JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 52(6):460-475, 2001

Modeling the Retrieval Process for an Information Retrieval System using an Ordinal Fuzzy Linguistic Approach

E.Herrera-Viedma

Artículo comentado
por
José Pino Díaz

(Alumno de Doctorado)

Resumen

El autor propone un Sistema de Recuperación de Información basado en un Modelo Difuso Lingüístico Ordinal.

Antecedentes

I.- Sistema de Recuperación de Información

Sistemas automáticos que se encargan del almacenamiento de documentos para su posterior recuperación ante las peticiones de los usuarios.

Los elementos principales de un SRI son:

- Base de Datos Documental
- Sistema de Formulación de Consultas
- Sistema de Evaluación de Consultas.

Los modelos que se utilizan para la formulación de consultas y la evaluación pueden ser:

- Modelo Booleano
- Modelo Vectorial
- Modelo Booleano Extendido
- Modelo Probabilístico
- Modelos Difusos

II.- Sistemas Difusos de Recuperación de Información

La Lógica Difusa (Teoría de Conjuntos Difusos) ha sido empleada para diseñar sistemas de recuperación de información.

Modelos Difusos de SRI son:

- SRI difuso básico Booleano
- SRI difuso con thesaurus de generalidad
- SRI difuso con thesaurus de sinonimia
- SRI difuso con consultas ponderadas con un solo peso
- SRI difuso con consultas ponderadas con múltiples pesos
- SRI difusos con consultas ponderadas numéricas, lingüísticas y lingüísticas ordinales.

En el Modelo Difuso básico se considera que cada documento es un conjunto difuso formado por la suma de los términos que lo indizan, cada uno de éstos caracterizado por su peso, con un valor entre 0 y 1.

Sistema de Recuperación de Información difuso lingüístico ordinal de Herrera-Viedma

El modelo de SRI difuso propuesto por Herrera-Viedma se caracteriza por:

- Consulta booleana ponderada con tres pesos sobre los términos; expresados los pesos por etiquetas pertenecientes a un conjunto ordenado de términos lingüísticos que se corresponden con diferentes valores de las variables lingüísticas Importancia y Relevancia.

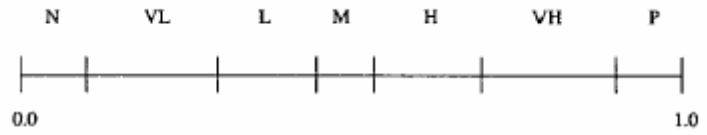

FIG. 1. A symmetrically distributed ordered set of seven linguistic terms.

$$S = \{s_0 = none, s_1 = very\ low, s_2 = low, s_3 = medium,$$
$$s_4 = high, s_5 = very\ high, s_6 = perfect\},$$

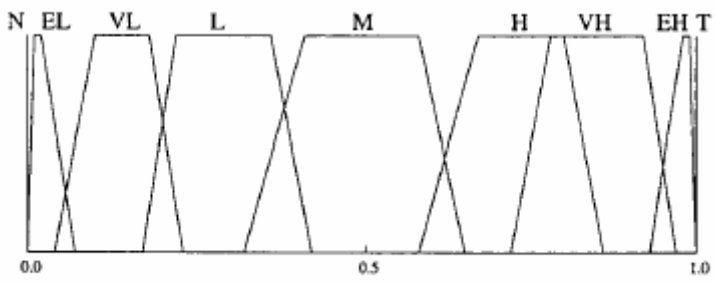

FIG. 3. A set of nine terms with its semantic.

Figure 3 *H*(*Importance*) = {*T*, *EH*, *VH*, *H*, *M*, *L*, *VL*, *EL*, *N*}

- Cada término de consulta es acotado por tres etiquetas lingüísticas.

$$(t_5, VH, M, VH) \quad (t_6, L, L, VL) \quad (t_7, H, L, H)$$

- La primera de ellas modula la semántica de umbral simétrica (etiqueta umbral a satisfacer por la etiqueta del término en la representación lingüística del documento)
- La segunda etiqueta modula la semántica cuantitativa (etiqueta que indica el número de documentos a recuperar para cada término)
- La tercera modula la semántica de importancia relativa entre los términos de la consulta.

- Sistema de Evaluación. Para proceder ala evaluación los documentos deberán estar representados mediante su forma lingüística (la función de translación *Label* asigna una etiqueta para cada peso de los términos). El sistema de evaluación sigue los siguientes pasos:

$$R_{d_1} = 0.7/t_5 + 0.4/t_6 + 1/t_7$$
$$R_{d_2} = 1/t_4 + 0.6/t_5 + 0.8/t_6 + 0.9/t_7$$
$$R_{d_3} = 0.5/t_2 + 1/t_3 + 0.8/t_4$$
$$R_{d_4} = 0.9/t_4 + 0.5/t_6 + 1/t_7$$
$$R_{d_5} = 0.7/t_3 + 1/t_4 + 0.4/t_5 + 0.8/t_9 + 0.6/t_{10}$$
$$R_{d_6} = 1/t_5 + 0.99/t_6 + 0.8/t_7$$
$$R_{d_7} = 0.8/t_5 + 0.02/t_6 + 0.8/t_7 + 0.9/t_8.$$

$$\downarrow$$

1. $R_{d_1} = H/t_5 + M/t_6 + T/t_7.$
2. $R_{d_2} = T/t_4 + M/t_5 + H/t_6 + VH/t_7.$
3. $R_{d_3} = M/t_2 + T/t_3 + H/t_4.$
4. $R_{d_4} = VH/t_4 + VL/t_6 + T/t_7.$
5. $R_{d_5} = H/t_3 + T/t_4 + M/t_5 + H/t_9 + M/t_{10}.$
6. $R_{d_6} = T/t_5 + EH/t_6 + H/t_7.$
7. $R_{d_7} = H/t_5 + EL/t_6 + H/t_7 + VH/t_8.$

1º) **Preprocesamiento de la consulta**: Consiste en poner la consulta en forma normal (DNF) con subexpresiones con no menos de dos átomos (cada átomo consta de un término seguido de sus tres etiquetas de ponderación)

$$q'^1 = (t_5, VH, M, VH) \vee (t_7, H, L, H),$$

$$q'^2 = (t_6, L, L, VL) \vee (t_7, H, L, H).$$

2º) **Evaluación de los átomos con respecto a la semántica de umbral simétrica**. La función de emparejamiento lingüístico g1 asigna mayor importancia a los términos de los documentos cuanto mayor es su

grado de satisfacción de la etiqueta de umbral semántico simétrico de los términos de la consulta.

$$g^1(s_a, s_b) = \begin{cases} s_0 & \text{if } s_b \geq s_{\mathcal{T}/2} \text{ and } s_a = s_0 \\ s_{t_1} & \text{if } s_b \geq s_{\mathcal{T}/2} \text{ and } s_0 < s_a < s_b \\ s_{t_2} & \text{if } s_b \geq s_{\mathcal{T}/2} \text{ and } s_b \leq s_a < s_{\mathcal{T}} \\ s_{\mathcal{T}} & \text{if } s_b \geq s_{\mathcal{T}/2} \text{ and } s_a = s_{\mathcal{T}} \\ s_{\mathcal{T}} & \text{if } s_b < s_{\mathcal{T}/2} \text{ and } s_a = s_0 \\ \text{Neg}(s_{t_1}) & \text{if } s_b < s_{\mathcal{T}/2} \text{ and } s_0 < s_a \leq s_b \\ \text{Neg}(s_{t_2}) & \text{if } s_b < s_{\mathcal{T}/2} \text{ and } s_b < s_a < s_{\mathcal{T}} \\ s_0 & \text{if } s_b < s_{\mathcal{T}/2} \text{ and } s_a = s_{\mathcal{T}} \end{cases}$$

$$i_1 = \text{Max}\left\{0, \text{round}\left(b - \frac{(b-a)}{\mathcal{H}}\right)\right\}$$
$$i_2 = \text{Min}\left\{\mathcal{T}, \text{round}\left(b + \frac{(a-b)}{\mathcal{H}}\right)\right\} \quad \mathcal{H} \in \{1, 2, 3, \ldots, b\}.$$

Para cada término de la consulta se obtiene un subconjunto de documentos, caracterizado cada uno de ellos por una etiqueta que indica el grado en el que ese documento satisface la etiqueta de umbral semántico.

$$M(t_5)_{VH} = VH/d_1 + H/d_2 + H/d_5 + T/d_6 + VH/d_7.$$

$$M(t_7)_{H} = T/d_1 + VH/d_2 + T/d_4 + H/d_6 + H/d_7.$$

$$M(t_6)_{L} = M/d_1 + M/d_2 + VH/d_4 + L/d_6 + VH/d_7.$$

3º) **Evaluación de los átomos con respecto a la semántica cuantitativa**. La función de emparejamiento lingüístico g2 reduce el número de documentos obtenidos por g1. La finalidad es obtener el mínimo número de documentos que satisfacen las restricciones lingüísticas expresadas por la etiqueta de semántica cuantitativa de los términos de la consulta.

$$g^2(\text{Supp}(M(t_i)_{w_i^1}), w_i^2, d_j) = \begin{cases} s_0 & \text{if } d_j \notin B^s \\ \mu_{M(t_i)_{w_i^1}}(d_j) & \text{if } d_j \in B^s \end{cases}$$

where B^s is the set of documents such that $B^s \subseteq \text{Supp-}(M(t_i)_{w_i^1})$, obtained according to the following algorithm:

1. $K = \#(\text{Supp}(M(t_i)_{w_i^1}))$.
2. REPEAT
   $M^K = \{s_q \in S : \mu_{s_q}(K/m) = \text{Sup}_v\{\mu_{s_v}(K/m)\}\}$.
   $s^K = \text{Sup}_q\{s_q \in M^K\}$.
   $K = K - 1$.
3. UNTIL $((w_i^2 \in M^{K+1})$ OR $(w_i^2 \geq s^{K+1}))$.
4. $B^s = \{d^{\sigma(1)}, \ldots, d^{\sigma(K+1)}\}$, such that $\mu_{M(t_i)_{w_i}}(d^{\sigma(h)}) \leq \mu_{M(t_i)_{w_i}}(d^{\sigma(l)}) \forall l \leq h$.

El resultado obtenido para cada término de la consulta es un subconjunto de documentos, acompañados cada uno de su etiqueta correspondiente, que satisfacen las semánticas de umbral y cuantitativa expresadas por el usuario en su ecuación de consulta para dicho término.

$$M(t_5)_{VH,VL} = T/d_6 \quad .$$

$$M(t_7)_{H,L} = T/d_1 + T/d_4.$$

$$M(t_6)_{L,L} = VH/d_4 + VH/d_7.$$

4º) **Evaluación de las subexpresiones y modelado de la importancia semántica**. Se utiliza la función g3 = LWD (operador lingüístico disyuntivo ponderado) o g3 = LWC (operador lingüístico conjuntivo ponderado). Esta función se aplica por separado a cada una de las subexpresiones en las que se divide la consulta. Para cada subexpresión y para cada uno de los documentos obtenidos en el paso anterior, la función empareja la etiqueta de la importancia semántica de cada término de la subexpresión de consulta con la etiqueta que el documento posee en el subconjunto de documentos obtenido para dicho término en el paso anterior.

$$q'^1 = (t_5, VH, M, VH) \vee (t_7, H, L, H),$$

$$q'^2 = (t_6, L, L, VL) \vee (t_7, H, L, H).$$

For example: $\mu_{M(q'^1)}(d_1) = \text{LWD}[(VH, N), (H, T)] = \text{MAX}\{\text{MIN}(VH, N), \text{MIN}(H, T)\} = H.$

Se obtienen subconjuntos de documentos, uno para cada subexpresión en la que se dividió la consulta. Los documentos que componen cada subconjunto, acompañados de su etiqueta, son aquellos que satisfacen las semánticas de umbral y cuantitativa y la importancia semántica expresada por el usuario en cada subexpresión en la que se ha dividido la consulta.

$$M(q'^1) = H/d_1 + H/d_4 + VH/d_6,$$

$$M(q'^2) = H/d_1 + H/d_4 + L/d_7.$$

5º) **Evaluación de la consulta completa**. Ahora, por último, se hace preciso evaluar la consulta completa; para ello tenemos el resultado de la evaluación de cada subexpresión de la consulta, obtenido en el paso anterior. Una última función de emparejamiento utilizada, la función g4 = MIN, combina los resultados de las evaluaciones de las subexpresiones para así obtener los documentos respuesta a la consulta formulada por el usuario.

$$M(q') = H/d_1 + H/d_4$$

Discusión

La semántica de umbral simétrico parte de la premisa de que un usuario cuando realiza una consulta a un sistema de recuperación de información quiere recuperar un conjunto de documentos con una presencia mínima aceptable o con una ausencia máxima aceptable del término de consulta. En un conjunto de etiquetas lingüísticas estaríamos, en el primer caso, a la derecha de f/2) o, en el segundo caso, a la izquierda de f/2.

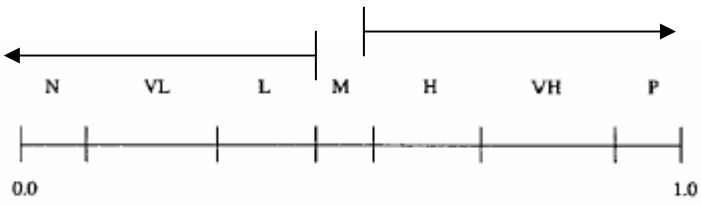

FIG. 1. A symmetrically distributed ordered set of seven linguistic terms.

No se plantea el caso de que un usuario desee recuperar los documentos que se sitúen en un intervalo dado.

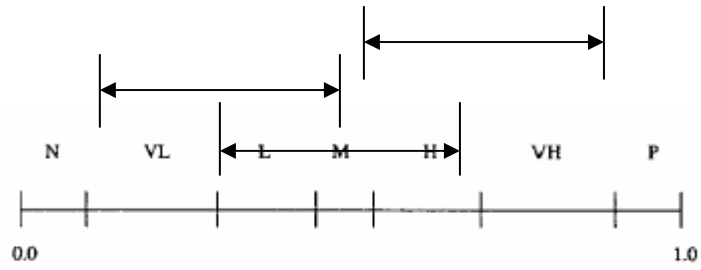

FIG. 1. A symmetrically distributed ordered set of seven linguistic terms.

Aunque se puede considerar que la premisa considerada por el autor no deja de ser un intervalo de documentos, entre N (0.0) y la máxima presencia deseada o entre la mínima presencia deseada y P (1.0).

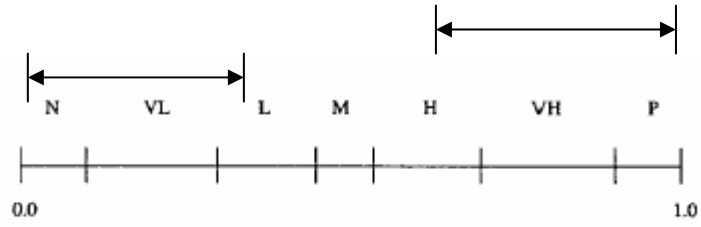

FIG. 1. A symmetrically distributed ordered set of seven linguistic terms.

En mi opinión se podría atender esta nueva exigencia del usuario (recuperar documentos dentro de un intervalo) mediante una función nueva de emparejamiento g1. Para ello, además, cada átomo de consulta debería expresarse con el término y cuatro etiquetas, las dos primeras indicarían los límites del intervalo (el límite inferior expresa la mínima exigencia a satisfacer y el límite superior, la máxima exigencia a satisfacer) y la tercera y la cuarta etiquetas serían las descritas en su artículo por el autor.

(ti, Sli, Sls, S3, S4)

La nueva función **g1(Sa, Sli-Sls)** debe contemplar todos los casos posibles.

- Sli = So
- Sli distinto de So
- Sls = Sf
- Sls distinto de Sf

Siendo Sli la etiqueta del límite inferior del intervalo, Sls la etiqueta del límite superior del intervalo, So igual a N (0.0), Sf igual a T (1.0) y Sa la etiqueta que acompaña al término en la representación del documento.

La función g1 dará mayor importancia (la etiqueta será de mayor importancia) a los documentos en los que la etiqueta que acompaña al término de consulta se sitúe en el intervalo [Sli, Sls] fijado por el usuario en su consulta. En definitiva se plantea una nueva semántica, la semántica de intervalo.

Bibliografía.

Castro Bonaño, J.M. (2002). Indicadores de Desarrollo Sostenible Urbano. Una Aplicación para Andalucía: Capítulos 4 (Métodos de Análisis Aplicados) y 5 (Análisis Empírico). Tesis Doctoral Facultad Ciencias económicas y Empresariales, Universidad de Málaga.

Galindo-Gámez, J. Conjuntos y Sistemas Difusos (Lógica Difusa y Aplicaciones). Operaciones con Conjuntos Difusos. Caracterización de Conjuntos Difusos: Entropía, Energía, Especificidad, Marcos de Conocimiento, Codificación/Decodificación. Relaciones Difusas. Lógica Difusa y Sistemas Basados en Reglas. Departamento de Lenguajes y Ciencias de la Computación Universidad de Málaga.

Herrera-Viedma, E. (2001). Modeling the retrieval process for an information retrieval using an ordinal fuzzy linguistic approach. JASIS, 52(6): 460-475.

Herrera, F., & Herrera-Viedma, E. (1997). Aggregation operators for linguistic weighted information. IEEE Transactions on Systems, Man, and Cybernetics, 27, 646–656.

Herrera, F., & Herrera-Viedma, E. (2000). Linguistic decision analysis: Steps for solving decision problems under linguistic information. Fuzzy Sets and Systems, 115, 67–82.

Herrera, F., Herrera-Viedma, & E., Martínez, L. (2000). A fusion approach for managing multi-granularity linguistic term sets in decision making. Fuzzy Sets and Systems 114, 43-58.

Herrera-Viedma, & E. Domínguez, J. (2001). SRI lingüístico difuso basado en el operador LOWA. Congreso ISKO España, Vol. 5.

Herrera-Viedma, E. (2001). Sistemas de Recuperación de Información Documental con T.I.A. Apuntes de la asignatura de Sistemas Expertos de Recuperación de Información de la Licenciatura de Documentación. Universidad de Granada.

http://www.imse.cnm.es/Xfuzzy/index_sp.html. Fuzzy Logic E-Book.