Un Modelo de Representación basado en Sintagmas para Recuperación de Información

Miguel Rodríguez H., Héctor Jiménez S. & David Pinto A.

Fac. de Ciencias de la Computación
B. Universidad Autónoma de Puebla
C.U. 72570, Puebla, México
mrodrigues11@hotmail.com, hjimenez@fcfm.buap.mx, dpinto@cs.buap.mx

1 Introducción

Hoy en día la demanda por información sigue creciendo al grado que la dependencia de las organizaciones e individuos sobre este recurso es crucial. Internet se ha convertido en el principal proveedor de información y por lo tanto, juega un papel estratégico en la toma de decisiones. Sin embargo, es deseable por todos los usuarios tener acceso a la información que satisface mejor sus necesidades. Así, la construcción de buenos buscadores en esta abrumadora cantidad de textos, imágenes y otro tipo de medios representa un reto ineludible. Estamos frente al problema de hacer búsquedas de información con base en los conceptos que contiene un texto y no las palabras.

En este trabajo se experimentó con una representación de textos altenativa a la tradicional, la cual tomó en cuenta el contenido semántico de los textos con el propósito de ser utilizada en la recuperación de información. La idea principal en este trabajo es emplear el "sentido de un sintagma". Se implantó un sistema que representa un documento considerando los sintagmas que contienen los términos índices del docuemento. Las consultas se representan de manera semejante a la representación de un documento y la búsqueda se realiza en forma semejante a como lo hace el modelo booleano de recuperación de información.

A continuación se describe de manera sintética el contenido de las secciones que integran este trabajo. En la sección 2 se describen los trabajos relativos a relaciones de sentido, lo cual conforma la base sobre la que el presente fue realizado. En la sección 3 se plantea el método de representación de documentos con relaciones de sentido, extraídas de un *corpus* y de EuroWordNet(EWN), además, se plantea la forma en que se realizan consultas al sistema. En la sección 4 se describen las pruebas realizadas con diferentes conjuntos de consultas y diccionarios que proveen relaciones de sentido. Al final se presentan las conclusiones de este trabajo seguidas del apéndice con parte de datos y resultados tomados de las pruebas.

2 Antecedentes

Se han revisado varios trabajos que inciden en la noción de "asociación de términos". Esta idea se presenta en [11] vinculada con las áreas de Psicología y

Lingüística. Además, a través del concepto de asociación de términos es posible hacer un planteamiento para explotar la semántica de los términos que aparecen en los textos. Por ejemplo, se ha visto que este concepto es útil en la construcción de tesoros [8], idea que efectivamente se ha comprobado en nuestro ámbito dentro del Laboratorio de Recuperación de Información de la FCC [3]. Los términos de asociación de primer orden, que son términos que rodean a una palabra en un texto, constituyen el punto de partida para representar la palabra en cuestión. Con este enfoque el consumo de recursos no es dispendioso, ya que únicamente es necesario un *corpus*; cuyo tamaño y balance influirá ciertamente en los resultados. Así, resulta atractivo investigar esta representación que es "más semántica" que la tradicional. Para ello han de realizarse experimentos diversos que respalden el valor de este enfoque.

Entre los experimentos realizados con esta representación podemos citar a aquél que clasificó frases nominales que ocurren en las respuestas a preguntas abiertas [9]. En este caso, se representaron las respuestas usando como *corpus* un diccionario del español. Tal trabajo se sustenta en la propuesta que hace J. García [7] sobre el sentido de un sintagma. La propuesta se basa en la combinación que los seres humanos hacemos de las propiedades asociadas a los términos, de ahí la fórmula que se emplea para representar un sintagma (ver ec. 3). Otro trabajo en el que subyace la representación basada en términos de primer orden es el efectuado para obtener el extracto de un texto [12]. Cabe mencionar que hay otras propuestas realizadas por el grupo de RI de la FCC. Por ejemplo, la descrita en [5], la cual propone una forma alternativa de elección de índices. Los trabajos antedichos nos orientan a realizar una prueba de la representación con términos de primer orden y el sentido de un sintagma para conocer su efectividad en la recuperación de información.

3 Representación de Textos

La propuesta que se hace para representar un documento es tomar cada oración que contenga índices y utilizar la idea expresada por "el sentido de un sintagma" [7]. Por tanto, requeriremos de un medio que proporcione una aproximación a las relaciones de sentido de cada término. Hemos considerado un *corpus* para efectuar dicha aproximación. Veamos a detalle este planteamiento.

Partimos de aproximar las relaciones de sentido de una palabra x con las palabras que ocurran en los contextos de x. Tomamos como contexto de una palabra las oraciones donde ocurra la palabra. Para lograr lo anterior emplearemos un *corpus* C, y denotemos con C' el *corpus* C preprocesado; es decir una vez eliminadas las palabras cerradas y truncando sus términos. Definimos entonces las palabras y del contexto de x según la relación xVy, donde V es:

$$V = \{(x, y)|x, y \quad \text{ocurren en} \quad S, \quad \text{con} \quad S \in C'\}. \tag{1}$$

Debido a que los contextos son oraciones, definimos una oración como una tupla:

$$S = (t_1, t_2, \ldots, t_k),$$

donde $t_1, t_2, \ldots, t_k$ son términos índices de un documento D. Debemos representar primeramente los términos índice de las oraciones con base en sus contextos obtenidos mediante la ecuación 1. Por tanto, nuestra representación $\overline{t}$ para un término t sería:

$$\overline{t} = \{y|(t, y) \in V\}. \tag{2}$$

Un documento D_i está compuesto de oraciones S_{ij}, donde i representa al documento al cual pertenece la oración y j el número de oración. Así, podemos ver al documento D_1 como:

$$D_1 = (S_{11}, S_{12}, \ldots, S_{1n}).$$

De esta forma la j-ésima oración del i-ésimo documento se denotará como:

$$S_{ij} = (t_{ij1}, \ldots, t_{ijn}).$$

Una vez obtenida la representación de los términos índice que integran las oraciones de un documento, mediante la ecuación 2, podemos obtener la representación $\overline{S_{ij}}$ de una oración S_{ij}, realizando la intersección de todas las combinaciones de los términos representados y uniendo todas estas intersecciones. Esto se logra mediante la siguiente expresión[7][9]:

$$\overline{S_{ij}} = \bigcup \overline{t_{ijk}} \cap \overline{t_{ijl}}, \quad k \neq l \tag{3}$$

Finalmentente, el documento se representará por la unión de todas las oraciones representadas:

$$\overline{D_i} = \bigcup \overline{S_{ij}}. \tag{4}$$

Una consulta q al sistema deberá hacer uso de una representación que considere el significado de las consultas, al igual que se hizo con los documentos. Para realizar una consulta $q = (q_1, q_2, \ldots, q_k)$, representamos los términos $q_1, q_2, \ldots, q_k$ de q, usando $\overline{q_1}, \overline{q_2}, \ldots, \overline{q_k}$. Esto es equivalente a representar una oración usando la ecuación:

$$\overline{q} = \bigcup (\overline{q_i} \cap \overline{q_j}), \quad i \neq j \tag{5}$$

Finalmente, los documentos más relevante se obtienen utilizando un criterio derivado del modelo booleano: un documento representado que comparta mayor número de términos con la consulta representada tendrá mayor relevancia.

4 Pruebas Realizadas

Los datos con los que se realizaron las pruebas consistieron de un glosario de física de 255 vocablos[1] y tres definiciones por vocablo. Para cada vocablo del glosario, la primera definición se almacenó en un archivo que se usó como *corpus*. La segunda defininición se almacenó en otro archivo usado como colección de

[1] Recurso amablemente proporcionado por el GIL-II-UNAM, a travé de G. Sierra M.

documentos; se tomó como documento cada definición para formar el archivo de la colección. La tercera definición del vocablo fue almacenada en un archivo que constituyó una colección de consultas al sistema para una de las pruebas efectuadas. Todos estos archivos fueron preprocesados en la forma que se ha mencionado.

Se realizaron dos experimentos. El primero empleó la colección de 255 preguntas descrita en el párrafo anterior. El segundo, una colección de 16 preguntas dadas por estudiantes de Ingeniería Mecánica de la BUAP[2]. Ambos utlizaron como fuente para proveer relaciones de sentido el *corpus* (255 definiciones de físisca) y relaciones provistas por EuroWordNet (EWN). En el primer caso se utilizaron relaciones alternativas extraídas de EWN: sinonimia, hiponimia, hiperonimia, y la combinación de todas ellas.

4.1 Consultas del Glosario

Para las 255 consultas se realizaron las pruebas con las relaciones de sinonimia, hiponimia, hiperonimia y los tres tipos de relaciones contenidas en un sólo archivo para la representación que exige el método. Los resultados promediados se muestran en la tabla 2.

Como puede observarse hay una mejora sustancial usando como soporte para la representación de sintagmas solamente empleando las relaciones de sinonimia. Aunque se nota, también mejor desempeño al usar relaciones de hiponimia con respecto a todas y las de hiperonimia. Es importante notar que el empleo de un *corpus* para proveer relaciones de sentido tiene el desempeño más bajo. La alta precisión en el nivel de evocación uno, al usar sinónimos, puede observarse en la columna "Posición de la Respuesta Correcta".

Todas estas observaciones deben aún comprobarse con mayor exhaustividad. Está por comprobarse, por ejemplo, si eligiendo términos de los contextos del *corpus* y filtrándolos con la medida información mutua hay una mejora. De igual forma, es necesario corroborar los resultados obtenidos con otras colecciones de textos.

4.2 Consultas dadas por Estudiantes

Se trata de 16 consultas que son descripciones del concepto "Tierra". La tabla 1 muestra estas consultas truncadas. Cada una de estas consultas fue hecha al sistema, utilizando el *corpus* y relaciones de sinonimia provenientes de EWN como apoyo a la representación por sintagmas. La tabla 3 muestra en resumen el promedio de los valores indicados para las dos formas de representación (*corpus* y relaciones de sinonimia), usando la fórmula 3.

Para estas consultas se realizó una prueba adicional, confrontando el modelo basado en sintagmas con el modelo vectorial. La tabla 4 presenta resultados promedios de las consultas realizadas con los dos modelos de recuperación de información. En ambos modelos se consideró el uso de sinónimos, y como se aprecia en el modelo vectorial el uso de sinónimos no provee mejora.

[2] Recurso amablemente proporcionado por L. Fraguela C. [6].

5 Discusión

En el presente trabajo se experimentó con una propuesta de reperesentación semántica de textos con el fin de ser utilizada en la recuperación de información. Se partió de la idea del "sentido de un sintagma" y se implantó un sistema que realiza la representación considerando los sintagmas que contienen los términos índice de los textos (aquellos términos que resultan al final del preprocesamiento del texto). Además, las consultas se representan de manera semejante y la búsqueda se efectúa como lo hace el modelo booleano de recuperación de información.

Las pruebas se apoyaron con tres colecciones cada una de 255 definiciones del área de física. La primera correspondió a la colección de documentos, la segunda al *corpus* y la tercera a la colección de consultas supervisadas. También se efecturaron 16 consultas truncadas para el concepto Tierra, obtenidas de estudiantes de la Facultad de Ingeniería Mecánica de la BUAP, apoyándose el sistema en un *corpus* y relaciones de sinonimia (extraídas de EWN).

Ademas de usar las relaciones de sinonimia, se utilizaron otras relaciones extraídas de EuroWordNet: hiperonimia, hiponimia y una colección que contenía a las tres anteriores.

Los resultados obtenidos muestran un desempeño absoluto bajo, debido principalmente al material empleado en el experimento. El número de elementos de la colección es de 255 y la mayoría de estos está compuesto de solamente una oración. Además, con respecto al número de documentos se utilizó un conjunto grande de consultas, cuyas características hacen difícil la recuperación exacta: por el origen coloquial de estas expresiones, el vocabulario de éstas es variado y, aunque se refieren al dominio, el tipo de lenguaje es informal y diferente al de los documentos. Aún así, consideramos importante realizar el experimento con este material. Todas estas observaciones afectan a ambos modelos probados.

En el caso del modelo basado en sintagamas, cuando se realiza la representación de los términos usando un corpus vemos que se introduce mucho "ruido" y por eso su desempeño es el más bajo. Como se espera, el empleo de sinónimos conlleva a una mejora, y al realizar la equivalencia entre términos el desempeño es casi el mismo al del modelo vectorial.

Pueden realizarse otras pruebas como la combinación de relaciones de sinonimia con hiperonimia, sinonimia con hiponimia e hiperonimia con hiponimia, aplicando la representación en diversos puntos del procesamiento dentro del Sistema de Recuperación de Información. Asimismo, experimentar con un corpus de mayor tamaño.

References

1. C.J. van Rijsbergen: *Information Retrival*. University of Glasgow, pp. 114 - 117. Second Edition, 1999.
2. Gerardo Sierra & John McNaught: "Natural Language System for Terminological Information Retrival". CICLing 2003, *LNCS* 2588, A. Gelbukh (Ed.) Springer Verlag, pp. 541-552, 2003.

3. Aceves, R., Castilla Y.: *Resultados del Laboratorio de Recuperación de Información de la FCC.* Facultad de Ciencias de la Computación. BUAP
4. Baeza-Yates Ricardo: *Modern Information Retrieval*,Addison Wesley.1999
5. Reyes A. Berenice, Moyotl H. Edgar: *Reducción de términos índice usando el punto de transición.* Facultad de Ciencias de la Computación. BUAP
6. Fraguela Cuesta L.: *Análisis de una Representación de Textos usando Lattices.* Tésis de Maestría en Ciencias de la Computación. BUAP. 2004
7. García Fajardo Josefina: *Estructura conceptual y comunicación* , El Colegio de México 1995.
8. Grefenstette Gregory: *Automatic Thesaurus Generation from Raw Text using Knowledge-Poor Techniques*,Rank Xerox Research Centre.
9. Jiménez Salazr H.,Morales Luna G.: *Domain membership and clasification methods* FCC. BUAP, CINVESTAV-IPN
10. Rodríguez Hernández M.: *Representación de Documentos utilizando Relaciones de Sentido.* Tesis de Licenciatura. Facultad de Ciencias de la Computación. BUAP. Junio 2004.
11. Ruge Gerda: *Combining corpus linguistics and human memory models for automatic tem association*,AI Group.Institut fur Informatik,Munchen. 1999
12. Salazar Martínez H.; Jimenez Salazar H. & Pinto A. D.: *Extraction: a corpus-based Approach.* Facultad de Ciencias de la de la Computación. BUAP

Datos y Resultados

Table 1. Consultas preprocesadas para el concepto Tierra.

Identificador	Consulta
q_1	tercer planet sistem sol vid diferent demas
q_2	mas efecto fuerz atraccion cuerpos
q_3	terc planet cercan Sol tard 24 hor gir propi eje rotacion cuenta divers cap recubr encuentr divers ecosistem encuentr divers tip biodivers
q_4	planet lug habit viv
q_5	planet habit Hombre unic vida
q_6	porcion material sol abund planet
q_7	convertir dia muramos
q_8	punt escencial lodo
q_9	planet sistem sol vivimos
q_{10}	primer cap cortez terrestr conoc tiempo
q_{11}	planet ocup terc lug cuatn posicion sistem sol Cuenta condicion propiedad favorec evol reproduccion vid organism
q_{12}	element natural capaz vida
q_{13}	nuev planet integr sistem solar
q_{14}	terc planet sistem sol satelit natural llam Luna
q_{15}	ecosistem conform plant animal element agua
q_{16}	terc enlac circuit

Table 2. Desempeño promedio del sistema usando diversas relaciones (255 consultas).

Relaciones	Promedio de Documentos Recuperados	Posición de la Respuesta Correcta	Precisión	Evocación	F_1
sinonimia	48.07843	7.050980	0.050846	0.560784	0.037055
hiperonimia	132.6352	26.77254	0.01768	0.73725	0.015286
hiponimia	113.7803	23.60392	0.023224	0.694117	0.018432
todas	180.1921	45.2	0.014089	0.835294	0.012144
corpus	219.6117	78.32156	0.0065	0.898039	0.011090

Table 3. Representación por corpus y sinónimos (16 consultas de estudiantes).

Representación	Documentos Recuperados	Posicion de la Respuesta Correcta	Precisión	Evocación	F_1
Corpus	151.125	41.875	0.00285	0.625	0.00568
Sinonimia	22.687	2	0.02444	0.5	0.04569

Table 4. Modelo vectorial *vs* sintagmas (16 consultas de estudiantes).

Modelo	Precisión	Evocación	F_1
Vectorial	0.021	0.5	0.041
Vectorial sinónimos	0.021	0.5	0.041
Sintagmas *corpus*	0.002	0.6	0.005
Sintagmas sinónimos	0.024	0.5	0.045