Procesamiento del lenguaje Natural, nº 41 (2008), pp. 173-180 recibido 7-05-2008; aceptado 16-06-2008

Funciones de Ranking basadas en Lógica Borrosa para IR estructurada*

Ranking Functions based on Fuzzy Logic for Structured IR

Joaquín Pérez-Iglesias
NLP Group at UNED
joaquin.perez@lsi.uned.es

Víctor Fresno
NLP Group at UNED
vfresno@lsi.uned.es

Jose R. Pérez-Agüera
UCM
jose.aguera@fdi.ucm.es

Resumen: Con el auge de los lenguajes de marcado se ha desarrollado un nuevo escenario en el campo de la recuperación de información centrado en los documentos que presentan una estructura, y asumiendo que ésta puede ayudar en el proceso de recuperación; es lo que se define como IR estructurada. Los modelos clásicos de IR se han aplicado a este problema adaptando sus funciones de ranking al considerar los campos en los que se estructura un documento, y estas adaptaciones se han realizado siempre asumiendo una independencia estadística entre estos campos. Este hecho fuerza a la elección o estimación de unos coeficientes de empuje que determinen los diferentes pesos que se quiere dar a cada uno de los campos considerados. En este trabajo se presenta una nueva función de ranking para IR estructurada, basada en lógica borrosa, que trata de modelar mediante conocimiento experto la relación existente entre campos.

Palabras clave: IR estructurada, Lógica borrosa, Funciones de Ranking.

Abstract: With the increase in the use of mark-up languages, a new scenario has raised into the IR field; this new scenario is focused on structured documents, and has been defined as structured IR. The classic IR models have been extended in order to be applied to this new scenario. Generally these adaptations have been carried on by weighting the fields that form the document structure, and making the assumption of statistics independence between fields. This assumption force to an estimation of the different boosts applied to every field. In this paper a new ranking function for structured IR is proposed. This new function is based on Fuzzy Logic, and its main aim is to model through heuristics and expert knowledge the relations between fields.

Keywords: Structured IR, Fuzzy Logic, Ranking Functions.

1. Introducción

En la actualidad, la Web se ha convertido en la mayor fuente de información disponible en el mundo y el acceso a toda esta información se realiza fundamentalmente por medio de motores de búsqueda web, sistemas de recuperación de información (IR) que permiten obtener una lista ordenada de documentos como respuesta a una necesidad de información (o consulta) expresada como un conjunto de términos.

Tradicionalmente, las tareas de IR se han realizado sobre documentos sin ningun tipo de estructura, de forma que desde el punto de vista del procesamiento de los documentos, no era posible distinguir entre diferentes partes del documento. Este tipo de recuperación sobre texto plano se denomina recuperación ad-hoc, y ha constituído el eje central de la investigación teórica y práctica en IR durante los últimos 40 años. Ahora bien, con el surgimiento de la Web y su lenguaje HTML, por un lado, y con el auge de los lenguajes de marcado como XML, por el otro, se ha generado un nuevo escenario de recuperación donde los documentos sí presentan de forma

* Este trabajo ha sido subvencionado parcialmente por el proyecto QEAVis-Catiex (TIN2007-67581-C02-01) del Ministerio de Ciencia e Innovación.

ISSN: 1135-5948

explicita una estructura que puede ser de utilidad de cara a la IR.

En el caso del XML la investigación se ha centrado en la reimplementación de los modelos clásicos de IR, de forma que la estructura puede ser tenida en cuenta en las funciones de ranking que permiten generar la ordenación de los documentos. En este sentido la competición INEX, al igual que TREC en su momento para IR ad-hoc, ha diseñado un marco de pruebas específico para IR estructurada, el cual está siendo de gran utilidad para la comunidad de investigadores en IR. Como producto de las sinergias creadas a partir de INEX, se han diseñado distintas estrategias a partir de los modelos clásicos de IR destinadas a la recuperación sobre documentos con estructura.

Desde el punto de vista de la teoría de IR los dos hitos más emblemáticos en este sentido han sido la adaptación del modelo de espacio vectorial a la IR estructurada (Schlieder y Meuss, 2002) y la adaptación del esquema de ponderación BM25 al tratamiento de la estructura de los documentos, denominado BM25F(Robertson, Zaragoza, y Taylor, 2004).

En el caso de la recuperación de documentos HMTL, mucho del trabajo realizado sobre XML puede ser fácilmente adaptado al tratamiento de la estructura específica de las páginas web. El principal factor diferenciador consiste en que si bien en XML podemos contar con un marcado directo de la estructura de los documentos, en HTML necesitamos inferir la estructura en función de un marcado meramente descriptivo, centrado en la definición de la apariencia con la que se deben mostrar los documentos en el momento de su visualización por medio de un navegador. Esta particularidad del HTML provoca que sea necesario un trabajo previo de definición de heurísticas que nos permita establecer relaciones entre la información de visualización del documento y su estructura subyacente.

Si nos fijamos en las funciones de ranking aplicadas a la IR ad-hoc, podemos ver que siempre asumen una independencia estadística, primero entre términos en la consulta, y después entre los documentos de la colección. De este modo, estas funciones se formalizan como sumatorios a los términos de la consulta de funciones de pesado que generalmente consideran, por un lado, la frecuencia del propio término en el documento y en la consulta y, por otro, su frecuencia inversa de documento, o IDF.

Al considerar la estructura del documento, estas funciones de ranking deberán incluir una variable añadida: los campos que representan las diferentes partes del documento. El tratamiento de esta estructura y su integración en las funciones de ranking clásicas se realiza asumiendo una independencia estadística entre campos, de modo que ahora la función general de ranking se formulará como un sumatorio a los términos de la consulta y para cada uno de los campos que se estén considerando. Esta independencia entre campos fuerza irremediablemente a la estimación o elección de una serie de valores que representen los pesos de cada uno de los campos dentro de la función de ranking. Es decir, para cada colección deben fijarse inicialmente unos coeficientes, conocidos como factores de boost o factores de empuje, que representarán la importancia que se quiere dar a cada campo dentro del valor final de ranking. Por tanto, esta función que debe ajustarse a la colección, en el caso de considerar la estructura del documento, deberá añadir nuevas variables, por lo que el problema de estimación se hace más complejo cuantos más campos sean considerados.

Sin embargo, esta independencia entre campos no siempre debería ser asumida. Cuando se trata de encontrar la relevancia de un documento respecto de una consulta, los campos a combinar no siempre deberían tratarse de modo independiente, como sucede en las funciones de ranking que encontramos en la literatura. El motivo es que a menudo la relevancia en un campo toma verdadera importancia en unión con otro. Por ejemplo, podría suceder que el título de un documento tuviera una componente retórica, de modo que los rasgos presentes en él no ayudaran a describir adecuadamente el contenido del mismo. Por este motivo, los rasgos presentes en el título deberían tener mayor relevancia si, además, aparecieran en otros campos del documento. Este tipo de consideraciones no se contemplan con funciones de ranking basadas en combinaciones lineales de factores, tales como las funciones usadas dentro del modelo de espacio vectorial aplicado a IR estructurada o el esquema BM25F. En estos casos, si un término aparece en el título, la componente relativa a ese campo título tomará un valor que se sumará siempre en el

cálculo final del valor de ranking, independientemente del valor que tomen el resto de campos.

En este trabajo se presenta una función de ranking aplicable a problemas de IR estructurada que sí tiene en cuenta las posibles dependencias entre los campos. La idea sobre la que se construye esta función es que con un sistema fuzzy (o borroso) podemos ser capaces de combinar conocimiento y experiencia en un conjunto de expresiones lingüísticas que manejan palabras en lugar de valores numéricos (W.G.J. Howells, 2001). En nuestro caso, con el objetivo de combinar la información de cada uno de los campos en que se estructura el documento. De este modo, un sistema de reglas fuzzy podría suponer un mecanismo más apropiado si tratáramos de combinar la información de diferentes campos en problemas de IR estructurada.

El resto de este artículo se estructura como sigue. En la Sección 2 se describe la extensión del modelo de espacio vectorial para documentos con estructura. En la Sección 3 se presenta el modelo propuesto, así como las bases teóricas que lo soportan. A continuación, en la Sección 4 se detalla la experimentación realizada y se realiza un análisis de los resultados obtenidos. Finalmente, en el Sección 5 se extraen las conclusiones y se sugieren posibles trabajos futuros.

2. Modelo de espacio vectorial para IR estructurada

A día de hoy todos los modelos clásicos cuentan con una adaptación de su función de ranking a documentos con estructura. Dentro del modelo de espacio vectorial, la función de ranking puede asumir distintas formas, en función del esquema SMART (Salton y Buckley, 1965) que estemos utilizando. En este trabajo partimos de la siguiente ecuación, que calcula la relevancia de un documento respecto una consulta:

$$score(d,q) = \sum_{t \in q} idf_t \cdot tf_t^d \qquad (1)$$

donde t representa los términos contenidos en la consulta q, idf_t la frecuencia inversa del término en la colección y tf_t^d la frecuencia de aparición del término en el documento d.

Teniendo en cuenta el uso de campos dentro de un documento, el recuento de la frecuencia tf deberá considerar esta posibilidad. Así:

$$tf_t^d = \sum_{c \in d} w_c \cdot tf_{tc}^d \qquad (2)$$

donde c se corresponde con cada uno de los campos que contiene el documento, w_c es el peso relativo asignado a cada uno de ellos y se corresponde con el factor de empuje utilizado para aumentar o disminuir la importancia de un campo frente a los demás y tf_{tc}^d se corresponde con la frecuencia del término en el campo y en el documento.

De esta forma se aplica el modelo de espacio vectorial teniendo en cuenta el peso de cada una de las unidades de las que se compone el documento, así como la frecuencia de los términos dentro de cada una de ellas.

3. Modelo fuzzy aplicado a la IR estructurada

La lógica borrosa (fuzzy logic) se ha mostrado como un marco de trabajo adecuado para capturar el conocimiento experto humano, aplicando heurísticas a la hora de resolver la ambigüedad inherente a procesos de razonamiento cualitativo. Esta es una característica importante, habida cuenta de que el objetivo principal de este trabajo es encontrar una función de ranking que combine información de los diferentes campos en los que se estructuran los documentos.

La lógica borrosa se contruye sobre el concepto de variable lingüística, variable que puede tomar como valor palabras del lenguaje natural y que se define a partir de conjuntos borrosos. Por otro lado, la Teoría de Conjuntos Borrosos (Zadeh, 1965) se basa en el reconocimiento de que determinados conjuntos poseen unos límites imprecisos. Estos conjuntos están constituidos por colecciones de objetos para los cuales la transición de "pertenecer" a "no pertenecer" es gradual.

Un conjunto borroso permite describir el grado de pertenencia de un objeto a una determinada clase. Dicho grado de pertenencia viene descrito por una función de pertenencia $\mu_F : U \rightarrow [0,1]$, siendo U el universo de discurso. Si el objeto $u \in U$ entonces $\mu_F(u)$ es su grado de pertenencia al conjunto borroso F.

La arquitectura básica de un sistema de inferencia fuzzy como el que se empleará en este trabajo se compone de tres etapas de procesamiento: borrosificación de entradas, aplicación de las reglas de inferencia que

constituyen la base de conocimiento del sistema, y una desborrosificación que permite obtener el valor final de ranking. Las funciones de pertenencia de los conjuntos borrosos se establecen también a partir del conocimiento experto aportado al sistema.

Para expresar la base de conocimiento se necesita una serie de reglas IF-THEN que describan el comportamiento que deberá tener el sistema de la manera más precisa posible, y donde se aporte el conocimiento experto. En nuestro caso estas reglas reflejarán el conocimiento heurístico que se tiene acerca de la relación existente entre los diferentes campos considerados dentro de un documento. En el proceso de inferencia se interpretan las reglas IF-THEN, asociando uno o más conjuntos borrosos de entrada (antecedentes) con un conjunto borroso de salida (consecuente). En un sistema de control borroso como el planteado en este trabajo estas reglas contienen varias entradas, correspondiente a cada uno de los campos considerados, y una única salida que se corresponderá con el valor asignado por la función de ranking.

Una vez fijada la base de conocimiento, los antecedentes de las reglas se combinarán a través de operadores de Unión (OR) o Intersección (AND) que pueden implementarse de muy diversas formas. De este modo, las operaciones definidas entre conjuntos borrosos permiten combinar los valores lingüísticos, expresando la base de conocimiento como afirmaciones condicionales borrosas.

Tras la obtención de los consecuentes para cada regla IF-THEN, y tras una etapa de agregación, se obtiene un conjunto agregado final, entrada a la ultima etapa del controlador, la desborrosificación, que realiza una correspondencia entre un conjunto borroso de salida con un punto concreto (salida nítida o "crisp") que representa el valor de relevancia. Una explicación más detallada de este proceso puede encontrarse en (Fresno, 2006).

Es importante asegurarse de que el uso de conjuntos borrosos producirá una representación más realista que si se emplearan las mismas variables pero definidas de una forma nítida. Como ya se ha indicado, el uso de heurísticas para la combinación de información de campo hace pensar que esta situación se puede dar y que una combinación borrosa podrá capturar mejor la información asociada a los campos considerados que una combinación lineal. En la sección 4 se describirá en detalle el modelo propuesto.

4. Experimentación

La idea de este trabajo es desarrollar un sistema basado en lógica borrosa capaz de asignar un valor de relevancia a cada término dentro de un documento, de modo similar a como lo haría una función de pesado clásica como TF-IDF. Por tanto, para la aplicación del modelo propuesto se debe construir un sistema capaz de representar el tipo de documento sobre el que se evaluará el modelo, y dentro de él, deberán establecerse un conjunto de reglas con las que cuantificar la relevancia de los términos del documento en función de su frecuencia en los diferentes campos presentes en el documento. Así, el sistema deberá modelar la estructura de los documentos, aplicando después heurísticas para combinar los campos establecidos.

Por tanto, el primer paso en el desarrollo del modelo consiste en la definición de las variables lingüísticas y de los conjuntos borrosos que modelen la estructura del tipo de documento. Para el caso de páginas web, basadas en HTML, se han definido tres campos:

- **'Título'**, que incluye el texto que se encuentra dentro de la etiqueta *TITLE*. En algunas páginas web, los términos que aparecen dentro de esta etiqueta se pueden considerar como muy relevantes. En cierto modo el título de un documento se puede entender como un resumen en pocas palabras del contenido del documento. La variable lingüística que representa el campo título se compone de dos conjuntos borrosos: 'Bajo' y 'Alto', como se muestra en la figura 1.
- **'Enfatizado'**, denominamos enfatizado al conjunto de términos que han sido resaltados de forma explícita por el creador del documento y de forma que destacan respecto del resto del contenido del documento. Así consideraremos que los términos que aparecen con una grafía que les hace resaltar sobre el resto poseen una mayor relevancia. Este subconjunto incluye los términos que aparecen entre las etiquetas: *B ,EM, U, STRONG, BIG I, H1, H2, H3, H4, H5, H6, CITE, DFN, BLOCKQUOT*. La variable lingüística que representa el campo énfasis se compone de dos conjuntos borrosos: 'Bajo' y 'Alto', como se muestra en

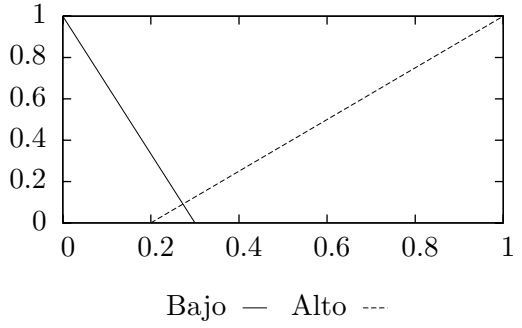

Figura 1: Variable Lingüística Título

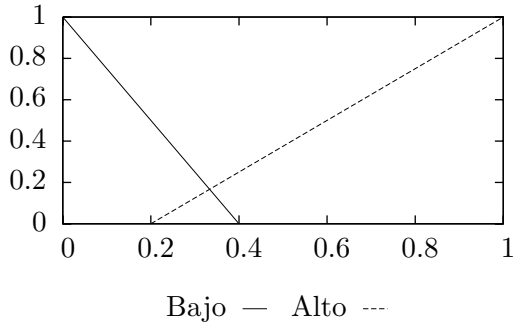

Figura 2: Variable Lingüística Enfatizado

la figura 2.

- **'Resto de contenido'**; este subconjunto incluye el resto de términos que aparezcan en el documento y la variable lingüística que lo representa se compone de tres conjuntos borrosos: 'Alta', 'Media' y 'Baja', como se muestra en la figura 3.

A continuación, se deben definir las funciones de entrada a cada una de las variables lingüísticas del sistema borroso. Estas funciones servirán para cuantificar el grado de

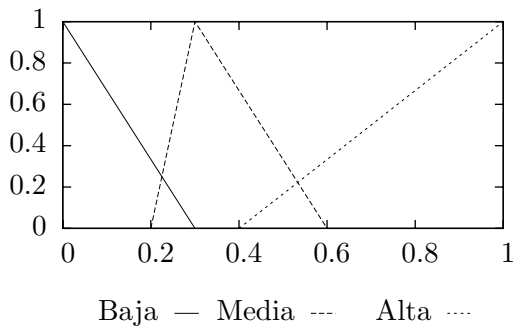

Figura 3: Variable Lingüística Resto

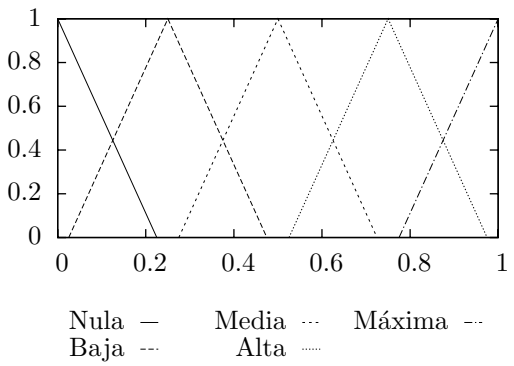

Figura 4: Variable Lingüística Resultado

pertenencia de un término a un conjunto borroso, y para ello se ha utilizado como base la frecuencia normalizada del término dentro de cada campo. Además, se ha decidido saturar la función tf resultante mediante el uso de la raíz cuadrada, de forma que se evite un crecimiento lineal de la relevancia de un término respecto a su frecuencia.

De este modo, esta función se define como:

$$f_c(t,d) = \sqrt{\frac{tf^d_{tc}}{tf^d_{max_c}}} \qquad (3)$$

donde $tf^d_{max_c}$ representa la frecuencia máxima de un término t en un campo c del documento d y para cada uno de los conjuntos borrosos.

A continuación, debe definirse la variable lingüística correspondiente a la salida del sistema. Esta variable se denominará 'Relevancia', representará la importancia de un término en el contenido de un documento, y se compondrá de cinco conjuntos borrosos:'Nula','Baja','Media','Alta' y 'Máxima', como se muestra en la figura 3.

Una vez definidas las variables lingüísticas y sus conjuntos borrosos, a continuación se presenta el conjunto de reglas que conforman la base de conocimiento y que se activarán a partir de las entradas a nuestro sistema, es decir, de las entradas a las variables lingüísticas de entrada y sus conjuntos borrosos. El total de reglas definidas es nueve y se muestran en el Cuadro 1.

El valor de relevancia total de un término se calculará a través del sistema borroso, de tal forma que combinen las frecuencias parciales de dicho término en cada uno de los campos, aplicando la base de conocimiento,

	Título		**Enfatizado**		**Resto**		**Relevancia**
IF	Alto	AND	-	AND	Alta	THEN	Máxima
IF	Alto	AND	-	AND	Media	THEN	Alta
IF	Alto	AND	Alto	AND	Baja	THEN	Alta
IF	Alto	AND	Bajo	AND	Baja	THEN	Media
IF	Bajo	AND	Alto	AND	Alta	THEN	Alta
IF	Bajo	AND	Bajo	AND	Alta	THEN	Media
IF	Bajo	AND	-	AND	Media	THEN	Baja
IF	Bajo	AND	Alto	AND	Baja	THEN	Baja
IF	Bajo	AND	Bajo	AND	Baja	THEN	Nula

Cuadro 1: Conjunto de reglas del sistema borroso global

y dando como resultado el valor de relevancia final del término para un documento.

Si denotamos nuestro sistema fuzzy como una función $Fuzzy(t,d)$ que estime la relevancia total del término t en el documento d, entonces

$$Fuzzy(t,d) = Fuzzy(f_{ti}(t,d), f_{en}(t,d), f_{re}(t,d))$$

teniendo que $Fuzzy(t,d) \in [0,1]$, donde f_{ti},f_{en},f_{re}, son las funciones que miden la relevancia parcial del término t en los campos, título, enfatizado y resto respectivamente.

De este modo, se define la relevancia total del siguiente modo: dada una necesidad informativa q compuesta por los términos $t_1, t_2, \ldots, t_n$ y una colección C formada por los documentos $d_1, d_2, d_3, \ldots, d_m$, el valor de relevancia R entre un documento d y una consulta q se calcula como

$$\sum_{t\ en\ q} Fuzzy(f_{ti}(t,d), f_{en}(t,d), f_{re}(t,d)) \quad (4)$$

Puede observarse que ahora ya no se tiene una combinación lineal entre los campos, como en el caso del VSM estructurado o BM25F, sino una función, en este caso borrosa, que combina globalmente la información del conjunto de campos seleccionados.

Dado que en el campo de la IR es bien sabida la utilidad del factor idf para corregir la información suministrada por la frecuencia de un término (Robertson, 2004), se añade el factor idf como corrector de la relevancia de un término en un documento en la ecuación 4.

La función de relevancia final que aplica nuestro sistema queda como sigue

$$\sum_{t\ en\ q} Fuzzy(f_{ti}(t,d), f_{en}(t,d), f_{re}(t,d)) * idf_t \quad (5)$$

siendo ésta la ecuación que representa el sistema propuesto.

4.1. Experimentación

En este apartado se describirán los diferentes experimentos realizados para la validación del modelo propuesto. Se comenzará describiendo la colección de pruebas utilizada, a continuación se detallará el modelo de espacio vectorial y sus extensiones con los que se realizará la evaluación de nuestra propuesta, para finalizar detallando los experimentos.

4.1.1. Colección Utilizada

La colección utilizada para la evaluación es un subconjunto de EuroGOV 2005, colección de páginas web en varios idiomas que se construyó para el *'track'* WebCLEF dentro del CLEF (Cross language Evaluation Forum) celebrado en el año 2005[1].

Para la experimentación realizada nos hemos ceñido al dominio de primer nivel UK, que contiene más de un 99 % de los documentos en inglés. El número total de documentos dentro de este dominio, que están en formato HTML, eliminando repeticiones y páginas vacías es de 58393. Por otro lado, el número de consultas y de los juicios de relevancia para este dominio es de 48.

4.1.2. Baseline

La función de relevancia utilizada para la evaluación de nuestro modelo se describe como: dada una necesidad informativa q compuesta por los términos $t_1, t_2, \ldots, t_n$ y

[1] http://www.clef-campaign.org/2005.html

una colección C formada por los documentos $d_1, d_2, d_3, \ldots, d_m$, el valor de relevancia R entre un documento d y una query q se calcula como:

$$R(q,d) = \sum_{t\ en\ q} tf_t^d * idf_t * norm(d) \qquad (6)$$

Extendiendo la ecuación 6 para documentos con una estructura formada por los campos $c_1, c_2, c_3, \ldots, c_k$.

$$R(q,d) = \sum_{t\ en\ q} idf_t * \sum_{c\ en\ d} tf_{tc}^d * norm(c) \quad (7)$$

donde

$$tf_{tc}^d = \sqrt{frecuencia \cdot w_c}$$
$$idf_t = 1 + \log\left(\frac{N}{df(t)\ +\ 1}\right)$$

siendo *frecuencia* el número de apariciones del término en el campo, N el número total de documentos de la colección C y w_c el factor de empuje asignado al campo c.

4.2. Medidas de Evaluación

A continuación se describen brevemente las medidas de evaluación que se utilizarán,

- MAP("Mean Average Precision"), trata de calcular una media de la precisión hallada a distintos niveles de cobertura. Más formalmente, sea $Rel = [d_1, d_2, \ldots, d_n]$ un conjunto de n documentos relevantes para una necesidad informativa q y sea R_j el conjunto de documentos recuperados antes de recuperar el j-ésimo documento relevante del conjunto Rel. MAP se define como la media aritmética del "Average Precision" sobre el conjunto Q.

$$MAP = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{n}\sum_{j=1}^{n} Precision(R_j)$$

- R-Prec("R-Precision"), se obtiene calculando la media aritmética de las distintas medidas de precisión a n, siendo n el número de documentos relevantes para la consulta q_i. Sea $Q = [q_1, q_2, \ldots, q_m]$ un conjunto de consultas, se define

$$R - Prec = \frac{1}{m} * \sum \frac{doc_{ret}}{n}$$

- MRR (Mean Reciprocal Rank), computa la media de los valores individuales alcanzados para cada consulta de la colección según la siguiente expresión:

$$MRR = \sum_{i=1}^{Q} \frac{1}{far_i}$$

siendo far el primer documento relevante recuperado para la consulta $q \in Q$.

4.3. Modelos Propuestos

Con el objetivo de validar el modelo propuesto, se han diseñado los siguientes experimentos

- **Baseline** En este caso, se evalúa sin tener en cuenta la estructura del documento o, lo que es lo mismo, asignando factores de peso equivalentes a cada campo.
- **VSM-I** En este caso se parte del sistema aplicado en el experimento anterior, pero ahora se asignan factores de ponderación a cada campo del siguiente modo: Título (5), Énfasis (2) y Resto (1). Se comprobó que estos valores representaban un máximo local.
- **VSM-II** A continuación, y partiendo como base del modelo anterior, se modificaron los factores de ponderación del siguiente modo: Título (10), Énfasis (5) y Resto (2,5); obteniéndose un nuevo máximo local.
- **Fuzzy** Finalmente se evalúa el sistema propuesto en este artículo, con las variables lingüísticas de entrada, los conjuntos borrosos y la base de conocimiento descritos en la sección 3.

Los resultados obtenidos se pueden observar en la siguiente tabla.

	Baseline	**VSM-I**	**VSM-II**	**Fuzzy**
MAP	0.455	0.462	0.474	**0.564**
R-Prec	0.368	0.368	0.388	**0.467**
MRR	0.466	0.479	0.490	**0.580**

Cuadro 2: Resultados

4.4. Análisis de Resultados

Los valores obtenidos para MAP y R-Prec nos permiten comprobar que la mejora obtenida gracias al uso de la estructura es consistente en términos de precisión y cobertura,

tanto en los enfoques analíticos como en la aproximación borrosa.

Por otro lado, la medida MRR muestra que no sólo se obtienen mejores resultados en todos los casos con el uso de la estructura sobre medidas clásicas de evaluación como MAP Y R-Prec, sino también sobre el orden en el que los documentos relevantes son recuperados, un factor de gran importancia en colecciones compuestas por un alto número de documentos como ocurre en el caso de la Web.

En el caso de VSM-I y VSM-II, se observa que la asignación de diferentes valores a los factores de empuje puede hacer variar la calidad de la recuperación, aunque en un grado bastante pequeño. La elección de los valores con los que se implementaron los modelos VSM-I y VSM-II se realizó tras una exploración exhaustiva del espacio de valores posibles, relativos a los factores de empuje de los campos 'Título', 'Enfatizado' y 'Resto'. Sin embargo, es reseñable el hecho de que la mejora conseguida por los métodos que realizan una combinación analítica, nunca suponen un aumento para las medidas de evaluación utilizadas por encima del 5 % sobre el baseline, mientras que la mejora conseguida por la función borrosa propuesta en este trabajo está en torno al 20 % para todas las medidas evaluación utilizadas.

4.5. Conclusiones y Trabajo Futuro

Al igual que se mostraba en (Robertson, Zaragoza, y Taylor, 2004), y considerando los resultados obtenidos mediante las distintas medidas de evaluación, se comprueba que el uso de la estructura de un documento mejora los resultados en problemas de IR. Este hecho puede observarse en el Cuadro 2, donde los resultados obtenidos por los modelos VSM-I y VSM-II mejoran los del baseline, único modelo de los evaluados que no utiliza estructura.

Si nos centramos en el análisis de los distintos métodos de recuperación que tienen en cuenta la estructura de los documentos, VSM-I, VSM-II y Fuzzy, podemos ver que el enfoque basado en lógica borrosa supone una mejora consistente, en función de las distintas medidas de evaluación consideradas, sobre los métodos basados en la simple combinación analítica de los campos. Esta mejora se manifiesta tanto en teŕminos de precisión como de cobertura. Asimismo, la ordenación de los documentos también se ve beneficiada por el enfoque borroso, mostrando que la interrelación entre campos es un problema a considerar cuando nos enfrentamos a una colección de documentos con estructura.

Como trabajo futuro destaca la comparación de la función borrosa con otros modelos de recuperación de información como BM25F, Modelos de Lenguaje y Divergence From Randomness. Por otro lado es necesario aplicar la función borrosa sobre otras colecciones de documentos estructurados que utilicen distintas formas de marcado al HTML, de forma que se pueda medir la viabilidad de nuestro enfoque sobre un rango amplio de esquemas de marcado. En este sentido, la colección INEX construída a partir de un marcado en XML de la Wikipedia conforma un marco de pruebas futuro de gran interés.

Bibliografía

Fresno, Víctor. 2006. *Representación Autocontenida de documentos HTML: una propuesta basada en combinaciones heurísticas de criterios.* Ph.D. tesis, Departamento de Ingeniería Telemática y Tecnología Electrónica, Universidad Rey Juan Carlos.

Robertson, Stephen. 2004. Understanding inverse document frequency: on theoretical arguments for idf. *Journal of Documentation*, 60.

Robertson, Stephen, Hugo Zaragoza, y Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. En *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, páginas 42–49, New York, NY, USA. ACM.

Salton, G. y C. Buckley. 1965. The SMART automatic documento retrieval system - an ilustration. *Communications of the ACM*, 8(6):391–398.

Schlieder, Torsten y Holger Meuss. 2002. Querying and ranking XML documents. *JASIST*, 53(6):489–503.

W.G.J. Howells, H. Selim. 2001. The autonomous document object (ado) model. En *Proceedings of the International Conference on Document Analysis and Recognition.*

Zadeh, L. A. 1965. Fuzzy sets. *Information and control*, 8:338–353.