Uso de Lattices para la Recuperación de Términos

Liset Fraguela C., Héctor Jiménez S.
Facultad de Ciencias de la Computación
B. Universidad Autónoma de Puebla
Edif. 135. Ciudad Universitaria, Puebla, Pue. CP 72570. México
Tel. (01222) 229 55 00 ext. 7212 Fax (01222) 229 56 72
lisetfraguela@hotmail.com, hjimenez@fcfm.buap.mx
Gerardo Sierra M.
Grupo de Ingeniería Lingüística
Universidad Nacional Autónoma de México
Torre de Ingeniería, Ciudad Universitaria, México, DF
GSierraM@iingen.unam.mx

1. Introducción

2. Antecedentes

La Teoría de Conceptos Formales fue introducida por R. Wille [2]. A partir de este planteamiento se han realizado varias aplicaciones en diferentes ámbitos [1] [**?**]. La motivación para utilizar esta teoría es el hecho de que un concepto se representa por una pareja cuyo primer componente es el conjunto de ejemplares que denota el concepto y la segunda componente es el conjunto de características que tienen todos los ejemplares del concepto. Además, consideramos los trabajos realizados en cuanto a la aproximación de conceptos [4][3]. Daremos a continuación una breve introducción de las bases en las descansa el presente trabajo.

Para precisar las ideas de concepto formal deberán considerarse las cotas superiores e inferiores de un subconjunto S de un conjunto ordenado P, asimismo, si existen, la mínima cota superior y máxima cota inferior de S: $supS$ e $infS$, también denotadas por $\bigvee S$ y $\bigwedge S$, respectivamente.

Dado un conjunto ordenado P, si $x \wedge y$ y $x \vee y$ existen para toda pareja $x, y \in P$, llamamos a P *lattice* (o *retícula*). Un contexto (G, M, φ) está formado por un conjunto de objetos, G, un conjunto de rasgos, M, y una correspondencia de G en M, φ. Se define así un concepto como la pareja (A, B), donde $A \subset G$ es llamado extento del concepto, $B \subset M$ es llamado intento del concepto. Las parejas (A, B) deben cumplir:

$$A = B^{'} = \{g \in G \mid \varphi(g) = m, \forall m \in B\} \quad (1)$$

$$B = A^{'} = \{m \in M \mid \varphi(g) = m, \forall g \in A\} \quad (2)$$

Una relación de orden parcial puede ser construida sobre los elementos del *lattice* de conceptos. Dados (A_1, B_1) y (A_2, B_2) decimos que (A_1, B_1) es *más particular* que (A_2, B_2), $(A_1, B_1) \leq (A_2, B_2)$, si y sólo si $A_1 \subseteq A_2$, o equivalentemente $B_2 \subseteq B_1$. Con lo anterior, en un contexto puede demostrarse que para cada pareja (A_1, B_1) y (A_2, B_2) existe el *sup* e *inf*, es decir L es un *lattice*. En un *lattice* L, x es *inf-irreducible* si para cada $y, z \in L$, $x = y \wedge z$ implica $x = y$ o $x = z$; dualmente, un elemento x es *sup-irreducible* si para cualesquier $y, z \in L$, $x = y \vee z$ implica $x = y$ o $x = z$.

Por otro lado, el enfoque de conjuntos rugosos [6] considera el empleo de relaciones de equivalencia o tolerancia para hacer aproximaciones a conjuntos. En un sistema de información (G, M), se considera G, un conjunto de objetos y M un conjunto de atributos. En (G, M) se definen las aproximaciones inferior y superior de $X \subset G$ como: $\underline{X}_B = \{x \in G | [x]_B \subset X\}$ y $\overline{X}_B = \{x \in G | [x]_B \cap X \neq \emptyset\}$, respectivamente, donde $[x]_B$ es la clase de equivalencia que induce $B \subset M$, las cuales se determinan por $[x]_B = \{y | x.b = y.b, \forall b \in B\}$. En el caso de las aproximaciones que se han intentado manejar en la teoría de conceptos formales hay variantes sobre el tipo de relación que induce las clases de los elementos de G, permitiendo, por ejemplo, relaciones de tolerancia (reflexivas y simétricas). Para este trabajo se empleó la aproximación propuesta por Keyun Hu *et.al.* [4], como veremos enseguida.

Dado el contexto (G, M, φ) y correspondiente *lattice* de conceptos L, se define $gR = \{m \in M | \varphi(g) = m\}$ para $g \in G$. Así, puede definirse una relación binaria J en G como: $g_1 J g_2$ si y sólo si $g_1 R \subseteq g_2 R$, donde $g_1, g_2 \in G$. Claramente, J es una relación de orden parcial en G. Denotamos las cotas superiores de g como la clase de orden parcial $[g]$, específicamente $[g] = \{g^\star \in G : g J g^\star\}$, y con P el conjunto de parejas $([g], [g]')$. Si SI es el conjunto de todos los elementos *sup-irreducibles* de L puede probarse que $P = SI$ [4]. Similarmente, dado $Rm = \{g \in G | \varphi(g) = m\}$, para un contexto (G, M, φ), la relación binaria K sobre M definida como: $m_1 K m_2$ si y solo si $Rm_1 \subseteq Rm_2$, donde $m_1, m_2 \in M$, permite referirnos a la clase parcial de m con $[m]$: $[m] = \{m^\star \in M : m K m^\star\}$. De manera semejante a los elementos *sup-irreducibles* tenemos que si II es el conjunto de todos los elementos *inf-irreducibles* de L y Q es el conjunto de todos los pares $([m]', [m])$, entonces $Q = II$.

Dados $B \subseteq M$, la aproximación inferior (A-Inf) y la aproximación superior (A-Sup) de B con respecto a L son:

$$B_\bullet = \cap\{B \subseteq B^* \mid (B^{*\prime}, B^*) \in II, \} \quad (3)$$

y

$$B^\bullet = \cup\{B^* \subseteq B \mid (B^{*\prime}, B^*) \in SI.\} \quad (4)$$

Con estas aproximaciones es posible encontrar en el lattice de conceptos elementos que aproximan a conjuntos de objetos o a conjuntos de atributos. La aplicación realizada considera a la consulta del usuario como un conjunto de atributos y espera obtener un conjunto de objetos que aproximen a la descripción dada.

3. Aplicación en un diccionario onomasiológico

El problema es: dada la descripción de algún concepto obtener las palabras clave cuya definición se aproxime a la descripción dada. Nuestra propuesta consiste en utilizar el de aproximación en lattices para aplicarlo a Sistemas de Recuperación de Información. Para la aplicación particular en un diccionario onomasiológico, se realizó una prueba con un listado de conceptos (glosario) de la Física[1], obtenido de los trabajos que desarrolla G.Sierra /J.McNaught[**?**].

Se consideraron los siguientes datos de entrada al sistema: dos glosarios del área de Física[2], uno de ellos tiene una definición para cada vocablo y el otro tres definiciones; una colección de consultas; y también fue utilizado un diccionario de sinónimos.

Un glosario está formado por una lista de registros, a la vez, cada registro está compuesto por el nombre del concepto o vocablo (D), que sería la entrada del diccionario, y a continuación el conjunto de palabras clave (T), que describen dicho vocablo. El vocabulario utilizado estará formado por las palabras clave que describen a todos los vocablos del glosario, eliminando de éste las palabras cerradas, y truncando el resto con un truncador de Porter adaptado al español.

Las clases parciales pertenecientes a los elementos de II se obtienen de una tabla inversa donde se representan por cada palabra clave de la definición, T_i, todos los vocablos donde ésta aparece. Las clases que pertenecen a los SI, por otra parte, se obtienen de una tabla donde aparece cada vocablo D_j representado por las palabras clave que componen su definición.

Partiendo de estas dos tablas se realizan las aproximaciones superior e inferior dada una consulta. Para la aproximación superior se utiliza la tabla del SI y para la aproximación inferior la tabla del II. Estas aproximaciones se obtienen realizando las operaciones que a continuación se indican.

3.1. Aproximación Inferior por atributo ($T_\bullet$)

Más detalladamente se realiza lo siguiente. En primer lugar hay que partir de una tabla inversa (II), que no es otra cosa que un listado de palabras clave -las cuales conforman el vocabulario del *corpus*- y asociados a cada una de estas palabras los vocablos en cuya definición ocurren dichas palabras clave. Luego, dada una consulta, que a su vez debe tener las mismas características que el texto (es decir, truncado, sin palabras repetidas y sin palabras cerradas), para cada palabra clave que compone dicha consulta se determina en qué definiciones

[1] Recurso amablemente brindado por el Grupo de Ingeniería Lingüística de la UNAM.
[2] G.Sierra /J.McNaught[**?**].

de vocablos aparece. El próximo paso será traslapar el resultado anterior y la Aproximación Inferior (ecuación 3) estará dada por aquellos vocablos que contengan en su descripción a las palabras clave de la consulta. La ecuación 3 se reexpresa como:

$$T_{\bullet} = \cap\{T \subseteq T^* \mid (T^{*'}, T^*) \in II\} \tag{5}$$

3.2. Aproximación Superior por atributo ($T^{\bullet}$)

A continuación se explicará con más detalle el método empleado para obtener dicha aproximación. Primeramente se parte de una tabla de vocablos (SI) representado con sus respectivas definiciones, donde el texto se encuentra truncado, sin palabras repetidas y sin palabras cerradas. La aproximación superior (ecuación 4) es la unión del conjunto de vocablos en el que las palabras clave de su definición están contenidas en la consulta.

Teniendo en cuenta que en muy pocos casos se van cumplir las contenciones exactas en ambas aproximaciones, se introduce el criterio de *mayoría*. Para la Aproximación Superior, éste consiste en aceptar como válido un vocablo donde la mayoría (la mitad del total más uno) de las palabras clave que lo definen aparezcan en la consulta. En el caso de la Aproximación Inferior, se refiere a que la mayoría de las palabras clave de la consulta estén contenidas en la definición de un vocablo.

Puede suceder que no se utilicen en la consulta las palabras clave que aparecen en las definiciones de los vocablos sino algún sinónimo de éstas, y es por ello que en una de las pruebas se hizo la representación de las palabras clave, tanto de las que componen las definiciones como de la consulta, con un diccionario de sinónimos.[3]

3.3. Método directo

Esta primera prueba fue realizada con el contexto de conceptos del área de la Física. Las consultas utilizadas versaron sobre nociones de Mecánica y Cinemática dadas por profesores de Ingeniería, estudiantes de computación y, otras, extraídas del libro de L.Landau /E.Lifshitz[5], las cuales se presentan en la Tabla 1.

[3]Ofrecido amablemente por el Grupo de Ingeniería Lingüística de la UNAM, G.Sierra /J.McNaught[8].

	Consulta (truncadas)	**Resp. Correcta (*RC*)**
C1	tercer planet sistem sol	*Tierra*
C2	choqu conserv cantidad movimient lineal energi	*Choque elástico*
C3	aplic fuerz provoc desplaz mism direccion	*Energía mecánica*
C4	movimient sistem describ line rect	*Movimiento lineal* *Movimiento rectilíneo* (*Rectilíneo*)
C5	cambi veloc tiemp	*Aceleración* (*Aceleración angular* y *Aceleración centrípeta*)

Tabla 1. Consultas utilizadas en la primera prueba.

Para las respuestas correctas de la Consulta C5 se encuentran entre paréntesis los hipónimos de *Aceleración*. Mientras que la consulta C4 tiene tres posibles respuestas correctas, dos de ellas son sinónimos y la que se encuentra entre paréntesis es un hiperónimo de las anteriores.

En cuanto a los resultados de esta prueba se puede mencionar el caso especial de la consulta C2, la cual no tiene resultado en la Tabla 2, se puede decir que en ocasiones el usuario no utiliza una descripción como las que aparecen en el glosario y, por lo tanto, no se encuentra la respuesta deseada. Esta consulta en particular fue obtenida en el libro L.Landau /E.Lifshitz[5].

Con respecto a los resultados de A-Inf se puede decir que la evocación aumentó en algunos casos (particularmente en la A-Inf de la consulta C5) y que la precisión se mantuvo igual.

Los resultados obtenidos en esta prueba con respecto a las Aproximación Superiores (A-Sup) fueron los esperados. Para las tres primeras consultas no hubo resultados puesto que es difícil que las descripciones de los vocablos estuvieran contenidas en la consulta.

3.4. Uso de un diccionario de sinónimos

En esta prueba se utilizó un diccionario de sinónimos[4] (41555 Kb) para representar las palabras clave de la consulta y de las definiciones mediante una palabra "equivalente". El diccionario de sinónimos es una lista de registros, cada uno está compuesto por el sinónimo representante (un total de 523) y a continuación el listado de sinónimos de dicho vocablo. La sustitución de las palabras clave, tanto de la consulta como de las definiciones, por su sinónimo representante (en caso que lo tenga) se lleva a cabo de la siguiente forma:

1. Buscar la palabra clave de la consulta o de la definición dentro de las palabras en el listado de sinónimos.

[4]Se trata de una parte de un diccionario de sinónimos, amablemente proporcionado por el Grupo de Ingeniería Lingüística de la UNAM.

2. Si se encuentra la palabra clave en el listado de sinónimos entonces la palabra clave se sustituye por el sinónimo representante del diccionario de sinónimos.

3. Si no se encuentra la palabra clave en dicho listado de sinónimos se mantiene ésta en la representación.

Hay que considerar que los sinónimos de una palabra pueden estar referidos a diferentes contextos. Es por ello que en algunas consultas representadas con sinónimos no resultan "legibles". Teniendo en cuenta esta representación las consultas quedaron como se muestra en la Tabla 2.

	Consulta equivalente (truncada)	**Resp. Correcta (*RC*)**
`C1eq`	tercer planet métod sol	*Tierra*
`C2eq`	colision manten porción tiemp lineal potenci	*Choque elástico*
`C3eq`	entreg virtud produc mov mism sent	*Energía mecánica*
`C4eq`	tiemp métod describ líne orifici	*Movimiento lineal* *Movimiento rectilíneo* (*Rectilíneo*)
`C5eq`	vuelt acel period	Aceleración (*Aceleración angular* y *Aceleración centrípeta*)

Tabla 2. Consultas utilizadas en el método de representación por sinónimos.

4. Análisis de resultados

Al comparar los resultados de los métodos Directo y Sinónimos se podría decir que el uso de un diccionario de sinónimos para la representación, tanto de las consultas como de los glosarios (secciones 4.2.1 y 4.2.2), aumenta levemente la precisión con respecto al método directo. Esta mejora se intuía antes de aplicar este segundo método, sencillamente porque una misma consulta puede ser expresada por varios usuarios de manera diferente, y esta diferencia es en parte por el uso de sinónimos. Por ejemplo, para la consulta `C3`, en nuestro caso, la expresamos de la siguiente manera: *aplicación de una fuerza que provoca desplazamiento en una misma dirección*, sin embargo otro usuario podría expresar lo mismo deciendo: *Generación de una fuerza que produce un movimiento en el mismo sentido*. En ambos casos estaríamos buscando que nuestro sistema nos devolviera como respuesta correcta *Energía mecánica o Trabajo mecánico*.

Aunque se puede considerar que hubo una mejora entre los resultados del método de representación de palabras claves usando un diccionario de sinónimos con respecto al método directo, sus resultados de A-Sup y A-Inf son muy parecidos. Por ejemplo, A-sup aumentó la precisión utilizando el glosario de tres

definiciones para cada vocablo en la consulta `C2eq` y la cantidad de resultados aumentó en las A-Inf de la consulta `C5eq`, al usar los dos glosarios y en la A-Sup de la consulta `C4eq`. Aparte de estos cambios, los demás resultados se mantuvieron igual a los del método anterior. Nos preguntamos, sin embargo, por qué no mejoraron aún más estos resultados. Un elemento que influye en dichas pruebas está relacionado con las entradas del diccionario de sinónimos utilizado. Otra limitación con la que contamos es que el diccionario de sinónimos utilizado en estas pruebas es un diccionario general y, por otra parte, no se considera el contexto que puede tener la palabra. Hay que tener en cuenta que partimos de dos glosarios donde como máximo contabamos con tres definiciones por vocablo. Sin embargo, siempre habrá más de tres formas de describir dichos vocablos.

	Directo		**Sinónimo**	
Consulta	**A-Inf**	**A-Sup**	**A-Inf**	**A-Sup**
`C1`	1/2	0/0	1/2	0/0
`C2`	0/0	0/0	0/0	0/0
`C3`	1/1	0/0	1/1	0/0
`C4`	1/2	2/2	1/2	2/2
`C5`	3/15	2/2	3/21	2/2

Tabla 3. Comparación de resultados de las A-Inf y A-Sup de los tres métodos en cuanto a precisión (mediante la relación $P = R/NRes$); usando glosario de una definición por vocablo.

	Directo		**Sinónimo**	
Consulta	**A-Inf**	**A-Sup**	**A-Inf**	**A-Sup**
`C1`	1/2	0/0	1/2	0/0
`C2`	0/0	0/0	0/0	1/1
`C3`	1/1	0/0	1/1	0/1
`C4`	1/2	2/2	1/2	2/2
`C5`	3/19	2/2	3/27	2/2

Tabla 4. Comparación de resultados de las A-Inf y A-Sup de los tres métodos en cuanto a precisión (mediante la relación $P = R/NRes$); usando glosario de tres definiciones por vocablo.

5. Conclusiones

Referencias

[1] Hans-Hermann Bock (Ed.): *Classification, Data Analysis, and Knowledge Organization*, North-Holland, Amsterdam, 1991.

[2] Bernhard Ganter & Rudolph Wille: *Formal Concepts Analyisis*, Springer Verlag, 1999.

[3] R. Kent: "Rough Concepts Analysis: A sythesis of rough set and formal concept analysis", *Fundamenta Informaticae* 27(1996), 169-181.

[4] Keyun Hu, Yuefei Sui, Yuchang Lu, Ju Wang & Chunyi Shi: "Concept Approximation in Concept Lattice". PAKDD 2001, *LNAI* 2035, Springer Verlag, pp. 167-173 2001.

[5] L. Landau & E. Lifshitz: *Curso Abreviado de Física Teórica, Mécanica y Electrodinámica*. Editorial Mir Moscú, 1979.

[6] Z. Pawlak: *Rough Sets -theoretical aspects of reasoning about data*, Kluwer, 1991.

[7] C.J. van Rijsbergen: *Information Retrival*. University of Glasgow, pp. 114 - 117. Second Edition, 1999.

[8] G. Sierra & J. McNaught: "Design of an onomasiological search system: A concept-oriented tool for terminology". *Terminology*. Vol. 6 (1), 2000.

[9] Gerardo Sierra & John McNaught: "Natural Language System for Terminological Information Retrival". CICLing 2003, *LNCS* 2588, Springer Verlag, pp. 541-552, 2003.