Recuperación de Información

hjimenez@fcfm.buap.mx, dpinto@cs.buap.mx

Héctor Jiménez Salazar & David Pinto Avendaño

octubre 2003

1 Introducción a la recuperación de información

Un sistema de recuperación de información (SRI) consiste básicamente de un conjunto de procesos interrelacionados que permiten obtener información de interés, a partir de una determinada consulta.

El concepto de recuperación de información es bastante simple, tal como se muestra en la figura 1. Sin embargo, los procesos involucrados en la determinacin de la relevancia de una consulta suelen ser complejos, especialmente cuando se está tratando con información que carece de una estructura definida.

Un SRI permite la recuperación de la información, previamente almacenada, por medio de la realización de una serie de consultas ("queries") a los documentos contenidos en la base de datos. Esta serie de preguntas se estructuran como sentencias formales de expresión de necesidades de información, y suelen venir expresadas por medio de un lenguaje de consulta. Un documento es un objeto de datos, de naturaleza textual generalmente, aunque la evolución tecnológica ha propiciado la adición de documentos multimedia, incorporándose al texto fotografías, ilustraciones gráficas, vídeo animado, audio, etc.

Un SRI debe soportar una serie de operaciones básicas sobre los documentos almacenados en el mismo, como son: introducción de nuevos documentos, modificación de los documentos almacenados y eliminación de los mismos. Debemos también contar con algún método de localización de los documentos (o con varios generalmente), para presentárselos posteriormente al usuario. Los SRI implementan estas operaciones en formatos muy diversos, lo que provoca una amplia diversidad en lo relacionado con la naturaleza de los mismos; regularmente es posible encontrar variaciones con respecto a los métodos de búsqueda y las técnicas de representación.

Figure 1: Esquema general de un SRI

Es importante entender que la recuperación de información no es realmente una necesidad nueva; la escritura es probablemente uno de los medios más antiguos de almacenar y transmitir el conocimiento, y es claro que a partir de cierto volumen de información se hace imprescindible un sistema que organice y de la posibilidad de localizar la información que pudiese ser requerida en cierto momento.

1.1 Explosión de la información

Las técnicas usadas para facilitar el acceso a la información se han mantenido practicamente si variación durante los últimos 200 años; básicamente hasta que arribaron las computadoras y las redes de comunicación; las primeras por su velocidad y capacidad de almacenamiento y las segundas por su característica de compartición de la información han marcado un parteaguas, dando pie a una explosión de la información que no puede ser afrontada sin un amplio conjunto de nuevas técnicas de almacenamiento, acceso, consulta y manipulación de esa información. En la actualidad, cualquier persona que tenga acceso a Internet puede almacenar información; es posible observar información de cualquier tipo escrita en prácticamente todos los idiomas del planeta. En la medida que las instituciones han descubierto las potencialidades de la autopista de la información (Internet) para hacer accesibles sus bases de datos textuales, ha crecido la importancia de la recuperación de este tipo de información.

El desarrollo de los sistemas automatizados de recuperación de información se inició con el objetivo de facilitar el manejo de la enorme cantidad de literatura científica surgida desde los años 40. No ha quedado restringida a este campo sino que se ha extendido a otras áreas: cualquier disciplina que base su trabajo en la utilización de documentos puede beneficiarse de las técnicas de recuperación de información textual. En los últimos 30 años se han desarrollado estructuras de datos eficientes para el almacenamiento de índices, sofisticados algoritmos de consulta, métodos de compresión e incluso hardware específico; más recientemente, se han aplicado técnicas de procesamiento del lenguaje natural en aspectos tales como la extracción de información, la formulación de consultas amigables y la generación de respuestas. Un componente importante de las técnicas de recuperación textual lo constituye la búsqueda de cadenas tanto exacta como aproximada -considerada por algunos autores tan básica como puedan serlo las operaciones aritméticas en otras áreas. También son importantes los métodos de construcción y manipulación de diccionarios -en general aceleran los procesos de búsqueda y reducen el tamaño de los índices. La construcción de diccionarios suele relacionarse con las técnicas de procesamiento del lenguaje natural.

Aunque existen volumenes de información almacenados en diversos lugares, es en Internet en donde podemos encontrar la mayor cantidad de documentos. De acuerdo a estadísticas obtenidas y que se muestran claramente en la tabla 1, el Internet no solamente sigue creciendo, sino además, el

Región de AMERICA	Población (2003)	Usuarios año 2000	Usuarios, Dato mas reciente	Crecimiento (200-2003)	% Población (Penetración)	(%) Tabla
América Central	40,182,800	505,000	1,316,000	160.6%	3.3%	0.6%
Norte América	424,881,000	110,784,200	201,380,066	81.8%	47.4%	86.5%
Sur América	359,595,300	14,292,100	28,075,767	96.4%	7.8%	12.3%
El Caribe	40,195,300	583,500	1,440,500	146.9%	3.6%	0.6%
Total AMERICA	864,854,400	126,164,800	232,212,333	84.1%	26.8%	100.0%

Table 1: Crecimiento de Internet en los últimos 3 años:

ritmo de éste es acelerado. La cantidad de usuarios existentes, tan solo en el continente americano sobrepasa los ochocientos sesenta y cuatro millones. Y de nuevo si pensamos que cada una de estas personas tiene la posibilidad de agregar información, entonces podremos entender la importancia que reviste el estudio de técnicas de recuperación de información, especialmente en este ambiente.

Los esfuerzos realizados en torno de recuperación de información en Internet se han visto mediante la generación de herramientas de búsqueda llamadas robots buscadores. Ejemplos clásicos de estas herramientas son: Altavista, Google, Yahoo, etc. La característica que cualquier usuario, que haya accedido a estas herramientas, puede observar es la baja precisión de los resultados. Este efecto se encuentra principalmente determinado por los requerimientos de velocidad en la respuesta de los mismos usuarios y en algunos casos por el deficiente análisis y construcción de índices para el acceso a la información relevante.

Está claro que existe una carencia en la existencia de herramientas para el manejo de la información que puedan lidiar no solamente con el volumen sino con otras características como el caso de la multilingualidad. Dichas herramientas deben ser generadas en breve, dado que una de las constantes actuales, es la búsqueda de información confiable en Internet para la toma de decisiones.

1.2 Prevalecencia de la información no estructurada

Existen, sin embargo, grandes retos por cumplir; si bien, la información ya está almacenada en Internet, la gran mayoría de ésta carece de una estructura definida que permita caracterizar los componentes internos de dicha información.

No solamente se trata de lidiar con documentos textuales, sino que además habrá que contemplar los archivos multimedia, que involucran, imágenes, archivos de audio, video. Tratar con digitalización de voz, ediciones electrónicas, entre otras, es ya el reto actual del tratamiento de información.

Cómo determinar en un documento sin estructura que cierta parte corresponde al nombre del autor del documento, o el título, la vista frontal del Taj Mahal, o la voz de un Presidente de un país?.

En el lenguaje humano tenemos palabras para describir prácticamente cualquier cosa, esto es derivado del hecho de que a pesar de tener un alfabeto finito, tenemos un lenguaje infinito. Pensemos por un momento... Si una

imagen digital (de tamaño indefinido) está caracterizada por pixeles, cada uno de los cuales tiene la posibilidad de tener un color definido a partir de tres colores básicos (computacionalmente hablando: rojo, verde y azul), entonces existirá una función de mapeo entre el lenguaje humano y este lenguaje definido por un alfabeto finito de colores?. En tal caso, será posible que dada una consulta en lenguaje natural, sea posible extraer una o más imágenes que concuerden con dicha consulta a partir de efectuar dicha función de mapeo?.

Han existido muchos esfuerzos para promover la escritura de documentos estructurados, sin embargo, no han fructificado y la realidad muestra que la tendencia en la escritura de documentos sin estructura predefinida es a la alta. La naturaleza humana trae consigo esta tendencia y aunque en la actualidad existen modelos definidos para almacenamiento estructurado, como es el caso de las bases de datos relacionales, aún es un misterio saber si es posible transformar cualquier documento no estructurado para que pueda ser almacenado en un sistema de base de datos (por ejemplo, relacional), el cual tendría claramente una estructura definida.

1.3 Representación en Bases de Datos Relacionales

La representación de la información en bases de datos ya se encuentra resuelta, ya que se almacena únicamente información previamente estructurada. Dicha información es almacenada en tablas que pueden ser consultadas posteriormente usando operaciones del álgebra relacional (producto cartesiano, selección, proyección, unión, etc).

Después de realizar una consulta, se devuelven los registros que satisfacen la misma. La localización es eficiente y el procedimiento es determinístico, obteniendose en todos los casos un 100 por ciento de precisión y evocación.

Por ejemplo, si se tienen almacenados todos los datos de los empleados de una empresa, es posible extraer todos aquellos empleados que trabajan en los departamentos ubicados en la ciudad de Puebla y cuyos sueldos rebasen en un 10 por ciento el promedio del sueldo de todos los empleados y que además hayan recibido una bonificación en los últimos dos años, etc. La localización de la información será eficiente y determinística, obteniendose todos aquellos empleados que cumplan con las características dadas (ni uno más, ni uno menos).

En los sistemas de bases de datos relacionales no existen deducciones, simplemente se llevan a cabo las operaciones del algebra relacional y se regresan los resultados.

1.4 Representación en Bases de Datos Deductivas

Para el caso de las bases de datos deductivas, se infiere información según una consulta. Para poder realizar esta tarea se hace necesario poseer un motor de inferencia, un conjunto de hechos y un conjunto de reglas. Uno de los sistemas más populares que son utilizados para realizar representación

en bases de datos deductivas es Prolog; este lenguaje posee todas las características y permite definir información extensional (tablas, hechos) e intensional (fórmulas que generan otra información o "conocimiento").

En las bases de datos deductivas se devuelven registros, o parte de ellos que satisfacen una consulta. Los registros devueltos son obtenidos mediante búsqueda y deducción en el dominio de la aplicación. Aunque uno de los inconvenientes es que dicha deducción, basada en mecanismos de inferencia, suele bajar o reducir el rendimiento del sistema.

1.5 Representación e Interfaces Humano-Computadora

Una de las motivaciones de usar el lenguaje natural ha sido la interacción con una computadora, y justamente los SRI presentan esa naturaleza. Las preguntas en un SRI van más alla de simplemente introducir palabras claves, ya que la riqueza del lenguaje conlleva a usar el lenguaje de una manera más amplia y precisa, con el fin de realizar peticiones a una computadora.

Uno de los objetivos es precisamente procesar el lenguaje natural, de tal manera que dada una consulta en lenguaje natural, se represente y posteriormente se introduzca (ya representada) para obtener la información relevante. Cabe recalcar, que los resultados se esperaría estuviesen también expresados en lenguaje natural.

1.6 Representación en Sistemas de Recuperación de Información

En los sistemas de recuperación de información, se representa información no estructurada, en este caso, el sistema devuelve los documentos más relevantes a una consulta.

Para representar cada documento, primero se extraen, de cada documento, el mínimo número de palabras que describan al documento; a estas palabras se les llaman las palabras representativas. Encontrar las palabras que describan mejor al documento puede ser una tarea difícil, especialmente, porque los puntos de vista de las personas pueden ser diversos con respecto a este tópico. Si preguntamos a una persona por las palabras representativas de un documento, entonces, su respuesta estará regularmente conducida por el dominio del tema que tenga dicha persona. En los sistemas de recuperación de información se suele utilizar el enfoque común basado en uso de frecuencias de ocurrencia de los términos, se usa también la morfología de los documentos, la combinación de estas dos técnicas, e incluso un enfoque no semántico.

Una vez que se han representado los documentos, entonces es posible aplicar funciones de similaridad, con la finalidad de encontrar aquellos documentos cuya representación sea más similar a la representación de la consulta. Cabe aclarar, que la consulta debe ser representada, usando el mismo procedimiento que se llevó a cabo con los documentos.

En un sistema de recuperación de información compiten la precisión y la evocación, de tal manera que regularmente a mayor precisión menor evocación y viceversa.

1.7 Modelo booleano

El modelo booleno permite representar y recuperar documentos mediante funciones de similaridad booleanas. En este caso, dada una consulta, el modelo regresa solamente aquellos documentos que empatan totalmente con la consulta. En este caso, cada documento es evaluado mediante una función booleana, y de ahí el nombre del modelo. Un documento no puede ser relevante a medias, o es relevante totalmente o simplemente no lo es.

La descripción formal del modelo booleano, incluyendo preproceso, representación y consulta, es dada a continuación:

Preproceso. Dado un texto D, denotemos con D' el que se obtiene por eliminar las *palabras cerradas* (preposiciones, artículos, etc.) y *lematizando* cada una de las restantes ("presidentas"→"presidente", "comprendámoslo"→ "comprender", etc.).

Definición de Índices. Sea $\mathcal{D} = \{D_1, \ldots, D_k\}$ una colección de documentos y $\mathcal{D}'$ la colección preprocesada. Sea $\mathcal{V} = \cup_i D'$ el vocabulario de la colección, y $\mathcal{V}_0 = [v_i]_i$ el vocabulario ordenado lexicográficamente. La *representación* de un texto D es el vector $\vec{D} = [d_i]_{i \leq n}$, donde $d_i = \begin{cases} 1 & \text{si } v_i \in D' \\ 0 & \text{si } v_i \notin D' \end{cases}$, con $n = \#\mathcal{V}_0$.

Búsqueda. Dada una consulta q formada por términos preprocesados, $q_1, \ldots, q_k$, se forma el vector $\vec{q} = [q_i]_{i \leq k}$. Los documentos recuperados bajo el modelo booleano son D_j tales que $\vec{D}_j \cdot \vec{q} \neq 0$.

1.8 Relevancia

La esencia de un SRI descansa en el concepto de relevancia. Esperamos que a una consulta se responda con documentos relevantes y sin documentos irrelevantes.

Una definición un tanto formal del concepto de relevancia puede ser dada de la siguiente manera:

Un *conjunto mínimo de premisas* (CMP) de un componente lingüístico es aquél en el que al eliminar un elemento, el componente ya no es una consecuencia del CMP. Por ejemplo si $d = \{a \wedge b, a \vee b, a, b\}$ el CMP es $\{a, b\}$ pues con $\{a\}$ o $\{b\}$ tenemos que: $\{a\} \not\models a \wedge b$ y $\{b\} \not\models b$.

Un documento es *relevante* a una consulta si algún componente de ésta pertenece al CMP del documento.

La lógica clásica puede describir bien esta situación siempre que estemos considerando como recuperación un "empate exacto".

Por ejemplo: Dados $D_1 = \{a, b\}$, $D_2 = \{b, c\}$, y $q = a \wedge b \wedge \neg c$, D_1 se recupera ya que $D_1 \models q$. (Obsérvese la "Close World Assumption" en sus versiones generalizada y extendida).

El concepto de relevancia es de suma importancia, ya que determina precisamente el hecho que un documento sea importante o de interés para un usuario, de acuerdo a una determinada consulta.

1.9 Coordinación

La exigencia de recuperar documentos relevantes y no irrelevantes es raramente cumplida en los sistemas reales. De hecho una consulta especificada por un usuario puede ser satisfecha por el sistema y para otro usuario la misma pregunta puede ser sólo un poco cubierta.

Esto es debido a que aunque dos usuarios distintos coloquen las mismas palabras como consulta, éstos pueden tener una concepción distinta de lo que desean obtener. Por ejemplo, si un usuario A, introduce la consulta "base de datos" y desea obtener información de lugares en donde existan bases de información almacenadas, éste tendría una concepción totalmente distinta de algún otro usuario que introdujera la consulta "base de datos" pensando en obtener una definición de base de datos.

Esta subjetividad se relaja con una medida del "nivel de coordinación". Veamos a continuación un ejemplo:

Notemos que con $D_1 = \{a, d\}$, $D_2 = \{b, c\}$, $D_3 = \{a, b, c\}$ y $q = \{a, b, c\}$, en el modelo booleano, esperamos D_3 como respuesta, pero D_2 sería más relevante que D_1. El nivel de coordinación es proporcional a los índices comunes a la consulta y a los documentos, i.e. proporcional a $\#(D \cap q)$.

2 El modelo de espacio vectorial

El modelo de espacio vectorial fué propuesto por Salton en su famoso libro "Information Retrieval...", este modelo plantea el uso de vectores matemáticos en dominios de múltiples dimensiones para representar cada documento. Un vector de representación de un documento estaría compuesto, en su versión más simple, de la frecuencia de ocurrencia de los términos del vocabulario del corpus, para ese documento en particular. Si cada documento tiene un vector como representación, entonces es posible calcular el producto interno, entre vectores y determinar si dos documentos son similares, a partir del valor del coseno del ángulo entre los dos vectores.

2.1 Modelo vectorial

La descripción formal del modelo de espacio vectorial, incluyendo preprocesamiento, representación y consulta, es dada a continuación:

Preproceso. Dado un texto D, denotemos con D' el que se obtiene por eliminar las *palabras cerradas* (preposiciones, artículos, etc.) y *lematizando* cada una de las restantes ("presidentas"$\rightarrow$"presidente", "comprendámoslo"$\rightarrow$ "comprender", etc.).

Vectores de índice. Sea $\mathcal{D} = \{D_1, \ldots, D_k\}$ una colección de documentos y $\mathcal{D}'$ la colección preprocesada. Sea $\mathcal{V} = \cup_i D'$ el vocabulario de la colección, y $\mathcal{V}_0 = [v_i]_i$ el vocabulario ordenado lexicográficamente. La *representación* de un texto D es el vector $\vec{D} = [d_i]_{i \leq m}$, donde $d_i = \begin{cases} 1 & \text{si } v_i \in D' \\ 0 & \text{si } v_i \notin D' \end{cases}$, y $n = \#\mathcal{V}_0$.

Asignación de pesos. Las componentes de cada vector $\vec{D}_i = [d_{i1}, \ldots, d_{in}]$ son ponderadas de la siguiente forma: $d_{ik} = tf_{ik} \cdot idf_k$ donde tf_{ik} es la frecuencia del término k en D_i, y idf_k está definido como $idf_k = \log_2(M) - \log_2(df_k) + 1$, siendo df_k el número de documentos que usan el término k, y M el número de documentos.

2.2 Similitud

La función de similitud sirve para determinar que tanto se parecen dos elementos. Es común considerar la función de similitud normalizada ($\in$[0,1]).

El índice de Jaccard $sim(D, q) = \#(D \cap q)/\#(D \cup q)$ se emplea en la representación booleana.

Para el modelo booleano también suele emplearse otra medida $sim(D, q) = \frac{2 \times \#(D_i \cap q)}{\#D_i + \#q}$, que es llamada el coeficiente de Dice.

Tal cual habíamos mencionado con anterioridad, en el caso de la representación vectorial se emplea el coseno del ángulo entre los vectores que representan a los documentos:

$$sim(\vec{D}_i, \vec{q}) = \frac{\sum_{k=1}^{m} d_{ik} q_k}{\sqrt{\sum_{k=1}^{m} d_{ik}^2 \cdot \sum_{k=1}^{m} q_k^2}}. \quad (1)$$

2.3 Ejemplos en AWK

A continuación se presentan ejemplos para generación de vocabulario, representación usando índices y consulta, usando el lenguaje AWK.

2.3.1 Obtención del vocabulario

```
awk '
# construye el vocabulario de un corpus.
# Se recomienda preprocesar el archivo de documentos de entrada.

{for(i=1;i<=NF;i++) vocabulario[$i]++}
END{for (x in vocabulario) print x, vocabulario[x]}
```

```
' $*
```

2.3.2 Representación usando vectorización (índices)

```
awk '
# Forma la matriz de documentos vs \'{\i}ndices
# El archivo de vocabulario es tomado del formato generado en el programa anterior
# Cada documento de entrada debe estar en un renglon del archivo de entrada.
# julio 18 2002 / hjs

BEGIN{lg2=log(2)}
FILENAME==voc {voca[$1]=1;next}

function noc(x,d) {return(gsub(x,x,d))} #ocurrencias de x en d

function suma(x, s) {for(k=1;k<=nd;k++)if(frec[x,k]!=0)s++; return(s)}

{doc[++nd]=$0}

END{for(x in voca)for(i=1;i<=nd;i++)frec[x,i]=noc(x,doc[i])
 for(x in voca){frec[x,nd+1]=suma(x);print x > "/dev/stderr"}
 for(x in voca)for(i=1;i<=nd;i++)mat[x,i]=frec[x,i]*(log(nd/frec[x,nd+1])/lg2+1)
 for(x in voca){printf("%s ",x)
  for(i=1;i<=nd;i++)printf("%5.3f ",mat[x,i]);printf("\n")}}
' voc=$1 $*
```

2.3.3 Consulta (búsqueda

```
awk '
# Construye el vector de una consulta y multiplica por la matriz de docs
# julio 18 2002 / hjs

FILENAME==matr {nd=NF-1;voca[$1]=++np
 for(i=2;i<=nd+1;i++)mat[np,i-1]=$i;next}

{q[++j]=$0; nq=split($0,a0)
 for(i=1;i<=nq;i++){if(a0[i] in voca)vq[voca[a0[i]]]=1;print a0[i],voca[a0[i]]}
 mult(vq,np,mat,nd,matq)
 for(i=1;i<=nd;i++)if(matq[i])print "relevancia",matq[i],"con el doc",i}
' matr=$1 $*
```

3 Evaluación de un SRI

En esta sección se presentan diversos tópicos relacionados con la evaluación de un sistema de recuperación de información. La idea fundamental es ser

capaz determinar que tan bueno o malo es el sistema. Regularmente se compara en base a los resultados obtenidos contra un conjunto de consultas previamente evaluadas (consulta supervisada).

3.1 Elementos de evaluación

SRI. Un *sistema de recuperación de información* (SRI) es una terna $S = \langle \mathcal{D}, Rep, \mathcal{E} \rangle$, donde $\mathcal{D}$ es una colección de documentos, $Rep : \mathcal{D} \to 2^{\mathcal{V}}$ una representación y $\mathcal{E} : \mathcal{V}^+ \to 2^{\mathcal{D}}$ una función de búsqueda.

Consulta. Una *consulta supervisada* Q es una pareja donde su primer componente es una expresión y el segundo un subconjunto de $\mathcal{D}$: $(q, r) \in \mathcal{V}^+ \times 2^{\mathcal{D}}$.

Precisión y Evocación. Para un SRI S y una consulta supervisada $Q = (q, r)$, se definen la *precisión* $P = \#(\mathcal{E}(q) \cap r)/\#\mathcal{E}(q)$, y la *evocación* $R = \#(\mathcal{E}(q) \cap r)/\#r$. Es común promediar estos valores para un conjunto de consultas supervisadas.

3.2 Coordinación, evocación y precisión

Coordinación. Así como se ha definido el nivel de coordinación para los términos que participan en una respuesta, en la evaluación consideramos el "ranking" de los documentos, i.e. los documentos ordenados descendentemente, de acuerdo al valor de similitud con la respuesta.

Gradación ("ranking") de una respuesta. Sea $S = \langle \mathcal{D}, Rep, \mathcal{E} \rangle$ un SRI, (q, r) una consulta supervisada con $\#r = nr$ y supongamos que la gradación de $\mathcal{E}(q)$ es $[D_{q_i}]_{i \leq n_q}$ con $\#\mathcal{E}(q) = n_q$. Si consideramos el primer elemento de esta lista, D_{q_j}, que ocurre en r, referimos al *nivel de evocación* 1.

Evaluación por niveles. Para el nivel de evocación 1 ($100 \times (1/nr)\%$), la precisión y evocación son $P = 1/j$ y $R = 1/n_r$. Para el nivel de evocación 2 ($100 \times (2/nr)\%$) $P = 2/j$ y $R = 2/n_r$, donde j es la posición en $[D_{q_i}]_i$ del segundo documento relevante ($\in r$), etc.

3.3 Valores promedio

Es posible tratar de determinar cual es el comportamiento en promedio de un SRI. El objetivo es observar que tan robusto es el sistema bajo distintas consultas, ya que podría suceder que el SRI se comportara muy bien para una determinada consulta, mientras que para una consulta distinta se comportase totalmente mal.

Para N_q consultas $(q_i, r_i)_i$: $P_i(\ell)$ es la precisión en el nivel ℓ de la consulta i, y la precisión promedio para el nivel ℓ:

$$\bar{P}(\ell) = \frac{1}{N_q} \sum_{i=1}^{N_q} P_i(\ell).$$

También es útil calcular niveles de evocación usando medidas estándar. Tradicionalmente, la gráfica de niveles se divide en 11 posiciones. **Los niveles de evocación estandar**, 0%, 10%, ..., 100%, para una consulta q_i, corresponden a tomar como nivel de evocación 1 el 10% de los documentos de r. Los valores de precisión se interpolan mediante:

$$P(\ell_j) = \underset{\ell_j \leq \ell \leq \ell_{j+1}}{P(\ell)}$$

Un ejemplo del uso de los niveles de evocación se presenta en la siguiente sección.

3.4 Ejemplo de evaluación

Considérense los documentos relevantes a una consulta q los siguientes:
`d1, d2, d3, d4, d5`

Y la respuesta a q dada por el sistema:
`d1, o1, o2, d2, o3, o4, o5, o6, o7,`
`d3, d4, o8, o9, d5, o10, o11, o12,o13.`

Para este ejemplo en particular, la precisión para el nivel de evocación 1 es 1, dado que d1 se encuentra en la posición 1 y es el único documento en el nivel. la precisión para el nivel de evocación 2 es 0.25, ya que el documento d2 se encuentra en la posición 4. La precisión para los siguientes niveles de evocación se calcula de la misma manera. Una gráfica de los resultados para este ejemplo puede verse en la figura 2.

4 Términos discriminantes

En esta sección hablaremos sobre la dimensionalidad del espacio que propone el modelo vectorial para la representación de documentos en un SRI. Con base a un breve análisis se motiva la reducción de los vectores que representan a los documentos. Finalmente, se presenta el experimento de Salton *et. al.* que permite identificar los términos dicriminantes para reducir el espacio vectorial en esta representación y se muestra un ejemplo del cálculo de estos términos.

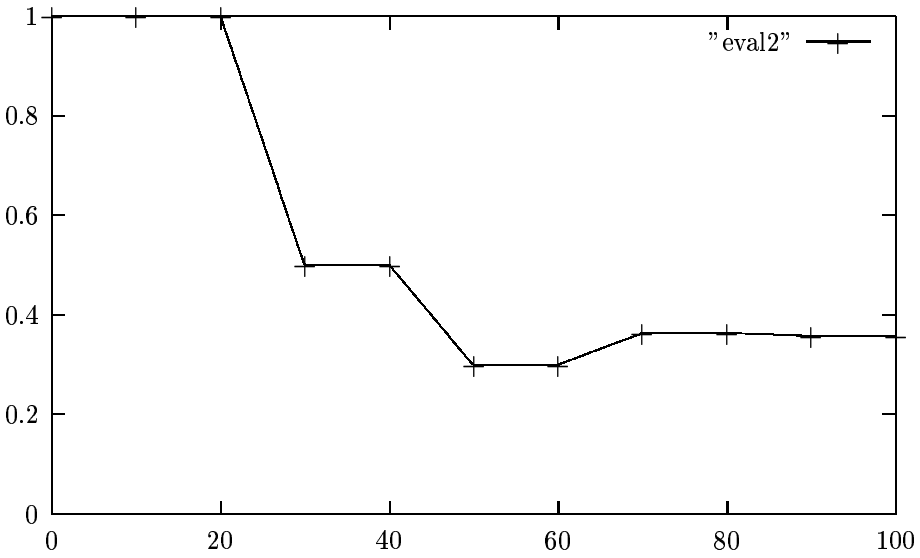

Figure 2: Gráfica de precisión y evocación por niveles

4.1 Dimensionalidad de la representación vectorial

Se ha dicho que el modelo vectorial, para representar documentos en un SRI, conduce a construir una matriz en la cual cada renglón represente documento un documento y cada columna a un índice. Claramente, el tamaño de esta matríz depende en forma crucial del número de términos. Por ejemplo, la dimensión del espacio vectorial para una colección de 500 documentos puede ser mayor a 10,000.

Un resultado importante que se debe tomar en cuenta, al trabajar con el modelo vectorial, es que más del 60% de los términos que conforman los índices son referidos por un único documento. Un porcentaje de términos, también alto, es empleado por dos documentos, etc. Lo cual nos lleva a concebir una matriz "rala". Esto es, muchos pesos son cero, pero además, muchos términos tienen un peso bajo; más que ayudar, su presencia es "ruidosa" en el proceso de identificación de documentos. La anterior afirmación obliga a realizar un análisis más detenido sobre la influencia de los términos como ídices de documentos.

4.2 Reducción de la dimensión usando términos discriminantes

De acuerdo con la definición del peso para cada uno de los componentes de un vector (documento) vemos que los vectores se limitan a ocupar el primer octante del espacio. Intuitivamente, podemos suponer que entre más densa sea la zona que ocupan los vectores será más difícil discernir sobre cuál o cuáles documentos son relevantes a una consulta. Salton, Wong & Yang [**?**],

realizaron un experimento para localizar los términos que al no ser incluidos en la representación de los vectores la densidad de éstos aumentaba (se "acercaban" unos a otros). Recíprocamente, también se encuentran términos que al no ser incluidos en la representación los documentos se dispersan. Los primeros son llamados *buenos* discriminantes ya que permiten identificar mejor a los documentos, mientras que los segundos son términos llamados *malos* discriminantes pues cuando se emplean en la representación la densidad de los vectores aumenta.

4.3 Cálculo del valor discriminante de un término

Enseguida se describe el experimento [**?**] para identificar los términos dicriminantes.

Sea $\mathcal{D}' = \{\vec{D}'_1, \ldots, \vec{D}'_M\}$ la colección de documentos vectorizada. El **centroide** de $\mathcal{D}'$ se define como

$$\vec{C} = \frac{1}{M} \sum_{i=1}^{M} \vec{D}'_i, \tag{2}$$

y la **densidad** de $\mathcal{D}'$ es

$$Q = \sum_{i} \text{sim}(\vec{C}, \vec{D}'_i). \tag{3}$$

El **valor discriminante** de un término k es $vd_k = Q_k - Q$, donde Q_k se interpreta como cálculo de Q sin el término k. Así, vd_k es alto cuando al eliminar k la densidad aumenta, o bien al considerarlo en la representación la dispersión aumenta. Esto significa que los términos con vd_k bajo no contribuyen a la identificación de documentos, pues cuando entre más densidad haya será más difícil discernir sobre ellos con la función de similitud.

4.4 Identificación de términos discriminantes

El experimento de Salton *et.al.*, en primer lugar, determina para cada término, k, el número de documentos que emplean a k, df_k, de tal manera que los términos pueden ser agrupados por el valor df_k; esto es, para 500 documentos tendríamos 500 clases. Luego, para cada término k calcula su valor discriminate dv_k. Finalmente, para cada clase con valor df_k se promedian los valores discriminantes.

El procedimiento antes expuesto fue realizado con un conjunto de 500 documentos. Debido a los cálculos exhaustivos en cada clase df se tomó solamente una muestra del 20%. La gráfica de la fig. 3 muestra las parejas $(df, \bar{dv})$, donde $df \in [1, 500]$ (clases del valor df_k), y $\bar{dv}$ es el promedio en dicha clase.

Como podemos ver en la figura 3 los términos k con $df_k \in [10, 100]$ tienen valores con rango de vd menor. Como el rango de una lista de valores pone al

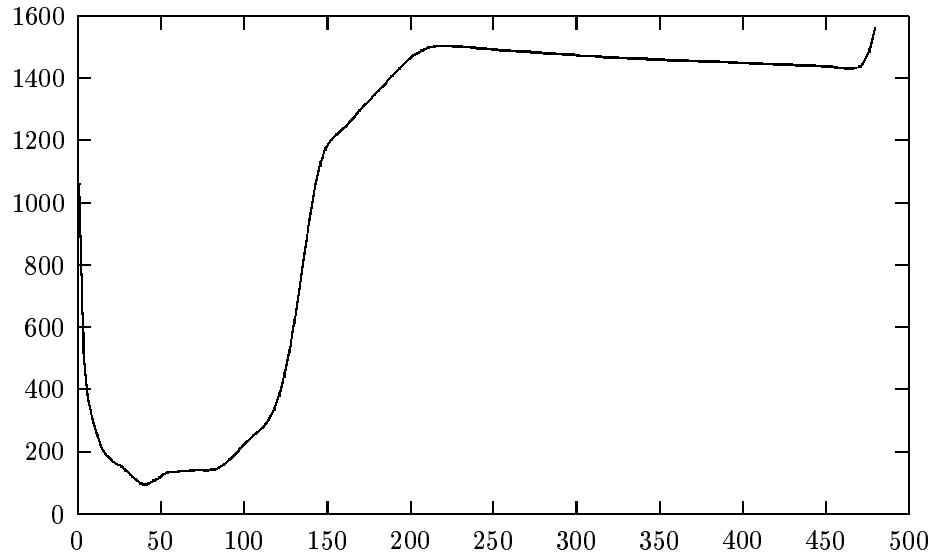

Figure 3: Gráfica de df_k *vs* $rango_{vd_k}$

inicio los más altos (los de menor rango serán los valores más altos), entonces estos términos tienen alto *vd*; es decir son buenos discriminantes. Además, era esperado que los términos con df_k alto (que muchos documentos los usan) no fueran buenos discriminantes, puesto que al aparecer en muchos dicumentos es difícil distinguir cuál o cuáles de ellos son relevantes para la consulta; estos términos tienen *dv* mayor a 400. Asimismo, hay muchos términos que son usados solamente por pocos documentos, de ellos podemos decir que no contribuyen determinantemente en la identificación de documentos; tienen una df_k baja; su *dv* es también mayor a 400 para $df_k < 10$. En suma, los términos de frecuencias medias en df_k son buenos discriminantes. Salton *et.al.* concluyeron que si M es el número de documentos, entonces un buen discriminante, k, es aquél que cumple $df_k \in [\frac{M}{100}, \frac{M}{10}]$. En nuestro caso, debido a particularidades del *corpus* obtuvimos que si k cumple $df_k \in [\frac{M}{50}, \frac{M}{5}]$ es un buen discriminante.

La conclusión es que eligiendo términos discriminantes para representar documentos reducimos la simensionalidad hasta el 25% [**?**]. Los términos restantes también son usados: los de baja frecuencia permiten constituir clases de palabras relacionadas con la temática del documento, y los términos de alta frecuencia es posible emplearlos en la identificación de frases que caracterizan a los textos que las contienen.

5 Empatamiento *vs* inferencia

Un SRI, lleva a cabo dos procesos: representación de información (documentos), y búsqueda. Esta sección está dedicada a introducir algunos de los

problemas que se enfrentan al tratar de generalizar el proceso de búsqueda.

Como sabemos, un SRI atiende consultas de un usuario entregando los documentos más relevantes para los elementos que ocurren en la consulta. Un SRI maneja en esencia una representación de los documentos y hace una búsqueda por **similitud**. La representación más empleada es la que por cada documento hay un conjunto de **índices**, alias "bolsas de palabras", por no considerar estructura mayor a la morfológica. Es indispensable el empleo de la estadística (frecuencias) para definir los índices que representan a un documento. Una metáfora de la búsqueda de los documentos es representar por un vector a cada documento, así como a la consulta. Los documentos más relevantes son aquellos cuyo *ángulo* respecto a la pregunta es menor que algún umbral; la similitud es el coseno del ángulo.

La esencia de un SRI descansa en el concepto de relevancia. Esperamos que a una consulta se responda con documentos relevantes y sin documentos irrelevantes. En el enfoque antedicho no hay referencia explícita a los significados de los términos que se emplean como índices y, sin embargo, esperamos que la respuesta a la consulta de un usuario sea relevante, esto es el contenido de los documentos que regresa el sistema tenga relación con los conceptos que aparecen en la consulta (recuérdese la definición del conjunto mínimo de premisas) Este hecho nos lleva a replantear el problema en el contexto del lenguaje natural.

La forma de abordar la recuperación de documentos, además, (información no estructurada) hereda algunos planteamientos que se han hecho en las bases de datos. Así como en las BD tradicionales hay una búsqueda que satisface una consulta, es natural pretender trasladar las ventajas que ofrecen las BD deductivas y atacar problemas del lenguaje natural que, como veremos, usa la inferencia de manera natural para comprender lo que un hablante expresa. Estos problemas, sin embargo, no son fáciles y aún siguen sin resolverse, pero señalan una pauta para el desarrollo de muchos sistemas de tratamiento del lenguaje natural.

5.1 Características del lenguaje natural

El hecho de que los manuales de gramática de las lenguas estén escritos en el mismo lenguaje que decriben es una muestra de la **universalidad** del lenguaje [Hajičova]. Los diversos *valores ilocutivos* (*saludar, inaugurar, explicar*, etc.) reiteran que en el lenguaje natural es posible expresar cualquier cosa. Esta universalidad se combina con dos características que dan paso a la creatividad. Esto es, las formas de interpretar y de expresar. En la creatividad, el caso más difícil, abudante e interesante es la **ambigüedad**. En el enunciado "La ley será vigente hasta mañana" por el conocimiento del contexto es que precisamos un significado (*ambigüedad léxica*). Este fenómeno no se limita a las palabras, además está presente en los sintagmas y, aún, en las formas de relacionar elementos del dominio. Por ejemplo, la famosa oración "El señor ve a los niños con un telescopio" tiene al menos dos

agrupamientos (*ambigüedad sintáctica*), y $(\forall x)(\exists y)(Cons(x) \rightarrow Prot(x,y))$ y $(\exists y)(\forall x)(Cons(x) \rightarrow Prot(x,y))$ son dos interpretaciones de "Los consejeros protestaron" por la predicación de conjunto-individual (*ambigüedad semántica*). La habilidad del ser humano para desambiguar tiene su base en el contexto y las "preferencias" Notemos que en "la ballena es mamífero" normalmente preferimos la predicación de conjunto pues raramente nos referimos a una ballena particular. Pero, como se aludió, también tenemos correspondencia entre muchas expresiones y un significado, lo que conocemos como **sinonimia** [Lyons].

Hay además, en el lenguaje natural, elementos implícitos que permiten, entre otras cosas, hacer inferencias. Tal es el caso de las *presuposiciones* e *implicaturas*. El "simple" uso de un artículo determinado nos lleva a una presuposición: "Rodrigo sobrevivió al temblor", el artículo de la contracción "al" permite inferir que "hubo un temblor" sin que esto se declare explícitamente [Levinson]. Nótese que se obtiene la misma inferencia negando el verbo principal: "Rodrigo no sobrevivió al temblor". La caracterizaciones de este fenómeno han sido variadas pero no satisfactorias. Si decimos que ($\mathcal{A}$ presupone $\mathcal{B}$) $\leftrightarrow ((\mathcal{A} \rightarrow \mathcal{B}) \wedge (\sim \mathcal{A} \rightarrow \mathcal{B}))$ tendríamos que $\mathcal{B}$ es una tautología.
En el diálogo (a) "¿Vas a servir té y café?" (b) "Café." la implicatura conversacional se apoya en un principio de cooperación [Grice], además de emplear una escala que permite alcanzar la respuesta sin ser explicitada.

En resumen, es requerido para enfrentar el problema de inferencia en un SRI proveer de información que ayude a resolver problemas como los antes expuestos. El sistema SMART, un sistema desarrollado en los años sesenta, aún es empleado para confrontar avances en la dirección señalada

6 Elementos de morfosintaxis

Una de los procesos más empleados en el tratamiento de textos es el etiquetamiento de las palabras con su parte de discurso (o categoría gramatical, *Part-Of-Speech* (POS)). Es necesario, por ejemplo en la tarea de *parsing* pero también puede ayudar a lematizar las palabras que componen un texto, o bien identificar frases de una manera relativamente sencilla.

Haremos en esta sección una presentación breve de cómo se compondría un etiquetador para el español, y de qué forma puede llevarse a cabo el análisis sintáctico dentro del contexto de PROLOG.

6.1 Etiquetamiento con partes del discurso

Esta tarea aparentemente sencilla, que en el caso de un lenguaje de programación (análisis léxico) se reduce al empleo de expresiones regulares, suele representar un problema de gran esfuerzo, debido a dos causas principales:

la ambigüedad y el desconocimiento pleno de algunas palabras.

Será presentado el diseño de un etiquetador [Jiménez-Morales] basado en el ordenamiento de los problemas que se enfrentan para determinar la parte de discurso de una palabra. El criterio toma en cuenta las palabras más frecuentes que ocurren en cualquier texto del espñol (proveniente de un estudio del español mexicano [Lara]). Asimismo, considera las regularidades e irregularidades conocidas para ciertas palabras, en su mayor parte verbos. Por último, los demás casos tratan los dos problemas antes aludidos mediante un método de clasificación tomado del aprendizaje automático: aprendizaje basado en memoria [Daelemans].

En la tabla **??** se muestran los recursos empleados en el etiquetador.

Tipo de palabra	**Recurso**
Palabras conocidas (más frecuentes)	Diccionarios
Palabras parcialmente conocidas	Reglas contextuales
Palabras ambiguas (más frecuentes)	Instancias (entrenamiento)
Ambigüedad `VERB/NOM`	Instancias (entrenamiento)
Palabras desconocidas	Instancias (entrenamiento)

Las instancias que refiere la tabla fueron extraídas del *Corpus del Español Mexicano Contemporáneo* (CEMC) [Lara].

Los recursos basados en competencias y su influencia en la tarea de etiquetamiento puede apreciarse en la tabla 3. La tabla muestra cuántas palabras del *CEMC* satisfacen el recurso.

Diccionario	**Tamaño**	**Tipo de palabra**	**Ejemplo**	**Ocur.** (%)
DICCD	184	definida frecuente	`el, y`	28.38
DICCA	163	ambigua frecuente	`la, que`	17.59
DICCS	11	puntuación	`<punto>`	11.20
DICCV	835	forma verbal frecuente	`tienen`	5.20
NOMP	3,337	nombre propio	`OCDE`	2.08
Total	**4,530**			**64.47**

De el CEMC también se determinan algunas reglas morfosintácticas. Esto es, se analiza si algunas palabras con POS ambigua puede ser determinado con base en su contexto. Aquellos contextos que determinan siempre el POS de esas palabras se eligen como condiciones de la regla que define el POS de la palabra.

Si $\Pr(tag(w) = m|C(w)) = 1$, donde $C(w)$ es un contexto predeterminado de w, entonces concluimos que

$$C(w) \Rightarrow (tag(w) = m).$$

Por ejemplo en el CEMC siempre se tiene que "la" precedido de "de" determina que "la" es artículo. Más específicamente una regla sería:

$$\Pr((v_{i-1}, m_i, v_{i+1}) \in \{el, al\} \times \{\text{NOM}\} \times \{de, del\} \\ |(v_{i-1}, v_{i+1}) \in \{el, al\} \times \{de, del\}) = 1$$

6.2 Clásulas definidas

Mostraremos el empleo de las cláusulas definidas para conseguir el análisis sintáctico de oraciones que han sido descritas por una gramática.

Recordaremos algunos conceptos básicos de la teoría de lenguajes formales.

Una **gramática** G es un cuádruple $\langle V_n, V_t, P, S\rangle$, donde V_n es un conjunto de símbolos no terminales, V_t es un conjunto de símbolos terminales, P es un conjunto de producciones (o reglas), y $S \in V_n$ el símbolo inicial.

Una **derivación** en G es una secuencia de reglas $R_1, \ldots R_k$ aplicadas a una expresión α_0, tal que R_1 deriva α_1 a partir de α_0: $\alpha_0 \Rightarrow \alpha_1$, $\alpha_1 \Rightarrow \alpha_2$, etc. Es decir, R_i es $\gamma_i \rightarrow \beta_i$, donde γ_i es una subcadena de α_{i-1} y β_i es una subcadena de α_i. Indicamos la derivación con $\alpha_0 \Rightarrow \alpha_1 \Rightarrow \ldots \Rightarrow \alpha_k$ o bien $\alpha_0 \Rightarrow^* \alpha_k$.

El **lenguaje** generado por una gramática G es el conjunto de expresiones que se derivan a partir del símbolo inicial: $L(G) = \{x|S \Rightarrow^* x\}$.
Un ejemplo de gramática que describe oraciones sencillas es:

$G_o = \langle\{\text{Oracion, Sujeto, Pred, Art, Nom, Adv, Verbo}\}$,
$\{\text{la, las, lo}, \ldots, \text{jose, buho, reloj}\ldots, \text{rapido}\ldots, \text{duerme}\ldots\}, P_o, \text{Oracion}\rangle$, donde P_o es:

Oracion	→	Sujeto Pred
Sujeto	→	Nom
Sujeto	→	Art Nom
Pred	→	Verbo
Pred	→	Verbo Adv
Art	→	el | la | ...
Nom	→	reloj | jose | ...
Adv	→	rapido | ...
Verbo	→	duerme | ...

Warren & Pereira introdujeron el concepto de DCG (*definite clause grammar*). Se trata de descripciones en lenguaje PROLOG adaptadas para expresar reglas gramaticales. Mediante un ejemplo veremos que simplemente se escriben con un formato las reglas gramaticales deseadas. Para P_o tenemos:

```
oracion(X,Y)            →  sujeto(X,X1), pred(X1,Y)
sujeto(X,X1)            →  nom(X,X1)
sujeto(X,X1)            →  art(X,X2), nom(X2,X1)
pred(X1,Y)              →  verbo(X1,Y)
pred(X1,Y)              →  verbo(X1,X2), adv(X2,Y)
art([el | X],X)         ...
nom([reloj | X],X)      ...
adv([rapido | X],X)     ...
verbo([duerme | X],X)   ...
```

En estas cláusulas se emplean dos variables. La primera representa un "flujo" de texto que es analizado por la regla correspondiente. La segunda es el texto "residual", el cual normalmente es pasado a la siguiente regla que se emplea en el mismo nivel de análisis. Las cláusulas permiten concentrarnos en otros aspectos del trabajo sintáctico y dejar de lado los problemas de implantación.

7 Conceptos formales

Paradójicamente, una carencia de la mayoría de los SRI es el empleo de la semántica del lenguaje natural para representar los textos.

Una forma de aproximarse al sentido de un texto es considerar una semántica composicional que retome las propiedades de los términos para construir el concepto subyacente. Las propiedades de los términos pueden ser accesibles vía las relaciones de sentido (proporcionadas por un *corpus*, una muestra del universo del discurso). Una propuesta de formalización del empleo de índices es considerar *conceptos formales*.

Consideremos $\Psi : \mathcal{V} \rightarrow \mathcal{V}^*$ como una función que accede a las relaciones

de sentido de un término del vocabulario $\mathcal{V}$.

En la teoría de conceptos formales el sentido se concentra en las propiedades comunes (intento) a todos los objetos instancias del concepto (extento). Formalemente para un concepto (A, B) se debe satisfacer $\cap\Psi(a)_{a\in A} = B$. Para un dominio se puede establecer una relación de orden, por ejemplo con

$$\text{astro} \rightsquigarrow (\{\text{tierra,luna,marte,...}\},\{\text{celeste,gravitación,...}\})$$

y

$$\text{estrella} \rightsquigarrow (\{\text{sol,alfa...}\},\{\text{luz-propia,celeste,gravitación,...}\})$$

tenemos que estrella$\leq$astro o bien, estrella$\rightarrow$astro.

La condición de propiedades comunes para todos los objetos de un concepto es restrictiva en aplicaciones, pues es posible que existan objetos que compartan rasgos con objetos de un extento pero no pertenezcan al intento. Veámoslo formalmente.

Siguiendo [Davey] y [Schneider], sea G un conjunto de objetos y M un conjunto de n rasgos ambos finitos, y sea ψ la correspondencia que asocia a cada objeto $g \in G$ un elemento de M. Representemos con $\Psi(g)$ el conjunto $\{m \in M|\psi(g) = m\}$. A la terna $\langle G, M, \psi\rangle$ la llamamos *contexto*. Un *concepto* en el contexto $\langle G, M, \psi\rangle$, es una pareja (A, B), compuesta del *extento* A, y del *intento* B, tal que $A \subset G$, $B \subset M$ y, además,

$$\bigcap_{a\in A} \Psi(a) = B. \tag{4}$$

Para dos conceptos, (A_1, B_1) y (A_2, B_2), definidos en $\langle G, M, \psi\rangle$, decimos que (A_1, B_1) es *más particular* que (A_2, B_2), $(A_1, B_1) \leq (A_2, B_2)$, o bien, que (A_2, B_2) es *más general* que (A_1, B_1), $(A_2, B_2) \geq (A_1, B_1)$, si se cumple que $A_1 \subset A_2$. Notemos también que si $A_1 \subset A_2$ entonces $B_2 \subset B_1$, ya que A_1 está contenido en A_2 se tiene que $\bigcap_{a\in A_2} \Psi(a) \subset \bigcap_{a\in A_1} \Psi(a)$. En otras palabras, $(A_1, B_1) \leq (A_2, B_2)$ sii $B_2 \subset B_1$. El conjunto de todos los conceptos definidos en un contexto $\langle G, M, \psi\rangle$, se denota $\mathfrak{B}\langle G, M, \leq\rangle$. Conviene reconocer la relación R_Ψ, que Ψ induce en el conjunto $G \times M$, como el establecimiento de $gR_\Psi m$ sii $m \in \Psi(g)$. De esta forma tendremos que para un concepto (A, B), el conjunto asociado a un extento A: $\{m \in M|gR_\Psi m, \forall g \in A\}$ es igual a B. Denotamos por A' al anterior conjunto y de manera semejante expresamos una operación sobre los subconjuntos de M: $B' = \{g \in G|gR_\Psi m, \forall m \in B\}$ para lo cual se satisface $B' = A$, siempre que $(A, B) \in \mathfrak{B}$. En general, para cualquier $A \subset G$ tendremos que $A \subset A''$, puesto que A' es el conjunto de rasgos comunes a todos los objetos de A, y A'' incluirá a todos los objetos

de G con rasgos A', así A'' incluye, en particular, a los objetos de A. De igual forma puede verse que $B'' \subset B, \forall B \subset M$. Observemos, además, que se cumplen las siguientes igualdades para un contexto $\langle G, M, \psi \rangle$: Si $A \subset G$ y $B \subset M$, tenemos que $A = A'''$, $B = B'''$, y si $\{A_j\}_{j \in J}$ y $\{B_j\}_{j \in J}$ son dos familias de subconjuntos de G y M entonces $(\bigcap_{j \in J} A_j)' = \bigcup_{j \in J} A'_j$ y $(\bigcap_{j \in J} B_j)' = \bigcup_{j \in J} B'_j$. Lo anterior conduce a enunciar el siguiente teorema: Sea $\langle G, M, \psi \rangle$ un contexto, entonces $\mathfrak{B}\langle G, M, \leq \rangle$ es un lattice completo en la cual las operaciones **sup** y **inf** están dadas por:

$$\bigvee_{j \in J}(A_j, B_j) = ((\bigcup_{j \in J} A_j)'', \bigcap_{j \in J} B_j), \bigwedge_{j \in J}(A_j, B_j) = (\bigcap_{j \in J} A_j, (\bigcup_{j \in J} B_j)''). \qquad (5)$$

Una aplicación interesante del uso de lattices orientado al uso de índices es la que aparece en [Pedersen].