenes totales $\leq_{qcl}$ sobre el conjunto de cláusulas de documentos es fiel si y sólo si satisface las propiedades siguientes:

- (PC1) Si $Mod(dcl_l) \subseteq Mod(qcl)$ y $Mod(dcl_m) \subseteq Mod(qcl)$, entonces no se cumple $dcl_l <_{qcl} dcl_m$, donde $<_{qcl}$ se define a partir de $\leq_{qcl}$ de la forma habitual.
- (PC2) Si $Mod(dcl_l) \subseteq Mod(qcl)$ y $Mod(dcl_m) \nsubseteq Mod(qcl)$, entonces se cumple $dcl_l <_{qcl} dcl_m$.
- (PC3) Si $qcl \equiv qcl'$, entonces $\leq_{qcl}=\leq_{qcl'}$.

Aunque en este caso las interpretaciones individuales no son accesibles, el propósito de la nueva definición de asignación fiel es trasladar al marco sintáctico restringido la definición en el marco de las interpretaciones lógicas. Esta definición es más débil que la original porque trata con un marco restringido donde solo se puede acceder a conjuntos de interpretaciones. La nueva definición captura el hecho que tratando sólo con conjuntos de interpretaciones, para el caso particular de documentos que tienen sólo un modelo, $\leq_{qcl}$ preserva las propiedades originales. Si PC1 y PC2 se cumplen, entonces para el caso particular de dcl_l y dcl_m que tienen sólo un modelo, $\leq_{qcl}$ se comporta como un pre-orden sobre interpretaciones lógicas que satisface PI1 y PI2. El pre-orden $\leq_{qcl}$ sobre el conjunto de cláusulas de documento es independiente de diferentes formulaciones lógicamente equivalentes de qcl. De nuevo, para el caso particular de documentos que sólo tienen un modelo, $\leq_{qcl}$ se comporta como un pre-orden sobre interpretaciones lógicas que satisface PI3.

La asignación de cláusulas de consulta en pre-órdenes totales $\leq_{qcl}$ cumple la nueva definición de asignación fiel. Por razones de espacio no daremos la prueba completa y sólo comentamos los pasos básicos de la misma. La prueba de PC1 y PC2 se basa en que para cualquier cláusula conjuntiva dcl_i, dcl_j y qcl, tal que $Mod(dcl_i) \subseteq Mod(qcl)$ y $Mod(dcl_j) \nsubseteq Mod(qcl)$, entonces $distqcl(dcl_i, qcl) = 0$ y $distqcl(dcl_j, qcl) > 0$. La prueba de PC3 se basa en que dos cláusulas conjuntivas equivalentes tienen el mismo conjunto de literales.

Si los documentos son fórmulas DNF, sus modelos son el resultado de la unión de los conjuntos de modelos de las cláusulas conjuntivas que componen el documento. Dado un documento di y una cláusula de consulta qcl, queremos definir una nueva medida de distancia que promedia las distancias entre las cláusulas de documento y la consulta, evitando el cómputo de cada modelo de di, lo que era necesario en la definición de la medida de distancia $distance$ dada en la sección (2.2).

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

donde las di_{cl_j} son las cláusulas conjuntivas componentes del documento di y $|di|$ es el número de estas cláusulas.

Ahora definimos una asignación de cláusulas de consulta qcl en pre-órdenes totales $\leq_{qcl}$ sobre el conjunto de documentos DNF, donde ahora $\leq_{qcl}$ se define como:

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

Ya que el pre-orden es inducido por la cláusula de consulta, lo denotamos por $\leq_{qcl}$. No debe confundirse con el pre-orden definido al principio de esta sección (3.1), que es un pre-orden sobre el conjunto de cláusulas de documento. Por razones similares a las presentadas anteriormente queremos dar una nueva definición de asignación fiel que tenga en cuenta que el orden es sobre documentos DNF (conjuntos de cláusulas conjuntivas) en lugar de sobre documentos con sólo una cláusula conjuntiva. La asignación de cláusulas de consulta qcl en pre-órdenes totales $\leq_{qcl}$ sobre el conjunto de documentos DNF es una asignación fiel si satisface:

- (PD1) Si $Mod(di) \subseteq Mod(qcl)$ y $Mod(dm) \subseteq Mod(qcl)$, entonces no se cumple $di <_{qcl} dm$, donde $<_{qcl}$ se define a partir de $\leq_{qcl}$ de la forma habitual.

- (PD2) Si $Mod(dl) \subseteq Mod(qcl)$ y $Mod(dm) \not\subseteq Mod(qcl)$, entonces se cumple $dl <_{qcl} dm$.
- (PD3) Si $qcl \equiv qcl'$, entonces $\leq_{qcl} = \leq_{qcl'}$.

En este caso las propiedades se satisfacen. De nuevo por cuestiones de espacio no damos una prueba completa y sólo los pasos básicos. La prueba de PD1 y PD2 se basa en que el conjunto de modelos de un documento DNF es la unión de los conjuntos de modelos de sus cláusulas constitutivas, y para cualquier documento DNF di, dj y cláusula conjuntiva qcl, tal que $Mod(di) \subseteq Mod(qcl)$ y $Mod(dj) \not\subseteq Mod(qcl)$, entonces $distanceqcl(di, qcl) = 0$ y $distanceqcl(dj, qcl) > 0$. La prueba de PD3 se basa en que dos cláusulas conjuntivas equivalentes tienen el mismo conjunto de literales.

3.2. Consultas DNF

En el caso de una consulta DNF, el modelo PLBR define la distancia de un documento con sólo una cláusula dcl a la consulta q_{dnf} como la distancia a la cláusula más próxima de la consulta.

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

donde las q_{cl_j} son las clásulas conjuntivas de q_{dnf} y $distqcl$ se definió en la sección (3.1).

Esta medida de distancia permite definir $\leq_{q_{dnf}}$, que es un pre-orden total sobre el conjunto de cláusulas de documento:

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

Se puede trasladar la definición de asignación fiel a pre-órdenes totales sobre el conjunto de cláusulas de documento inducidos por consultas DNF. Una asignación funcional de fórmulas proposicionales q_{dnf} en pre-órdenes totales $\leq_{q_{dnf}}$ es fiel si satisface las tres propiedades:

- Si $Mod(dcl_l) \subseteq Mod(q_{dnf})$ y $Mod(dcl_m) \subseteq Mod(q_{dnf})$, entonces no se cumple $dcl_l <_{q_{dnf}} dcl_m$, donde $<_{q_{dnf}}$ se define a partir de $\leq_{q_{dnf}}$ de la forma habitual.
- Si $Mod(dcl_l) \subseteq Mod(q_{dnf})$ y $Mod(dcl_m) \not\subseteq Mod(q_{dnf})$, entonces se cumple $dcl_l <_{q_{dnf}} dcl_m$.
- Si $q_{dnf} \equiv q'_{dnf}$, entonces $\leq_{q_{dnf}} = \leq_{q'_{dnf}}$.

La asignación de consultas DNF en pre-órdenes totales $\leq_{q_{dnf}}$ sobre el conjunto de cláusulas de documento no es fiel. Puede probarse a partir de un ejemplo. Considérese el alfabeto proposicional $L = \{a, b\}$, las cláusulas de documento $dcl_1 = a$, $dcl_2 = b$ y $dcl_3 = \neg a \wedge b$, y las consultas $q_{dnf} = (\neg a \wedge b) \vee (a \wedge b)$ y $q'_{dnf} = b$. Observe que $dcl_2 \models q_{dnf}$, esto es $Mod(dcl_2) \subseteq Mod(q_{dnf})$, pero la distancia de dcl_2 a cada una de las cláusulas de la consulta es 0.5 porque hay un literal en cada cláusula de la consulta (a, $\neg a$) que no aparece en la cláusula del documento. Por tanto $distqdnf(dcl_2, q_{dnf}) > 0$. Ya que también se cumple $dcl_3 \models q_{dnf}$ pero $distqdnf(dcl_3, q_{dnf}) = 0$, la primera propiedad no se cumple; esto es, se cumple $dcl_3 <_{q_{dnf}} dcl_2$. La segunda propiedad no se cumple porque $dcl_2 \models q_{dnf}$ y $dcl_1 \not\models q_{dnf}$ pero $distqdnf(dcl_2, q_{dnf}) = 0{,}5$ y $distqdnf(dcl_1, q_{dnf}) = 0{,}5$; entonces no se cumple $dcl_2 <_{q_{dnf}} dcl_1$. La tercera propiedad no se cumple porque $q_{dnf} \equiv q'_{dnf}$ pero $dcl_2 \not<_{q_{dnf}} dcl_1$ y $dcl_2 <_{q'_{dnf}} dcl_1$.

Si los documentos son fórmulas DNF, puede definirse una nueva medida de distancia, $distanceqdnf$, que promedia las distancias entre las cláusulas constitutivas de un documento y la consulta tal y como se hizo en la definición $distanceqcl$ en la sección (3.1). Ya que la asignación de cláusulas de consulta en pre-órdenes totales $\leq_{q_{dnf}}$ no es fiel, sólo podemos concluir que $distqdnf$ y $distanceqdnf$ son buenas medidas de distancia en tanto en cuanto están basadas en $distqcl$ y en la distancia de Dalal. En la sección (5) reconsideraremos este punto para tratar algunas alternativas a los problemas causado por las consultas DNF.

4. Incorporación de información sobre términos en la medida de distancia

Es posible incorporar información sobre los términos en el modelo lógico PLBR pero manteniendo el formalismo proposicional. Para hacerlo modificamos las medidas de distancia pero manteniendo el formalismo proposicional. Existen muchas alternativas en la forma de hacer esas modificaciones que reflejan diferentes heurísticas o intuiciones. En esta sección presentamos algunas de estas alternativas que preservan la propiedad de fidelidad. Restringimos el estudio al caso básico de documentos que tienen sólo un modelo porque es suficiente para ilustrar los conceptos importantes.

La distancia de Dalal es una medida de distancia entre interpretaciones lógicas. Ya que en el modelo PLBR, las letras proposicionales representan términos de indexación, podemos modificar la distancia de Dalal para reflejar la intuición de que los términos no contribuyen de la misma manera a la distancia. Podemos definir una nueva medida de distancia $dist'(I,J)$ entre interpretaciones lógicas donde las letras proposicionales para las que la interpretación es diferente contribuyen en función de su *frecuencia inversa de documento (idf)*: $dist'(I,J) = \sum_t idf(t)$, donde t recorre el rango de letras proposicionales cuya interpretación es diferente en I y J.

La distancia de Dalal, $Dist$ y $Dist'$ son medidas de distancia o desacuerdo. Por esta razón sólo los términos en los que difieren documento y consulta contribuyen a la distancia. Observe que este modelo es distinto al modelo de espacio vectorial de RI donde se obtiene directamente una medida de similaridad. Una ventaja del modelo PLBR es que podemos considerar diferentes heurísticas para incorporar también información de los términos comunes en la medida de distancia. Podemos definir otra modificación de la distancia de Dalal, $dist''(I,J)$:

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

donde ahora t recorre el rango de las letras proposicionales cuya interpretación es la misma en I y J. Recuerde que $dist'(I,J)$ contiene la contribución de los términos diferentes.

Para el caso de documentos que tienen sólo un modelo, podemos definir una nueva medida de distancia $Dist'$ ($Dist''$) de forma análoga a como se hizo la definición de $Dist$ en la sección (2.1), pero basada en $dist'$ ($dist''$) en lugar de $dist$, y un nuevo pre-orden basado en $Dist'$ ($Dist''$) de forma análoga a como se hizo la definición de $\leq_q$. Es fácil comprobar que para las dos heurísticas presentadas, las nueva asignaciones funcionales de consultas en pre-órdenes totales $\leq_q$ conservan las propiedades que definen una asignación fiel siempre que se asegure que $idf > 0$ para todo término, lo cual se consigue con las formas logarítmicas habituales con las que se usa esta información.

5. Discusión y trabajo futuro

En esta sección introducimos varias cuestiones que componen una agenda de tareas pendientes para una mejor caracterización teórica del modelo PLBR.

La primera es el uso de formas DNF minimales para las consultas (minimales con respecto al número de cláusulas conjuntivas). Un *implicante* D de una fórmula f es una conjunción de literal tales que $D \models f$ y D no contiene literales complementarios. Un *implicante primo* de f es un implicante D tal que para cualquier otro implicante D' de f, $D' \not\subset D$. Una representación minimal de una fórmula f es una disyunción de algunos de sus implicantes primos. En el ejemplo de la sección (3.2), $q_{dnf} = (\neg a \wedge b) \vee (a \wedge b)$ podría reducirse a su forma minimal ($q'_{dnf} = b$) evitando los problemas causados por la forma no minimal. Para fórmulas monótonas (fórmulas que sólo tienen literales positivos, o, equivalentemente, la función Booleana correspondiente es monótona), existe una única representación DNF minimal. Para una formúla arbitraria, existen en general varias formas minimales, lo que exigiría investigar si se preserva la propiedad de fidelidad en estas varias formas. En cualquier caso, el uso de formas minimales siempre sería interesante por cuestiones de eficiencia, aunque cabe recordar que el problema general de obtener una forma minimal DNF es NP-completo, por lo que debe establecerse la limitación real en el contexto específico de aplicación en RI.

Las medidas de distancia que manejamos, ya que están basadas en la distancia de Dalal, son distancias entre interpretaciones lógicas. Por esta razón la incorporación de *idf*, que es una información global e independiente del documento, se hizo de forma natural en el modelo PLBR. Debe aún investigarse como incorporar información como la de frecuencia de término en el documento, que en otros modelos permite mejorar la recuperación, sin dañar las propiedades básicas del formalismo.

6. Conclusions

En este trabajo hemos visto como la relación entre revisión de conocimiento con cambio mínimo y órdenes entre interpretaciones lógicas, permite una cierta justificación epistémica del *ranking* de documentos en la tarea de recuperación. El estudio se ha hecho teniendo en cuenta las restricciones sintácticas que posibilitan un *ranking* eficiente y considerando la extensión del formalismo básico que puede tratar con información global de los términos como la frecuencia inversa de documento.

Agradecimientos

Este trabajo está financiado por el Ministerio de Ciencia y Tecnología y FEDER, proyecto TIC2002-00947. El segundo autor está financiado en parte por el Ministerio de Ciencia y Tecnología y FEDER, bajo el programa Ramón y Cajal.

Referencias

[1] C.E. Alchourrón, 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] F. Crestani, M. Lalmas, and C.J (eds.) van Rijsbergen. *Information Retrieval, Uncertainty and Logics: Advanced Models for the Representation and Retrieval of Information.* Kluwer Academic, Norwell, MA, 1998.

[3] 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.

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

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

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

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

[8] 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.

[9] 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.

[10] 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 2003.

[11] 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.

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

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

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