# Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado

por

**Rosa María Ortega Mendoza**

Tesis sometida como requisito parcial para obtener el grado de

**Maestra en Ciencias en la Especialidad de Ciencias Computacionales**

en el

**Instituto Nacional de Astrofísica, Óptica y Electrónica**

Supervisada por:

**Dr. Luis Villaseñor Pineda**
Investigador Titular del INAOE

**Dr. Manuel Montes y Gómez**
Investigador Titular del INAOE

Tonantzintla, Puebla
Diciembre 2007

# Resumen

Hoy en día, gracias a la Web, disponemos de una inmensa cantidad de textos electrónicos. Dada la disponibilidad y el fácil acceso a estos textos, ha surgido el interés por manipularlos de manera automática para extraer información relevante. La información extraída puede ser aprovechada para crear o enriquecer recursos léxicos. Generalmente, este tipo de recursos contiene conocimiento acerca de las palabras de un idioma. Típicamente, para construir automáticamente estos recursos se proponen métodos que extraen relaciones semánticas a partir del texto.

El presente trabajo de investigación se ubica dentro de la construcción automática de recursos léxicos. En particular, se centra en la construcción de un catálogo de hipónimos. Básicamente, el método propuesto se basa en el uso de patrones para abordar la extracción automática de hipónimos en textos no estructurados.

Tradicionalmente, los métodos que usan patrones para resolver el problema tienden a incluir información morfológica o sintáctica en la definición de los patrones. Sin embargo, en este trabajo se evita el uso de este tipo de información. Por lo tanto, los patrones se encuentran definidos en un nivel exclusivamente léxico. Esto propicia que el método sea independiente tanto del idioma como del dominio, pues se evita el uso de herramientas lingüísticas propias de un idioma (por ejemplo: etiquetadores, analizadores sintácticos, etc.); pero se favorece la extracción de información incorrecta (parejas de palabras que no representan una relación de hiponimia). Para enfrentar este inconveniente, se proponen dos enfoques que permiten estimar la confianza de las parejas hipónimo-hiperónimo extraídas.

Finalmente, para mostrar la utilidad del método propuesto se evaluó la precisión del catálogo de hipónimos resultante. Los resultados obtenidos son alentadores y muestran la factibilidad de usar patrones léxicos para extraer automáticamente hipónimos a partir de textos no estructurados.

# Abstract

Nowadays, thanks to the Web, we dispose of a huge number of electronic texts. Given the availability and easy access to these texts, it has emerged an interest for manipulating them in an automatic way with the aim to extract prominent information. The extracted information can be used to create or to enrich lexical resources. In general, this type of resources contains knowledge about the language's words. Typically, it proposes methods that extract semantic relationships from texts for building automatically these resources.

The present investigation work is located inside the automatic construction of lexical resources. In particular, this work is focused on the construction of a hyponyms catalog. Basically, the proposed method is based on the use of patterns to treat the automatic extraction of hyponyms in non-structured texts

Traditionally, methods that use patterns to solve the problem involve morphological or syntactic information in the patterns' definition. In contrast with these methods, we work without this type of information. Therefore, the patterns are defined exclusively at a lexical level. This way, the proposed method achieves language independence and domain independence. In addition, the use of linguistic tools characteristic of a language is avoided (for example: taggers, syntactic analyzers, etc.). However, the extraction of incorrect information is favored. The proposed method confronts this inconvenience by applying two approaches in order to estimate the confidence of the extracted hyponym-hypernym couples.

Finally, for showing the utility of the proposed method we evaluated the precision of the obtained catalog. The achieved results are encouraging and they show the feasibility of using lexical patterns to extract automatically hyponyms from non-structured texts.

# Agradecimientos

Un agradecimiento muy especial a mis asesores de tesis, Luis Villaseñor Pineda y Manuel Montes y Gómez, quienes siempre me brindaron apoyo, motivación, consejos y valiosas enseñadas.

Mis sinceros agradecimientos al Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) por las facilidades proporcionadas durante mis estudios de maestría.

También expreso mi gratitud al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo económico otorgado durante mis estudios de maestría a través de la beca No. 201730.

Finalmente, agradezco a mi familia por su apoyo constante e incondicional.

# Dedicatoria

*A mis queridos padres,*
*quienes en todo momento me brindan amor,*
*apoyo, palabras de aliento y motivación*
*para realizar mis sueños.*

*A Ricardo, el amor de mi vida,*
*por su apoyo, comprensión, por compartir mis ideales*
*y sobre todo, por el amor que siempre me ha brindado.*

# Contenido

# Lista de figuras

# Lista de tablas

# Capítulo 1

# Introducción

## 1.1 Motivación

La expresión escrita se ha convertido en un valioso medio de manifestación del lenguaje. Habitualmente, utilizamos el texto como un intermediario para comunicar nuestras ideas, sentimientos, y sobre todo, el conocimiento que hemos adquirido. Más aún, desde la aparición de la computadora, los textos en formato electrónico han llegado a considerarse una de las formas principales de intercambio de información, gracias a las posibilidades que ofrece la Web.

Actualmente, en el intento por manipular automáticamente los textos electrónicos se realizan tareas como: recuperación de información, clasificación de textos y traducción automática. Estas tareas pueden apoyarse en el uso de recursos léxicos tales como: diccionarios, ontologías, tesauros, etc. Estos recursos, tienen por objetivo principal establecer explícitamente una o varias relaciones entre las palabras de un idioma. Gracias a esta información, es posible mejorar el desempeño de estas aplicaciones relacionadas con el tratamiento automático de textos.

Así pues, la valiosa contribución de los recursos léxicos ha sido motivo de interés para estudiar y proponer métodos que permitan construir dichos recursos, ya sea de manera manual o automática.

En un principio, los recursos léxicos fueron construidos de manera manual, demandando importantes esfuerzos. Desafortunadamente, es prácticamente imposible plasmar en un solo recurso todas las relaciones semánticas entre las palabras de un idioma. Un ejemplo representativo de este tipo de recursos es WordNet (Miller, 1999), en el cual se han invertido incontables horas hombre en su construcción y, a pesar de ello, aún no se ha logrado cubrir todas las acepciones posibles de los términos del idioma inglés.

Dados estos inconvenientes, se han buscado alternativas menos ambiciosas que conduzcan a la construcción automática o semi-automática de estos recursos. Generalmente, esta construcción se limita a tratar dominios definidos o relaciones semánticas específicas. Entre las relaciones semánticas tratadas se pueden mencionar la sinonimia, antonimia, meronimia e hiponimia.

En particular, el presente trabajo de investigación se centra en la extracción de hipónimos. Un hipónimo es aquella palabra que posee todos los rasgos

semánticos de otra palabra más general -su hiperónimo- pero que añade en su definición otros rasgos semánticos que la especializan diferenciándola de la segunda. Por ejemplo, gato es un hipónimo de felino.

En la actualidad, existen varios métodos que han sido propuestos para extraer automáticamente instancias de la relación de hiponimia. Entre los enfoques más conocidos que abordan este problema destacan los enfoques basados en agrupamiento y los enfoques basados en patrones. Las características de cada uno de estos enfoques se mencionan en el capítulo tres.

Sin duda, uno de los trabajos más conocidos dentro los enfoques basados en patrones es el realizado por Hearst (1992), quien fue pionera en el uso de patrones para extraer instancias de relaciones semánticas. Ella propone un esquema basado en el uso de patrones léxico-sintácticos[1]. Desde entonces, ese trabajo ha sido la base de una variedad de investigaciones relacionadas con la extracción automática de hipónimos. De ahí que la información sintáctica sea ampliamente utilizada. Sin embargo, los métodos que usan información sintáctica dependen de herramientas lingüísticas como: etiquetadores de partes de la oración, analizadores morfológicos y sintácticos, etc. Dichas herramientas están limitadas a tratar idiomas específicos. Por tanto, la adaptación de estos métodos a otros idiomas, no es una tarea fácil. Dadas las limitaciones producidas por el uso de información sintáctica, surgen dos preguntas que han motivado el desarrollo del presente trabajo de tesis: ¿Es indispensable incluir información sintáctica para extraer

[1] En esta tesis, se reconocen dos niveles para tratar el texto previamente al descubrimiento de patrones. El "nivel léxico" hace referencia a una simple separación de unidades léxicas (*tokens*). El "nivel sintáctico" es un análisis más profundo, que puede incluir el reconocimiento de etiquetas de partes de la oración (etiquetas POS), un análisis morfológico, reconocimiento de entidades nombradas o propiamente, un análisis sintáctico.

automáticamente hipónimos?, en otras palabras, ¿Es suficiente el uso de información léxica?

Para responder las interrogantes anteriores, en esta tesis se propone un método que se circunscribe únicamente al uso de información léxica. De esta forma se evita la dependencia hacia herramientas lingüísticas específicas. Esto permite aplicar el método a diferentes idiomas sin realizar grandes cambios.

## 1.2 Descripción del problema

Como se mencionó anteriormente, el presente trabajo de investigación se enfoca en la extracción automática de hipónimos en textos no estructurados (textos escritos en prosa natural). Específicamente, se propone un método basado en patrones para adquirir automáticamente hipónimos de la Web. El uso de patrones se sustenta en la observación de que existen convenciones, frases o estilos que las personas repiten para expresar hipónimos dentro de un texto. Generalizando dichas convenciones o frases se pueden formar patrones de extracción. Estos patrones, al ser aplicados[2] a una colección de textos, permiten recuperar un conjunto de hipónimos.

Tradicionalmente, los métodos automáticos que usan patrones para resolver el problema consideran patrones expresados a nivel léxico-sintáctico. En contraste con esos métodos, el método propuesto emplea exclusivamente patrones expresados en un nivel léxico.

[2] Para aplicar los patrones, éstos son convertidos en expresiones regulares donde alguna variable actúa como un comodín que extrae fragmentos de texto.

Por supuesto, trabajar a nivel léxico sin considerar información sintáctica conlleva varios retos. En primer lugar, es difícil detectar los componentes de una oración (por ejemplo, artículos, verbos, entidades, etc.). En cambio, con información sintáctica, estos componentes sí podrían ser detectados, ofreciendo la posibilidad de crear patrones con un alto grado de generalización. Por otra parte, debido a la naturaleza de la información léxica, después de aplicar los patrones léxicos a una colección se pueden obtener numerosos falsos hipónimos.

Para afrontar estos retos, el método propuesto toma en consideración un gran número de patrones léxicos, lo que permite compensar su pobre grado de generalización. La adquisición de este gran número de patrones se logra a través de una técnica de minería de texto. Posteriormente, se estima la confianza de las parejas hipónimo-hiperónimo extraídas con la finalidad de determinar con mayor certeza los hipónimos de la colección, distinguiéndolos de aquellos falsos hipónimos. Para enfrentar este problema, el método contempla un proceso iterativo de estimación de la confianza de las parejas hipónimo-hiperónimo. El proceso de estimación se basa en la siguiente idea, la cual fue planteada con base en la observación de las parejas hipónimo-hiperónimo extraídas: las parejas confiables son extraídas por varios patrones diferentes, así mismo, los patrones confiables permiten extraer varias instancias confiables.

## 1.3 Objetivos

A continuación se presentan los objetivos propuestos en este tema de tesis.

### 1.3.1 Objetivo general

- Desarrollar un método para descubrir automáticamente hipónimos a partir de textos no estructurados (c.f. sección 2.3).

### 1.3.2 Objetivos específicos

- Definir y aplicar un método de minería de texto que permita obtener patrones léxicos de extracción para la relación de hiponimia.
- Desarrollar un método para determinar la confianza de los patrones léxicos de extracción obtenidos.
- Desarrollar un método para determinar la confianza de las parejas hipónimo-hiperónimo extraídas.

## 1.4 Organización de la tesis

El contenido de esta tesis está estructurado de la siguiente forma:

En el capítulo 2 se presentan los conceptos básicos que introducen al lector dentro del contexto de este trabajo de investigación. Primero, se presenta una introducción a los recursos lingüísticos y su clasificación. Allí se toma especial interés en la descripción y construcción automática de recursos léxicos (un tipo de recursos lingüísticos). Después, se describen las relaciones semánticas más comunes que han sido extraídas automáticamente e incluidas en recursos léxicos. Posteriormente, se trata el problema de la extracción de relaciones semánticas en textos no estructurados. Finalmente, se exponen las medidas utilizadas para evaluar el

desempeño general de un sistema relacionado con el tratamiento automático de textos.

En el capítulo 3 se presenta una revisión general de los trabajos enfocados en la extracción automática de hipónimos. Principalmente se examinan los métodos basados en diccionarios, los métodos basados en agrupamiento y los métodos basados en patrones.

En el capítulo 4 se presentan las características del método propuesto para extraer automáticamente hipónimos en textos no estructurados de la Web. Además, se describe la estructura general del método propuesto. Básicamente el método se compone de dos etapas. La primera etapa está orientada a recuperar de la Web, un conjunto de parejas hipónimo-hiperónimo. Por su parte, la segunda etapa está orientada a ordenar las parejas hipónimo-hiperónimo de acuerdo con su confianza.

En el capítulo 5 se detalla la primera etapa del método propuesto, la cual, como se mencionó anteriormente, está orientada a recuperar un conjunto de parejas hipónimo-hiperónimo. Este conjunto de parejas conforman el catálogo de hipónimos.

En el capítulo 6 se ubica la contribución principal de esta tesis. Allí se desarrolla la segunda etapa del método propuesto, la cual está orientada a ordenar (de acuerdo con su confianza) las parejas hipónimo-hiperónimo que conforman el catálogo de hipónimos. Es decir, aquellas parejas hipónimo-hiperónimo con mayor probabilidad de ser correctas se ubicarán en las primeras posiciones del catálogo. Para ello, se necesita estimar la confianza de cada pareja hipónimo-hiperónimo. Principalmente se exponen dos enfoques de estimación. El primer enfoque está basado en el uso de *información mutua*. Por su parte, el segundo está basado en el uso de la

*medida F*. Básicamente, la contribución principal queda resumida en una nueva propuesta para estimar la confianza de las parejas hipónimo-hiperónimo. Dicha propuesta se basa en una adaptación a la *medida F*.

Finalmente, en el capítulo 7 se exponen las conclusiones y direcciones futuras de este trabajo de tesis.

# Capítulo 2

# Conceptos básicos

El objetivo de este capítulo es introducir y familiarizar al lector con los conceptos básicos relacionados con el presente trabajo de investigación. Se espera que el contenido de este capítulo contribuya a lograr una mayor comprensión del tema.

## 2.1 Recursos lingüísticos

La construcción de recursos lingüísticos ha ganado importancia en el campo de la investigación de lingüística computacional y tratamiento automático de textos, pues estos recursos han enriquecido tareas relacionadas con esas

disciplinas. Particularmente, el método que se propone en esta tesis permite construir un catálogo de hipónimos, el cual puede ser considerado como un recurso lingüístico, específicamente un recurso léxico. Por ello, en esta sección se describen formalmente este tipo de recursos.

El término *recursos lingüísticos* se refiere a conjuntos de datos del lenguaje en formato legible por máquina. Estos recursos son usados en la construcción, mejoramiento o evaluación de sistemas del lenguaje natural. No obstante, el término puede ser extendido para incluir herramientas de software cuyo objetivo es preparar, coleccionar, administrar o usar otros recursos (Godfrey y Zampolli, 1997).

Los recursos lingüísticos pueden ser clasificados en tres categorías (Gellerstam, 1995): corpus, herramientas y recursos léxicos. Estas tres categorías serán descritas brevemente a continuación, prestando especial atención a los recursos léxicos.

**Corpus**

En (Sinclair, 1991) se define un corpus como una colección de textos en lenguaje natural, elegida para caracterizar un estado o variedad de un lenguaje. Sin embargo, en la actualidad existen diversos tipos de corpus, entre los que destacan: corpus del lenguaje escrito y corpus del lenguaje hablado. En cualquier caso, un corpus actúa como repositorio de información la cual puede ser manipulada para extraer conocimiento.

**Herramientas**

Actualmente existe una variedad de herramientas lingüísticas que ayudan a analizar los textos electrónicos. Entre las más comunes se encuentran:

- *Etiquetadores de partes de la oración (etiquetadores POS*)[3]. Estas herramientas están enfocadas a reconocer unidades léxicas dentro de un texto. Dichas unidades léxicas son referenciadas con etiquetas correspondientes a su categoría gramatical (por ejemplo, nombres, verbos, pronombres). Además, los etiquetadores POS incluyen, generalmente, una fase de lematización consistente en obtener la forma base de la palabra, es decir, el lexema. Por ejemplo, para la frase *"El jaguar es un felino*" , un etiquetador POS obtendrá una expresión como la siguiente:

  *[ART(el), NC(jaguar), VB(ser), ART(un), NC(felino)]*

  En la expresión anterior, *ART*, *NC* y *VB* son etiquetas[4]. Por su parte, las palabras entre paréntesis hacen referencia al lema de cada unidad léxica.

  En la actualidad existen diversos etiquetadores POS, entre los cuales se pueden mencionar: TreeTagger (Schmid,1994), TnT Tagger (Brants, 2000), SEPE (Jiménez y Morales, 2002), etc.

- *Analizadores morfológicos*. La información que proporcionan estas herramientas hace referencia a elementos morfológicos (i.e. número, género, tiempo). Por ejemplo, el análisis morfológico de la palabra "felinos" corresponde a la siguiente expresión:

  *Lema (felino) NC M PL*

  que indica que se trata de un sustantivo común (*NC*) plural (PL), de género masculino (*M*) y cuya forma base (*Lema*) es *felino*.

  Entre los analizadores morfológicos se pueden citar: SPOST (Farwell et al.,1995) y FreeLing (Carreras et al., 2004).

[3] Algunos etiquetadores POS pueden integrar un análisis morfológico
[4] Ejemplos de etiquetas POS: NC = nombre común, VB = verbo, ART = artículo

- *Analizadores sintácticos*. Estas herramientas producen un análisis sintáctico completo de una sentencia. Dicho análisis representa las relaciones sintácticas entre las diferentes unidades léxicas.

  Como ejemplos de analizadores sintácticos se pueden mencionar: LoPar (Graham et. al., 1980), TACAT (Castellón et al., 1998), etc.

Básicamente, estas herramientas existen para idiomas específicos y quizá existan idiomas que no dispongan de este tipo de herramientas. De ahí que los métodos que usan estas herramientas necesiten realizar algunos cambios considerables si desean adaptarse a otros idiomas. En contraste, el método propuesto solo requiere de una separación de unidades léxicas (*tokens*).

**Recursos léxicos**

Estos recursos contienen un conjunto de palabras válidas en un lenguaje. Así mismo, pueden contener propiedades lingüísticas, el significado de las palabras y/o relaciones entre las palabras o grupos de palabras. A la fecha se pueden encontrar diversos recursos léxicos, por ejemplo listas de palabras, tesauros, ontologías, banco de términos, glosarios, etc.

Hoy en día, la construcción de este tipo de recursos ha ganado importancia gracias a las distintas aplicaciones de estos recursos dentro del procesamiento del lenguaje natural. Entre las aplicaciones más comunes se encuentran las que se mencionan a continuación.

## Aplicaciones de los recursos léxicos

Dentro de la diversidad de aplicaciones de los recursos léxicos en el procesamiento del lenguaje natural, se pueden mencionar las siguientes:

- En recuperación de información, recursos léxicos como las ontologías, son usados para expandir consultas o peticiones. Es decir, los términos de la consulta se enriquecen con algunas palabras relacionadas a dichos términos. Regularmente, estas palabras son obtenidas de ontologías. El resultado de esta expansión de peticiones ha reflejado un incremento en el recuerdo (más documentos son recuperados), pero también se ha visto un decremento en la precisión (más documentos incorrectos son recuperados) (Mitkov, 2003).

- La extracción de información está basada en la identificación de entidades y extracción de hechos sobre esas entidades. El proceso es frecuentemente guiado por una plantilla que debe ser llenada. La semántica de la plantilla puede estar ligada a una ontología. Mientras que, para el reconocimiento de entidades y hechos se han usado recursos como WordNet o EuroWordNet.

- La generación automática de resúmenes puede utilizar relaciones de sinonimia e hiponimia para conectar sentencias de palabras que son diferentes, pero que están relacionadas.

- En desambiguación del sentido de las palabras, recursos como enciclopedias, tesauros y bases de conocimiento léxico, son utilizados para encontrar un sentido a cada ocurrencia de la palabra en consideración; eligiendo finalmente el sentido más conveniente para cada palabra según el contexto tratado.

Si bien esta lista no es una lista exhaustiva de las aplicaciones de este tipo de recursos, es un medio para manifestar su importancia en las tareas del procesamiento del lenguaje natural. No obstante, al pasar del tiempo, las tareas del procesamiento del lenguaje natural exigen apoyarse en recursos actualizados, especializados hacia un dominio en particular o simplemente recursos más extensos. Lo anterior ha motivado a los investigadores a diseñar métodos que construyan este tipo de recursos de manera automática. Generalmente, estos métodos automáticos se abocan a tratar especialmente, una o más relaciones semánticas. Existen varias relaciones semánticas que han sido objeto de este tipo de métodos, y de ellas se hablará en la siguiente sección.

## 2.2 Relaciones semánticas

Las relaciones semánticas relacionan las palabras de acuerdo a su significado. Existen varias relaciones semánticas, entre las más comunes para cuya extracción automática se han propuesto una gran variedad de métodos se encuentran:

### Sinonimia

Este tipo de relación se establece entre dos palabras diferentes, pero que en un cierto contexto tienen el mismo significado. Por ejemplo, las palabras *"elegir*" y *"escoger"* establecen una relación de sinonimia.

La extracción automática de sinónimos ha sido tratada por varios trabajos, de los cuales aquí sólo citamos algunos: (Lin et al., 2003), (Baroni y Bisi, 2004) y (Turney, 2001).

**Antonimia**

Esta relación se establece entre dos palabras cuyo significado es opuesto. Por ejemplo, las palabras *"rápido"* y "*lento*" son consideradas antónimos.

Generalmente, es difícil diseñar métodos automáticos para extraer este tipo de relación. De ahí que a la fecha exista una escasa cantidad de trabajos enfocados a extraer automáticamente instancias de esta relación. Entre ellos se pueden citar: (Lucero et al., 2004) y (Schwab et al., 2002).

**Meronimia**

Se trata de la relación que se da entre las *partes* y los *todos*, como en "*rueda"* y "*coche"*. Esta relación generalmente sigue el siguiente patrón: *"X es una parte de Y"* (Crystal, 2000).

Entre los trabajos que tratan este tipo de relación se encuentran: (Girju et al., 2003), (van Hage, 2006) y (Berland y Charniak, 1999).

**Hiponimia**

Se llama hiponimia a la relación de inclusión semántica de un término en otro. Por ejemplo, *roble* es un hipónimo de *árbol*. Así mismo, entre las palabras *"jaguar"* y *"felino"* se presenta una relación de hiponimia. Consecuentemente, la hiperonimia es la relación inversa a la hiponimia. Entonces, la hiperonimia es la relación de un término que abarca a otros semánticamente (Alcaraz y Martínez, 1997).

En particular, el método que se propone en este trabajo de investigación se enfoca en descubrir automáticamente hipónimos. Este tema también ha sido

tratado por trabajos tales como: (Hearst, 1992), (Caraballo, 1999) y (Pantel y Pennacchiotti, 2006) (ver capítulo 3 para una descripción más completa de los trabajos que tratan esta relación).

La relación de hiponimia es muy importante porque permite estructurar la información en categorías, de esta manera se facilita su búsqueda y entendimiento. A demás ella proporciona un nivel de generalización que permite definir relaciones entre datos en una forma concisa y abstracta sin tener que enumerar todos los casos concretos existentes en la relación. Además, este tipo de relación es la columna (*backbone*) de recursos como las ontologías (Cimiano, 2006). De la importancia de este tipo de relación nació el interés de desarrollar este tema de investigación. Es importante señalar que en este trabajo la información resultante no está organizada en una forma jerarquizada, más bien se descubren instancias de la relación de hiponimia entre dos términos, lo cual es el principio de la construcción de recursos más complejos como lo son las ontologías.

Ahora bien, existen métodos que extraen automáticamente instancias de relaciones semánticas, a través del tratamiento y análisis de textos estructurados (por ejemplo, diccionarios, glosarios, etc.). Otros métodos, trabajan con textos no estructurados. Dado que el método propuesto manipula textos no estructurados, en la siguiente sección se describirá este tipo de extracción.

## 2.3 Extracción de conocimiento en texto no estructurado

En la actualidad, muchos trabajos relacionados con el tratamiento automático de textos, usan la Web como corpus de datos, pues entre más grande sea el número de textos a manipular, más evidencia existirá y más información

podrá ser extraída. Básicamente, a través de la Web se puede encontrar una enorme cantidad de textos pertenecientes a diversos dominios. De manera general, los textos pueden ser clasificados de la siguiente manera (Tustison, 2004):

- *Texto estructurado*. Es almacenado en un formato riguroso (que permita diferenciar unas partes de otras en un documento). Por ejemplo: una base de datos, una tabla, diccionarios o glosarios. Cuando se conoce previamente la estructura de los datos, la extracción de información puede resultar favorecida. En la figura 1 se muestra un ejemplo de texto estructurado.

| Palabra | Hipónimos | Hiperónimo |
|---|---|---|
| felino | tigre, jaguar, etc. | mamífero |
| insecto | hormiga, abeja, etc. | artrópodo |

**Figura 1. Ejemplo de texto estructurado**

- *Texto semi-estructurado*. Este tipo de texto contiene algún tipo de estructura, pero no la suficiente para ser tratado como texto estructurado. Por lo tanto, se encuentra en un punto intermedio (por ejemplo, documentos html, sgml, xml). En estos textos la ubicación de las palabras sirve como base para extraer información. Por ejemplo, en la figura 2, previamente se puede conocer la estructura ordenada para los términos de la figura 1. Entonces, para este ejemplo los hiperónimos más generales estarán ubicados entre los encabezados *<h1></h1>*; entre los encabezados *<h2></h2>* se encontrarán términos más específicos; entre los encabezados *<h3></h3>* se encontrará la definición de los términos inmediatos superiores y de ella se pueden extraer ejemplos o hipónimos. Con esa estructura, el proceso de extracción se facilita en gran medida.

```
<html>
<head>
<title>Animales</title>
</head>
<body>
<h1 align="center">Mamíferos</h1>
<h2>Felinos</h2>
<h3>Los felinos son cazadores sigilosos. Por
ejemplo, el tigre, el jaguar, etc. </h3>
<h1 align="center">Artrópodos</h1>
<h2>Insectos</h2>
<h3>Los insectos se caracterizan por sus antenas.
Por ejemplo la hormiga, abeja, etc. </h3>
</body>
</html>
```

**Figura 2. Ejemplo de texto semi-estructurado**

- *Texto no estructurado*. Este tipo de texto no posee una estructura definida, es decir, es texto libre de formato. Ninguna parte del contenido tiene más importancia que otra. De ahí que la extracción de información en este tipo de textos no es una tarea sencilla y generalmente implica una etapa de pre-procesamiento. El texto no estructurado consiste generalmente de prosa en forma natural (i.e. documentos de texto creados con algún procesador de textos, el cuerpo de un mensaje de correo electrónico, etc.). La figura 3 muestra un ejemplo de texto no estructurado.

*Los felinos son mamíferos cazadores sigilosos. Entre los felinos más conocidos se encuentran: el jaguar, el tigre, el puma, etc.*

*...*

*Los insectos son artrópodos caracterizados por tener antenas. Por ejemplo, una hormiga es un insecto con antenas acodadas y patas largas. Una abeja es otro ejemplo de un insecto*

**Figura 3. Ejemplo de texto no estructurado**

En particular, en esta tesis se trabaja con textos[5] no estructurados. Por ello, a continuación se describen algunos detalles del proceso de extracción en este tipo de textos.

A saber, todo texto escrito en lenguaje natural es potencialmente útil para extraer instancias de alguna relación semántica. Por ejemplo, la frase "*El **fútbol** es el **deporte** favorito de toda mi familia"* ha sido escrita en lenguaje natural expresando la preferencia de un individuo. Esta frase es considerada un texto del tipo no estructurado; y a pesar de su simplicidad, se tiene la posibilidad de extraer conocimiento mediante su manipulación o tratamiento. Así pues, las palabras de la frase anterior pueden ser relacionadas para inferir[6] que: *fútbol* es un hipónimo de la palabra *deporte.*

Aunque en el ejemplo anterior parece sencillo extraer la relación semántica, realmente no es fácil extraerla automáticamente. Primero, porque el texto debe ser tratado para identificar las unidades léxicas y después, porque el contexto de cada palabra debe ser analizado (con alguna técnica, por ejemplo, el uso de patrones, técnicas de agrupamiento, etc) para inferir el conocimiento semántico. De hecho, hay casos donde existe una relación semántica, pero es más complicado extraerla. Por ejemplo, en la frase siguiente no es una tarea sencilla inferir que *fútbol* es un *deporte*: "*Los niños que practican algún deporte tienden a mejorar su salud. Por ejemplo, aquellos niños que practican fútbol consiguen fortalecer sus huesos con más facilidad".* La dificultad radica en que la palabra *deporte* y la palabra *fútbol* aparecen en oraciones diferentes. En estas condiciones, una persona podría deducir que *fútbol* es un *deporte*, ya que puede relacionar el contenido de cada oración. Sin embargo, una máquina requiere un análisis complejo para

[5] De aquí en adelante, cuando se mencione la palabra "texto" o "textos" se hará referencia a textos no estructurados

[6] Una forma de inferir que fútbol es un deporte es mediante la aplicación del patrón *"<hipónimo> es el <hiperónimo>"*

relacionar el contenido de cada oración (por ejemplo, un análisis de dependencias).

Los ejemplos anteriores muestran que aunque el texto no estructurado es la forma más natural para expresarse de manera escrita, no es trivial tratar este tipo de información de manera automática. Uno de los problemas a los que se enfrentan los métodos automáticos que extraen relaciones semánticas, radica en que todas las partes que componen el texto tienen la misma importancia. De hecho, una computadora sólo toma el texto como una secuencia de caracteres, sin distinción de verbos, artículos, nombres de organizaciones, etc. Para enfrentar este problema, algunos trabajos usan herramientas lingüísticas para pre-procesar el texto (i.e. etiquetadores POS, analizadores morfológicos, analizadores sintácticos).

Pues bien, este trabajo de investigación se pretende evitar el uso de herramientas lingüísticas para extraer hipónimos en textos no estructurados, de tal forma que la etapa de pre-procesamiento del texto consiste en una simple separación de unidades léxicas.

Por otro lado, una vez que los sistemas realizan la extracción de alguna relación semántica, miden su desempeño usando medidas como las que se presentan a continuación.

## 2.4 Medidas de evaluación del desempeño de un sistema

Cuando se construye automáticamente un catálogo de instancias de relaciones semánticas se necesitan mecanismos que permitan evaluar su

calidad. Para la evaluación, se utilizan las medidas: precisión, recuerdo (*recall)*[7] y medida F.

**Precisión**

Bajo el contexto de este trabajo al igual que en (Girju et al., 2003) la precisión se estima aplicando la fórmula 1. En (Maynard et al., 2006) se presenta una definición más general de este término.

$$P = \frac{|relaciones\ correctas\ recuperadas|}{|relaciones\ recuperadas|} \quad (1)$$

**Recuerdo**

Nuevamente bajo el contexto de esta tesis, el recuerdo se estima a través de las relaciones recuperadas como se indica en la fórmula 2.

$$R = \frac{|relaciones\ \ correctas\ recuperadas|}{|relaciones\ correctas|} \quad (2)$$

Ahora bien, en la ecuación 2 se observa la necesidad de conocer previamente el número de relaciones correctas. Sin embargo, es muy difícil conocer cuántas relaciones correctas existen en una colección de documentos y todavía más, cuando la colección utilizada es la Web. Por ejemplo, en el contexto de esta tesis, se recupera un catálogo de hipónimos tomando la Web como corpus de datos. Ahora, si se quisiera medir el recuerdo general del catálogo, se necesitaría saber cuántos hipónimos

[7] Existen varias acepciones para referirse a este término en español. Entre ellas se tienen: recuerdo, alcance, cobertura, evocación y recubrimiento. En particular, en esta tesis se utilizará la acepción *recuerdo*.

existen en la Web o en todos los documentos sobre los que se realizó la extracción, lo cual es prácticamente imposible. De ahí que muchos trabajos se enfoquen a evaluar la calidad de su catálogo únicamente a través de la precisión.

**La medida F**

La medida F ($F_\beta$) fue introducida por Rijsbergen (1979) y surge como una forma de combinar la precisión y el recuerdo en una sola medida. Para calcularla se usa la ecuación 3.

$$F_\beta = \frac{(\beta^2 + 1.0) \times P \times R}{\beta^2 \times P + R} \quad (3)$$

Donde $\beta$ toma valores correspondientes a números reales no negativos y representa la importancia relativa existente entre el recuerdo y la precisión. Por lo tanto, si el recuerdo y la precisión tienen la misma importancia, se debe usar un valor de $\beta = 1.0$. Pero si el recuerdo es más importante que la precisión se utiliza un valor de $\beta > 1$. Por el contrario, si el recuerdo es menos importante que la precisión, $\beta < 1$. Por ejemplo, si el recuerdo es dos veces más importante que la precisión, se debe emplear un valor de $\beta = 2$. Ahora, si el recuerdo tiene la mitad de importancia que la precisión se necesita un valor de $\beta = 0.5$. Usualmente se maneja un valor de $\beta = 1$, dando igual importancia a las dos medidas y donde $F_\beta$ se considera la media armónica de la precisión y el recuerdo.

# Capítulo 3

# Trabajo relacionado

A la fecha existen trabajos de investigación dedicados a la extracción de relaciones semánticas entre palabras (Hearst, 1992; Caraballo, 1999; Fleischman et al., 2003, Pantel et al., 2004; Ravichandran et al., 2004; Pantel et. al., 2006). Se han propuesto diferentes métodos dependiendo de la relación semántica a extraer. En particular, esta tesis se centra en la relación de hiponimia, también conocida como relación *is-a.*

Los trabajos que tratan la relación de hiponimia usan diversas técnicas para realizar la tarea. Entre los enfoques más comunes se encuentran: los métodos basados en diccionarios, los métodos basados en agrupamiento y

los métodos basados en patrones. En este capítulo se describen estos métodos.

### 3.1 Métodos basados en diccionarios

Los métodos basados en diccionarios asumen que los diccionarios en formato legible por máquina contienen conocimiento explícito de manera estructurada. Este conocimiento puede ser extraído recuperando instancias de relaciones semánticas, entre ellas la hiponimia.

Estos métodos tuvieron su auge hace casi dos décadas, y parten de la siguiente idea: el hiperónimo de una palabra puede aparecer en la primera frase nominal de la definición de la misma. Entonces, estos métodos están enfocados a encontrar la palabra que actúa como hiperónimo. Por ejemplo, en la siguiente frase:

***primavera***[8] "La estación entre invierno y verano en la cual aparecen flores"

se puede extraer *"estación"* como hiperónimo de "*primavera"* por estar en la primera frase nominal de la definición. En consecuencia se establece una relación de hiponimia entre los términos "*primavera"* y "*estación"*

Este tipo de métodos han sido explorados en: (Dolan et al., 1993), (Alshawi,1987) y (Calzolari, 1984), con la finalidad de recuperar automáticamente instancias de la relación de hiponimia.

Por otro lado, debido a la estructura generalmente regular de los diccionarios, las precisiones obtenidas son relativamente altas. Por ejemplo,

[8] Ejemplo tomado de (Dolan et al., 1993)

en (Dolan et al., 1993) se cita una precisión de 87%, Alshawi obtiene una precisión de 77% y Calzolari menciona haber obtenido una precisión de más del 90%.

Ahora bien, estos métodos son muy precisos; sin embargo, presentan fuertes desventajas. Por ejemplo, no contemplan términos específicos de un dominio. Lo anterior, se debe a que los diccionarios son, casi siempre, recursos muy generales que tratan términos comúnmente utilizados en varios dominios. Este inconveniente propició el interés en la exploración de otros enfoques para la tarea de extracción de hipónimos, algunos de estos enfoques se describen en las siguientes secciones.

### 3.2 Métodos basados en agrupamiento

Los métodos basados en agrupamiento toman como base la hipótesis de Harris citada por Cimiano (2006), la cual indica que “las palabras similares comparten contextos similares”. Bajo este enfoque las palabras se caracterizan por su contexto y éstas son agrupadas de acuerdo con la similitud entre contextos. La relación de hiponimia queda manifiesta entre cada elemento del grupo (i.e. cada palabra) y la etiqueta asociada al grupo.

Existen varios trabajos basados en agrupamiento y enfocados a extraer hipónimos automáticamente. Entre estos trabajos, se pueden citar los que a continuación se mencionan.

En (Pereira et al., 1993) se usó agrupamiento para construir una jerarquía no etiquetada de palabras. La técnica de agrupamiento utilizada en ese trabajo es descendente (*top-down*).

Riloff y Shepherd (1997) formaron grupos de nombres usando frases en aposición y conjunciones. El sistema que ellos proponen toma como entrada un conjunto de palabras semillas para cada categoría predefinida. Enseguida, relacionan frases nominales a las categorías mediante el uso de estadísticas de co-ocurrencia. Entonces, la salida del sistema es una lista de palabras asociadas a cada categoría. Posteriormente, Roark y Charniak (1998), tomando como base el trabajo de Riloff y Shepherd, desarrollaron un método que usa conjunciones, frases en aposición, listas y nombres compuestos para crear grupos de palabras relacionadas. Sin embargo, ellos usan estadísticas diferentes a las de Riloff y Shepherd.

En (Cimiano et al., 2004) se puede encontrar un estudio comparativo de las técnicas de agrupamiento más comúnmente utilizadas para aprender relaciones taxonómicas desde texto. Es decir, relaciones donde cada grupo está relacionado a otro grupo más general.

Por otro lado, la dificultad principal del enfoque basado en agrupamiento surge al intentar determinar automáticamente una etiqueta apropiada para cada uno de los grupos obtenidos (Pantel y Pennacchiotti, 2006). Para resolver este problema, Caraballo (1999) utilizó conjunciones que involucran la palabra *"other"* (como en *"X,Y, and other Zs"*) para etiquetar los grupos. En (Ravichandran et al., 2004) se presenta una sección que estudia un modelo que utiliza dependencias sintácticas para inducir los nombres o etiquetas de las clases semánticas generadas previamente por el algoritmo CBC (*Clustering by Committee*).

Generalmente estos métodos obtienen precisiones más altas cuando tratan corpus grandes, que cuando tratan corpus pequeños. Por ejemplo, en (Ravichandran et al., 2004) se reporta una precisión de 14.6 % cuando se trabajó con un corpus de 15 MB. Pero también se reporta una precisión de

64.9% cuando se trabajó con un corpus de 6 GB[9]. En ese trabajo, la evaluación fue realizada sobre una lista aleatoria de 50 instancias (hipónimos). Por cada instancia se extrajeron los tres conceptos (hiperónimos) más frecuentes, los cuales fueron calificados como correctos o incorrectos.

La ventaja de los métodos basados en agrupamiento es que permiten identificar relaciones de hiponimia incluso cuando no aparecen explícitamente en el texto, por ejemplo es posible encontrar hipónimos de una palabra aunque aparezcan en párrafos diferentes y alejados al de dicha palabra. Sin embargo, generalmente fallan cuando se desea producir grupos coherentes partiendo de corpus pequeños (menos de 100 millones de palabras) (Pantel y Pennacchiotti, 2006).

### 3.3 Métodos basados en patrones

Los métodos basados en patrones se apoyan en la idea de que existen frases, convenciones o estilos que las personas repiten al momento de relacionar un hipónimo con su hiperónimo dentro de un texto. Dichas frases o convenciones pueden generalizarse en forma de patrones. Estos patrones permiten extraer instancias de la relación de hiponimia al aplicarse sobre una colección de textos.

Los primeros intentos bajo este enfoque utilizaron patrones construidos manualmente. Es decir, después de observar la forma en que se describen y relacionan conceptos dentro de un texto, un experto identificaba y formaba un conjunto de patrones léxico-sintácticos usados comúnmente para

[9] En ese trabajo se reportan precisiones para tres sistemas (basado en co-ocurrencias, basado en patrones aplicando un filtro de máxima entropía y basado en patrones sin filtro). Las precisiones citadas aquí corresponden al sistema basado en co-ocurrencias.

introducir una pareja *hipónimo-hiperónimo*. Con el paso del tiempo, surgieron varias propuestas para crear automáticamente patrones de extracción de hipónimos. En las siguientes secciones se describirán los trabajos que han sido propuestos. Primero, se describirán los trabajos que hacen uso de patrones construidos manualmente. Posteriormente, se describirán los trabajos que hacen uso de patrones descubiertos automáticamente.

**Trabajos basados en patrones construidos manualmente**

El uso de patrones para la extracción de relaciones semánticas fue propuesto por (Hearst, 1992). En ese trabajo se presenta un método que utiliza, inicialmente, tres patrones léxico-sintácticos construidos manualmente para extraer hipónimos. Los tres patrones utilizados son mostrados en la siguiente lista[10]:

- *$NP_0$ such as {$NP_1$, $NP_2$ …, (and | or )} $NP_n$*
- *such NP as {NP,}*{(or |and)} NP*
- *NP {,NP}* {,} or other NP*

En estos patrones, cuando una relación de hiponimia es descubierta entre dos frases nominales, se realiza un proceso de lematización para relacionar únicamente las formas bases de las palabras.

Hearst plantea, por primera vez, un interesante mecanismo para descubrir semi-automáticamente nuevos patrones. Es decir, a partir de los patrones mencionados anteriormente se realiza un proceso, que aunque manual, permite extraer nuevos patrones. A través de este mecanismo Hearst descubre tres patrones más:

[10] En los patrones *NP* = Frase nominal; * = operador en expresiones regulares, el cual expresa 0 o más casos de la expresión que señalan.

- *NP{, NP}*{,} and other NP*
- *NP{,} including {NP , }*{or | and} NP*
- *NP{,} especially {NP ,}*{or | and} NP*

De acuerdo con Hearst, los patrones que descubrió satisfacían los siguientes requisitos:

- Ocurrir frecuentemente y en una gran variedad de géneros de texto.
- Indicar la relación de interés.
- Ser reconocidos sin o con poco conocimiento pre-codificado.

Después de la propuesta de Hearst, han surgido varios trabajos de investigación que aprovechan patrones creados manualmente para extraer hipónimos. Por ejemplo:

Mann (2002) aprovechó el uso de patrones léxicos-POS creados por observación para obtener instancias de la relación de hiponimia entre nombres propios.

En (Fleischman et al., 2003) se propone un esquema similar al de Mann. Pero se extiende el número de patrones utilizados en relación al trabajo de Mann. Además, también se integran técnicas de aprendizaje automático para filtrar las instancias extraídas.

Generalmente, los patrones creados de forma manual son muy precisos. No obstante, el lenguaje es tan variado y existen numerosas formas de introducir dos palabras que mantienen una relación de hiponimia. Entonces, se requeriría mucho esfuerzo si se intentara construir manualmente los patrones suficientes para describir la relación de hiponimia. Este inconveniente ha motivado el desarrollo de investigaciones enfocadas a descubrir patrones

automáticamente. En los siguientes párrafos se discuten algunos trabajos que utilizan patrones de extracción de hipónimos descubiertos de forma manual.

**Trabajos basados en patrones construidos automáticamente**

A partir de las ideas de Hearst, otros trabajos exploraron métodos automáticos para el descubrimiento de patrones. Básicamente, estos métodos parten de un conjunto de parejas de palabras que mantienen la relación deseada, conocidas comúnmente como *semillas*. Enseguida, se identifican los fragmentos de texto en donde aparecen dichas semillas y, finalmente, se generalizan dichos fragmentos para obtener los patrones de extracción. En los siguientes párrafos se mencionan algunos ejemplos de trabajos que utilizan métodos de este tipo.

Inicialmente, Pasca (2004) parte de patrones elegidos manualmente, los cuales resume en el siguiente patrón:

*<[StartOfSent] X [such as|including] N [and | , | .]>*.[11]

Con este patrón se extrae un conjunto de instancias de la relación de hiponimia. Enseguida, utiliza un mecanismo que permite descubrir nuevos patrones. De esta manera, los patrones nuevos aumentan el número de las instancias extraídas.

En (Pantel et al., 2004) se presenta un algoritmo para aprender automáticamente y a gran escala, patrones léxico-sintácticos en múltiples niveles (nivel léxico y nivel POS). En ese trabajo se menciona la importancia de delimitar los patrones léxicos a través de etiquetas de partes de la oración.

[11] Donde *StartOfSent* = Inicio de la sentencia; *X* = Frase nominal; *N* = Nombre;

En (Pantel y Pennacchiotti, 2006) se presenta un algoritmo para extraer relaciones semánticas. Este algoritmo toma como entrada un conjunto de semillas de una relación semántica en particular, para extraer instancias pertenecientes a esa relación. El mecanismo iterativo que se maneja en ese trabajo, permite aprender patrones léxico-sintácticos en cada iteración. En consecuencia, se aumenta el conjunto de instancias a la salida. Por último, en ese trabajo evalúan la confianza de las instancias y de los patrones con base en la información mutua que existe entre patrones e instancias de la relación.

Ahora bien, la ventaja de los métodos que se basan en patrones es que son muy confiables. Por ejemplo, en (Pantel y Pennacchiotti, 2006) se reporta una precisión de 85% sobre una muestra aleatoria de 20 instancias de un total de 200 instancias[12]. Sin embargo, la desventaja principal de estos métodos es que necesitan un corpus muy grande para encontrar suficientes patrones de todas las formas posibles que describen una relación de hiponimia (Cimiano, 2006).

En particular, el presente trabajo de investigación se ubica dentro de este grupo de trabajos. Específicamente, se propone un método que se basa en el uso de patrones descubiertos automáticamente. La característica distintiva de este trabajo es el uso de patrones expresados en un nivel exclusivamente léxico, pues en los trabajos mencionados anteriormente se usan patrones en un nivel léxico-sintáctico.

Así pues, el método propuesto está orientado a extraer hipónimos para los términos de un vocabulario definido previamente. El método inicia con un conjunto de semillas (parejas *hipónimo-hiperónimo*) que permiten descubrir

[12] En ese trabajo se presentan los resultados de diferentes experimentos, con distintas colecciones de documentos. La precisión citada corresponde a la precisión más alta que fue reportada para tratar la relación de hiponimia.

un conjunto de patrones léxicos de extracción de hipónimos. Enseguida, los patrones se aplican sobre la Web para extraer un conjunto de posibles hipónimos (para los términos del vocabulario). Finalmente, se estima la confianza de que las parejas mantengan una relación de hiponimia.

Ahora bien, es importante mencionar que a lo largo de este capítulo se mostraron ciertas precisiones reportadas en algunos trabajos. Sin embargo, ellas no representan un parámetro de comparación adecuado. Primero, porque cada trabajo evalúa su método con colecciones diferentes. Algunos hacen uso de la Web, y otros utilizan colecciones cerradas (por ejemplo, colecciones de noticias). Además, la mayoría de los trabajos, reportan la precisión obtenida en una muestra aleatoria. Sin embargo, el tamaño de la muestra difiere en cada trabajo, y no todos realizan un muestreo estadístico para elegir el tamaño adecuado de la misma. En concreto, las condiciones de evaluación son distintas. De ahí que para comprobar el alcance de nuestro método se haya definido un punto de referencia propio.

# Capítulo 4

# Método propuesto

En este capítulo se presenta una descripción general del método propuesto para extraer automáticamente hipónimos en texto no estructurado. Primero, se presentan las características que distinguen al método propuesto de otros que han abordado este problema. Posteriormente, se presenta la arquitectura general del método. Dicha arquitectura se compone de dos etapas. La primera etapa está enfocada a recuperar un conjunto de parejas hipónimo-hiperónimo. Por otra parte, la segunda etapa está orientada a definir esquemas que permitan estimar la confianza de las parejas. Para brindar más claridad al lector, la primera etapa se describe en el capítulo 5. Por su parte, la segunda etapa se detalla en el capítulo 6.

## 4.1 Características del método propuesto

El método que se propone en el presente trabajo de investigación aborda el problema de la extracción automática de parejas *hipónimo-hiperónimo* a partir de textos no estructurados tomados de la Web. La idea es formar un catálogo de hipónimos relacionado a un vocabulario predefinido. Para ello, el método se basa en el uso de patrones.

Generalmente, los trabajos basados en el uso de patrones para extraer hipónimos automáticamente se han enfocado al descubrimiento de patrones léxico-sintácticos. Este tipo de patrones tienen un alto nivel de generalización. Es decir, con un único patrón se pueden capturar muchas de las formas posibles que expresan una relación de hiponimia. Por ejemplo, el siguiente patrón léxico sintáctico:

*NP {NP,} * {,} y otros NP*[13]

encierra en la etiqueta *NP* un abundante conjunto de las maneras posibles de crear una frase nominal, y también se tiene la posibilidad de lematizar las frases para trabajar sólo con las formas base de las palabras. De ahí que los patrones léxico-sintácticos tengan un nivel de generalización muy grande.

Así mismo, los patrones léxico-sintácticos tienden a ser precisos. Es decir, tienen una alta capacidad de extraer correctamente un par de palabras que mantengan la relación deseada. Sin embargo, no están exentos de extraer información incorrecta. Por ello, algunos métodos automáticos que trabajan con patrones léxico-sintácticos (por ejemplo: (Pantel y Pennacchiotti, 2006) y (Ravichandran y Hovi, 2002)) evalúan la confiabilidad de sus patrones y únicamente usan aquellos más confiables. Además, la construcción de estos

[13] Patrón tomado de (Hearst, 1992) y traducido al español

patrones no es sencilla, pues se depende de herramientas lingüísticas (por ejemplo, analizadores sintácticos, etiquetadores de partes de la oración, etc.).

Ahora bien, el método que se propone en este trabajo de investigación trata con patrones expresados en un nivel exclusivamente léxico. Construir este tipo de patrones es simple y no se necesita un fuerte conocimiento del idioma. Tampoco se depende de herramientas como etiquetadores o analizadores sintácticos, pues no incluyen información morfológica, ni tampoco sintáctica. Básicamente, estos patrones sólo surgen de relacionar palabras (unidades léxicas). Sin embargo, estos patrones tienen un nivel de generalización pobre. Por ejemplo, considere los siguientes patrones léxicos:

*los (palabra)$^{+}$ y otros (palabra)* $^{+}$[14]
*las (palabra)* $^{+}$ *y otros (palabra)* $^{+}$
*unas (palabra)* $^{+}$ *y otros (palabra)* $^{+}$
*unos (palabra)* $^{+}$ *y otros (palabra)* $^{+}$

Los patrones léxicos anteriores son sólo algunos de los expresados por el patrón léxico-sintáctico: *NP {NP,} * {,} y otros NP.* Dado que el nivel de generalización de los patrones léxicos es pobre, éstos declaran explícitamente algunas de las formas posibles de construcción de una frase nominal.

Pues bien, para afrontar el problema de la pobre generalización de los patrones léxicos se necesita descubrir un gran número de ellos, ya que los patrones léxicos son más específicos y no tienen una alta capacidad de extracción. Entonces, al descubrir un gran número de patrones léxicos se trata de capturar un gran número de las formas posibles que expresan una

[14] En expresiones regulares el signo + indica 1 o más ocurrencias. En este caso, 1 o más palabras.

relación de hiponimia. No obstante, esto impacta en la precisión del catálogo final, pues es posible descubrir patrones muy amplios (patrones que se aplican numerosas veces en un texto) con baja precisión. Para resolver este problema, el método propuesto itera los procesos de estimación de confianza de los patrones y estimación de confianza de las parejas. De esta manera, una pareja hipónimo-hiperónimo será considerada pertinente si varios patrones la extraen, y de igual manera, un patrón será adecuado mientras mayor número de parejas correctas recupere. Este proceso iterativo permite estimar la confianza de los patrones con base en la evidencia determinada en cada ciclo. Cabe resaltar que bajo este esquema es posible discernir la información pertinente extraída con patrones amplios, sin necesidad de aplicar sólo los patrones con mayor confianza. Inclusive, el proceso iterativo trata de aprovechar todos los patrones (tanto amplios como específicos) para enriquecer la información que ayudará a estimar la confianza de las parejas.

## 4.2 Arquitectura del método propuesto

En la figura 4 se ilustra la arquitectura general del método propuesto. Se observa que el método consta de tres fases organizadas en dos etapas. La etapa 1 permite recuperar el conjunto parejas hipónimo-hiperónimo (por convención, a partir de este momento, se utilizará el término *tupla* para referirse a cada pareja hipónimo-hiperónimo) que conformarán el catálogo de hipónimos. Para ello, contempla las fases: *descubrimiento de patrones* y *extracción de tuplas*. Por su parte, la etapa 2 ordena las tuplas del catálogo estimando su confianza. Esta etapa consta de una única fase: *ordenamiento de tuplas*. En ella se desarrollan esquemas de estimación que permiten determinar la confianza de las tuplas. Ahora bien, como ya se ha mencionado, la etapa 1 se describe detalladamente en el capítulo 5. Por otra parte, la etapa 2 se expone en el capítulo 6.

A partir de la arquitectura mostrada, se observa que el método parte de un conjunto de semillas, las cuales representan una relación de hiponimia. Con estas semillas se descubre un conjunto de patrones léxicos de extracción de hipónimos. Posteriormente, estos patrones son enviados a la Web para encontrar un conjunto de tuplas (parejas *hipónimo-hiperónimo)*. Estas tuplas están relacionadas a un determinado vocabulario. Entonces, el conjunto de tuplas extraídas conforma el catálogo de hipónimos. Finalmente, las tuplas son ordenadas de acuerdo con su confianza, de manera que aquellas tuplas con más probabilidad de ser correctas (es decir, que realmente expresen una relación de hiponimia) se localicen en las primeras posiciones del catálogo.

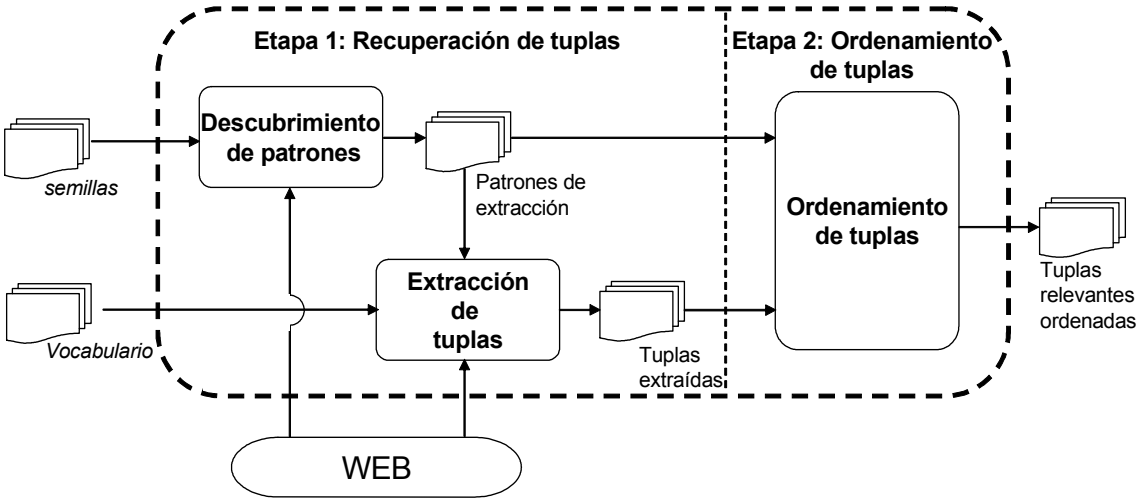


**Figura 4. Arquitectura general del método propuesto**

Cabe mencionar que el uso de patrones léxicos propicia que el método propuesto funcione para cualquier dominio. Por supuesto, si en la Web no se encuentran datos relacionados al dominio, no se recuperará la información requerida (hipónimos).

# Capítulo 5

# Recuperación de tuplas

En este capítulo se describe la etapa 1 del método propuesto, la cual está orientada a recuperar un conjunto de hipónimos. Esta etapa se compone de dos fases: descubrimiento de patrones y extracción de tuplas. En la primera fase se descubre un conjunto de patrones de extracción, a partir de un conjunto de semillas. Después, en la segunda fase se extrae un conjunto de tuplas mediante la aplicación de los patrones sobre la Web. Al final de la etapa 1 se tendrá un conjunto de hipónimos relacionados a un vocabulario específico. En la parte final de este capítulo se muestran los resultados de los experimentos realizados correspondientes a la etapa 1.

## 5.1 Descubrimiento de patrones

En la Web podemos encontrar una gran cantidad de textos pertenecientes a diversos dominios. Cada uno de estos textos puede introducir en su contenido algunas instancias de la relación de hiponimia. Generalmente, dichas instancias son expresadas según el estilo de redacción del propio autor. Si bien el estilo difiere en cada persona, es posible encontrar ciertas regularidades que permitan generalizar las convenciones o frases que expresan una relación de hiponimia. En esta tesis, esta generalización se logra a través del descubrimiento de un conjunto de patrones léxicos.

Idealmente, el conjunto de patrones captura la mayoría de las convenciones que introducen la relación de hiponimia entre dos palabras dentro de los textos. Ahora bien, en esta tesis para descubrir los patrones léxicos, se adaptó el método descrito en (Denicia et al., 2006). Más concretamente, los patrones fueron descubiertos a través de los módulos mostrados en la figura 5 y los cuales serán descritos a continuación.

### 5.1.1 Selección de semillas

El primer módulo para extraer el conjunto de patrones léxicos consiste en seleccionar un conjunto de semillas (ver figura 5). Las semillas ayudarán a reunir un conjunto de ejemplos, los cuales muestran el uso de la relación de hiponimia dentro de los textos.

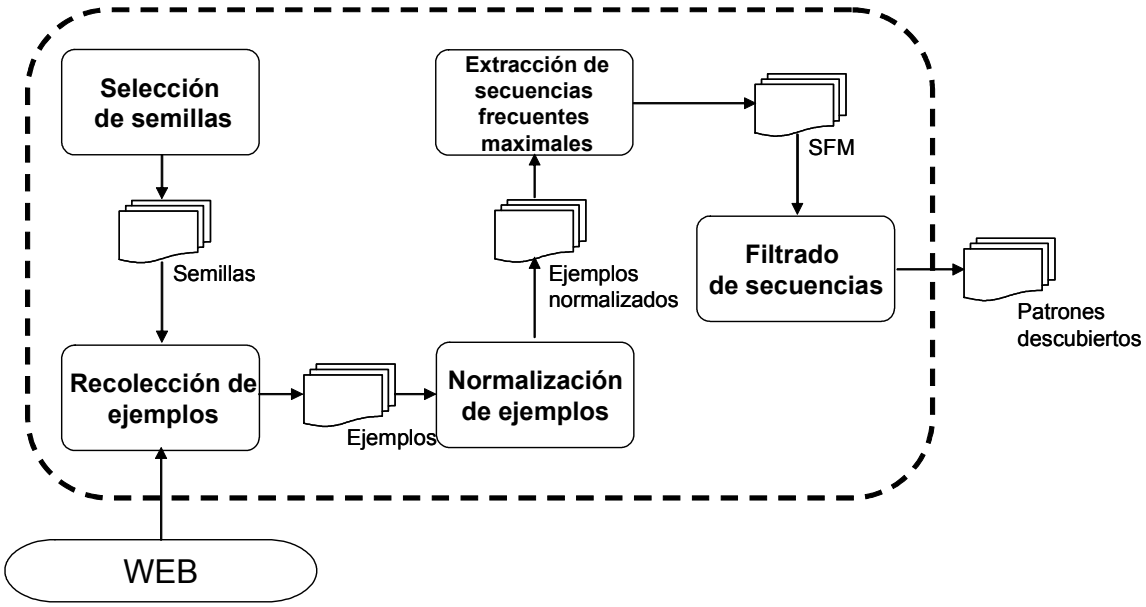


**Figura 5. Módulos de la fase: descubrimiento de patrones**

Ante todo, se debe señalar que las semillas se componen de dos términos, los cuales exhiben una relación de hiponimia. En otras palabras, las semillas conservan la siguiente estructura:

*hipónimo-hiperónimo*

Entonces, las parejas de palabras citadas en la tabla 1 representan ejemplos de semillas para la relación de hiponomia.

**Tabla 1. Ejemplos de semillas para la relación de hiponimia**

| ***Ejemplos de semillas*** |
|---|
| *agua-líquido* |
| *gato-animal* |
| *manzana-fruta* |

Por otro lado, es importante mencionar que el proceso de selección de semillas se realiza de forma manual. Lo anterior, debido a la necesidad de recuperar ejemplos verídicos de la relación de hiponimia. Sin embargo, y a pesar de la importancia de seleccionar manualmente buenas semillas, no se cuenta con un estudio preliminar orientado a elegir las semillas más adecuadas para la tarea. No obstante, en este trabajo de investigación se usaron los siguientes criterios de selección:

1. Identificar semillas frecuentes, es decir, semillas que permitan recuperar una gran cantidad de ejemplos del uso de la relación de hiponimia. Por ejemplo, en la tabla 2 se muestran algunas posibles semillas. Así mismo, se señala el número de documentos que pueden ser recuperados de la Web (mediante un motor de búsqueda) usando dichas semillas. Con base en este criterio, la semilla *águila-ave* resulta más adecuada en comparación con las otras dos. En contraste, *agamí-ave* no parece una semilla apropiada para la tarea, pues a través de ella, no se podría obtener un gran número de ejemplos.

**Tabla 2. Ejemplo de selección de semillas frecuentes**

| Semilla | Doc. Recuperados |
|---|---|
| águila - ave | 379,000 |
| garza - ave | 228,000 |
| agamí - ave | 220 |

2. Usar semillas pertenecientes a distintos dominios. De esta manera se evita la especialización de los patrones a un dominio específico. En la tabla 3 se muestra un ejemplo de un conjunto de semillas que cumplen este criterio por pertenecer a dominios diferentes.

**Tabla 3. Ejemplo de selección de semillas de diversos dominios**

| Semilla | Doc. Recuperados |
|---|---|
| águila - ave | 379,000 |
| bulimia - enfermedad | 409,000 |
| arroz - cereal | 1,200,000 |

Además de los criterios mencionados anteriormente, es recomendable contemplar semillas de género tanto masculino como femenino. Por ejemplo, las siguientes semillas siguen esta recomendación: *gato-animal*(género masculino) y *manzana-fruta*(género femenino). Así mismo, se recomienda integrar combinaciones de número(plural y singular). Por ejemplo, gatos(*plural*)-animales(*plural*) y manzana(*singular*)-frutas(*plural*). Estas recomendaciones son muy importantes cuando se trabaja a nivel léxico, pues en este nivel, no hay una generalización de las variantes morfológicas de una palabra, en una sola etiqueta (sustantivo, artículo, verbo, etc.). En cambio, esta generalización sí existe cuando se trabaja a nivel sintáctico.

En concreto, los criterios de selección de semillas tienen la finalidad de capturar, en tanto como sea posible, la mayor diversidad de patrones de extracción de hipónimos.

Por otro lado, como ya se ha mencionado, las semillas seleccionadas se usarán para recolectar un conjunto de ejemplos del uso de la relación de hiponimia. La recolección de ejemplos es el segundo paso (módulo) en la fase *descubrimiento de patrones* y será descrito a continuación.

### 5.1.2 Recolección de ejemplos

En este módulo se recuperan de la Web, fragmentos de texto que muestran cómo las personas relacionan un hipónimo con su hiperónimo dentro de los textos. Estos fragmentos de texto, en este documento, han sido nombrados *ejemplos del uso de la relación de hiponimia*.

Ahora bien, la razón principal del uso de la Web como repositorio de información se atribuye a la inmensa cantidad de documentos que en ella se pueden encontrar y que están relacionados con diversos dominios. Además, ya que las semillas pertenecen a varios dominios, el uso de la Web es apropiado para la tarea. Por otro lado, entre mayor sea el número de ejemplos recolectados, mayor será la evidencia para descubrir patrones.

Básicamente, para obtener los ejemplos, todas las semillas son pasadas a la Web como consultas mediante un motor de búsqueda. Al finalizar este paso, se tendrá un conjunto de ejemplos[15] como los que se muestran en la tabla 4.

**Tabla 4. Ejemplos del uso de la relación de hiponimia para la semilla águila-ave**

| **Ejemplos del uso de la relación de hiponimia** |
|---|
| *"El **águila** es un **ave** extraordinariamente territorial…"* |
| *"El **águila** como **ave** de los dioses…"* |
| *"El **águila** es un **ave** de rapiña muy grande y muy poderoso…"* |
| *"Como el **águila** es el **ave** suprema cazadora del cielo…"* |
| *"El **águila** era un **ave** mítica entre los aztecas…"* |

[15]En la presente tesis, los ejemplos corresponden a los *snippets (segmentos de texto)* que muestran los resultados de una consulta a la Web a través de un motor de búsqueda.

Después de obtener los ejemplos, éstos deben ser preparados para el descubrimiento de patrones a través de una normalización, la cual es explicada en la siguiente sección.

### 5.1.3 Normalización de ejemplos

Una vez recolectados los ejemplos del uso de la relación de hiponimia, éstos son normalizados con la finalidad de facilitar el descubrimiento de patrones de extracción de hipónimos.

Como se ha mencionado anteriormente, los ejemplos recuperados se encuentran relacionados con las semillas. Para normalizarlos, cada término de la semilla que está presente en los ejemplos, es cambiado por la etiqueta *<hipónimo>* o por la etiqueta *<hiperónimo>* según corresponda. Por ejemplo, en la tabla 5 se exhibe el resultado de la normalización de los ejemplos del uso de la relación de hiponimia presentados en la tabla 4.

**Tabla 5. Normalización de ejemplos**

| **Ejemplos del uso de la relación de hiponimia** |
|---|
| *"El <hipónimo> es un <hiperónimo> extraordinariamente territorial…"* |
| *"El <hipónimo> como <hiperónimo> de los dioses…"* |
| *"El <hipónimo> es un <hiperónimo> de rapiña muy grande y poderoso…"* |
| *"Como el <hipónimo> es el <hiperónimo> suprema cazadora del cielo…"* |
| *"El <hipónimo> era un <hiperónimo> mítica entre los aztecas…"* |

Básicamente, la normalización prepara a los ejemplos para la extracción de secuencias frecuentes maximales que permitan inferir patrones. El proceso de extraer secuencias se expone a continuación.

### 5.1.4 Extracción de secuencias frecuentes maximales

Este módulo tiene como objetivo encontrar las secuencias de palabras frecuentes que introducen una relación de hiponimia. Para lograrlo, se aplica una técnica de minería de texto sobre los ejemplos normalizados. En esta tesis se utilizó un algoritmo para extraer secuencias frecuentes maximales (SFM's). Particularmente, se trabajó con la implementación presentada en (García-Hernández et al., 2006).

Para comprender el concepto detrás de las secuencias frecuentes maximales, es necesario describir primero los siguientes conceptos.

Se asume que $D$ es un conjunto de textos cuya cardinalidad está denotada por $|D|$ (un texto puede representar un documento completo o incluso una sola sentencia), y cada texto consiste de una secuencia de palabras. A continuación reproducimos las siguientes definiciones de Ahonen-Myka (2002):

*Definición 1. Sobre las subsecuencias.*
Una secuencia $p = a_1 \ldots a_k$ es una subsecuencia de una secuencia $q$ si todos los ítems $a_i$, $1 \leq i \leq k$, ocurren en $q$ y además ocurren en el mismo orden que en $p$. Si una secuencia $p$ es una subsecuencia de una secuencia $q$, entonces se dice que $p$ ocurre en $q$.

*Definición 2. Sobre las secuencias frecuentes*
Una secuencia $p$ es frecuente en $D$ si $p$ es una subsecuencia de por lo menos $\beta$ textos de $D$, donde $\beta$ es un umbral de frecuencia dado.

*Definición 3. Sobre las secuencias frecuentes maximales*
Una secuencia $p$ es una secuencia frecuente maximal en $D$ si no existe ninguna secuencia $p'$ en $D$ tal que $p$ sea una subsecuencia de $p'$ y $p'$ sea frecuente en $D$.

Ahora bien, el problema de encontrar las secuencias frecuentes maximales de una colección de documentos puede formularse como: Dada una colección de textos $D$ y un valor entero arbitrario de $\beta$ tal que $1 \leq \beta \leq |D|$, enumerar todas las secuencias frecuentes maximales en $D$ con un umbral $\beta$.

El umbral $\beta$ depende del número de los ejemplos normalizados. Es decir, si el número de ejemplos normalizados es muy grande (por ejemplo 100,000), $\beta$ no debe ser muy pequeño (como el valor 2 o el valor 3), pues se obtendrían SFM's muy específicas, las cuales no podrían ser patrones adecuados porque serían poco aplicables. Por otro lado, si se elige un valor de $\beta$ muy grande (por ejemplo, mayor al 50% del número de ejemplos) se tendrían SFM's muy cortas (pronombres, artículos, etc.) que no podrían considerarse patrones.

De acuerdo con lo anterior, se observa que las SFM's poseen características que benefician la extracción de patrones. Primero, las secuencias de este tipo conservan el orden de las palabras de acuerdo con el orden mostrado en el texto. Segundo, permiten generar secuencias de diferentes tamaños. Tercero, las secuencias no están restringidas a una longitud específica, facilitando el descubrimiento de patrones con longitudes diferentes y sin un tamaño límite.

Entonces, al aplicar el algoritmo de extracción de secuencias frecuentes maximales obtenemos cadenas como las que se muestran en la tabla 6.

**Tabla 6. Ejemplos de secuencias frecuentes maximales**

| Longitud | Frecuencia | Secuencia |
|---|---|---|
| 2 | 11 | : no |
| | 42 | es muy |
| | 26 | <hipónimo> - |
| 3 | 36 | <hipónimo> (<hiperónimo>) |
| | 10 | el tratamiento <hiperónimo> |
| | 28 | <hiperónimo> : <hipónimo> |
| 5 | 63 | el <hipónimo> es una <hiperónimo> |
| | 12 | de <hiperónimo> como <hipónimo> y |
| | 17 | <hiperónimo> de la <hipónimo> |
| 6 | 15 | la <hipónimo> es el <hiperónimo> de |
| | 11 | el <hipónimo> es un <hiperónimo> que |
| | 11 | el <hipónimo> es el único <hiperónimo> |

Las secuencias resultantes podrían representar patrones de extracción de hipónimos. Sin embargo, sólo son consideradas como patrones hasta el momento que pasan satisfactoriamente un proceso, el cual ha sido nombrado *filtrado* en este documento. Este proceso es explicado a continuación.

### 5.1.5 Filtrado de secuencias frecuentes maximales

El último módulo del descubrimiento de patrones involucra la aplicación de un filtro a las secuencias frecuentes maximales previamente obtenidas.

Ante todo, se debe tener presente que no todas las secuencias frecuentes maximales pueden representar un patrón léxico adecuado para la extracción de hipónimos. Por ejemplo, en la tabla 6 podemos encontrar secuencias como "*: no", "*es muy" etc. Estas secuencias no son apropiadas para funcionar como patrones de extracción, porque no contienen las etiquetas <hipónimo> <hiperónimo>. Por otro lado, estas etiquetas actúan como un comodín para extraer información, por ello, la carencia de las mismas en un patrón no permitiría extraer información.

En concreto, el grupo de patrones léxicos de extracción de hipónimos es conformado por todas aquellas secuencias frecuentes maximales que satisfagan los siguientes tres criterios:

1. Que contengan ambas etiquetas en su definición (<hipónimo> y <hiperónimo>), pues la etiqueta <hipónimo> actúa como un comodín para obtener expresiones aspirantes a ser hipónimos. Por su parte la etiqueta <hiperónimo> será reemplazada por los términos del vocabulario (ver sección 5.2.2). De ahí la necesidad de exigir que se conserven ambas etiquetas.

2. Que no contengan signos de puntuación. Este segundo punto del filtro consiste en retener únicamente aquellas secuencias que no contengan signos de puntuación en su definición. Si bien las secuencias con signos de puntación constituirían patrones útiles en otras tareas (por ejemplo, búsqueda de respuesta usando patrones sobre colecciones cerradas) para el método propuesto no tienen relevancia. Lo anterior, es generado porque se utiliza la Web para formar el catálogo de hipónimos y generalmente, los motores de búsqueda no reconocen los signos de puntuación en sus consultas. Para ejemplificar este punto retomemos la tabla 6. Se puede observar

que algunas de las secuencias que no pasan este punto del filtro son: *“<hiperónimo> : <hipónimo>”* , *“ : no”, “<hipónimo> -”, <hipónimo> (<hiperónimo>)”.*

3. Que contengan fronteras entre etiquetas. Es decir, que satisfagan las expresiones de la tabla 7:

**Tabla 7. Expresiones para los patrones de extracción de hipónimos**

| **Expresiones** |
|---|
| *<frontera izquierda><hipónimo><frontera centro><hiperónimo>* |
| *<hiperónimo><frontera centro><hipónimo><frontera derecha>* |

Las expresiones anteriores surgieron, en el contexto de la presente tesis, como un mecanismo para delimitar cuándo inicia y termina la cadena de caracteres que representará al hipónimo. Otra forma de hacerlo es usar etiquetadores de entidades nombradas. Sin embargo, el método trabaja con patrones a nivel léxico evitando el uso de herramientas lingüísticas. Por tanto, se decidió resolver el problema utilizando fronteras. Entonces, en las expresiones anteriores las etiquetas: *<frontera izquierda>*, *<frontera centro>* y *<frontera derecha>* son secuencias de caracteres que funcionan como fronteras y delimitan el inicio y final de los hipónimos.

Es importante señalar que, de acuerdo con el método, no se necesitan fronteras para delimitar al término hiperónimo, pues este término estará especificado por los términos del vocabulario (ver sección 5.2.2). Para mayor claridad, se presenta el ejemplo de la figura 6. En esa figura se consideró el patrón: “*el <hipónimo> es el único <hiperónimo>”*. Entonces, con base en la figura 6 se puede observar que las fronteras: izquierda y centro, delimitan a la cadena *“lince*

*ibérico"* como un hipónimo para *"felino"*, donde *"felino"* actúa como frontera izquierda del patrón.

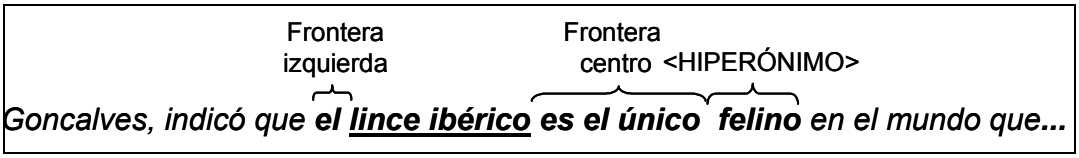


**Figura 6. Función de las fronteras de los patrones léxicos**

Finalmente, todas aquellas secuencias que pasan satisfactoriamente el filtro son conocidas como *patrones léxicos de extracción de hipónimos*. Por ejemplo en la tabla 8 se muestran los patrones léxicos de las SFM's de la tabla 6.

**Tabla 8. Patrones léxicos de extracción obtenidos de la tabla 6**

| No. | Patrones léxicos |
|---|---|
| 1 | el <hipónimo> es una <hiperónimo> |
| 2 | de <hiperónimo> como <hipónimo> y |
| 3 | la <hipónimo> es el <hiperónimo> de |
| 4 | el <hipónimo> es un <hiperónimo> que |
| 5 | el <hipónimo> es el único <hiperónimo> |

## 5.2 Extracción de tuplas

Esta es la segunda fase del método propuesto. La finalidad de esta fase es: extraer las tuplas que conformarán al catálogo de hipónimos, el cual está relacionado a un vocabulario predefinido. Para lograr lo anterior, básicamente los patrones léxicos de extracción son aplicados a una colección de documentos.

De manera general, en la figura 7 se presentan los módulos necesarios para realizar la extracción de tuplas. Se observan dos elementos de entrada: el vocabulario del dominio y un conjunto de patrones de extracción. Los patrones corresponden a aquellos descubiertos en la fase anterior. Por otro lado, la siguiente sección introduce la función del vocabulario.

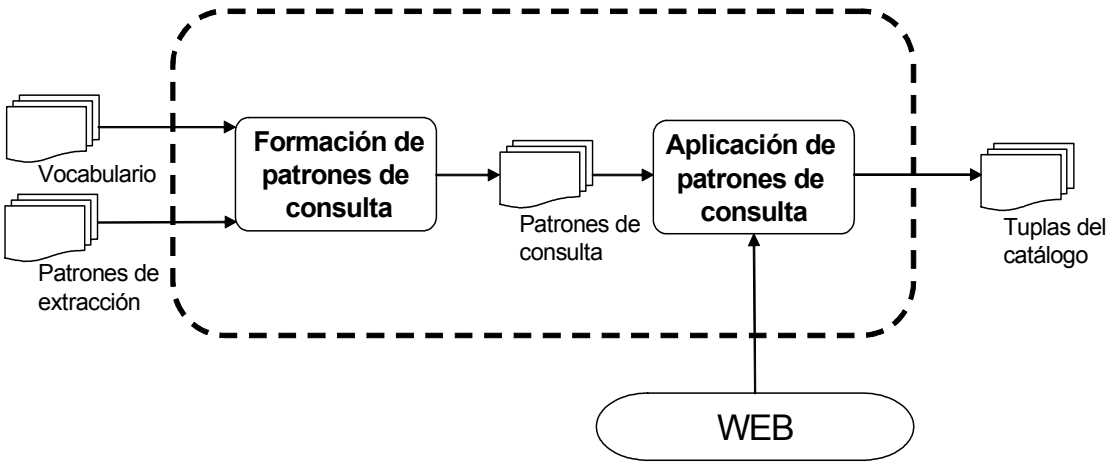


**Figura 7. Módulos de la fase: extracción de tuplas**

### 5.2.1 Sobre el vocabulario del dominio

El vocabulario del dominio es un conjunto de términos o conceptos para los cuales se desea descubrir hipónimos. El dominio se refiere al área de aplicación donde podemos ubicar el vocabulario, por ejemplo un dominio médico, deportes, etc.

El vocabulario ejerce principalmente dos funciones. La primera consiste en orientar el método al dominio sobre el cuál se trabajará. De tal manera que, para cambiar el dominio es suficiente cambiar el vocabulario y continuar con el módulo de extracción de tuplas. De esta manera se evita la necesidad de extraer nuevamente patrones. Por otra parte, el vocabulario permite delimitar

patrones, es decir, los términos del vocabulario funcionarán como frontera (izquierda o derecha) en la definición de los patrones (ver figura 6, donde el término *felino* funciona como frontera derecha).

Una vez que se ha explicado la función del vocabulario, se explicará, en las siguientes secciones, cada uno de los módulos que hacen posible la extracción de tuplas.

### 5.2.2 Formación de patrones de consulta

El primer módulo para extraer las tuplas que conforman el catálogo de hipónimos consiste en formar patrones de consulta.

Para formar los patrones de consulta es necesario "aterrizar" cada uno de los patrones léxicos de extracción con los términos del vocabulario. Es decir, cada uno de los términos del vocabulario actuará como hiperónimo en la definición del patrón. Entonces, la etiqueta <*hipónimo*> actuará como un comodín extrayendo posibles hipónimos para los términos del vocabulario.

Para ilustrar este proceso, considere la palabra *felino* como término del vocabulario y el siguiente patrón: "*el <hipónimo> es el <hiperónimo> que".* Establecidas estas circunstancias, el patrón de consulta se obtiene reemplazando la etiqueta <*hiperónimo*> por el término *felino*. Como resultado se obtiene el siguiente patrón de consulta:

*"el <hipónimo> es el felino que*".

Para ejemplificar más, en la tabla 9 se muestran los patrones de consulta formados con el término *felino* y los ejemplos de los patrones mostrados en la tabla 8.

**Tabla 9. Ejemplo de patrones de consulta para los patrones de la tabla 8**

| No. | Patrones de consulta |
|---|---|
| 1 | el <hipónimo> es una ***felino*** |
| 2 | de ***felino*** como <hipónimo> y |
| 3 | la <hipónimo> es el ***felino*** de |
| 4 | el <hipónimo> es un ***felino*** que |
| 5 | el <hipónimo> es el único ***felino*** |

Ahora bien, los patrones resultantes del reemplazo de los términos del vocabulario en los patrones de extracción representan el conjunto de *patrones de consulta*. Reciben este nombre porque, como se verá en la siguiente sección, ejercerán el papel de consultas para recuperar, a través de la Web, hipónimos candidatos.

### 5.2.3 Aplicación de los patrones

Una vez aterrizados los patrones con los términos del vocabulario, se hace necesario el uso de una colección de documentos para extraer posibles hipónimos. Si bien se podrían utilizar colecciones de texto cerradas para extraer parejas *hipónimo-hiperónimo*, no es del todo conveniente. El inconveniente se propicia por la relación que deben mantener la colección y el vocabulario, es decir, ambos deben estar asociados. De lo contrario no sería posible extraer hipónimos para los términos del vocabulario.

Además, se debe tener presente que el vocabulario puede ser modificado requiriendo una colección de textos siempre relacionada a éste. Por esta

razón, en esta tesis se emplea la Web para recuperar las tuplas del catálogo. Dicho de otra manera la Web actúa como un corpus a medida para el vocabulario (pues contiene documentos casi para cualquier dominio).

Ya establecidas las razones que sustentan el uso de la Web, se procederá a explicar cómo se aplican los patrones. Primeramente, cada patrón de consulta es enviado a la Web a través de Google. Para cada patrón de consulta se analizan varios extractos de texto. Entonces, para cada patrón se obtiene un par de secuencias de palabras provenientes de los extractos de texto. La primera secuencia corresponde al concepto hipónimo. Por su parte, la segunda secuencia corresponde al concepto hiperónimo.

Por ejemplo, el patrón aterrizado *"el <hipónimo> es el felino que*" es enviado a la Web. Ahí puede emparejarse con un extracto de texto como el de la figura 8. Entonces, la pareja *lince ibérico-felino* es recuperada por el patrón.

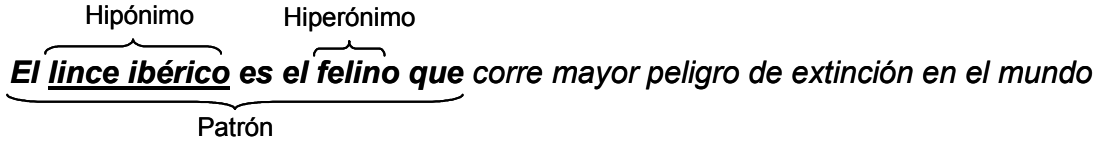


**Figura 8. Ejemplo de la aplicación de los patrones de consulta**

Después de aplicar cada uno los patrones a la Web se recuperan parejas como las siguientes:

Jaguar-felino,
Ocelote-felino,
Fuego- felino

donde los términos: *Jaguar, Ocelote y Fuego,* corresponden a los hipónimos y el término *felino* corresponde al concepto hiperónimo (previamente definido en el vocabulario). Como ya se había mencionado, por convención cada pareja *hipónimo-hiperónimo* es nombrada "*tupla*". El conjunto de tuplas conforma el *catálogo de hipónimos*. En la tabla 10 se muestra un ejemplo de un catálogo de hipónimos para un vocabulario con los términos: *felino* y *enfermedad*.

**Tabla 10. Ejemplo de un catálogo de hipónimos**

| No. | Hipónimo | Hiperónimo |
|---|---|---|
| 1 | *obesidad* | *enfermedad* |
| 2 | *cáncer* | *enfermedad* |
| 3 | *diabetes* | *enfermedad* |
| 4 | *tuberculosis* | *enfermedad* |
| 5 | *tiroides* | *enfermedad* |
| 6 | *osteoporosis* | *enfermedad* |
| 7 | *causa* | *enfermedad* |
| 8 | *neumonía* | *enfermedad* |
| 9 | *amor* | *enfermedad* |
| 10 | *mundo* | *enfermedad* |
| 11 | *origen canino* | *felino* |
| 12 | *ocelote* | *felino* |
| 13 | *margay* | *felino* |
| 14 | *puma* | *felino* |
| 15 | *pájaro* | *felino* |
| 16 | *lince* | *felino* |
| 17 | *fuego* | *felino* |
| 18 | *lince ibérico* | *felino* |
| 19 | *jaguar* | *felino* |
| 20 | *ocelote* | *felino* |

Ahora bien, es realmente importante señalar que el catálogo resultante está propenso a contener información incorrecta. Recordemos que se están

utilizando patrones léxicos, los cuales no consideran el uso de información sintáctica. Lo anterior favorece la extracción de tuplas incorrectas. Para ser más explícitos retomemos el ejemplo del catálogo de la tabla 10. Si bien en este catálogo se observan tuplas correctas como *osteoporosis-enfermedad, jaguar-felino, etc.* También se observan varias tuplas incorrectas como son: *tiroides-enfermedad, fuego-felino, mundo-enfermedad, etc.*

El problema anterior puede resolverse estimando la confianza (de ser correctas) de cada una de las tuplas, pero de ello se hablará en el capítulo 6.

Pues bien, la *aplicación de los patrones* es el último módulo que se realiza en la etapa 1 del método propuesto. Así que en la siguiente sección se muestran los resultados obtenidos en esta etapa.

## 5.3 Resultados experimentales

El método propuesto se probó de manera general realizando algunos experimentos. En esta sección se muestran únicamente los resultados obtenidos en la primera etapa (recuperación de tuplas). Los resultados de la segunda etapa (ordenamiento de tuplas) serán analizados en el capítulo 6. Así pues, en las siguientes secciones se muestran los resultados obtenidos de la etapa: *recuperación de tuplas*.

Para descubrir los patrones se consideraron 25 semillas. Por cada semilla, se recuperaron de la Web 500[16] segmentos de texto, para lo cual se utilizó la API del motor de búsqueda Google[17]. Por otra parte, el vocabulario empleado

[16] Por razones técnicas se recuperaron 500 segmentos de texto, pero se debe recordar que un mayor número de segmentos agregaría confiabilidad a la información extraída.
[17] Se utilizó la API de Google (http://www.google.com), por su gran cantidad de documentos indexados y la facilidad de uso de la misma.

para la extracción de tuplas se conformó de cinco términos[18]: *banco*, *enfermedad*, *felino*, *profesión* y *rocas*.

### 5.3.1 Resultados obtenidos

#### Sobre el descubrimiento de patrones

En la tabla 11 se resumen los resultados generados en el descubrimiento de patrones. Es posible notar que con 25 semillas se obtuvieron 43 patrones léxicos de extracción. Quizá pueden parecer pocos los patrones descubiertos en relación al número de semillas utilizado. Sin embargo, es necesario recordar que estos patrones no contienen signos de puntuación. Por consiguiente, el número de patrones adecuados para el método se ve reducido.

**Tabla 11. Resultados sobre el descubrimiento de patrones**

| Semillas | Ejemplos de uso | SFM's | Patrones encontrados |
|---|---|---|---|
| 25 | 500 por cada semilla | 4623 | 43 |

La tabla 12 muestra algunos ejemplos de los 43 patrones léxicos de extracción de hipónimos. La lista de los 43 patrones se muestra en el *apéndice A*. En general, dentro de la diversidad de patrones, se tienen desde patrones muy amplios hasta patrones muy específicos. Por ejemplo, el patrón "*<hiperónimo> de <hipónimo> y"* es muy amplio porque puede extraer muchas tuplas (ya que puede aplicarse un gran número de veces); sin

[18] Los términos elegidos son diferentes a las semillas. Se utilizaron términos muy distantes entre sí (no pertenecen al mismo dominio) con objeto de probar la generalidad de los patrones obtenidos.

embargo, su precisión es baja porque la mayoría de las tuplas extraídas son incorrectas. En cambio, el patrón "*la <hipónimo> es el único <hiperónimo> natural"* es muy específico porque podría extraer pocas tuplas, pero en su mayoría correctas, es decir, tiene alta precisión pero poco recuerdo.

**Tabla 12. Ejemplos de los patrones léxicos de extracción de hipónimos obtenidos**

| Ejemplos de patrones léxicos de extracción de hipónimos |
|---|
| el <hipónimo> es el único <hiperónimo> |
| el <hipónimo> es un <hiperónimo> que |
| los <hipónimo> y otros <hiperónimo> |
| el <hipónimo> es uno de los <hiperónimo> más |
| que la <hipónimo> es una <hiperónimo> |
| la <hipónimo> es una <hiperónimo> que |
| el <hipónimo> es un <hiperónimo> de |
| el <hipónimo> es el <hiperónimo> que |
| las <hipónimo> son una <hiperónimo> |
| la <hipónimo> es el <hiperónimo> de |
| <hiperónimo> de <hipónimo> y |
| la <hipónimo> es el único <hiperónimo> natural |
| el uso de la <hipónimo> como <hiperónimo> |
| el <hipónimo> es una <hiperónimo> |
| la <hipónimo> como una <hiperónimo> |

Ahora bien, se pueden observar varios patrones que presentan una forma similar. Es decir, patrones donde únicamente cambian los artículos que unen las etiquetas: *<hiperónimo>* y *<hipónimo>.* Por ejemplo, los siguientes patrones:

*el <hipónimo> es un <hiperónimo> que*
*la <hipónimo> es una <hiperónimo> que*

*el <hipónimo> es el <hiperónimo> que*

Lo anterior surge automáticamente como una solución para abarcar todas las variantes morfológicas de un patrón sintáctico, pues a nivel léxico no se tiene la posibilidad de generalizar todos estos artículos en una etiqueta única. Por ejemplo, a nivel léxico-sintáctico, los patrones anteriores podrían ser generalizados en un solo patrón:

*[ART] <hipónimo> es [ART] <hiperónimo> que*

Sin embargo, como ya se ha mencionado, el incluir información sintáctica origina la dependencia de herramientas sintácticas.

**Sobre la extracción de tuplas**

Una vez que se han formado los patrones de consulta, éstos son aplicados en la Web. Los resultados de la aplicación de los patrones se muestran en la tabla 13. De acuerdo con la misma tabla, con poca información (aproximadamente 8.6 Mb.) se obtuvieron 13,626 tuplas. Como es de imaginar mientras más datos se tengan, mejores serán los resultados alcanzados. Sin embargo, aún con poca información se obtuvieron resultados muy alentadores (ver capítulo 6).

**Tabla 13. Resultados de la extracción de tuplas**

| Ejemplos para cada patrón | Tamaño (Mb) de todos los ejemplos | Tuplas en el catálogo |
|---|---|---|
| 700 | 8.6 | 13,626 |

Por otro lado, las 13,626 tuplas que conforman el catálogo de hipónimos están organizadas de la forma mostrada en la tabla 14.

**Tabla 14. Organización de tuplas del catálogo construido**

| Hiperónimo | Tuplas |
|---|---|
| banco | 1,993 |
| enfermedad | 3,459 |
| felino | 307 |
| profesión | 4,364 |
| rocas | 3,503 |

Ahora bien, en la tabla 15 se muestra un fragmento del catálogo obtenido. Puede observarse que existe información incorrecta. Por ejemplo, las tuplas: *Cerebros-Banco*, *Caja-Banco*, *Fuego-Felino*, *Fé-Profesión*, *Origen Plutónico y-Rocas* y *Agua-Rocas*. Si bien estas tuplas son en su mayoría palabras relacionadas, no expresan una relación de hiponimia. Por ello, surge la necesidad de estimar la confiabilidad de las tuplas. Dicho de otro modo, se quiere estimar qué tan confiable es una tupla para establecer una relación de hiponimia. Este problema será tratado en el capítulo siguiente (capítulo 6).

## 5.4 Discusión

Cuando se utilizan patrones para extraer hipónimos siempre existe una porción de información que no es confiable (Fleischman et al., 2003). Por ello, generalmente los métodos que hacen uso de patrones incluyen una etapa de filtrado (o restricción de características) cuyo objetivo es encontrar patrones más apropiados para la tarea de extracción de hipónimos.

**Tabla 15. Fragmento del catálogo obtenido**

| Hipónimo | Hiperónimo |
|---|---|
| HSBC | Banco |
| Cerebros | Banco |
| Caja | Banco |
| Standard Bank | Banco |
| Tuberculosis | Enfermedad |
| Cáncer | Enfermedad |
| Amor | Enfermedad |
| Jaguar | Felino |
| Fuego | Felino |
| Fé | Profesión |
| Contaduría | Profesión |
| Biblioteconomía | Profesión |
| Origen plutónico y | Rocas |
| Rocas sedimentarias | Rocas |
| Agua | Rocas |

Por ejemplo, Pasca (2004) aplica patrones para extraer un conjunto de hipónimos y el nombre de sus categorías (en esta tesis las categorías corresponden a los hiperónimos) a partir de un corpus. Aunque él utiliza patrones que usan etiquetas POS, determina algunas condiciones que una frase debe cumplir para ser considerada el nombre de la categoría. Por ejemplo, restringe el nombre de una categoría a una palabra en plural. Además, el nombre de la categoría puede contener adjetivos no informativos como: otros, varios, muchos, etc. Entonces, él sólo toma la parte informativa. Lo anterior es un ejemplo de una etapa de filtrado dentro de la extracción de hipónimos. Sin embargo, aún con este filtrado, existe la posibilidad de extraer información incorrecta.

En (Pantel et al., 2004) para extraer hipónimos se utilizan patrones que incluyen etiquetas POS. En ese trabajo, para enfrentar el problema de encontrar los límites de los conceptos que actuarán como hipónimos e hiperónimos, se incluyen etiquetas de diferente anchura[19]. En particular en esta tesis, para delimitar los conceptos que actuarán como hiperónimos e hipónimos se seleccionan únicamente expresiones que contienen fronteras (ver sección 5.1.5).

Generalmente, después de la etapa de filtrado, aún se tiene la posibilidad de extraer información incorrecta. De ahí que existan métodos que busquen evaluar la calidad de los patrones o de las mismas tuplas extraídas (Pantel et al., 2006; Blohm y Cimiano, 2006; Ortega-Mendoza et al., 2007), de esta manera determinan con mayor certeza el conjunto de parejas extraídas más confiables. En esta tesis, también se determina la calidad de los patrones y de las tuplas (ver capítulo 6).

Por ejemplo, en (Fleischman et al., 2003) después de aplicar los patrones se usan técnicas de aprendizaje automático para filtrar las instancias extraídas a través de los patrones. En ese trabajo se explica que el "ruido" (instancias incorrectas) se debe a los patrones amplios y también a los errores que presentan los etiquetadores POS. Además, aún cuando en ese trabajo se utiliza información léxico-sintáctica se puede extraer información incompleta (nombres de hiperónimos incompletos). De ahí, que en ese trabajo se filtren las instancias extraídas.

En (Pantel et al., 2006) se determina la calidad de los patrones y de las tuplas en un proceso recursivo. El proceso de evaluación puede verse como

[19] Combinaciones de varias etiquetas POS, por ejemplo: JJ NN, JJ NN NN, JJ NN NN NN, etc. Donde JJ= Adjetivo; NN= Nombre común

un proceso de filtrado, el cual determina cuáles son las tuplas más confiables dentro del conjunto de tuplas extraídas.

Así pues, el uso de patrones léxico-sintácticos no garantiza la confiabilidad de la información extraída. Tampoco el uso de información léxica tiene esta garantía. Por ello, se han desarrollado técnicas de filtrado y evaluación (como las descritas previamente) para tratar de generar patrones más confiables o para determinar un conjunto de instancias confiables dentro del conjunto de instancias extraídas. En el siguiente capítulo se mostrarán esquemas de ordenamiento para determinar la confiabilidad de las instancias.

# Capítulo 6

# Ordenamiento de tuplas

Este capítulo se enfoca en estimar la confianza de las tuplas del catálogo de hipónimos. El objetivo es ubicar a las tuplas con mayor probabilidad de ser correctas en las primeras posiciones del catálogo. En este capítulo se encuentran las contribuciones más importantes de este trabajo de investigación. Básicamente se proponen dos enfoques para estimar la confianza de los patrones y de las tuplas. El primero se apoya en el uso de información mutua. Por su parte, el segundo enfoque explora el uso de la medida F para determinar la confianza de las tuplas y los patrones.

## 6.1 Introducción a la fase de ordenamiento de tuplas

Como se mencionó en el capítulo anterior, el catálogo creado está propenso a contener varias tuplas incorrectas. El problema anterior podría resolverse utilizando un experto que separe manualmente las tuplas correctas de aquellas incorrectas. Sin embargo, sería una tarea costosa en tiempo y esfuerzo, debido a la gran cantidad de tuplas resultantes en el catálogo. Inclusive, la diversidad de dominios que pueden ser tratados dificultaría el proceso manual de evaluación de tuplas. De ahí la necesidad de crear métodos automáticos que permitan resolver el problema estimando la confianza de las tuplas resultantes.

En esta tesis se proponen dos enfoques automáticos para resolver el problema anterior. Estos enfoques parten de dos observaciones. En primer lugar se observó que una tupla puede ser extraída por uno o varios patrones. En segundo lugar se observó que cada patrón puede extraer una o varias tuplas. Teniendo estas observaciones en mente, se formularon los siguientes criterios generales:

**Criterios Generales**

1. Una tupla es más confiable mientras mayor sea la cantidad de patrones que la extraen.

2. Un patrón es más confiable mientras mayor sea la cantidad de tuplas confiables que extraiga.

Los criterios anteriores constituyen un punto de partida para encontrar un método de estimación de la confianza de las tuplas y de los patrones. Bajo estos criterios, se sugiere un esquema donde la confianza de las tuplas

ayude a determinar la confianza de los patrones[20], y así mismo, que la confianza de los patrones ayude a estimar la confianza de las tuplas[21].

Para empezar la etapa de ordenamiento, las tuplas del catálogo son pasadas a través de un filtro. En la siguiente sección se exponen los detalles de dicho filtro.

## 6.2 Filtro inicial del catálogo

Retomando el primer criterio general mencionado anteriormente, se puede intuir que aquellas tuplas extraídas por únicamente un patrón tienen probabilidades muy pequeñas de representar una tupla correcta. Si esas tuplas permanecieran en el catálogo, probablemente al final, el método las ubicaría en las últimas posiciones del catálogo. Sin embargo, también existe la posibilidad de que el método se vea afectado por estas tuplas (que desde un inicio se sabe que tienen pocas probabilidades de ser correctas), ya que existen varias tuplas que son extraídas por un único patrón. De ahí que se prefiera aplicar un filtro inicial al catálogo.

Básicamente, el filtro consiste en eliminar todas las tuplas obtenidas por un un patrón. Como resultado, el catálogo estará conformado únicamente por aquellas tuplas extraídas por dos o más patrones. La sección siguiente muestra los resultados obtenidos después de aplicar el filtro al catálogo.

[20] Básicamente, en esta tesis se utilizan dos métricas para obtener la confianzas de un patrón, *información mutua* y una adaptación a la *medida F*.
[21] En esta tesis, la confianza de una tupla se obtiene usando dos enfoques: basado en *información mutua* y basado en *la medida F*.

### 6.2.1 Resultados del filtro inicial

El filtro fue aplicado al catálogo obtenido en los experimentos del capítulo anterior. Después de eliminar todas las tuplas que son extraídas únicamente por un patrón, el catálogo se conformó de 851 tuplas. Dichas tuplas están organizadas como se muestra en la tabla 16:

**Tabla 16. Organización de las tuplas del catálogo después de aplicar el filtro inicial**

| Hiperónimo | Tuplas |
|---|---|
| *banco* | 193 |
| *enfermedad* | 307 |
| *felino* | 9 |
| *profesión* | 226 |
| *rocas* | 116 |
| Total | 851 |

Una vez filtrado el catálogo, las tuplas de éste serán ordenadas de acuerdo con los dos enfoques que se proponen en este capítulo y cuya arquitectura se presenta a continuación.

## 6.3 Estructura general de los enfoques propuestos

Como ya se ha mencionado, en este capítulo se proponen dos enfoques para ordenar las tuplas del catálogo:

1. *Enfoque centrado en el uso de información mutua*. Este enfoque se caracteriza por el uso de *información mutua.* Esta medida ya ha sido usada en varios trabajos (Pantel y Ravichandran, 2004; Blohm y

Cimiano, 2006; Pantel y Pennacchiotti, 2006). Sin embargo, en esos trabajos se estima la confianza de patrones léxico-sintácticos. En cambio, en este trabajo de investigación, esta medida estima la confianza de patrones léxicos.

2. *Enfoque centrado en el uso de la medida F*. En este enfoque se propone un esquema que utiliza una adaptación[22] de la medida F para estimar la confianza de las tuplas y de los patrones.

Ambos enfoques siguen la misma estructura y únicamente difieren en las métricas utilizadas para estimar la confianza de los patrones y de las tuplas. En concreto, la estructura que siguen estos enfoques se ilustra en la figura 9.

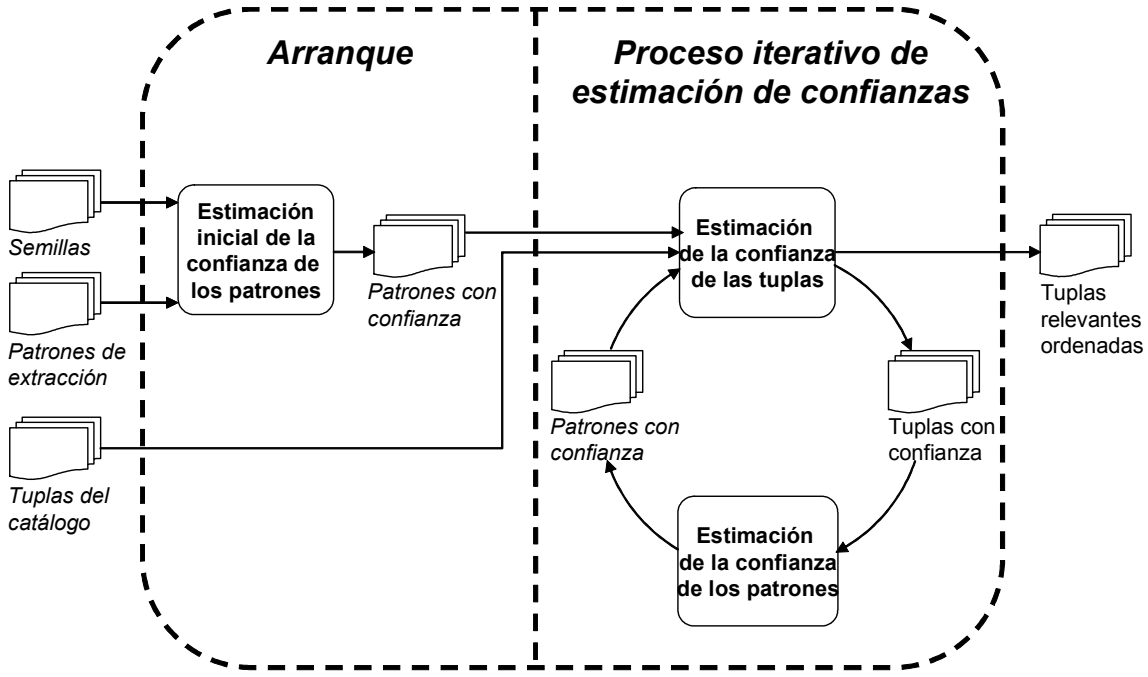


**Figura 9. Estructura general de los enfoques propuestos para ordenar las tuplas**

[22] En este trabajo de tesis, se propone una métrica basada en la medida F, la diferencia con esta medida es la forma de estimar la precisión y el recuerdo (ver sección 6.4.1 y 6.5).

De la figura 9, se observa que la estructura general de los enfoques propuestos consta de dos módulos. El primer módulo llamado *"Arranque"* permite estimar la confianza inicial a los patrones de extracción. En principio, la confianza de los patrones es obtenida usando información derivada del conjunto de semillas, pues se debe recordar que los patrones fueron descubiertos a partir de las semillas. Por ello, se considera que las semillas pueden aportar información valiosa para estimar la confiabilidad de los patrones.

El segundo módulo *"Proceso iterativo de estimación de confianzas"* busca estimar la confianza de las tuplas del catálogo iterando entre la estimación de la confianza de las tuplas y la estimación de la confianza de los patrones. La idea detrás de este proceso iterativo es el enriquecimiento de información o evidencia en cada iteración con respecto a la iteración anterior. En concreto, se espera que la precisión del catálogo incremente en cada iteración. Ahora bien, en este punto la información de las semillas deja de ser relevante, pues ya ha sido tomada en cuenta para asignar una confianza inicial a los patrones de extracción. Finalmente, el proceso iterativo termina cuando se cumple una cierta condición de paro (ver sección 6.6).

Así pues, los dos enfoques que se presentarán a continuación (sección 6.4 y 6.5) son análogos y únicamente difieren en la forma de estimar la confianza de las tuplas y de los patrones.

## 6.4 Enfoque centrado en el uso de información mutua

Este primer enfoque se caracteriza por el uso de una medida tradicional para estimar la confianza de patrones y tuplas. Esta medida es llamada *información mutua*.

Es importante mencionar que Pantel y Pennacchiotti (2006) exploran ideas similares a las presentadas en este *enfoque centrado en el uso de información mutua*. Sin embargo, el esquema que ellos proponen difiere del que se propone en esta tesis en varios aspectos como los siguientes:

- En primer lugar, el presente trabajo se apoya únicamente en el uso de patrones léxicos. En cambio, en el trabajo que se realiza en (Pantel y Pennacchiotti, 2006) se utilizan patrones léxico-sintácticos. Ahora bien, prescindir de información sintáctica genera un incremento sustancial del número de patrones descubiertos y aplicados.

- En segundo lugar, en (Pantel y Pennacchiotti, 2006) se utiliza la información extraída para determinar el conjunto de patrones más apropiado con el cual, finalmente, se construirá el catálogo definitivo. En otras palabras, en ese trabajo se establece un conjunto inicial de patrones y posteriormente, se agregan nuevos patrones (los más confiables) en cada iteración. En nuestro caso, se realizan iteraciones únicamente para determinar la confianza de los patrones y de las tuplas. Es decir, el conjunto de patrones se establece desde el primer paso y se conserva a lo largo del *proceso iterativo de estimación de confianzas*. Durante cada iteración se determinan nuevas confianzas, tanto para patrones como para las tuplas del catálogo, hasta alcanzar un ordenamiento descendente (de acuerdo con las confianzas).

- Mientras en (Pantel y Pennacchiotti, 2006) se seleccionan y mantienen únicamente las tuplas con mejor confianza en cada iteración, en esta tesis, el conjunto de tuplas en cada iteración permanece íntegro desde el inicio. Únicamente al finalizar *el proceso iterativo de estimación de confianzas* es cuando se podría seleccionar y conservar sólo las tuplas más confiables (con mayor confianza) del catálogo (ver sección 6.6).

- En este trabajo se expone el uso de dos medidas para obtener la confianza de las tuplas y de los patrones (información mutua y una adaptación de la medida F). Mientras en el trabajo de Pantel y Pennacchiotti sólo se usa información mutua.

Finalmente, cada uno de los módulos del enfoque centrado en el uso de información mutua (*arranque y proceso iterativo de estimación de confianzas*) serán explicados detalladamente en las siguientes secciones.

### 6.4.1 Arranque

Este módulo se encarga de asignar una confianza inicial a los patrones. Aquí, se asume que la confianza inicial de los patrones tiene un alto grado de relevancia para determinar la precisión del catálogo. En particular, en esta tesis se experimentó con dos alternativas para asignar inicialmente la confianza de cada uno de los patrones. La primera alternativa se basa en el uso de *información mutua*. La segunda alternativa hace referencia al uso de la *medida F*. En las siguientes secciones se explica a detalle cómo se desarrollan las dos alternativas mencionadas.

**Alternativa 1: Usando información mutua**

Esta alternativa asigna confianzas a los patrones de acuerdo al grado de asociación que presentan los mismos con el conjunto de semillas. Entonces, un patrón fuertemente asociado a las semillas tiene alta probabilidad de ser un patrón confiable (con una alta confianza). Comúnmente se utiliza información mutua para calcular el grado de asociación entre patrones y tuplas. Por ello, en esta primera alternativa se experimenta con esta medida.

Antes de explicar cómo se aplicó esta medida es necesario explicar qué es *información mutua.*

De acuerdo a Cover y Thomas (1991) la información mutua *pmi* es una métrica que mide la cantidad de información que una variable aleatoria contiene acerca de otra variable aleatoria. Información mutua también es considerada como la reducción de la incertidumbre de una variable aleatoria debido al conocimiento de otra. Más concretamente, la información mutua es comúnmente usada para medir la fuerza de asociación entre dos eventos $u$ y $v$, como se muestra en la ecuación 4.

$$pmi(u,v) = \log \frac{P(u,v)}{P(u)\ P(v)} \tag{4}$$

Es importante señalar que esta medida ya ha sido explorada en la extracción automática de relaciones semánticas. Especialmente es aplicada para medir la fuerza de asociación entre patrones y tuplas (Pantel y Ravichandran, 2004; Blohm y Cimiano, 2006; Pantel y Pennacchiotti, 2006). De hecho, esta medida ha sufrido adaptaciones de acuerdo a las necesidades de los investigadores. Por supuesto, siempre se busca encontrar una adaptación que mejore los resultados. En particular, en (Blohm y Cimiano, 2006) se distinguen las siguientes tres ecuaciones (5-7) para calcular información mutua entre un patrón $p$ y una tupla $i = (x, y)$:

$$pmi_1(p,t) = \frac{|x,p,y|}{|x,*,y|} \tag{5}$$

$$pmi_2(p,t) = \log \frac{|x,p,y|}{|*,p,*||x,*,y|} \tag{6}$$

$$pmi_3(p,t) = \log \frac{|x,p,y||*,p,*|}{|x,p,*||*,p,y|} \quad (7)$$

En las ecuaciones 5-7, $|x,p,y|$ es la frecuencia con la cual un patrón $p$ extrae a la tupla $t$ formada por el hipónimo $x$ y el hiperónimo $y$. Por su parte el símbolo "$*$" actúa como un comodín. Es decir, $|x,*,y|$ es la frecuencia de aparición de la tupla extraída por cualquier patrón y formada por el hipónimo $x$ y el hiperónimo $y$. Entonces, $|*,p,*|$ corresponde al número total de tuplas que extrajo el patrón $p$. Análogamente, $|x,p,*|$ la frecuencia de aparición del patrón $p$ extrayendo las tuplas formada por el hipónimo $x$ y cualquier hiperónimo. Similarmente, $|*,p,y|$ indica la frecuencia de aparición del patrón $p$ extrayendo tuplas que contengan cualquier hipónimo y el hiperónimo $y$.

Particularmente, en esta tesis, se trabajó con estas tres formas de calcular la información mutua entre patrones y tuplas.

Por otra parte, de acuerdo con los criterios generales, los patrones con mayor confianza son aquellos que permiten extraer un mayor número de tuplas confiables (con mayor confianza). Entonces, capturando estos criterios la confianza de un patrón $p$ denotada como $c_\pi(p)$ se estima mediante la ecuación 8.

$$c_\pi(p) = \frac{\sum_{t \in T'} \left( \frac{pmi(p,t)}{\max_{pmi}} \times c_\sigma(t) \right)}{|T|} \quad (8)$$

Donde:

- $T$ representa el conjunto de tuplas (en el *arranque* las tuplas bajo consideración son las semillas).

- $T'$ representa el conjunto de tuplas extraídas por un patrón $p$.
- $\max_{pmi}$ es el valor máximo de información mutua de todos los patrones con todas las tuplas.
- La función $pmi$, hace referencia a cualquiera de las funciones: $pmi_1$, $pmi_2$ ó $pmi_3$.
- $c_\sigma(t)$ se refiere a la confianza de una tupla $t$

Por su parte, el valor de la función $c_\sigma(t)$ dentro del módulo de *arranque* es igual a 1, es decir, $c_\sigma(t)=1$ para cualquier tupla $t$ perteneciente al conjunto de semillas. Lo anterior, porque las semillas han sido consideradas tuplas altamente confiables debido a su elección manual.

**Alternativa 2: Usando la medida F**

La alternativa basada en el uso de información mutua es un tipo de estimación probabilística. Dado que tenemos un corpus pequeño (formado por los ejemplos recopilados de la Web), la alternativa basada en el uso de información mutua tiende a favorecer a patrones con alta precisión o con alto recuerdo, pero no necesariamente da preferencia a esos patrones que muestran un balance entre ambas medidas. Este inconveniente motivó el desarrollo de una segunda alternativa, la cual, trata de rectificar este problema. Esta segunda alternativa se basa en el uso de la medida F para estimar inicialmente la confianza de los patrones. A continuación se describe más detalladamente el uso de esta medida.

En el capítulo 2 se presentó la definición formal de precisión ($P$), recuerdo ($R$) y la medida F ($F_\beta$), para la evaluación general del desempeño de un sistema de extracción de relaciones semánticas. Sin embargo, en este

momento no se evaluará el sistema de forma general. Más bien, con base en estas mismas funciones (ver ecuaciones 1, 2 y 3) se tratará de estimar la precisión, el recuerdo y la medida F para cada patrón $p$, es decir, $P_\pi(p)$, $R_\pi(p)$ y $F_\pi(p)$ respectivamente.

Para definir la forma en que se estimará el recuerdo y la precisión de un patrón, primero se debe considerar que los patrones pueden extraer una gran cantidad de tuplas. Dentro del conjunto de éstas se pueden hallar algunas tuplas que pertenezcan al conjunto de las semillas, tal como se muestra en la figura 10.

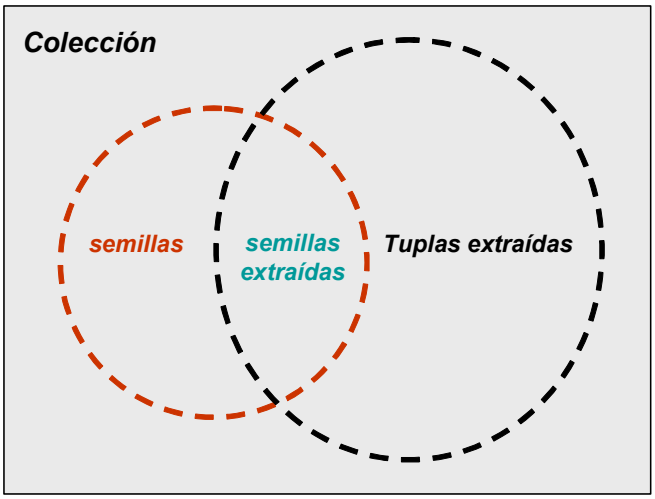


**Figura 10. Precisión y recuerdo de un patrón**

Así pues, la precisión de un patrón *p,* es decir $P_\pi(p)$, es estimada como el cociente del número de semillas entre el número total de tuplas extraídas por el patrón *p* (ver ecuación 9).

$$P_\pi(p) = \frac{|semillas\ extraídas\ por\ p|}{|tuplas\ \ extraídas\ por\ p|} \qquad (9)$$

Por su parte, el recuerdo de un patrón *p*, es decir $R_\pi(p)$, es estimado como la porción de semillas extraídas por el patrón p (ver ecuación 10).

$$R_\pi(p) = \frac{|semillas\ extraídas\ por\ p|}{|semillas|} \quad (10)$$

Ante todo, tanto la precisión como el recuerdo son importantes para determinar la confianza de un patrón. Idealmente, un patrón sería perfecto si tuviera alto recuerdo y al mismo tiempo, alta precisión. Sin embargo, generalmente estas dos medidas están inversamente relacionadas. Por ejemplo, un patrón con alto recuerdo como " *<hiperónimo> de <hipónimo> y* " generalmente recupera muchas tuplas, pero en su mayoría incorrectas. Es decir, tiene alto recuerdo pero baja precisión. Por otro lado, un patrón que es muy específico como: *" la <hipónimo> es el único <hiperónimo> natural* " recupera muy pocas tuplas, pero generalmente recupera tuplas correctas.

Entonces, ambas medidas son importantes para determinar la confianza de un sistema. Pero se necesita una sola medida que ayude a comparar la confianza de los patrones. Dado que la medida F ofrece la posibilidad de combinar en un valor escalar la precisión y el recuerdo, ha sido elegida para estimar la confianza inicial de los patrones. Básicamente, la confianza de un patrón $p$, la cual ha sido denotada como $c_\pi(p)$, se obtiene mediante la ecuación 11, la cual representa una adaptación[23] a la tradicional medida F.

$$c_\pi(p) = \frac{F_\pi(p)}{\max_{\forall q \in P}\{F_\pi(q)\}}, \text{ donde} \quad (11)$$

[23] Se le asignó el nombre de *adaptación* porque la precisión y el recuerdo son determinados de manera diferente a la forma tradicional.

$$F_{\pi}(p)=\frac{2\cdot P_{\pi}(p)\cdot R_{\pi}(p)}{P_{\pi}(p)+R_{\pi}(p)} \quad (12)$$

Donde $\max_{\forall l\in P}\{F_{\pi}(l)\}$ es un factor de normalización.

Una vez que los patrones han recibido una confianza inicial, su confianza es aprovechada para asignar un valor de confianza a cada tupla. Esta asignación se realiza dentro de un proceso iterativo, el cual es descrito a continuación.

### 6.4.2 Proceso iterativo de estimación de confianzas

El proceso iterativo de estimación de confianzas busca estimar la confianza de los patrones con base en la confianza de las tuplas y viceversa. Aquí se distinguen dos pasos similares y dependientes: La estimación de confianza de las tuplas y la estimación de confianza de los patrones. Para mayor claridad, estos dos pasos se explicarán por separado en las siguientes secciones.

#### 6.4.2.1 Confianza de tuplas

Para estimar la confianza de las tuplas se toma en cuenta la confianza de los patrones obtenida previamente. Para estimar la confianza de las tuplas también considera el uso de información mutua. Más formalmente, la confianza de una tupla $t$, denotada como $c_{\sigma}(t)$ se obtiene a través de la ecuación (13).

$$c_{\sigma}(t)=\frac{\sum_{p\in P'}\left(\frac{pmi(p,t)}{\max_{pmi}}\cdot c_{\pi}(p)\right)}{|P|} \quad (13)$$

Donde $c_{\pi}(p)$ hace referencia a la confianza del patrón $p$ obtenida en el paso inmediato anterior. Al mismo tiempo recuerde la siguiente notación:

- $T$ representa el conjunto de tuplas (en el *proceso iterativo de estimación de confianzas* las tuplas bajo consideración son las tuplas del catálogo).
- $P$ representa el conjunto total de patrones.
- $P'$ representa el conjunto de patrones que permiten extraer la tupla $t$.
- $\max_{pmi}$ es el máximo punto de información mutua entre todos los patrones y todas las tuplas.
- La función $pmi$ hace referencia a cualquiera de las funciones: $pmi_1$, $pmi_2$ ó $pmi_3$.

### 6.4.2.2 Confianza de patrones

Similarmente al arranque, la confianza de cada uno de los patrones se obtiene utilizando información mutua. Entonces, la confianza de un patrón *p* ($c_{\pi}(p)$) se determina a través de la ecuación 8 (página 74), la cual se muestra nuevamente para mayor claridad:

$$c_{\pi}(p)=\frac{\sum_{t\in T'}\left(\frac{pmi(p,t)}{\max_{pmi}}\cdot c_{\sigma}(t)\right)}{|T|}$$

Pero se debe recordar que la confianza de las tuplas ($c_{\sigma}(t)$) corresponde a aquella obtenida en el paso inmediato anterior de las tuplas del catálogo (no de las semillas).

### 6.4.3 Resumen del enfoque centrado en el uso de información mutua

En resumen, en el enfoque centrado en el uso de información mutua se presentaron dos alternativas para realizar el *arranque*. La primera centra su atención en el uso de información mutua. La segunda se basa en el uso de una adaptación de la medida F. Por su parte, el *proceso iterativo de estimación de confianzas* está basado únicamente en el uso de información mutua. Por último, para efectos de claridad, en la tabla 17 se muestra un formulario que resume las ecuaciones involucradas en este enfoque.

**Tabla 17. Formulario general del enfoque centrado en el uso de información mutua**

| **Arranque** | |
|---|---|
| *Alternativa 1* | *Alternativa 2* |
| $c_{\pi}(p)=\frac{\sum_{t\in T'}\left(\frac{pmi(p,t)}{\max_{pmi}}\cdot c_{\sigma}(t)\right)}{\lvert T\rvert}$ | $c_{\pi}(p)=\frac{F_{\pi}(p)}{\max_{\forall l\in P}\{F_{\pi}(l)\}}$ |
| **Proceso iterativo de estimación de confianzas** | |
| *Confianza de tuplas* | *Confianza de patrones* |
| $c_{\sigma}(t)=\frac{\sum_{p\in P'}\left(\frac{pmi(p,t)}{\max_{pmi}}\cdot c_{\pi}(p)\right)}{\lvert P\rvert}$ | $c_{\pi}(p)=\frac{\sum_{t\in T'}\left(\frac{pmi(p,t)}{\max_{pmi}}\cdot c_{\sigma}(t)\right)}{\lvert T\rvert}$ |

### 6.4.4 Resultados

En esta sección se muestran los resultados obtenidos cuando se utiliza el *enfoque centrado en el uso de información mutua* para ordenar el catálogo de hipónimos. Para valorar el desempeño de este enfoque de ordenamiento, se adaptó un esquema de evaluación utilizado en (Pasca, 2004)[24]. Específicamente, en el presente trabajo de tesis se evaluó manualmente la precisión de las 200 primeras tuplas del catálogo (las que tienen mayor confianza).

Por otro lado, para realizar los experimentos se utilizó el catálogo resultante después de aplicar el filtro inicial (ver sección 6.2). Particularmente, en el *proceso iterativo de estimación de confianzas* se trabajó con tres iteraciones, únicamente para comparar los dos enfoques de ordenamiento. Sin embargo, en la sección 6.6 se menciona una descripción más detallada del *criterio de paro* para el proceso iterativo.

Se realizaron dos grupos de experimentos. Estos dos grupos de experimentos difieren únicamente en la alternativa utilizada para realizar el *arranque*. Entonces, el primer grupo de experimentos (sección 6.4.4.1) muestra los resultados obtenidos empleando información mutua en el arranque. En cambio, el segundo grupo (sección 6.4.4.2) muestra los resultados conseguidos empleando la adaptación de la medida F. Por último, en la sección 6.4.4.3 se realiza una breve comparación entre ambas alternativas.

[24] En ese trabajo se evalúan las instancias extraídas más confiables de cada categoría (hiperónimo).

#### 6.4.4.1 Resultados usando información mutua en el arranque

Aquí se muestran los resultados obtenidos al usar información mutua para asignar inicialmente a los patrones valores de confianza. En esta sección se describen los resultados alcanzados utilizando las tres formas que fueron mencionadas para obtener la información mutua entre patrones y tuplas: $pmi_1$, $pmi_2$ y $pmi_3$.

**Primera iteración**

La figura 11 expresa las precisiones obtenidas en la primera iteración. De esa figura, se puede observar que las curvas de precisión para $pmi_2$ y $pmi_3$ tienden a caer conforme se toma un mayor número de tuplas. De modo que para las 200 primeras tuplas del catálogo se obtienen precisiones de 44.50% y 27.50%, respectivamente. Aún cuando ambas curvas presentan una caída similar, claramente se nota la superioridad en precisión de $pmi_2$, contra $pmi_1$ y $pmi_3$.

Ahora bien, la curva correspondiente a $pmi_1$ no presenta una caída. Sin embargo, las precisiones que reporta se ubican por debajo de las reportadas por $pmi_2$ y $pmi_3$.

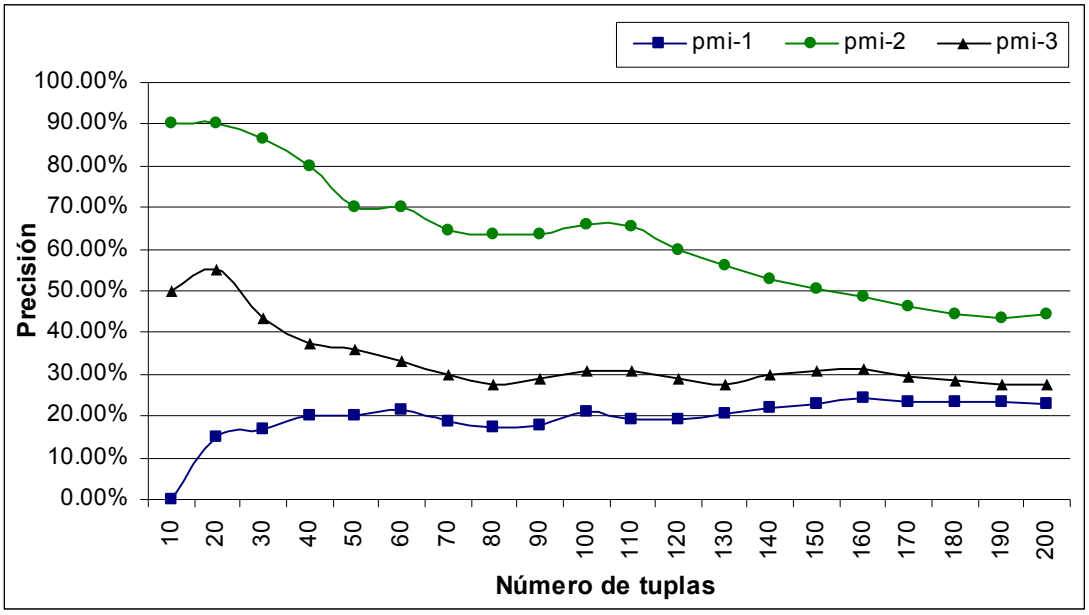


**Figura 11. Resultados en la 1ra. Iteración con información mutua en el arranque**

### Segunda iteración

Después de realizar la segunda iteración se obtuvieron los resultados reportados en la figura 12. Al igual que en la primera iteración, se observa la superioridad en precisión de la curva correspondiente a $pmi_2$ .

Ahora bien, la curva correspondiente a $pmi_2$ presenta un aumento en la precisión comparándola con su curva en la primera iteración. Por ello, se cree que en cada iteración se obtiene una mejor estimación de la confiabilidad de las tuplas del catálogo. De ahí que exista un aumento en la precisión final del catálogo. Las curvas pertenecientes a $pmi_1$ y $pmi_3$ no presentan un cambio sobresaliente con respecto a la primera iteración.

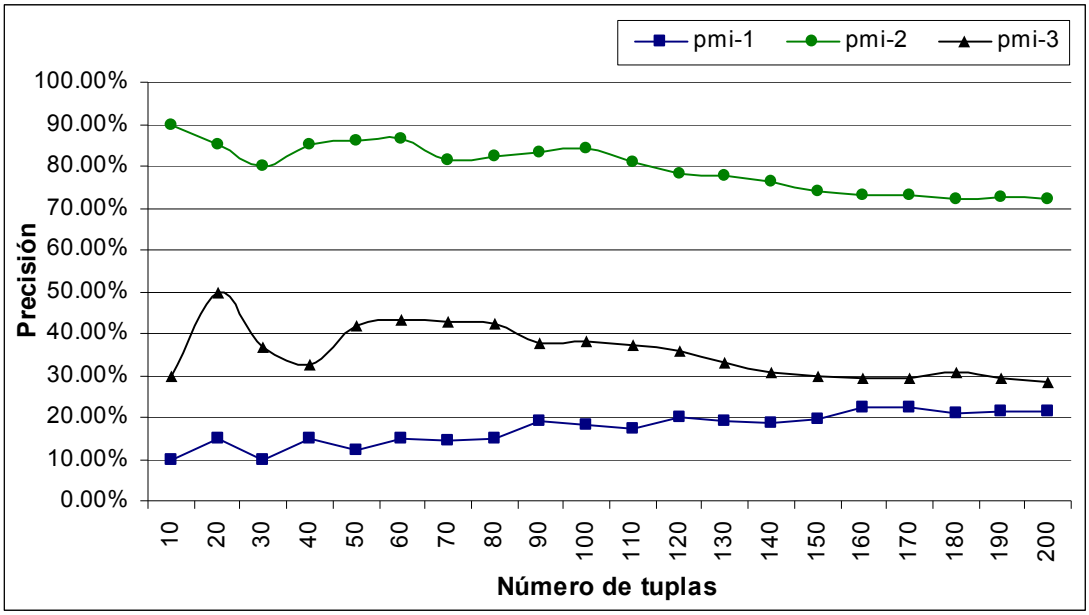


**Figura 12. Resultados en la 2da. Iteración con información mutua en el arranque**

### Tercera Iteración

Los resultados de la tercera iteración se muestran en la figura 13. En esta iteración sobresalen los siguientes aspectos:

- Las precisiones obtenidas empleando $pmi_2$ son, por mucho, superiores a las precisiones obtenidas usando $pmi_1$ y $pmi_3$.

- Por su parte, las curvas correspondientes a $pmi_1$ y $pmi_3$ presentan un ligero incremento en precisión con respecto a la iteración anterior. Pero aun con este incremento, se mantienen por debajo de $pmi_2$.

- Se ha alcanzado una precisión de 100% para las 30 primeras tuplas del catálogo utilizando $pmi_2$.

- La curva $pmi_2$ presenta una ligera caída conforme aumenta el número de tuplas evaluadas. Sin embargo, también es importante señalar, que a pesar de la caída, siempre se mantiene una precisión mayor al 75%. Finalmente, se consigue una precisión de 79.50% para las primeras 200 tuplas del catálogo.

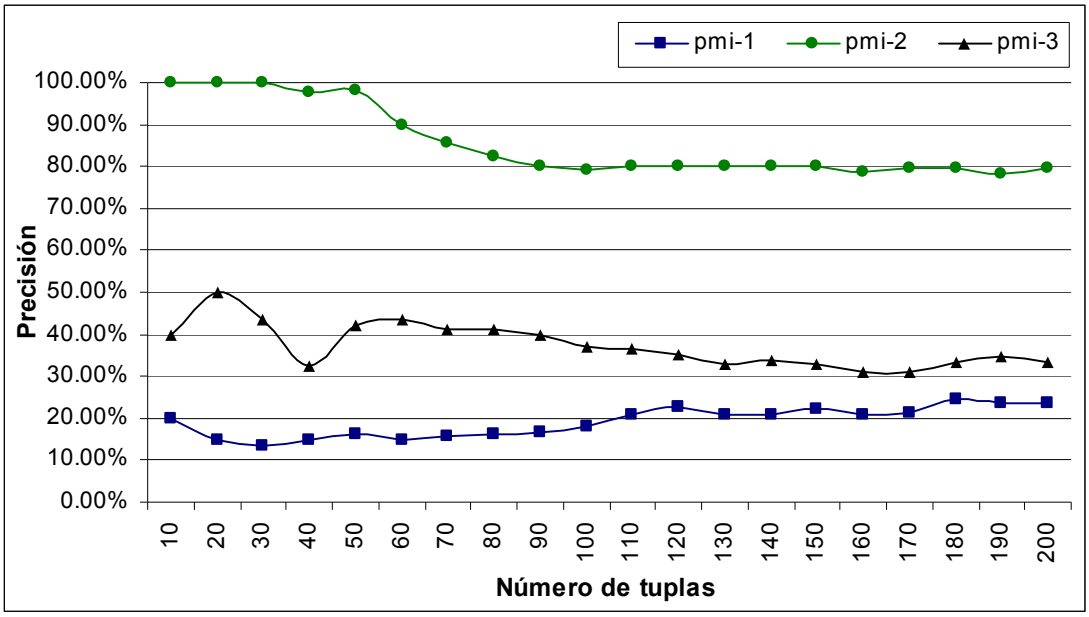


**Figura 13. Resultados en la 3ra. Iteración con información mutua en el arranque**

### Comparación de resultados sobre las tres iteraciones usando $pmi_2$

Los resultados presentados en la sección anterior evidenciaron el alza de la precisión del catálogo cuando se utiliza $pmi_2$, en comparación con $pmi_1$ o $pmi_3$. Por lo tanto, se puede concluir que $pmi_2$ se adapta mejor al método que se propone. En la tabla 18 se resume el comportamiento del método en las tres iteraciones realizadas usando $pmi_2$. De esa tabla, se observa que en cada iteración se percibe un incremento en precisión con respecto a la

iteración anterior. De tal forma que en la tercera iteración se alcanza un 79.50% de precisión para las primeras 200 tuplas del catálogo.

**Tabla 18. Resultados con $pmi_2$ usando información mutua en el arranque**

| No. de tuplas | 1ra. iteración | 2da. iteración | 3ra. iteración |
|---|---|---|---|
| 10 | 90.00% | 90.00% | 100.00% |
| 20 | 90.00% | 85.00% | 100.00% |
| 30 | 86.67% | 80.00% | 100.00% |
| 40 | 80.00% | 85.00% | 97.50% |
| 50 | 70.00% | 86.00% | 98.00% |
| 60 | 70.00% | 86.67% | 90.00% |
| 70 | 64.29% | 81.43% | 85.71% |
| 80 | 63.75% | 82.50% | 82.50% |
| 90 | 63.33% | 83.33% | 80.00% |
| 100 | 66.00% | 84.00% | 79.00% |
| 110 | 65.45% | 80.91% | 80.00% |
| 120 | 60.00% | 78.33% | 80.00% |
| 130 | 56.15% | 77.69% | 80.00% |
| 140 | 52.86% | 76.43% | 80.00% |
| 150 | 50.67% | 74.00% | 80.00% |
| 160 | 48.75% | 73.13% | 78.75% |
| 170 | 46.47% | 72.94% | 79.41% |
| 180 | 44.44% | 72.22% | 79.44% |
| 190 | 43.68% | 72.63% | 78.42% |
| 200 | 44.50% | 72.00% | 79.50% |

### 6.4.4.2 Resultados usando la medida F en el arranque

En esta sección se presentan los resultados del segundo grupo de experimentos. La característica distintiva de este grupo, con respecto a los experimentos del grupo anterior, es el uso de la adaptación de la medida F en el arranque. No obstante, en el *proceso iterativo de estimación de confianzas*, se usaron las tres opciones mencionadas para obtener la información mutua entre patrones y tuplas ( $pmi_1$ , $pmi_2$ , $pmi_3$ ).

**Primera iteración**

En este espacio se presentan las precisiones obtenidas de la evaluación del catálogo en la primera iteración. Más concretamente, en la figura 14 se muestra el comportamiento de las curvas correspondientes a $pmi_1$, $pmi_2$ y $pmi_3$.

En la figura 14 sobresale un aspecto importante. En efecto, para las tres curvas de precisión se obtienen porcentajes mayores o iguales al 60%. De esta forma se demuestra la conveniencia de usar la adaptación de la medida F en el arranque. Inclusive, aunque se observa una caída en las tres curvas, la magnitud de la caída de $pmi_2$ es más pequeña (menos del 50%) que la magnitud de la caída presentada en la primera iteración cuando se utiliza información mutua en el arranque.

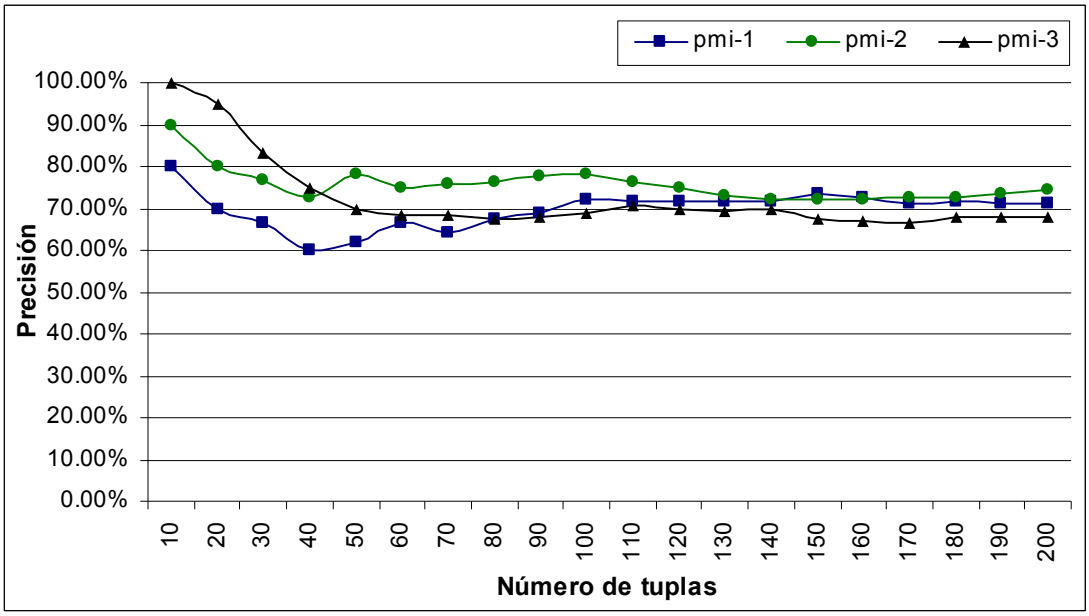


**Figura 14. Resultados en la 1ra. Iteración usando la medida F en el arranque**

**Segunda iteración**

La figura 15 presenta los resultados generados en la segunda iteración. A partir de esta figura, se observa un mayor nivel de inestabilidad en cada una de las curvas, con respecto a las curvas de la primera iteración. La inestabilidad puede atribuirse al uso de información mutua en el proceso iterativo del enfoque. Así también, se observan precisiones cercanas al 100%, específicamente en las primeras tuplas del catálogo.

Por otro lado, la curva correspondiente a $pmi_1$ presenta un comportamiento singular, pues aún cuando parece ser ascendente en precisión, la precisión para las primeras tuplas está por debajo de las otras dos curvas. Por lo tanto, se puede suponer que el comportamiento de $pmi_2$ es más adecuado para la tarea, pues con $pmi_2$ se obtiene una precisión mayor al 75% sobre toda la extensión de la curva. Además, se obtienen altas precisiones cuando se evalúan las tuplas. Finalmente, se asegura una precisión de 78.50% para las primeras 200 tuplas del catálogo.

**Tercera iteración**

Los resultados producidos en la tercera iteración, se ilustran en la figura 16. En esta iteración se marca definitivamente la diferencia entre las tres curvas de precisión. En particular, se observa la superioridad de la curva $pmi_2$, pues se han conseguido precisiones de hasta 100% para las primeras 30 tuplas del catálogo. Por otra parte, aun cuando la curva $pmi_2$ cae conforme aumenta el número de tuplas evaluadas, siempre se mantiene con precisiones mayores al 78%.

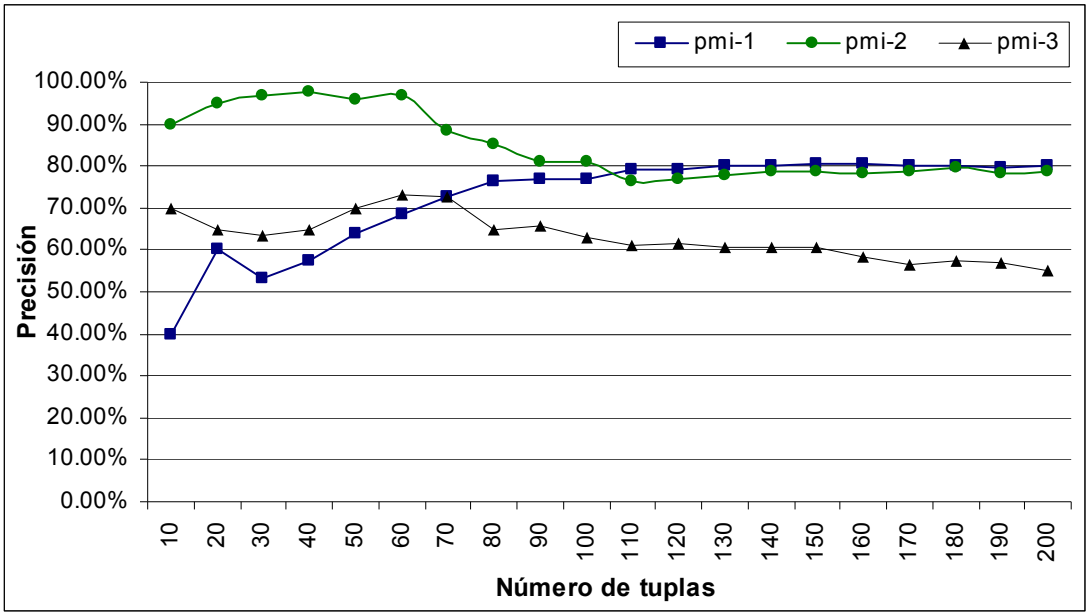


**Figura 15. Resultados en la 2da. Iteración usando la medida F en el arranque**

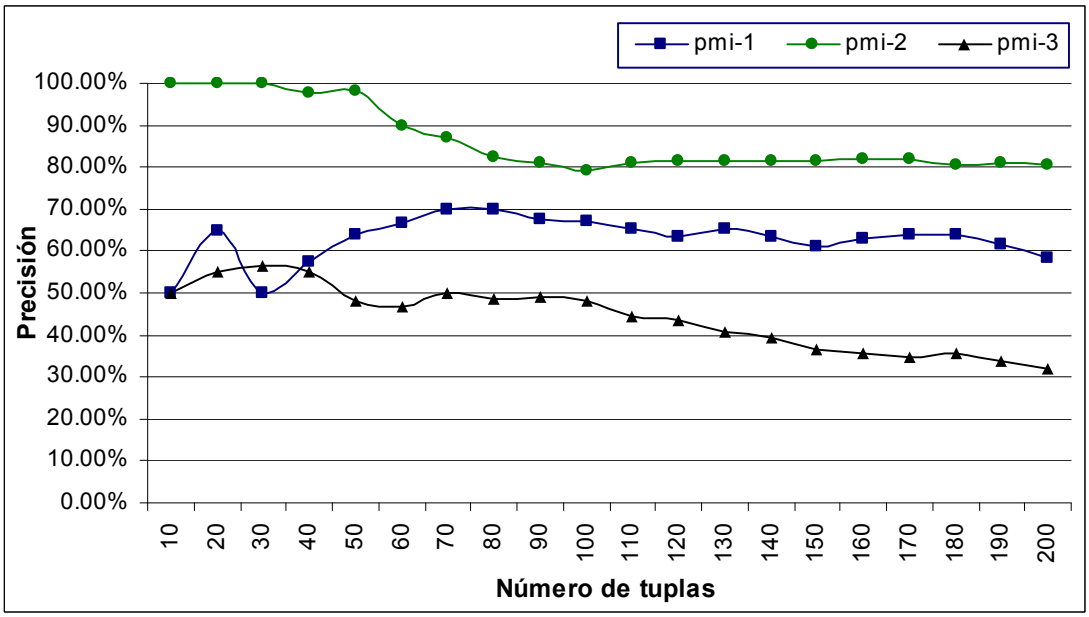


**Figura 16. Resultados en la 3ra. Iteración usando la medida F en el arranque**

### Comparación de resultados sobre las tres iteraciones usando $pmi_2$

Evidentemente, en todas las iteraciones realizadas se distingue la precisión alcanzada por el uso de $pmi_2$. En la tabla 19, se muestra un resumen de su comportamiento. En esa tabla se observa una ligera caída de precisión conforme aumenta el número de tuplas evaluadas. También se observa un comportamiento casi similar entre la segunda y tercera iteración. Sin embargo, sí existe un ligero incremento de precisión en la tercera iteración, obteniendo un 80.50% para las 200 primeras tuplas del catálogo.

**Tabla 19. Resultados con $pmi_2$ usando la medida F en el arranque**

| No. de tuplas | 1ra. iteración | 2da. iteración | 3ra. iteración |
|---|---|---|---|
| 10 | 90.00% | 90.00% | 100.00% |
| 20 | 80.00% | 95.00% | 100.00% |
| 30 | 76.67% | 96.67% | 100.00% |
| 40 | 72.50% | 97.50% | 97.50% |
| 50 | 78.00% | 96.00% | 98.00% |
| 60 | 75.00% | 96.67% | 90.00% |
| 70 | 75.71% | 88.57% | 87.14% |
| 80 | 76.25% | 85.00% | 82.50% |
| 90 | 77.78% | 81.11% | 81.11% |
| 100 | 78.00% | 81.00% | 79.00% |
| 110 | 76.36% | 76.36% | 80.91% |
| 120 | 75.00% | 76.67% | 81.67% |
| 130 | 73.08% | 77.69% | 81.54% |
| 140 | 72.14% | 78.57% | 81.43% |
| 150 | 72.00% | 78.67% | 81.33% |
| 160 | 71.88% | 78.13% | 81.88% |
| 170 | 72.35% | 78.82% | 81.76% |
| 180 | 72.78% | 79.44% | 80.56% |
| 190 | 73.68% | 78.42% | 81.05% |
| 200 | 74.50% | 78.50% | 80.50% |

#### 6.4.4.3 Comparación de las alternativas de arranque

Las dos alternativas presentadas para realizar el arranque difieren únicamente en la forma de estimar inicialmente la confianza de los patrones. La primera alternativa basa la estimación de la confianza inicial de los patrones en el uso de información mutua. Por su parte, la segunda alternativa estima inicialmente la confianza de los patrones con base en la medida F. Ahora bien, ambas alternativas usan información mutua en el *proceso iterativo de estimación de confianzas.* Por ello, en ambas se experimentó con las tres opciones mencionadas para obtenerla ( $pmi_1$, $pmi_2$ y $pmi_3$ ).

Básicamente, los resultados de la sección anterior mostraron:

- La ganancia en precisión usando $pmi_2$ en ambas alternativas. De ahí que se concluya que $pmi_2$ se adaptó mejor a nuestro problema.

- La ganancia en precisión cuando se utiliza la adaptación de la medida F para asignar una confianza inicial a los patrones. Entonces, la medida F se ha adaptado mejor al problema, y tiene características que pueden ser explotadas.

- La "prontitud" con la cual se obtienen resultados favorables cuando se utiliza la adaptación de la medida F. Se observa que desde la primera iteración se están alcanzando precisiones de 74.50% para las 200 tuplas evaluadas. En cambio, con información mutua en la primera iteración se obtiene 44.50%.

Así pues, se deduce que la confianza inicial de los patrones es un factor relevante para determinar la precisión final del catálogo. Por otro lado, el impacto de precisión obtenido cuando se usa la adaptación de la medida F propicia la idea de un segundo enfoque que aproveche la ventaja que ofrece la medida, ya que la medida F evita favorecer a aquellos patrones que tienen un alto recuerdo pero una precisión casi nula. En su lugar, favorece a aquellos que guardan un balance entre precisión y recuerdo. Sin embargo, esta medida también brinda la posibilidad de controlar la variación de la importancia del recuerdo sobre la precisión en un momento determinado.

## 6.5 Enfoque centrado en el uso de la medida F

En esta sección se presenta el segundo enfoque propuesto para ordenar las tuplas del catálogo. Este enfoque fue diseñado para aprovechar las bondades que ofrece la medida F. Sin embargo, antes de comenzar a explicar el desarrollo de este enfoque, primero se definirá la notación y los conceptos principales que serán utilizados en esta sección.

### 6.5.1 Definición de conceptos

Para explicar con más claridad este enfoque, se han definido algunos términos, los cuales han resultado del desarrollo de esta investigación.

**Ámbito de una tupla**

En este documento, se definió al término *ámbito de una tupla* como el conjunto de patrones que extraen dicha tupla, es decir, $P'$ y las frecuencias

de aparición del patrón y la tupla, es decir $|x, p, y|$. Por ejemplo, en la tabla[25] 20 se muestra el ámbito de la tupla $t_{10}$=(diabetes, enfermedad). Para este ejemplo (tabla 20), el conjunto de patrones que extraen la tupla $t_{10}$, está determinado por $P'=\{p_{1,}p_{4,}p_{5,}p_{11,}p_{14,}p_{15}\}$, entonces el *ámbito de la tupla* indica que el patrón $p_{1,}$ y el patrón $p_4$ extrajeron a la tupla $t_{10}$ 5 veces, análogamente el patrón $p_5$ extrajo a dicha tupla 4 veces, el patrón $p_{11,}$ la extrajo 28 veces, el patrón $p_{14,}$ 58 veces y finalmente, el patrón $p_{15}$ la extrajo dos veces.

**Tabla 20. Ámbito de la tupla $t_{10}$ (*diabetes , enfermedad)***

| $P'$ | $\|x, p, y\|$ |
|---|---|
| $p_1$ | 5 |
| $p_4$ | 5 |
| $p_5$ | 4 |
| $p_{11}$ | 28 |
| $p_{14}$ | 58 |
| $p_{15}$ | 2 |

**Relevancia en el ámbito de la tupla**

En esta tesis, se utiliza el término *relevancia en el ámbito de la tupla*, denotada por $ra_\sigma$, para capturar el grado de importancia que tiene un determinado patrón dentro del ámbito de una determinada tupla. El valor de la relevancia en el ámbito de la tupla está ubicado en el intervalo [0,1]. De tal forma que el patrón que extraiga a la tupla con la mayor frecuencia, obtiene un valor de relevancia igual a 1.

[25] La definición o descripción de los patrones mostrados en la tabla 20, se muestran en la tabla 22.

En concreto, la relevancia de una tupla $t=(x, y)$ y un patrón $p$ en el ámbito de la tupla, es decir, $ra_{\sigma}(t, p)$, está dada por la ecuación (14).

$$ra_{\sigma}(t, p)=\frac{|x, p, y|}{max\{|x, *, y|\}} \quad (14)$$

Donde $max\{|x, *, y|\}$ es la máxima frecuencia de aparición de la tupla con cualquier patrón. En otras palabras, es el máximo número de veces que algún patrón extrajo la tupla $t$.

Entonces, para el ejemplo de la tabla 20, la relevancia del patrón $p_{15}$ (en el ámbito de la tupla $t_{10}$: (*diabetes, enfermedad)* está dado por: $ra_{\sigma}(t_{11}, p_{15})=\frac{2}{58}$. Y la relevancia del patrón $p_{14}$ en el ámbito de la misma tupla corresponde a $ra_{\sigma}(t_{10}, p_{14})=\frac{58}{58}=1$. En tales casos y de acuerdo con la descripción del término *relevancia en el ámbito de la tupla,* el patrón $p_{14}$ es más relevante que el patrón $p_{15}$, lo cual indica que dicho patrón extrajo a la tupla con mayor frecuencia que el resto de los patrones.

**Ámbito de un patrón.**

En este documento también se introdujo el término *ámbito de un patrón.* Este ámbito está formado por aquellas tuplas que extrae un patrón $p$ , es decir, $T'$, y además, la frecuencia de aparición de una tupla $t=(x, y)$ y el patrón, es decir, $|x, p, y|$. Por ejemplo, la tabla 21 muestra el ámbito del patrón $p_{11}$: "*que la <hipónimo> es una <hiperónimo>".* En este ejemplo se indica que el patrón $p_{11}$ extrajo a la tupla $t_9$ 21 veces, así como a la tupla $t_{10}$ 28 veces, a la tupla $t_{12}$ 2 veces y finalmente, a la tupla $t_{14}$ la extrajo una sola vez.

**Tabla 21. Ámbito de $p_{11}$: *que la <hipónimo> es una <hiperónimo>***

| $T'$ | $\|x, p, y\|$ |
|---|---|
| $t_9$ | 21 |
| $t_{10}$ | 28 |
| $t_{12}$ | 2 |
| $t_{14}$ | 1 |

**Relevancia en el ámbito del patrón**

De forma análoga a la *relevancia en el ámbito de la tupla*, en este documento se utilizó el término *relevancia en el ámbito del patrón,* denotada por $ra_\pi$, para capturar el grado de importancia que tiene una determinada tupla dentro del ámbito de un determinado patrón.

Concretamente, para una tupla $t=(x, y)$ y un patrón $p$ la importancia de la tupla $t$ en el ámbito del patrón $p$, es decir, $ra_\pi(t, p)$, está determinada por la ecuación 15.

$$ra_\pi(t, p)=\frac{|x, p, y|}{\max\{|*, p, *|\}} \quad (15)$$

Donde $\max\{|*, p, *|\}$ , es la máxima frecuencia con la que el patrón extrajo alguna tupla.

Para el ejemplo de la tabla 21, la tupla $t_{10}$:(*diabetes, enfermedad*) tiene la mayor relevancia (valor de 1) en el ámbito del patrón $p_{11}$, es decir, $ra_\pi(t_{10}, p_{11})=\frac{28}{28}=1$.

**Patrones-semilla**

Durante los módulos de este enfoque se hace referencia al término *patrones-semilla*. En esta tesis se utiliza este término para hacer referencia al subconjunto de patrones que actuarán como semillas en un momento determinado. Si bien reciben el nombre de *patrones-semilla*, no significa que sean elegidos manualmente como las semillas que toma como entrada el método (ver sección 5.1.1). Más bien, son elegidos automáticamente. Y reciben ese nombre únicamente por analogía con las semillas iniciales, pues en un momento determinado se supondrá que ese conjunto de patrones es confiable.

Ahora bien, para elegir el conjunto de patrones-semilla, el cual es denotado como $PS$, se siguen los siguientes pasos:

1. Determinar la cardinalidad ($s_\pi$) del conjunto $PS$. Es decir, determinar el número de patrones que formarán el conjunto de patrones-semilla.

2. Ordenar descendentemente, de acuerdo con su confianza, el conjunto completo de patrones $P$. Por el momento, el lector no debe preocuparse por conocer cómo se obtiene la confianza de cada uno de los patrones, ya que en las siguientes secciones se mostrará con detalle dicho proceso. Brevemente y de manera general, para obtener la confianza de los patrones se propone una métrica basada en la medida F (para más detalles, ver sección 6.5.3) y un esquema propuesto de estimación de confianzas, el cual será mostrado en la sección 6.5.4.2.

3. Tomar los $s_\pi$ patrones con las confianzas más altas para formar el conjunto $PS$.

Todos aquellos patrones restantes forman el conjunto de patrones ordinarios, el cual está denotado por $PO$.

**Sobre la cardinalidad $s_\pi$.** Este valor representa el número de elementos del conjunto $PS$, por lo tanto $s_\pi = |PS|$. Si bien puede elegirse arbitrariamente un número (mayor que $0$ y menor que $|P|$. Por ejemplo: 10, 20 ó 50), sería preferible realizar un estudio sobre el valor adecuado para $s_\pi$. Sin embargo, aún cuando en este trabajo de tesis, no se realizó un estudio minucioso para elegir adecuadamente este valor, se propone lo siguiente: el valor $s_\pi$ corresponde al número de patrones que sobrepasan el promedio de las confianzas del conjunto entero de patrones ($P$). Entonces, el conjunto de patrones-semilla puede escribirse como: $PS = \{ p \in P | c_\pi(p) \geq (prom_\pi) \}$ donde $prom_\pi$ es el promedio de las confianzas de los patrones. Por ejemplo, para la tabla 22, el valor de $s_\pi$ es igual a 8, por ser ocho los patrones que tiene confianza mayor al promedio general (0.46).

Entonces, para los patrones de la tabla 22, el conjunto de patrones-semilla está determinado por el conjunto: $PS = \{ p_{12}, p_9, p_{11}, p_{14}, p_2, p_5, p_{10}, p_8 \}$. Por otra parte, el conjunto de patrones ordinarios está determinado por: $PO = \{ p_4, p_3, p_{13}, p_1, p_7, p_{15}, p_6 \}$.

Básicamente, los patrones-semilla tienen las mejores confianzas porque guardan un mejor equilibrio entre precisión y recuerdo. Con esta aseveración

se puede suponer que aquellas tuplas que son extraídas por estos patrones tienen una alta probabilidad de ser tuplas correctas.

**Tabla 22. Ejemplo de selección de *patrones-semilla***

| | Num. | | Patrón | Confianza |
|---|---|---|---|---|
| *semillas* | 1 | $p_{12}$ | el <hipónimo> es el único <hiperónimo> | 1.00 |
| | | $p_9$ | el <hipónimo> es un <hiperónimo> que | 0.90 |
| | . | $p_{11}$ | que la <hipónimo> es una <hiperónimo> | 0.73 |
| | . | $p_{14}$ | la <hipónimo> es una <hiperónimo> que | 0.64 |
| | . | $p_2$ | el <hipónimo> es un <hiperónimo> de | 0.63 |
| | | $p_5$ | de la <hipónimo> como <hiperónimo> de | 0.62 |
| | | $p_{10}$ | el <hipónimo> es el <hiperónimo> que | 0.57 |
| | $s_\pi$ | $p_8$ | de <hiperónimo> como <hipónimo> y | 0.45 |
| | | $p_4$ | la <hipónimo> es la <hiperónimo> | 0.31 |
| | | $p_3$ | el <hipónimo> es la <hiperónimo> | 0.22 |
| | | $p_{13}$ | de <hipónimo> o <hiperónimo> | 0.18 |
| | | $p_1$ | la <hipónimo> como una <hiperónimo> | 0.16 |
| | | $p_7$ | de <hipónimo> y <hiperónimo> | 0.06 |
| | | $p_{15}$ | la <hipónimo> una <hiperónimo> | 0.06 |
| | | $p_6$ | <hiperónimo> de <hipónimo> y | 0.05 |

### Tuplas-semilla

Análogamente al proceso de selección de patrones-semilla, aquí se seleccionará un conjunto de tuplas-semilla, el cual es denotado como $TS$. A su vez, el conjunto de tuplas ordinarias estará denotado por $TO$. La tabla 23, muestra un ejemplo de selección de tuplas-semilla. En concreto, para la selección se realiza lo siguiente:

1. Determinar la cardinalidad ($s_\sigma$) del conjunto $TS$.

2. Ordenar descendentemente, de acuerdo con la confianza de las tuplas. La forma de estimar la confianza de las tuplas se mostrará más adelante en la sección 6.5.4.1.

3. Tomar las $s_\sigma$ tuplas con las confianzas más altas para formar el conjunto $TS$.

**Tabla 23. Ejemplo de selección de *tuplas-semilla***

| | Num. | Tupla | Confianza |
|---|---|---|---|
| *semillas* | 1 | $t_9$ | (tuberculosis, enfermedad) | 1 |
| | | $t_8$ | (cáncer, enfermedad) | 0.86 |
| | . | $t_{10}$ | (diabetes, enfermedad) | 0.82 |
| | . | $t_{13}$ | (bulimia, enfermedad) | 0.41 |
| | . | $t_1$ | (puma, felino) | 0.35 |
| | | $t_3$ | (jaguar, felino) | 0.33 |
| | | $t_6$ | (lince ibérico, felino) | 0.14 |
| | $s_\sigma$ | $t_{12}$ | (neumonía, enfermedad) | 0.12 |
| | | $t_{11}$ | (periodontitis, enfermedad) | 0.09 |
| | | $t_4$ | (ocelote, felino) | 0.08 |
| | | $t_{14}$ | (insuficiencia cardiaca, enfermedad) | 0.07 |
| | | $t_{15}$ | (cardiopatía isquémica, enfermedad) | 0.06 |
| | | $t_7$ | (origen canino, felino) | 0.02 |
| | | $t_5$ | (fuego, felino) | 0.01 |
| | | $t_2$ | (pájaro, felino) | 0.01 |

**Sobre la cardinalidad $s_\sigma$.** Este valor representa la cardinalidad del conjunto *TS*. Si bien, dicho valor puede ser un número arbitrario, en este trabajo de tesis se propuso un criterio. Este criterio establece que el valor de $s_\sigma$ corresponde al número de tuplas cuya confianza sobrepasa la siguiente suma: $(prom_\sigma + n\delta)$, donde $prom_\sigma$ representa al promedio general de las

confianzas de las tuplas, $n$ es una constante que indica el número de veces que la desviación estándar ($\delta$) será sumada al promedio. Así pues, el conjunto de tuplas-semilla puede escribirse como: $TS = \{t \in T \mid c_\sigma(t) \geq (prom_\sigma + n\delta)\}$.

En efecto, este criterio difiere del criterio propuesto para elegir el valor de la cardinalidad de los patrones-semilla ($s_\pi$), pues se sabe que el número de tuplas en el catálogo es mucho más grande que el número de patrones. Por ello, el número de tuplas que pueden sobrepasar el promedio podría ser muy grande. En tal caso, la finalidad de usar tuplas-semilla podría perderse, pues recordemos que las semillas son un conjunto selectivo que presume de alta confiabilidad. Por lo tanto, para resolver este problema se añade a la condición, la suma de la desviación estándar $n$ veces.

### 6.5.2 Arquitectura del enfoque

Ya se había mencionado que los enfoques de ordenamiento propuestos en esta tesis constan de dos módulos (ver figura 9, página 69):

1. *Arranque*
2. *Proceso iterativo de estimación de confianzas.*

Pues bien, este enfoque sigue esa misma arquitectura. Por supuesto, los cambios con respecto al enfoque anterior radican en las medidas utilizadas para estimar las confianzas de las tuplas y de los patrones. Estos módulos serán descritos en las siguientes secciones.

### 6.5.3 Arranque

Este enfoque inicia estimando inicialmente la confianza de los patrones a través de una adaptación de la medida F. Ahora bien, esta confianza ($c_\pi$) se obtiene con la ecuación (11) (página 77) la cual se muestra nuevamente para efectos de claridad.

$$c_\pi(p) = \frac{F_\pi(p)}{\max\limits_{\forall q \in P} \{F_\pi(q)\}}$$

Esta ecuación es retomada porque en el enfoque centrado en el uso de información mutua (ver sección 6.4.1) se observó un mejor desempeño cuando se usó la adaptación de la medida F para asignar la confianza inicial de los patrones.

### 6.5.4 Proceso iterativo de estimación de confianzas

Nuevamente, en este proceso se presentan dos elementos: la confianza de tuplas y la confianza de los patrones. Éstos serán explicados en las siguientes secciones.

#### 6.5.4.1 Confianza de tuplas

Básicamente, para estimar la confianza de las tuplas del catálogo, en este enfoque, se realizan los siguientes pasos:

1. Elegir el conjunto de patrones-semilla (PS). Este conjunto proporcionará información que ayudará a determinar la confianza de las tuplas.

2. Asignar un valor de confianza a cada tupla del catálogo a través de la ecuación 16.

$$c_{\sigma}(t)=\frac{\left(\lambda_1 \cdot f_{\sigma}(t)+\lambda_2 \cdot v_{\pi}(t)\right)\cdot e_{\sigma}(t)}{\max\limits_{\forall q \in T}\left\{c_{\sigma}(q)\right\}} \quad (16)$$

Ahora bien, la fórmula anterior conjuga cuatro componentes:

- $f_{\sigma}$, una adaptación de la medida F para tuplas.
- $v_{\pi}$, una función de valoración de los patrones.
- $e_{\sigma}$, un valor de premio.
- $\max\limits_{\forall q \in T}\left\{c_{\sigma}(q)\right\}$, un factor de normalización.

En la misma fórmula $\lambda_1$ y $\lambda_2$ representan valores escalares, los cuales reflejan la importancia de los componentes. Por lo tanto, $\lambda_1 + \lambda_2 = 1$. Por ejemplo, si $\lambda_1 > \lambda_2$ se está considerando más relevante el primer componente (la medida F) que el segundo componente (la valoración de los patrones). Pues bien, recordando que el objetivo de este enfoque es explotar el uso de la medida F, se sugiere dar una relevancia mayor del 50% al primer componente, es decir, $\lambda_1 > 0.5 > \lambda_2$. Inclusive, como se verá más adelante, *la valoración los patrones* surgió como un componente necesario para enriquecer el uso de la medida F.

En los siguientes párrafos se explicará cada uno de los componentes que integran la ecuación 16. Lo anterior permitirá comprender el significado implícito en dicha ecuación.

**Primer Componente: $f_\sigma$, una adaptación de la medida F para tuplas**

La función $f_\sigma$ es una adaptación de la medida F, como se muestra en la siguiente ecuación:

$$f_\sigma(t)=\frac{(\beta^2+1)\cdot P_\sigma(t)\cdot R_\sigma(t)}{\beta\cdot P_\sigma(t)+R_\sigma(t)} \quad (17)$$

Donde:

- $P_\sigma$ (ver ecuación 18), indica la porción de patrones-semilla que extraen la tupla $t$.
- $R_\sigma$ (ver ecuación 19) es el cociente del número de patrones-semilla que extraen la tupla $t$ y el conjunto de patrones-semilla.

$$P_\sigma(t)=\frac{|PS'|}{|P'|} \quad (18)$$

$$R_\sigma(t)=\frac{|PS'|}{|PS|} \quad (19)$$

Ahora bien, idealmente se asume que mientras más grande sea el número de patrones-semilla que extraen una tupla, más confiable será dicha tupla. De ahí que $\beta=2$ en la ecuación 17, lo cual indica que la porción de semillas extraídas ($R_\sigma$) es más relevante que la precisión ($P_\sigma$). Para justificar la relevancia del recuerdo sobre la precisión, se presenta el siguiente caso:

Para $|PS|=4$, una tupla $t$ extraída por únicamente un patrón, el cual es un patrón-semilla, obtendría precisión igual a 1, es decir, $P_{\sigma}(t)=1$. En cambio, el recuerdo de la tupla sería menor a uno, es decir, $R_{\sigma}(t)<1$. De lo anterior se observa que el recuerdo es más relevante, ya que captura adecuadamente la importancia de extraer un mayor número de semillas, sin prestar tanta atención al número de patrones ordinarios que extraen la tupla. No obstante, la información proporcionada por los patrones ordinarios será considerada en el segundo componente, el cual se explicará más adelante.

La tabla 24 muestra un ejemplo del resultado de estimar la confianza de las tuplas con la función $f_{\sigma}$.

**Segundo Componente: $v_{\pi}$, valoración de los patrones**

Una vez que la función $f_{\sigma}$ es calculada para todas las tuplas del catálogo (ver tabla 24), se observan dos situaciones:

1. **Confianzas nulas**. Se obtienen para aquellas tuplas que son extraídas únicamente por patrones ordinarios. Lo anterior indica que: el valor de $f_{\sigma}$ para una tupla extraída por varios patrones ordinarios y ningún patrón-semilla, es igual a 0. Sin embargo, probablemente la confianza de la tupla incrementaría si se considerara el número de patrones ordinarios que extraen la tupla en suma con la confianza de esos patrones.

   Por ejemplo, en la tabla 24 se presentan confianzas nulas para la tupla: (insuficiencia cardiaca, enfermedad), la cual es correcta pero tiene un valor $f_{\sigma}=0$ porque *es* extraída por los patrones: $p_1, p_{13}, p_{15}$ que no forman parte del conjunto de patrones-semilla (de acuerdo con la tabla 22).

**Tabla 24. Ejemplo de las tuplas cuya confianza es estimada con la medida F**

| No. | Hipónimo | Hiperónimo | P' | \|PS'\| | $f_\sigma$ |
|---|---|---|---|---|---|
| $t_3$ | jaguar | felino | $p_9, p_{10}, p_{12}$ | 3 | 0.43 |
| $t_1$ | puma | felino | $p_2, p_9, p_{12}$ | 3 | 0.43 |
| $t_{13}$ | bulimia | enfermedad | $p_1, p_8, p_{11}, p_{14}$ | 3 | 0.42 |
| $t_{10}$ | diabetes | enfermedad | $p_1, p_4, p_5, p_{11}, p_{14}, p_{15}$ | 3 | 0.39 |
| $t_8$ | cáncer | enfermedad | $p_2, p_3, p_7, p_8, p_9, p_{13}$ | 3 | 0.39 |
| $t_9$ | tuberculosis | enfermedad | $p_1, p_4, p_5, p_7, p_{11}, p_{14}, p_{15}$ | 3 | 0.38 |
| $t_6$ | lince ibérico | felino | $p_2, p_{10}$ | 2 | 0.29 |
| $t_4$ | ocelote | felino | $p_3, p_{10}$ | 1 | 0.15 |
| $t_{11}$ | periodontitis | enfermedad | $p_4, p_{11}, p_{13}$ | 1 | 0.14 |
| $t_{12}$ | neumonía | enfermedad | $p_4, p_7, p_{14}, p_{15}$ | 1 | 0.14 |
| $t_2$ | pájaro | felino | $p_7, p_{13}$ | 0 | 0.00 |
| $t_7$ | origen canino | felino | $p_7, p_{13}$ | 0 | 0.00 |
| $t_{14}$ | insuficiencia cardiaca | enfermedad | $p_1, p_{13}, p_{15}$ | 0 | 0.00 |
| $t_5$ | fuego | felino | $p_6, p_{13}$ | 0 | 0.00 |
| $t_{15}$ | cardiopatía isquémica | enfermedad | $p_4, p_6, p_7$ | 0 | 0.00 |

2. **Empate de confianzas**. Se obtienen para aquellas tuplas que son extraídas por el mismo número de patrones-semilla y el mismo número de patrones ordinarios; sin importar el valor de confianza de los patrones que extrajeron la tupla. Por ejemplo, en la tabla 24, las tuplas: *(jaguar, felino*) y (*puma, felino*), presentan el mismo valor para la función $f_\sigma$. No obstante, la confianza de los patrones que extraen cada una de estas tuplas es diferente.

Así pues, la *valoración de los patrones* ($v_\pi$) intenta resolver los problemas marcados por las dos situaciones anteriores.

Ante todo, la función $v_{\pi}$ determina qué tanta confianza aportan los patrones que extraen una tupla. Básicamente, la aportación de un patrón $p$ para una tupla $t$ está determinada por tres elementos:

1. Confianza del patrón, $c_{\pi}(p)$. Idealmente, se considera que entre más alta sea la confianza del patrón, la confiabilidad de la tupla será enriquecida.

2. Relevancia en el ámbito de la tupla, $ra_{\sigma}(t,p)$. Aquí se trata de capturar el grado de importancia que tiene el patrón en el ámbito de la tupla. De tal manera que el patrón más relevante para una tupla gana el derecho de aportar mayor porción de su confianza, que aquellos patrones que no son relevantes.

3. Relevancia en el ámbito del patrón $ra_{\pi}(t,p)$. Aquí se trata de capturar la importancia de la tupla en el ámbito del patrón. De tal manera que un patrón aportará una mayor porción de su confianza a las tuplas más relevantes de su ámbito.

Con base en lo anterior, se puede deducir que aquel patrón que tenga una relevancia de 1 en el ámbito de la tupla y así mismo, una relevancia de 1 en el ámbito del patrón, aporta su confianza completa.

Entonces, la aportación ($a_{\pi}$) de un patrón $p$ denotada por $a_{\pi}(p)$, está determinada con la fórmula 20.

$$a_{\pi}(p)=c_{\pi}(p)\cdot ra_{\sigma}(t,p)\cdot ra_{\pi}(t,p) \qquad (20)$$

Concretamente, la valoración de los patrones para una tupla $t$, es decir, $v_{\pi}(t)$ está determinada por la suma de las aportaciones de cada uno de los patrones que extraen la tupla (ver ecuación 21). Definitivamente, entre más alta sea esta suma, la confianza de la tupla incrementará.

$$v_{\pi}(t) = \sum_{p \in P'} a_{\pi}(p) \tag{21}$$

**Tercer componente: $e_{\sigma}$, un valor premio**

Para explicar este componente, primero es necesario recordar el primer criterio general planteado al inicio de este capítulo. Este criterio manifiesta que las tuplas extraídas por un gran número de patrones tienen una alta probabilidad de ser correctas. Pues bien, este criterio queda reflejado a través de un premio ($e_{\sigma}$). Entonces, el valor del premio para una tupla $t$, es decir $e_{\sigma}(t)$, corresponde al número de patrones que extraen la tupla $t$. En concreto, este valor se obtiene con la ecuación 22.

$$e_{\sigma}(t) = |P'| \tag{22}$$

**Cuarto componente: $\max_{\forall q \in T} \{c_{\sigma}(q)\}$, un factor de normalización**

Este factor de normalización, asegura que las tuplas siempre mantengan confianzas dentro de un intervalo [0,1].

#### 6.5.4.2 Confianza de patrones

El procedimiento para estimar la confianza de los patrones, es realmente análogo al procedimiento seguido para estimar la confianza de las tuplas.

1. Elegir el conjunto de tuplas-semilla ($TS$ ).

2. Asignar un valor de confianza a cada patrón a través de la ecuación 23.

$$c_\pi(p)=\frac{\left(\left(\lambda_1 \cdot f_\pi(p)\right)+\left(\lambda_2 \cdot v_\sigma(p)\right)\right)\cdot e_\pi(p)}{\max\limits_{\forall q\in P}\left\{c_\pi(q)\right\}} \quad (23)$$

De manera similar a la estimación de la confianza de las tuplas, la estimación de la confianza de los patrones presenta cuatro componentes:

- $f_\pi$ , una adaptación de la medida F para patrones.
- $v_\sigma$ , una función de valoración de las tuplas.
- $e_\pi$ , un valor de castigo.
- $\max\limits_{\forall q\in P}\left\{c_\pi(q)\right\}$ , un factor de normalización.

Cada uno de estos componentes será explicado a continuación.

**Primer componente: $f_\pi$ , una adaptación de la medida F para patrones**

La función $f_\pi$ es una adaptación de la medida F. La ecuación 24 describe esta función.

$$f_\pi(p)=\frac{(\beta^2+1)\cdot P_\pi(p)\cdot R_\pi(p)}{\beta\cdot P_\pi(p)+R_\pi(p)} \text{ donde } \beta=2 \qquad (24)$$

Nuevamente, se sabe que un patrón es más confiable si extrae un mayor número de tuplas-semilla. De ahí que $\beta=2$, lo que indica que es más relevante la porción de semillas extraídas (recuerdo) que la precisión.

En la tabla 25, se muestra un ejemplo de la aplicación de la función $f_\pi$ al conjunto de patrones de la tabla 22.

**Tabla 25. Ejemplo de los patrones cuya confianza es estimada con la medida F**

| No. | Patrón | T' | \|TS'\| | $f_\pi$ |
|---|---|---|---|---|
| $p_{14}$ | la <hipónimo> es una <hiperónimo> que | $t_9, t_{10}, t_{12}, t_{13}$ | 4 | 0.56 |
| $p_2$ | el <hipónimo> es un <hiperónimo> de | $t_1, t_6, t_8$ | 3 | 0.43 |
| $p_9$ | el <hipónimo> es un <hiperónimo> que | $t_1, t_3, t_8$ | 3 | 0.43 |
| $p_{11}$ | que la <hipónimo> es una <hiperónimo> | $t_9, t_{10}, t_{11}, t_{13}$ | 3 | 0.42 |
| $p_1$ | la <hipónimo> como una <hiperónimo> | $t_9, t_{10}, t_{13}, t_{14}$ | 3 | 0.42 |
| $p_{15}$ | la <hipónimo> una <hiperónimo> | $t_9, t_{10}, t_{12}, t_{14}$ | 3 | 0.42 |
| $p_4$ | la <hipónimo> es la <hiperónimo> | $t_9, t_{11}, t_{12}, t_{15}$ | 3 | 0.41 |
| $p_7$ | de <hipónimo> y <hiperónimo> | $t_2, t_7, t_8, t_9, t_{12}, t_{15}$ | 3 | 0.39 |
| $p_{12}$ | el <hipónimo> es el único <hiperónimo> | $t_1, t_3$ | 2 | 0.29 |
| $p_8$ | de <hiperónimo> como <hipónimo> y | $t_8, t_{13}$ | 2 | 0.29 |
| $p_5$ | de la <hipónimo> como <hiperónimo> de | $t_9, t_{10}$ | 2 | 0.29 |
| $p_{10}$ | el <hipónimo> es el <hiperónimo> que | $t_3, t_4, t_6$ | 2 | 0.29 |
| $p_3$ | el <hipónimo> es la <hiperónimo> | $t_4, t_8$ | 1 | 0.15 |
| $p_{13}$ | de <hipónimo> o <hiperónimo> | $t_2, t_5, t_7, t_8, t_{11}, t_{14}$ | 1 | 0.13 |
| $p_6$ | <hiperónimo> de <hipónimo> y | $t_5, t_{15}$ | 0 | 0.00 |

**Segundo componente: $v_\sigma$, valoración de las tuplas**

Al aplicar el primer componente ($f_\pi$) a los patrones, puede observarse la reincidente aparición de las situaciones:

1. Confianzas nulas.
2. Empate de confianzas.

Para resolver estas situaciones surge este segundo componente, el cual contempla las aportaciones de las confianzas de cada una de las tuplas extraídas por un patrón.

Así pues, la aportación de confianza de una tupla $t$, denotada por $a_\sigma$, es decir $a_\sigma(t)$, está determinada por tres elementos:

1. Confianza de la tupla $c_\sigma(t)$.
2. Relevancia en el ámbito de la tupla $ra_\sigma(t,p)$.
3. Relevancia en el ámbito del patrón $ra_\pi(t,p)$.

Concretamente, $a_\sigma$ para una tupla $t$ se obtiene aplicando la ecuación 25.

$$a_\sigma(t)=c_\sigma(t)\cdot ra_\sigma(t,p)\cdot ra_\pi(t,p) \qquad (25)$$

Ahora bien, los patrones amplios tienen probabilidades altas de extraer muchas tuplas ordinarias e incluso tuplas-semilla. Para evitar favorecer este tipo de patrones, se normaliza tanto la aportación de las tuplas-semilla como la aportación de las tuplas ordinarias. En concreto, la valoración de las tuplas ($v_\sigma$) depende de la información proporcionada por las tuplas-semilla y

también de la información proporcionada por los tuplas ordinarias que extrajo el patrón, es decir, $TS'$ y $TO'$ respectivamente. Específicamente, el valor de $v_\sigma$ para un patrón $p$ se obtiene aplicando la ecuación 26.

$$v_\sigma(p) = \frac{\sum_{t \in TS'} a_\sigma(t)}{|TS'|} + \frac{\sum_{i \in TO'} a_\sigma(i)}{|TO'|} \quad (26)$$

**Tercer componente: $e_\pi$, castigo a patrones amplios**

Típicamente, aquellos patrones que extraen un gran número de tuplas generalmente extraen tuplas incorrectas (patrones amplios). Lo anterior indica que estos patrones no son muy confiables, principalmente por la abundancia de tuplas incorrectas que pueden extraer. Entonces, para no favorecer a estos patrones se propuso incluir un castigo, denotado por $e_\pi$ y el cual se calcula con la fórmula 27.

$$e_\pi(p) = \frac{1}{|T'|} \quad (27)$$

**Cuarto componente: $\max_{\forall q \in P} \{c_\pi(q)\}$, un factor de normalización**

La finalidad de la normalización es mantener la confianza de los patrones en un intervalo [0,1].

### 6.5.5 Resumen del enfoque centrado en el uso de la medida F

Para mayor claridad, en tabla 26 se resumen las ecuaciones utilizadas en el *enfoque centrado en el uso de la medida F*.

**Tabla 26. Formulario general del enfoque centrado en el uso de la medida F**

| **Arranque** | |
|---|---|
| $c_\pi(p)=\frac{F_\pi(p)}{\max_{\forall q\in P}\{F_\pi(q)\}}$ | |
| **Proceso iterativo de estimación de confianzas** | |
| *Confianza de tuplas* | *Confianza de patrones* |
| $c_\sigma(t)=\frac{(\lambda_1\cdot f_\sigma(t)+\lambda_2\cdot v_\pi(t))\cdot e_\sigma(t)}{\max_{\forall q\in T}\{c_\sigma(q)\}}$ | $c_\pi(p)=\frac{((\lambda_1\cdot f_\pi(p))+(\lambda_2\cdot v_\sigma(p)))\cdot e_\pi(p)}{\max_{\forall q\in P}\{c_\pi(q)\}}$ |

### 6.5.6 Experimentos y resultados

El catálogo obtenido después del filtro inicial, es el catálogo utilizado para probar este segundo enfoque de ordenamiento. Se trabajó con los 43 patrones resultantes de los experimentos en el capítulo 5.

Para seleccionar los patrones-semilla se utilizó un valor de $s_\pi=15$, ya que son 15 los patrones que inicialmente sobrepasaron el promedio de la confianza de todos los patrones. Por otra parte, para seleccionar las tuplas-semilla se usó un valor de $s_\sigma=20$, con $n=3$. Es decir, fueron 20 las tuplas cuya confianza sobrepasó la suma del promedio y tres veces la desviación estándar.

A su vez, para determinar la confianza de los patrones y de las tuplas se usó la siguiente configuración de parámetros: $\lambda_1=0.75$ y $\lambda_2=0.25$ en las fórmulas 17 y 23.

### 6.5.6.1 Resultados experimentales

En la tabla 27 y en la figura 17 se presentan los resultados obtenidos al aplicar el enfoque *centrado en el uso de la medida F* para ordenar las tuplas del catálogo. Aunque existen diferencias en el comportamiento entre iteraciones sucesivas, realmente no existe una gran diferencia en precisión. Sin embargo, gracias a las iteraciones se logra aumentar la precisión obtenida en la primera iteración de 79.50% a 82.50% en la tercera iteración.

**Tabla 27. Resultados del enfoque centrado en el uso de la medida F**

| No. de tuplas | 1ra. iteración | 2da. iteración | 3ra. iteración |
|---|---|---|---|
| 10 | 100.00% | 100.00% | 100.00% |
| 20 | 95.00% | 100.00% | 100.00% |
| 30 | 93.33% | 93.33% | 93.33% |
| 40 | 90.00% | 95.00% | 95.00% |
| 50 | 86.00% | 88.00% | 90.00% |
| 60 | 85.00% | 88.33% | 86.67% |
| 70 | 85.71% | 84.29% | 85.71% |
| 80 | 85.00% | 85.00% | 86.25% |
| 90 | 85.56% | 84.44% | 87.78% |
| 100 | 85.00% | 86.00% | 87.00% |
| 110 | 84.55% | 85.45% | 85.45% |
| 120 | 85.00% | 85.00% | 85.83% |
| 130 | 86.15% | 86.15% | 84.62% |
| 140 | 85.00% | 84.29% | 84.29% |
| 150 | 84.00% | 82.67% | 82.67% |
| 160 | 82.50% | 81.88% | 82.50% |
| 170 | 81.76% | 81.76% | 82.94% |
| 180 | 81.67% | 81.11% | 81.11% |
| 190 | 81.05% | 81.05% | 82.11% |
| 200 | 79.50% | 80.00% | 82.50% |

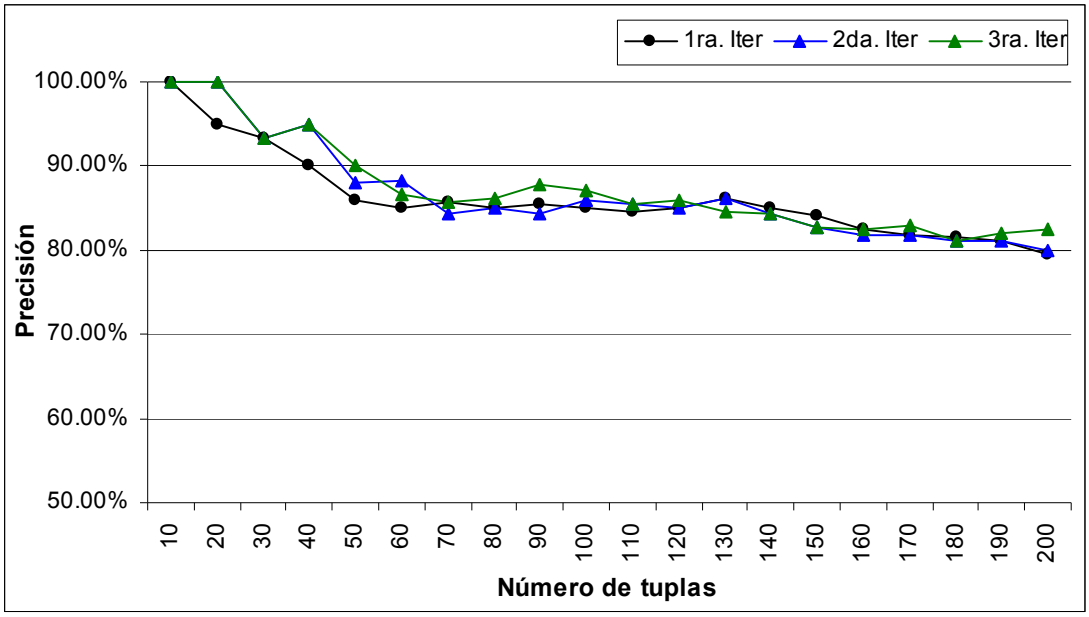


**Figura 17. Precisión del enfoque centrado en el uso de la medida F**

**Tabla 28. Ejemplo de las tuplas ordenadas**

| Hipónimo | Hiperónimo |
|---|---|
| tuberculosis | enfermedad |
| caries | enfermedad |
| obesidad | enfermedad |
| política | profesión |
| enfermería | profesión |
| gripe aviar | enfermedad |
| diabetes | enfermedad |
| medicina | profesión |
| psicología | profesión |
| abogacía | profesión |
| cáncer | enfermedad |
| psoriasis | enfermedad |
| docencia | profesión |
| depresión | enfermedad |
| osteoporosis | enfermedad |
| esquizofrenia | enfermedad |
| lepra | enfermedad |
| anorexia | enfermedad |
| artritis reumatoide | enfermedad |
| epoc | enfermedad |
| ingeniería | profesión |
| epilepsia | enfermedad |
| arquitectura | profesión |
| neumonía | enfermedad |
| educación | profesión |
| homosexualidad | enfermedad |
| prostitución | profesión |
| jaguar | felino |
| adicción | enfermedad |
| malaria | enfermedad |

## 6.6 Sobre el criterio de paro y umbral de corte

Se considera que en cada iteración del *proceso iterativo de estimación de confianzas* la precisión del catálogo incrementará. Sin embargo, es necesario fijar un número de iteraciones máximas, es decir, una condición de paro. Aunque no se realizó un estudio exhaustivo sobre el número de iteraciones adecuadas. Se propone lo siguiente:

Dejar de iterar cuando la diferencia de los promedios de las confianzas de las tuplas ( $prom_{\sigma}$ ) entre iteraciones sucesivas sea menor que cierto umbral. Es decir, el p*roceso iterativo de estimación de confianzas* termina cuando se cumpla la siguiente condición: $|prom_{\sigma}^{k+1} - prom_{\sigma}^{k}| <= \mu$ , donde $k$ es el número de iteración y $\mu$ es una constante que actúa como umbral.

En otro asunto, una vez que el método dejado de iterar se tendrá como resultado el catálogo de hipónimos ordenado, el puede ser aplicado a diversas tareas del procesamiento del lenguaje natural. En algún momento podrían solicitarse únicamente las $m$ tuplas más confiables. El término $m$ representa un **umbral de corte del catálogo**, el cual depende directamente de la aplicación donde será utilizado el catálogo.

## 6.7 Discusión

En este capítulo se presentaron dos enfoques diferentes cuya finalidad es ordenar las tuplas del catálogo. Ambos enfoques tienen la misma estructura y únicamente difieren en las medidas utilizadas para estimar la confianza de los patrones y de las tuplas. El primer enfoque se basa en el uso de

información mutua. Por su parte, el segundo enfoque se basa en el uso de la medida F.

En el primer enfoque (enfoque centrado en el uso de información mutua) se presentaron dos alternativas para estimar inicialmente la confianza de los patrones. La primera alternativa usa información mutua. Por su parte, la segunda alternativa usa una nueva propuesta basada en la medida F. Ésta última generó resultados superiores a la primera alternativa. De hecho, la segunda alternativa es la pieza que motivó el desarrollo del segundo enfoque, en el cual se encuentra la contribución principal de esta tesis. Ahora bien, los resultados obtenidos permiten deducir la importancia de estimar la confianza de los patrones de manera adecuada. Particularmente, se observó que la confianza inicial de los patrones repercute en la precisión final del catálogo. Finalmente, los resultados reflejan la factibilidad de la nueva medida (adaptación de la medida F), de ahí que en el segundo enfoque se haya decido explorar dicha medida.

En el segundo enfoque, enfoque centrado en el uso de la medida F, las variaciones de precisión obtenidas entre iteraciones sucesivas no son tan grandes. Es decir, las tres curvas de precisión de los resultados de cada iteración presentan comportamientos muy similares. Además, se observó que desde la primera iteración se alcanzan resultados favorables. Entonces, se puede decir que el *enfoque centrado en el uso de la medida F* alcanza rápidamente una precisión favorable; posteriormente, el proceso iterativo ayuda a afinar dicha precisión. En contraste, el *enfoque centrado en el uso de información mutua* confía altamente en el uso de iteraciones para obtener resultados comparables con los resultados del *enfoque centrado en el uso de la medida F*.

La figura 18 presenta una comparación de ambos enfoques. Para el enfoque centrado en el uso de información mutua (primer enfoque) se graficó la curva de precisión en la tercera iteración usando información mutua en el *arranque*. Por otro lado, para el enfoque centrado en el uso de la medida F (segundo enfoque) se graficó la curva de precisión en la tercera iteración.

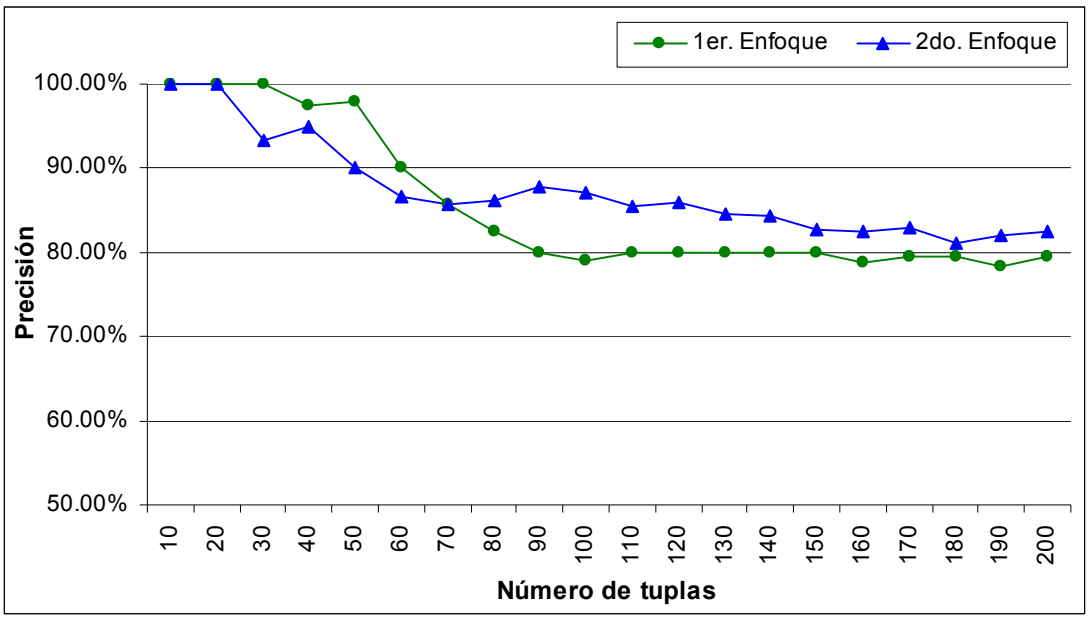


**Figura 18. Comparación de los enfoques de ordenamiento**

Básicamente, en la figura 18 se observa que ambos enfoques alcanzan un 100% para las 20 primeras tuplas del catálogo. Finalmente, para las 200 tuplas evaluadas, el enfoque centrado en el uso de la medida F obtuvo una precisión de 82.50%. En cambio, el enfoque centrado en el uso de información mutua obtuvo una precisión de 79.50%, precisión que obtuvo el enfoque centrado en el uso de la medida F pero en la primera iteración. Entonces, de manera general, se puede deducir que el enfoque centrado en el uso de la medida F obtuvo mejores resultados con un menor número de iteraciones.

Por otro lado, de acuerdo con los experimentos realizados el enfoque centrado en el uso de la medida F genera mejores resultados. Sin embargo, en este enfoque se definieron varios parámetros que utilizan valores constantes (por ejemplo el parámetro $\beta$, el cual corresponde al número de veces que el recuerdo es más relevante que la precisión). Todos esos parámetros podrían afinarse para mejorar el rendimiento del método. Es decir, mediante un estudio más detallado, podrían encontrarse valores más adecuados que permitan mejorar la precisión final del catálogo recuperado.

# Capítulo 7

# Conclusiones y trabajo futuro

## 7.1 Conclusiones

Se presentó un método automático para extraer hipónimos relacionados a un vocabulario predefinido. El método propuesto se basa en el uso de información léxica y prescinde de información morfológica o sintáctica.

El método parte de un conjunto de semillas seleccionadas manualmente, para después descubrir instancias de la relación de hiponimia. Posteriormente se proponen dos enfoques para evaluar las tuplas del

catálogo, de forma que se distingan aquellas tuplas confiables de aquellas que no lo son.

Los resultados demostraron la posibilidad de usar patrones léxicos para extraer hipónimos de texto no estructurado. Por consiguiente, no se requiere del uso de etiquetados o analizadores sintácticos para descubrir los patrones o para aplicarlos.

Se presentaron dos enfoques de ordenamiento de las tuplas del catálogo, cada uno de ellos con características específicas. El primer enfoque tiene como base el uso de información mutua para ordenar las tuplas del catálogo. El segundo enfoque se caracteriza por el uso de la medida F.

El primer enfoque permite observar la relevancia del valor de la confianza inicial de los patrones. Pues dicha confianza puede impactar en la precisión del catálogo. Por otro lado, el uso de iteraciones enriquece la precisión final del catálogo.

Ahora bien, en el primer enfoque se presentaron dos alternativas para asignar a los patrones una confianza inicial. La primera alternativa se basa en el uso de información mutua. Por su parte, la segunda alternativa usa una adaptación a la tradicional medida F. De manera general, en los resultados experimentales se resalta el uso de esta última medida sobre el uso de información mutua. Con esta medida se producen precisiones más altas desde la primera iteración. En cambio, con información mutua la precisión inicia abajo y va aumentando con el número de iteraciones. Prácticamente, cuando el método usa la adaptación propuesta de la medida F tiene una tendencia a alcanzar resultados favorables en menos iteraciones. De ahí nació la inquietud de realizar un segundo enfoque que buscara explorar y aprovechar el uso de la medida F.

El segundo enfoque *centrado en el uso de la medida F* fue diseñado y propuesto para tomar la medida F como la base para estimar la confianza de los patrones y tuplas del catálogo. A partir de los resultados se percibió un comportamiento propicio para desarrollar la tarea, pues desde la primera iteración se obtienen resultados favorables y mejores a los obtenidos con el e*nfoque centrado en el uso de información mutua*. Inclusive, la precisión entre iteración e iteración aumenta; pero no tan drásticamente, ya que a partir la primera iteración se obtienen resultados muy relevantes.

Básicamente, el *enfoque centrado en el uso de información mutua* representa un medio factible para ordenar las tuplas del catálogo. Sin embargo, el enfoque *centrado en el uso de la medida F* obtuvo mejores resultados desde la primera iteración.

## 7.2 Trabajo futuro

Este trabajo de investigación consiguió resultados interesantes. No obstante, existen varios aspectos sobre los cuales se desea trabajar a futuro, entre ellos se tienen:

- Probar la escalabilidad del método. En los experimentos mostrados en esta tesis, se trabajó con un vocabulario formado de cinco términos. Entonces, para probar la escalabilidad del método, se planea realizar algunos experimentos que involucren el uso de un vocabulario que contenga un mayor número términos. Se cree que el método trabajará sin problemas porque al aumentar el tamaño del vocabulario, se recopilará más información y en consecuencia se tendrá más evidencia para determinar la confiabilidad de las tuplas. Es decir, el método puede beneficiarse de la abundancia de la información para determinar la confiabilidad de las tuplas.

- Aumentar el conjunto de patrones. Con un mayor número de patrones, se tendría la capacidad de extraer más información, la cual puede ayudar a determinar la confiabilidad de las tuplas. También se tendría la capacidad de aumentar el número de tuplas en el catálogo. Una forma de aumentar automáticamente el número de patrones consiste en adjuntar un módulo, el cual, después de aplicar el método, tome las tuplas del catálogo más confiables y las utilice como semillas. Con estas semillas el método sería aplicado nuevamente. De esta manera se descubrirán nuevos patrones, los cuales tienen la ventaja de estar especializados al dominio tratado en el vocabulario. Los patrones nuevos pueden ser sumados al conjunto de patrones y como resultado, se aumenta el conjunto de patrones.

- Evaluar el desempeño del método en el descubrimiento de instancias de otras relaciones semánticas. Inicialmente, se propuso un método para extraer hipónimos. Sin embargo, el método propuesto puede ser adaptado para tratar otras relaciones semánticas (por ejemplo meronimia, relación que se da entre las *partes* y los *todos*). Esta acción es posible gracias al uso de semillas para descubrir automáticamente los patrones. Se considera que cambiando las semillas, según la relación semántica a tratar, el método sería capaz de construir un catálogo que contenga instancias de dicha relación.

- Adaptar y evaluar el método para trabajar con otros idiomas. Los resultados mostraron que el método es adecuado tratando con el idioma español. Ahora bien, debido al uso de patrones a nivel léxico, el método puede ser adaptado de manera sencilla para aplicarse a otros idiomas, principalmente idiomas cuya morfología sea similar a la del español. Por esta razón, se planea evaluar el método con otro idioma, por ejemplo, el idioma Inglés.

# Bibliografía


- Ahonen-Myka H. (2002). Discovery of frequent word sequences in text source. *Proceedings of the ESF Exploratory Workshop on Pattern Detection and Discovery.* London, UK. 180-189.

- Alcaraz E. y Martínez M. A. (1997). Diccionario de lingüística moderna. Ariel, S.A., Barcelona, España.

- Alshawi H. (1987). Processing dictionary definitions with phrasal pattern hierarchies. *Computational Linguistics*. 195-202.

- Baroni M. y Bisi S. (2004). Using cooccurrence statistics and the web to discover synonyms in a technical language. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC).*

Lisbon, Portugal. 1725-1728.

- Berland M. y Charniak E. (1999). Finding parts in very large corpora. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. Maryland, USA. 57-64.

- Blohm S. y Cimiano P. (2006). Learning patterns from the web. Evaluating the evaluation functions. *Ontologies in Text Technology: Approaches to Extract Semantic KnowledInformation*. Osnabrück, Germany.

- Brants T.(2000). TnT - A statistical part-of-speech tagger. *Proceedings of the 6th conference on Applied Natural Language Processing*. Seattle, Washington. 224-231.

- Calzolari N. (1984). Detecting patterns in a lexical data base. *Proceedings of the 22nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Stanford, California. 170-173.

- Caraballo S. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD. 120-126.

- Carreras X., Chao I. Padró y Padró M. (2004) FreeLing: An open-source suite of language analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*. Lisbon, Portugal. 239-242.

- Castellón I., Civit M., Atserias J. (1998), Syntactic parsing of unrestricted Spanish text. *1st International Conference on Language Resources and Evaluation*. Granada, España. 603-610.

- Cimiano P. (2006) Ontology learning and population from text, algorithms, evaluation and applications. Springer. New York, USA.

- Cimiano P., Hotho A. y Staab S. (2004). Comparing conceptual, divisive and agglomerative clustering for learning taxonomies from text. *Proceedings of the 16th European Conference on Artificial Intelligence (ECAI)*. 435-439.

- Crystal D. (2000). Diccionario de lingüística y fonética. Octaedro. Barcelona.

- Denicia C., Montes M., Villaseñor L. y García R. (2006). A text mining approach for definition question answering*. 5th International Conference on Natural Language Processing*. Turku, Finland. 76-86.

- Dolan W., Vanderwende, L. y Richardson, S. (1993). Automatically deriving structured knowledge bases form online dictionaries. *Proceedings of the Pacific Association for Computational Linguistic (PACLING).* 5-14.

- Farwell D., Helmreich S y Casper M. (1995). SPOST: a Spanish part-of-speech tagger. *Procesamiento del Lenguaje Natural*. 42-57.

- Fleischman M., Hovy E. y Echihabi A. (2003). Offline strategies for online question answering: answering questions before they are asked. *Proceedings of the 41st Annual Meeting on Association For Computational Linguistics.* Sapporo, Japan. 1-7

- García-Hernández R., Martínez-Trinidad F. y Carrasco-Ochoa A. (2006). A new algorithm for fast discovery of maximal sequential patterns in a document collection*. International Conference on Computational*

*Linguistics and text Processing.* Mexico City, Mexico. 514-523.

- Gellerstam, M. (1995). Lexical resources and their application. *Proceedings of the 1st Trans-European Language Resources Infrastructure (TELRI) Seminar on Language Resources for Language Technology*. Tihany, Hungary. 57-64.

- Girju R., Badulescu A. y Moldovan D. (2003). Learning semantic constraints for the automatic discovery of part-whole relations. *Proceedings of the 2003 Conference of the North American Chapter of the Association For Computational Linguistics on Human Language Technology.* Edmonton, Canada. 1-8.

- Godfrey J. y Zampolli A. (1997). Language resources: overview. *Survey of the State of the Art in Human Language Technology*, R. Cole, Ed. Cambridge Studies. Natural Language Processing Series. Cambridge University Press, New York, NY. 381-384.

- Graham S., Harrison M., y Ruzzo W. (1980). An improved context-free recognizer. *ACM Transactions on Programming Languages and Systems*. 2(3). 415-462.

- Hearst M. (1992). Automatic acquisition of hyponyms from large text corpora. *Proceedings of the 14th International Conference on Computational Linguistics (COLING).* Nantes, France. 539-545.

- Jiménez H. y Morales G. (2002). Sepe: A POS tagger for Spanish. *Proceedings of the 3rd international Conference on Computational Linguistics and intelligent Text Processing*. Mexico City, Mexico. 250-259.

- Lin D., Zhao S., Qin L. y Zhou M. (2003). Identifying synonyms among distributionally similar words. *International Joint Conference of Artificial Intelligence (IJCAI-2003)*. Acapulco, México. 1492-1493.

- Lucero C., Pinto D. y Jiménez H. (2004). A tool for automatic detection of antonymy relations. *Workshop on Herramientas y Recursos Lingüísticos para el Español y el Portugués (IBERAMIA)*. *Proceedings of Workshops on Artificial Intelligence*. Puebla, Mexico. 273-281.

- Mann G.S. (2002). Fine-grained proper noun ontologies for question answering. *International Conference On Computational Linguistics On SemaNet: Building and Using Semantic Networks*. Taipei, Taiwan. 1-7

- Maynard D., Peters W. y Li Y. (2006). Metrics for evaluation of ontology–based information extraction. *Proceedings of the 4th International Workshop on Evaluation of Ontologies for the Web (EON)*, Edinburgh, UK.

- Miller G. (1999). WordNet: An online lexical database. *International Journal of Lexicography*, 3(4). 235-312.

- Mitkov R. (2003). The Oxford handbook of computational linguistics. Oxford University Press Inc. Oxford, NewYork.

- Ortega-Mendoza R., Villaseñor-Pineda L. y Montes-y-Gómez M. (2007). Using lexical patterns for extracting hyponyms from the Web. MICAI 2007, Ags., México.

- Pantel P. y Pennacchiotti M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. *Proceedings of Conference on Computational Linguistics / Association for Computational*

*Linguistics (COLING/ACL)*. Sydney, Australia. 113-120.

- Pantel P. y Ravichandran D. (2004). Automatically labeling semantic classes. *Proceedings (HLT-NAACL).* Boston, Massachusetts, USA. 21-328.

- Pantel P., Ravichandran D. y Hovy E. (2004). Towards terascale knowledge acquisition. *Proceedings of the International Conference on Computational Linguistics*, Geneva, Switzerland. 771–777.

- Pasca M. (2004). Acquisition of categorized named entities for Web search*. Proceedings of the 13th ACM international conference on Information and knowledge management.* Washington, D.C, USA. 137-145.

- Pereira F., Tishby N. y Lee L. (1993). Distributional clustering of english words. *Proceedings of the 31st Annual Meeting of the Association for Computational Lingistics (ACL)*. 183-190.

- Ravichandran D. y Hovy E. (2002). Learning surface text patterns for a question answering system. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, PA. 41-47.

- Ravichandran D., Pantel P. y Hovy E. (2004). The terascale challenge. *Proceedings of KDD Workshop on Mining for and from the Semantic Web*. Seattle, WA. 1-11.

- Riloff E. y Shepherd J. (1997). A Corpus-based approach for building semantic lexicons. *Proceedings of the 2nd Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 117-124.

- Roark B. y Charniak E. (1998). Noun-phrase co-occurrence statistics for semi-automatic semantic lexicon construction. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*. 1110-1116.

- Schmid H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*. Manchester, UK. 44-49.

- Schwab D., Lafourcade M., y Prince V. (2002). Antonymy and conceptual vectors. *Proceedings of Computational Linguistics*. Taipei, Taiwan. 904-910.

- Sinclair J. (1991). *Corpus, concordance, collocation.* Oxford University Press, Oxford, New York.

- Sundblad H. (2002). Automatic acquisition of hyponyms and meronyms from question corpora. *Proceedings from the Workshop on Natural Language Processing and Machine Learning for Ontology Engineering at ECAI2002*. Lyon, France.

- Turney P. D. (2001). Mining the Web for synonyms: PMI-IR versus LSA on TOEFL. P*roceedings of the 12th European Conference on Machine Learning*. Freiburg, Germany. 491-502.

- Tustison C.A (2004). Logical form identification for medical clinical trials, Master's thesis. Department of Linguistics and English Language. Brigham Young University.

- van Hage W. R. , Kolb H. P. y Schreiber A. Th. A method for learning

part-whole relations. *Proceedings of the 5th International Semantic Web Conference (ISWC)*. Athens, GA, USA. 723-735.

# Apéndice A

En este apéndice se muestra la lista de patrones descubiertos. Esta lista se muestra en la siguiente página.

**Tabla 29. Lista de patrones descubiertos**

| No. | Patrón |
|---|---|
| 1 | el <hipónimo> es el único <hiperónimo> |
| 2 | el uso de la <hipónimo> como <hiperónimo> |
| 3 | el <hipónimo> es uno de los <hiperónimo> más |
| 4 | de la <hipónimo> como <hiperónimo> de |
| 5 | de las <hipónimo> como <hiperónimo> |
| 6 | las <hipónimo> son una <hiperónimo> |
| 7 | el <hipónimo> es un <hiperónimo> que |
| 8 | el <hipónimo> es el <hiperónimo> que |
| 9 | de <hiperónimo> como <hipónimo> y |
| 10 | la <hipónimo> es un <hiperónimo> |
| 11 | la <hipónimo> una <hiperónimo> |
| 12 | las <hipónimo> son <hiperónimo> que |
| 13 | el <hipónimo> es un <hiperónimo> de |
| 14 | la <hipónimo> es la <hiperónimo> |
| 15 | la <hipónimo> es una <hiperónimo> que |
| 16 | la <hipónimo> como una <hiperónimo> |
| 17 | que la <hipónimo> es una <hiperónimo> |
| 18 | el <hipónimo> es una <hiperónimo> |
| 19 | la <hipónimo> es el <hiperónimo> de |
| 20 | de <hipónimo> y otras <hiperónimo> |
| 21 | del <hipónimo> como <hiperónimo> |
| 22 | el <hipónimo> es la <hiperónimo> |
| 23 | <hiperónimo> de <hipónimo> de |
| 24 | de <hipónimo> y <hiperónimo> |
| 25 | <hiperónimo> de <hipónimo> y |
| 26 | de <hipónimo> o <hiperónimo> |
| 27 | los <hipónimo> son <hiperónimo> |
| 28 | de <hipónimo> como <hiperónimo> de |
| 29 | el <hipónimo> y las <hiperónimo> |
| 30 | de los <hipónimo> y <hiperónimo> |
| 31 | de los <hipónimo> y los <hiperónimo> |
| 32 | la <hipónimo> es el único <hiperónimo> natural |
| 33 | <hiperónimo> de la actividad <hipónimo> y el deporte |
| 34 | la anorexia y la <hipónimo> son <hiperónimo> |
| 35 | de <hipónimo> y otros <hiperónimo> |
| 36 | el <hipónimo> es el <hiperónimo> de mayor longevidad |
| 37 | los <hipónimo> y otros <hiperónimo> |
| 38 | facultad de <hiperónimo> de la actividad <hipónimo> y |
| 39 | la <hipónimo> y otros <hiperónimo> |
| 40 | las <hipónimo> marinas son <hiperónimo> |
| 41 | el <hipónimo> es el <hiperónimo> interno más |
| 42 | licenciado en <hiperónimo> de la actividad <hipónimo> y del deporte |
| 43 | el <hipónimo> es el <hiperónimo> más grande del cuerpo |