UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO

UNIDAD ACADÉMICA PROFESIONAL TIANGUISTENCO

DETECCIÓN DE FRAGMENTOS DE TEXTO COMO CANDIDATO A HIPERVÍNCULO

TESIS

QUE PARA OBTENER EL GRADO DE

MAESTRA EN CIENCIAS DE LA COMPUTACIÓN

PRESENTA:

MARCELA CAMACHO AVILA

TUTOR ACADÉMICO:

DR. RENÉ ARNULFO GARCÍA HERNÁNDEZ

TUTORES ADJUNTOS:

DRA. YULIA NIKOLAEVNA LEDENEVA

DR. JOSÉ LUIS TAPIA FABELA

TIANGUISTENCO, ESTADO DE MÉXICO ENERO 2015

UAEM | Universidad Autónoma del Estado de México

DICTÁMEN DE AUTORIZACIÓN Y OBTENCIÓN DE GRADO DE MAESTRÍA

Tianguistenco, Méx. , a 9 de enero de 2015

Título del proyecto:

Detección de fragmentos de texto como candidato a hipervínculo.

Tesista:

Lic. Marcela Camacho Avila

Dictamen: **No. de revisión:** 9

- ☐ **Rechazado**
- ☐ **Sujeto a modificaciones**
- ☐ **Aceptado, condicionado**
- ■ **Aceptado**

UAP TIANGUISTENCO

COORDINACIÓN DE LA MAESTRÍA EN CIENCIAS DE LA COMPUTACIÓN

Observaciones generales:

Aceptado para la impresión
Aceptado para la defensa de grado

Tutor Adjunto	Tutor Académico	Tutor Adjunto
Dra. Yulia Nikolaevna Ledeneva	**Dr. René Arnulfo García Hernández**	**Dr. José Luis Tapia Fabela**

MCPR2015 submission 9
MCPR2015 [mcpr2015@easychair.org]

Enviado el: miércoles, 21 de enero de 2015 09:39 a.m.
Para: Marcela Camacho Avila

Dear authors,

We have successfully received your paper:

  Authors : Marcela Camacho Avila, René Arnulfo García Hernández and Yulia Nikolaevna Ledeneva
  Title   : Detection of fragments of text and hyperlink candidate
  Number : 9

The paper was submitted by Marcela Camacho <mcamachoa@uaemex.mx>.

Thank you for submitting to MCPR2015.

Best regards,
EasyChair for MCPR2015.

Agradecimientos

Un cordial agradecimiento a mi asesor Dr. René Arnulfo García Hernández quien con su conocimiento, experiencia y sobre todo por su gran paciencia para llevar a buen término este trabajo.

A la Universidad Autónoma del Estado de México y en particular a la Unidad Académica Profesional de Tianguistenco por todas las facilidades prestadas durante mi estancia académica.

A mis profesores y revisores por su conocimiento, experiencia y amistad.

A mis compañeros de maestría por su entusiasmo y apoyo.

Finalmente se agradece al Consejo Nacional de Ciencia y Tecnología (CONACYT) por el apoyo otorgado a través de la beca para estudios de maestría número 480912

Dedicatorias

A Dios por la oportunidad de vivir esta gran experiencia.

A mis hermanas por todo el apoyo que me han brindado desde siempre.

A Esteban y Carolina por todo el amor, apoyo y comprensión que siempre he recibido de su parte.

Resumen

El enriquecimiento de información en un documento ha permitido comunicar de mejor manera el mensaje que un autor desea expresar. En el caso de documentos electrónicos, el enriquecimiento de información se ha dado al incorporar al texto formatos, imágenes, audios, videos e hipervínculos hacia otros documentos. En particular, la hipervinculación de documentos electrónicos ha generado la WEB, una red de documentos relacionados entre sí, en la cual es posible navegar entre ellos de acuerdo a la necesidad de información del lector de manera que al elegir un hipervínculo se pueda ampliar la información sobre éste. La hipervinculación de documentos es una tarea de edición e investigación que debe hacer el autor de manera manual para incorporar tal característica a su documento. Normalmente, los hipervínculos de un documento se dirigen hacia documentos de la misma colección al cual pertenece el documento, puesto que otras colecciones al estar en otros sitios pueden cambiar o eliminar su dirección electrónica lo cual desvincularía al documento.

La hipervinculación de documentos en la WEB es una tarea dinámica de búsqueda y edición por parte del autor puesto que depende de los documentos contenidos en la colección. Por lo que al agregar o eliminar un documento de una colección se podría afectar los hipervínculos de un documento. Hoy en día es posible ver colecciones de documentos WEB fuertemente hipervinculadas como lo es Wikipedia, pero también hay ejemplos colecciones débilmente o nulamente hipervinculadas como las noticias. En el caso de Wikipedia se encontró que cada documento tiene 49 hipervínculos en promedio, es decir para 20

documentos habría 980 hipervínculos. En este sentido, según la agencia de noticias Notimex genera 200 noticias por día en promedio y si a cada noticia se le construyeran 49 hipervínculos se tendrían que generar 9800 hipervínculos por día; lo cual es prácticamente imposible de manera manual.

Como un paso previo a la hipervinculación automática, en este trabajo se investigó si hay patrones en el texto que el humano sigue para hacer un hipervínculo. Para la experimentación se utilizaron 10,000 documentos seleccionados aleatoriamente de la colección Wikipedia 2008 en español. De acuerdo a la experimentación, es posible ver que hay patrones valiosos ya que por un lado se repiten frecuentemente y por otro lado aunque son pocos están presentes en varios experimentos, alcanzando un F-measure de 51%.

Abstract

Enriching information in a document has allowed communicating the message of an author in better way. In the case of electronic documents, enriching information was given to incorporate text formats, images, audios, videos and hyperlinks to other documents. In particular, hyperlinked electronic documents have generated the WEB, a network of interrelated documents, in which is possible to navigate between them according to the information needs of the reader in the manner that when choosing a hyperlink a piece of information could be expanded. Document hyperlinking is a task of editing and searching that the author should do manually to add this feature for such documents. Normally, hyperlinks in a document addressed to documents from the same collection to which the document belongs, since other collections to be in other sites can change or remove its electronic address which dissociate the document.

The hyperlinking of documents on the web is a dynamic task of searching and editing by the author since it depends on the documents in the collection. Adding or removing a document of a collection could affect the hyperlinks in a document. Today, in the WEB it is possible to see collections strongly hyperlinked as is Wikipedia, but there are also examples of collections weakly or with none hyperlinked as news. As part of this this research, we found that each document of Wikipedia has an average 49 hyperlinks, i.e. for 20 documents there were 980 hyperlinks. In this sense, according to Notimex news agency generates an average of 200 news per day, and if for each new were have to be built 49 hyperlinks then 9800 hyperlinks would have to be generated per day; which is practically impossible manually.

As a previous step of the automatic hyperlinking, in this thesis we investigated whether there are patterns in the text that human follows for doing a hyperlink. For experimentation, 10,000 randomly selected documents in Spanish Wikipedia 2008 collection were used. According to experimentation, we can see that there are valuable patterns because for on one side these patterns are frequently repeated and, for the other side, although are few, these patterns are present in several experiments; reaching an F-measure of 51%.

Contenido

Figuras

Tablas

Capítulo 1

Antecedentes

El lenguaje natural se fue desarrollando de acuerdo a sus propias reglas, en cada época se fue adecuando según las necesidades de la comunicación humana. Con el surgimiento de la escritura en papel, la transmisión de ideas fue más rápida y eficiente. Por siglos el lenguaje escrito ha sido una forma importante de comunicación que también permite transmitir conocimiento sin la necesidad de hacerlo personan a persona. El modo más natural de comunicación para un ser humano es hablar y escuchar, no escribir y leer. Tenemos que escribir y leer porque de esa forma realizamos las tareas principales de procesamiento de información: búsqueda y comparación. [Gelbukh.2010]

Durante la historia de la humanidad, la mayor parte del conocimiento se ha comunicado, guardado y manejado en la forma de lenguaje natural. En su forma escrita el lenguaje se encuentra en documentos, libros o artículos y con la aparición de la computadora, en forma

electrónica, o sea, digital. En este sentido, las computadoras se han convertido en una ayuda enorme para el procesamiento del conocimiento. Sin embargo, lo que es conocimiento para nosotros —los seres humanos— no lo es para las computadoras. Para ellas son sólo archivos, secuencias de caracteres y nada más. Una computadora puede copiar un texto electrónico, respaldarlo, transmitirlo, borrarlo; pero no puede buscar las respuestas a las preguntas en el texto, ni hacer inferencias lógicas sobre su contenido, ni generalizar, ni resumir —es decir, hacer todo lo que las personas normalmente hacemos con el texto. La computadora depende completamente del ser humano y de lo que uno ponga en ella. Es por eso que resulta muy complicado enseñarle el lenguaje a una computadora, ya que requiere de esfuerzos enormes.

El área que se encarga de habilitar a las computadoras para entender el texto, en función del enfoque práctico o teórico, tiene varios nombres: *procesamiento de lenguaje natural, procesamiento de texto, tecnologías de lenguaje o lingüística computacional* [Gelbukh.2010]. El área del procesamiento del lenguaje natural trata de procesar el texto por su sentido y no como un archivo binario. La aplicación del procesamiento de lenguaje natural más obvia y quizá más importante en la actualidad, es la búsqueda de información (llamada también recuperación de información).

La WEB y en particular las bibliotecas digitales contienen una cantidad enorme de conocimiento que puede dar respuestas a muchísimas preguntas que tenemos, por lo que es primordial recuperar la información más importante de acuerdo a la consulta expresada por el usuario. Las técnicas más usadas actualmente para la recuperación de información implican la búsqueda por palabras clave: se buscan los textos que contienen las palabras que el usuario teclea. Es decir, la representación formal usada es el conjunto de las cadenas de letras (palabras), usualmente junto con sus frecuencias en el texto (número de ocurrencias). Cuando varios fragmentos de textos están relacionados pueden constituir un hecho o evento el cual difícilmente puede ser recuperado con las técnicas de recuperación

de información, puesto que descomponen el texto en palabras aisladas. Para tal búsqueda se utilizan las técnicas de extracción de información.

Los textos electrónicos pueden ser asociados mediante un vínculo electrónico o hipervínculo, tal como se hace en la construcción de diccionarios, donde las palabras son descritas haciendo uso de otras palabras, generando así una red de palabras unidas por un vínculo de referencia.

Un *hipervínculo* es un punto dentro de un documento de hipertexto el cual vincula a otro documento [RFC1983], que consiste de una o más palabras diferenciadas por un formato diferente al resto del párrafo y, que al dar clic sobre ella con el ratón, permite navegar a un documento diferente que amplía la información de las palabras del hipervínculo [WikiEs.2013], [Scott.2013]. Un hipervínculo es expresado con una etiqueta de ancla la cual llama al documento destino usando un localizador uniforme de recursos (URL) [Chakrabarti.2003] [WikiEs.2013]. Por lo que entonces al construir un hipervínculo se deben identificar tanto el ancla (una o más palabras) como el documento destino con el fin de construir una red de conocimiento [Wei.2007].

El construir hipervínculos es una manera de enriquecimiento de texto. Otras maneras son: el tamaño del texto, el color del texto, etc. [RFC1896] y a su vez es una manera de enriquecer colecciones de documentos.

A su vez, un *hipertexto* es un documento el cual contiene hipervínculos a otros documentos [RFC1983]. Los hipertextos son escritos en diferentes lenguajes, por ejemplo HTML, XML o GXML.

El hipervínculo, y por consecuencia el hipertexto, dan origen a la *World Wide Web* (Red informática mundial) mejor conocida como la *WEB* o la *WWW* (por sus siglas en ingles), siendo la WEB un sistema distribuido de información basada en hipertexto [RFC1983] y siendo el navegar la acción que se da sobre ella. Esto es, navegamos de un hipertexto a otro

en la WEB por medio del hipervínculo, entrando a diferentes colecciones de documentos (hipervínculos externos) o a la misma colección (hipervínculos internos).

La WEB revolucionó entre otras cosas, el hábito de lectura de periódicos y la consulta en enciclopedias de conocimiento, ya que ahora es más frecuente que los lectores consulten las ediciones digitales de periódicos y enciclopedias. En la actualidad, los sitios WEB que contienen los hipertextos, han aumentado exponencialmente, originando que los textos electrónicos aumenten de igual manera. La WEB nos permite navegar entre estos textos gracias al hipervínculo, generando así el mayor repositorio de documentos escritos. En la figura 1.1 se puede apreciar una gráfica que describe el crecimiento de enero del 2001 con 27,539,210 sitios WEB a enero del 2015 con 876,812,666 sitios WEB, los datos fueron tomados de la página WEB Netcrafr [Netcrafr.2015].

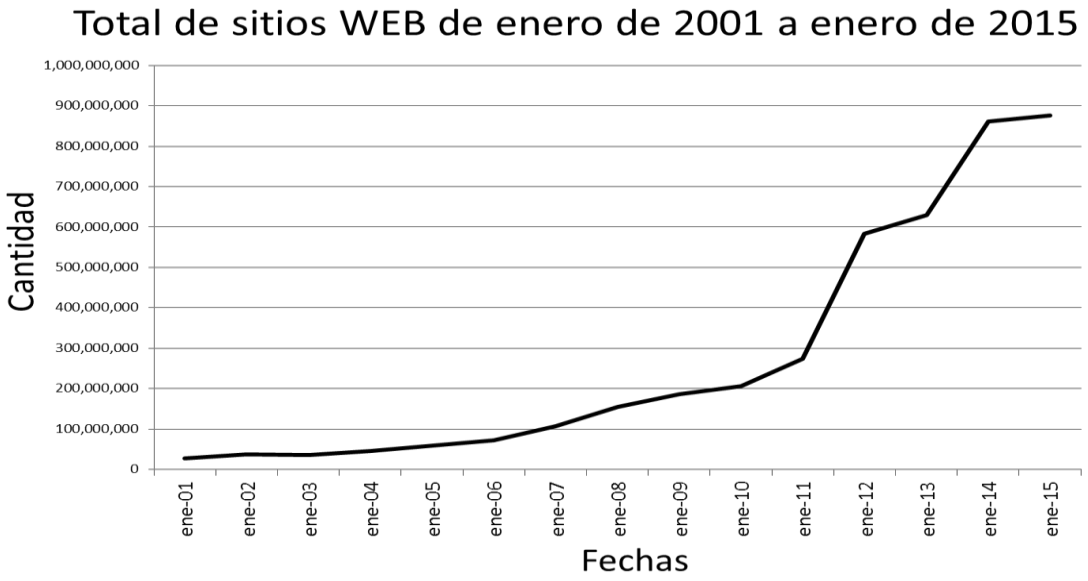

Figura 1. 1 Total de sitios WEB de enero de 2001 a enero de 2015 [Netcrafr.2015].

Por ejemplo, una colección de documentos que crece de manera acelerada es la enciclopedia Wikipedia, que es la enciclopedia digital más grande existente en la WEB. Está escrita en 287 idiomas, cuenta con más de 34 millones de artículos. En la tabla 1.1 podemos visualizar los 20 idiomas con mayor número de artículos de Wikipedia. Por ejemplo, el idioma inglés consta de más de 4 millones de artículos, siendo la edición más grande. Por su parte, la edición en español contiene más de un millón de artículos, ocupando el décimo sitio, esto a fecha de 15 enero de 2015.

Tabla 1. 1 20 lenguajes de las ediciones de Wikipedia [WikiEn.2015]

Las 20 ediciones mayores de Wikipedia

IDIOMA	NÚMERO DE ARTÍCULOS	PORCENTAJE
Inglés	4,695,719	13.7
Sueco	1,953,020	5.7
Holandés	1,807,040	5.3
Alemán	1,800,093	5.3
Francés	1,580,794	4.6
Waray-waray	1,259,032	3.7
Cebuano	1,208,485	3.5
Ruso	1,179,537	3.4
Italiano	1,169,427	3.4
Español	1,150,301	3.4
Vietnamita	1,111,945	3.2
Polaco	1,086,825	3.2
Japonés	942,482	2.8
Portugués	862,108	2.5
Chino	806,228	2.4
Ucranio	547,875	1.6
Catalán	447,426	1.3
Persa	440,940	1.3
Noruego	404,537	1.2
Finlandés	364,214	1.1
Otros	9,398,307	27.5
Total	**34,216,335**	**100.0**

En particular, Wikipedia en español tiene 13 años en la WEB y cuenta con más de 1 millón de artículos y más de 3 millones de usuarios. Para lograr esto, cuenta con 520 colaboradores frecuentes, que son los que realizan 100 artículos por mes.

Haciendo una revisión de algunos de los artículos de Wikipedia en español, del año 2008, se encontró que de una colección de 10 mil artículos seleccionados aleatoriamente se localizaron alrededor de 493 mil hipervínculos (en promedio 49 hipervínculos por documento), todos ellos realizados por sus colaboradores.

En específico, Wikipedia trabaja con un hipervínculo de color rojo (no en color azul como normalmente lo hacen sus colaboradores), que es aquel que hace referencia a un documento que nos informa que no hay artículo que amplié la información, ya sea porque se ha borrado el documento destino o porque no se ha construido, invitando a los usuarios a realizar dicho documento. Lo que nos deja ver que hay un problema en Wikipedia en español, porque aún con 520 colaboradores y sus 13 años en la WEB, no puede construir todos los hipervínculos necesarios en sus artículos.

Los periódicos digitales no se salvan de este problema, puesto que sus noticias no contienen hipervínculos entre ellas. Por ejemplo, uno de los periódicos digitales más visitados en la WEB, es el periódico MILENIO [Merca20.2013], el cual genera alrededor de 200 noticias por día, que es el doble de los documentos que Wikipedia genera en un mes. Para dimensionar este problema se puede utilizar el número de hipervínculos que hay por artículo en Wikipedia y multiplicarlo por el número de noticias que MILENIO genera. Es decir, se tendrían que generar alrededor de 9,800 hipervínculos por un sólo día de noticias del periódico MILENIO, lo cual se ve prácticamente imposible de hacer con pocos humanos de manera manual.

La construcción de hipervínculos es un proceso muy complicado para los autores de los hipertextos. Quizá por ello los autores de noticias no construyen hipervínculos, ya que es imposible saber de la existencia de todos los documentos escritos con anterioridad en su colección. De igual modo, los colaboradores de Wikipedia pueden tener dificultad de

identificar qué documentos vincular con su documento y por ello quizá exista el hipervínculo rojo o quizá por ello hay documentos con pocos o sin hipervínculos.

El costo para que los hipervínculos estén al día es muy alto, los autores no pueden estar al tanto de todos los hipervínculos posibles para su documento. En otras palabras, la construcción de hipervínculos es un problema en la colección dinámica porque los hipervínculos dependen de los textos disponibles.

Para eliminar el esfuerzo humano necesario para construir hipervínculos correctos, para reducir la posibilidad de hipervínculos erróneos, y para mantener hipervínculos al día; se necesitan mecanismos de construcción de hipervínculos de manera automática.

Recordemos que según la definición de hipervínculo, este consiste de una o más palabras que al darles clic llama a un documento destino; pero no hay una definición lingüista de cómo es un hipervínculo. Por esta razón, en la construcción automática de hipervínculos, primero debemos encontrar la manera de identificar esas palabras que pudieran tener un hipervínculo hacia otro documento. Algunos investigadores manejan la premisa de que esas palabras son entidades, pero no hay un estudio de ello.

Con las técnicas del procesamiento automático, se ha trabajado el problema de la construcción de hipervínculos en una colección de textos, este fue el caso del foro INEX 2007[Wei.2007], cuyo objetivo fue la construcción de hipervínculos en Wikipedia. Algunos trabajos en este foro parten de la premisa de que, las palabras donde se construyen los hipervínculos son conceptos, nombres de personas, nombres geográficos o de instituciones, esto es entidades. Otros parten de las palabras contenidas en los títulos, haciendo un catálogo de palabras. De tal forma que no se ha hecho un estudio formal acerca de las palabras que contienen un hipervínculo sin partir de una premisa.

Dado que se tiene una gran colección de documentos en la WEB (Wikipedia) y dado que se quiere identificar cómo son esas palabras a las que los autores les construyen un hipervínculo, se pretende identificar los patrones que los autores siguen en la construcción

de hipervínculos. Existen diferentes trabajos de investigación para identificar patrones que el humano sigue al escribir. Por ejemplo el trabajo de investigación referente a la construcción de métodos para la extracción de información basados en patrones léxicos [Orta.2008] donde se identifican patrones léxicos incorporando algunos mecanismos que facilitan la selección y el etiquetamiento manual de patrones de extracción (obtenidos con la técnica de secuencias frecuentes maximales). Otro ejemplo en la extracción de patrones es la tarea de responder a preguntas de definición, en particular existe un trabajo de investigación [Denicia.2007] que lo hace mediante el descubrimiento de patrones léxicos con la técnica de secuencias frecuentes maximales. Sin embargo no se han aplicado en la extracción de patrones para la construcción de hipervínculos.

1.1 Planteamiento del problema

Este trabajo de investigación consiste en identificar si hay patrones en el texto que el humano sigue para la construcción de hipervínculos y qué tan valiosos serían para su aplicación en la construcción de hipervínculos en una colección de documentos dada y bajo un dominio.

1.2 Hipótesis

Es posible que a partir de la detección de patrones léxicos en los hipervínculos creados por el humano, se puedan detectar fragmentos de texto candidato a hipervínculo en texto plano, replicando los patrones léxicos encontrados.

1.3 Delimitación del problema

En este trabajo de tesis:

- No se pretende analizar o enriquecer toda la WEB, ya que sería un proceso muy costoso y complicado, por lo cual se iniciará con el estudio de pequeñas colecciones.
- Sólo se analizará un dominio específico, que es Wikipedia 2008.
- No se va a garantizar que los resultados se puedan aplicar a otros dominios o lenguajes.

1.4 Objetivos de la tesis

El objetivo de este trabajo es construir un método para detectar fragmentos de texto como candidatos a hipervínculo en una colección de documentos.

Como objetivos específicos se busca:

- Identificar qué características tienen los hipervínculos en una colección de documentos.
- Identificar si los autores de hipertexto siguen algún patrón para detectar fragmentos de textos que puedan tener hipervínculo.
- Mostrar si el contexto ayuda a determinar que fragmento de texto puede ser un hipervínculo.
- Determinar a qué fragmentos de texto se les puede genera un hipervínculo.

1.5 Estructura de la tesis

El contenido del documento se detalla a continuación:

En el capítulo 2 se presentan los conceptos que introducen al lector dentro del contexto de este trabajo de investigación. Primero los relacionados al descubrimiento de conocimiento en bases de datos (KDD) y las fases que lo componen; haciendo énfasis en la fase de Minería de datos. Después revisaremos los conceptos relacionados con el descubrimiento de conocimiento en texto (KDT) y su relación con la Minería WEB y las Secuencias Frecuentes Maximales. Finalizando con las medidas de evaluación que nos permitirán hacer un análisis de los resultados obtenidos en el presente trabajo.

En el capítulo 3 se describen algunos trabajos que abordan el tema de la construcción automática de hipervínculos, para después analizar algunos trabajos que abordan el descubrimiento de conocimiento en texto; todos ellos relacionados con este trabajo de investigación.

En el capítulo 4 se describe de manera general y especifica el método que se está proponiendo para la detección de fragmentos de texto como candidato a hipervínculo, basado en el proceso KDT.

En el capítulo 5 se presentan varios experimentos realizados con diferentes colecciones de documentos tomadas de Wikipedia en español 2008.

Finalmente en el capítulo 6 se exponen las conclusiones y el trabajo futuro que se desprende de este trabajo de tesis.

Capítulo 2

Marco teórico

En este capítulo se introducen los conceptos básicos para que el lector pueda familiarizarse con el presente trabajo de investigación. En la sección 2.1, se muestra la definición del proceso de descubrimiento de conocimiento en bases de datos (KDD), así como de las fases que lo componen, ya que es el modelo a seguir para descubrir conocimiento en el presente trabajo. En la sección 2.2 basándose en el proceso KDD y en la definición del proceso de descubrimiento de conocimiento en texto, se propone un modelo KDT. Además se introduce la definición de minería web de contenido ya que es la tarea que se va utilizar para cumplir los objetivos. Por último, en la sección 2.3 se introduce la definición de secuencias frecuentes maximales ya que es la técnica de minería a utilizar.

2.1 Descubrimiento de conocimiento en bases de datos (KDD)

En un inicio las bases de datos tenían como propósito principal guardar y organizar la información de una empresa o institución. Sin embargo, los datos guardados ahí representaban un conocimiento sobre la empresa o institución, por lo que se empezó a explotar esos datos primero con fines estadísticos. No obstante había diferentes formas y formatos de la información, por lo que surgieron así diferentes herramientas y técnicas para analizar datos y extraer conocimiento útil desde la información disponible. Estas herramientas y técnicas son utilizadas en lo que hoy se conoce como el "Proceso de descubrimiento de conocimiento en bases de datos" (*Knowledge Discovery in Databases*, KDD).

Se define el *KDD* como: el proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y comprensibles a partir de los datos [Hernández.2007]. El proceso KDD realiza la selección, limpieza y transformación de los datos con el objetivo de poder analizarlos para extraer patrones y modelos adecuados; para posteriormente evaluar e interpretar esos patrones para producir finalmente conocimiento. De esta manera, el conocimiento puede utilizarse para resolver problemas.

KDD es un proceso iterativo ya que la salida de alguna de las fases puede volver a pasos anteriores y a menudo son necesarias varias iteraciones para extraer conocimiento de calidad. Es interactivo porque el experto en el dominio del problema debe ayudar en la preparación de los datos y en la validación del conocimiento extraído [Fayyad.1996].

En la figura 2.1 se muestra el proceso KDD, organizado en cinco fases con un flujo descendente, el cual no limita a regresar a cualquiera de las fases anteriores para reconsiderar o hacer cambios en los datos o en los procesos propios de cada fase. Por ejemplo, una vez terminada la fase de evaluación es posible regresar a la fase de integración

y recopilación si es que se detectan fallas en el momento de integrar los datos. Así mismo, una vez realizada la fase de minería de datos es posible regresar a la fase de selección, limpieza y transformación al detectar que hay fallas en la limpieza. Todo ello dependerá del objetivo de utilizar los procesos.

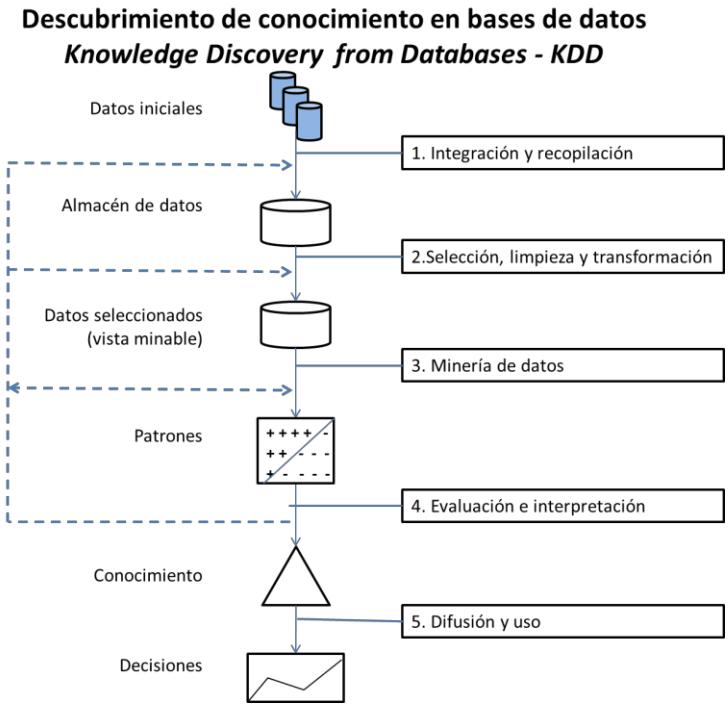

Figura 2. 1 Fases del proceso de descubrimiento de conocimiento en bases de datos, KDD [Hernández.2007].

A continuación se describe en qué consiste cada fase.

2.1.1.-Fase de integración y recopilación

En esta fase se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas, así como salvar todos los obstáculos que se presenten para la obtención de los datos, como pueden ser los formatos o dispositivos en que se encuentren almacenados los datos, las restricciones de acceso a los datos y hasta la capacitación que tenga el usuario de los datos. [Hernández.2007][Witten.2011].

2.1.2.-Fase de selección, limpieza y transformación

A continuación se transforman todos los datos a un formato común, frecuentemente mediante un almacén de datos que consiga unificar de manera operativa toda la información recogida, detectando y resolviendo las inconsistencias [Hernández.2007] [Witten.2011].

En esta fase se eliminan o se corrigen los datos incorrectos y se decide la estrategia a seguir con los datos incompletos. Además se proyectan los datos para considerar únicamente aquellas variables relevantes, con el objetivo de hacer más fácil la tarea de Minería y para que los resultados de la misma sean más útiles [Hernández.2007] [Witten.2011].

2.1.3.-Fase de minería de datos

La *minería de datos* se define como el proceso de extraer conocimiento útil y comprensible, previamente desconocido, desde grandes cantidades de datos almacenados en distintos formatos [Hernández.2007]. Es decir, la tarea fundamental de la minería de datos es el descubrimiento y extracción de patrones (modelos) inteligibles a partir de datos. Para que este proceso sea efectivo debería ser automático o semiautomático (asistido) y el uso de los

patrones descubiertos debería ayudar a tomar decisiones más seguras que reporten, por tanto, algún beneficio a la organización [Witten.2011].

El objetivo de esta fase es producir nuevo conocimiento que pueda utilizar el usuario. Esto se realiza construyendo un modelo basado en los datos recopilados para este efecto. El modelo es una descripción de los patrones y relaciones entre los datos que pueden usarse para hacer predicciones, para entender mejor los datos o para explicar situaciones pasadas. Para ello es necesario tomar una serie de decisiones [Hernández.2007] [Witten.2011].

- Determinar qué tipo de tarea de minería es el más apropiado
- Elegir el tipo de técnica de minería a utilizar que se adapte mejor a la tarea de minería y al modelo que deseamos obtener.
- Elegir el algoritmo de minería que resuelva la tarea y obtenga el tipo de modelo que estamos buscando.

Dentro de las *tareas de la minería de datos* tenemos, por ejemplo, la clasificación, la regresión, el agrupamiento, las correlaciones y las reglas de asociación [Hernández.2007] [Fayyad.1996].

Un caso especial de las reglas de asociación son las *reglas de asociación secuencial*, que se usa para determinar patrones secuenciales en los datos. Estos patrones se basan en secuencias temporales de acciones y difieren de las reglas de asociación en que las relaciones entre datos se basan en el tiempo [Hernández.2007].

Dado que la minería de datos es un campo interdisciplinario [Fayyad.1996], existen diferentes técnicas utilizadas para esta fase: técnicas de inferencia estadística, árboles de decisión, redes neuronales, inducción de reglas, algoritmos genéticos, aprendizaje bayesiano, programación lógica inductiva, entre otros. Cada uno de estos incluyen diferentes algoritmos y variaciones de los mismos, así como otro tipo de restricciones que hacen que la efectividad del algoritmo dependa del dominio de aplicación, no existiendo un método universal para todo tipo de aplicación [Hernández.2007].

La minería de datos se distingue porque no obtiene información extensional (datos) sino intencional (conocimiento) [Hernández.2007].

Por lo tanto, dos son los retos de la minería de datos. Por un lado, trabajar con grandes cantidades de datos, procedentes mayoritariamente de sistemas de información, con los problemas que ello conlleva (ruido, datos ausentes, intratabilidad, volatilidad de los datos…). Por otro lado usar técnicas adecuadas para analizar los mismos y extraer conocimiento novedoso y útil. En muchos casos la utilidad del conocimiento minado esta intimamente relacionado con la comprensibilidad del modelo inferido. No debemos olvidar que, generalmente, el usuario final no tiene por que ser un experto en las técnicas de minería de datos, ni tampoco puede perder mucho tiempo interpretando los resultados. Por ello es importante hacer que la información descubierta sea comprensible para el usuario final.

La relación entre KDD y minería de datos: Kdd es el proceso global de descubrir conocimiento útil desde las bases de datos mientras que la minería de datos se refiere a la aplicación de los métodos de aprendizaje y estadísticos para la obtención de patrones y modelos. Esto es, la minería de datos es una fase del Proceso KDD [Fayyad.1996].

2.1.4.-Fase de evaluación e interpretación

Se evalúan y analizan los patrones y si es necesario se vuelve a fases anteriores. Medir la calidad de los patrones descubiertos por un algoritmo de minería de datos no es un problema trivial, ya que esta medida puede atañer a varios criterios, algunos de ellos bastante subjetivos. Idealmente los patrones descubiertos deben tener tres cualidades: precisos, comprensibles e interesantes.

Para entrenar y probar el modelo se parten los datos en dos conjuntos: el conjunto de entrenamiento y el conjunto de prueba. Esta separación es necesaria para garantizar que la validación de la precisión del modelo es una medida independiente. Si no se usan conjuntos diferentes de entrenamiento y prueba, la precisión del modelo será sobreestimada, es decir, tendremos estimaciones muy optimistas.

En los modelos predictivos, el uso de esta separación entre entrenamiento y prueba es fácil de interpretar. Por ejemplo, para una tarea de clasificación, después de generar el modelo con el conjunto de entrenamiento, este se puede usar para predecir la clase de los datos de prueba. Entonces, la razón de precisión, se obtiene dividiendo el número de clasificaciones correctas por el número total de instancias. La precisión es una buena estimación de cómo se comportará el modelo para datos futuros similares a los de la prueba.

2.1.5.- Fase de difusión y uso

Se hace uso del nuevo conocimiento y se hace partícipe de él a todos los posibles usuarios

2.2 Descubrimiento de conocimiento en texto (KDT)

En la actualidad una gran cantidad de información no aparece en una estructura de base de datos, sino en colecciones de documentos que surgen de varias fuentes, como los correos electrónicos, los hipertextos, los documentos de texto, los archivos HTML, etc.

La *minería de texto* es similar a la minería de datos excepto porque las herramientas de la minería de datos está diseñada para datos estructurados de una base de datos, y la minería de texto trabaja con conjunto de datos (texto) no estructurados o semi-estructurados.

Basándonos en el esquema KDD, en [Feldman.2007] y en [Hernández.2007] proponemos el esquema para el descubrimiento de conocimiento en texto KDT en la figura 2.2. Este proceso, al igual que el KDD, consta de cinco fases para la obtención de conocimiento; como datos iniciales se tiene una colección de documentos, los cuales se integran y recopilan para tener una colección inicial de documentos con características similares. Se realiza una selección, limpieza y transformación para obtener una colección minable de documentos. En la fase de minería de texto se obtiene como resultado un conjunto de patrones los cuales se evalúan e interpretan para obtener un conocimiento útil y novedoso. Por ultimo este conocimiento se difunde y aplica, siendo la quinta fase del proceso KDT.

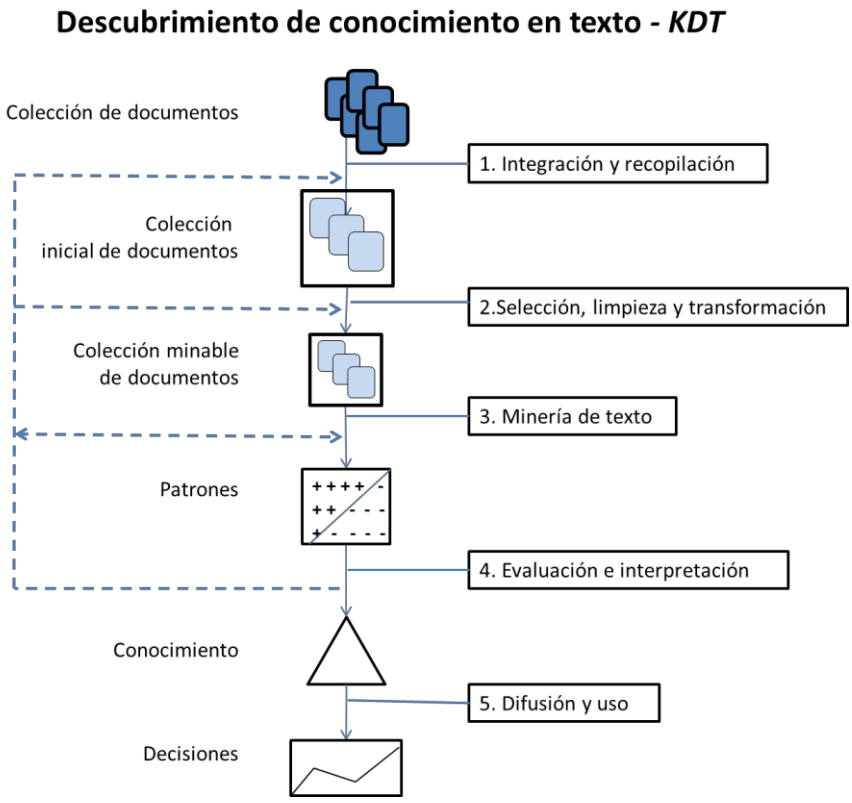

Figura 2. 2. Descubrimiento de conocimiento en texto [Hernández.2007] [Feldman.2007].

La *minería de textos* pretende extraer información útil a partir de una fuente de datos a través de la identificación y exploración de patrones interesante, donde las fuentes de datos son colecciones de documentos, y patrones interesantes se encuentran no entre los registros de una base de datos formalizados pero si en los datos de texto no estructurados en los documentos de estas colecciones [Feldman.2007].

Por lo que el *descubrimiento de conocimiento en texto* (KDT) se refiere al proceso de la extracción de información interesante y no trivial de texto no estructurado o semi-estructurado [Feldman.1995].

2.2.1 Minería WEB

La World Wide WEB (o simplemente la WEB), es una colección de billones de documentos con hipervínculos a otros documentos. Estos documentos o hipertextos, son escritos en diversos lenguajes, y tratan de diferentes temas esenciales para el ser humano [Chakrabarti.2003].

La minería WEB consiste en extraer información a partir de documentos de la WEB. Las técnicas de minería WEB difieren de la minería de datos, ya que la WEB es un repositorio de gran tamaño donde los documentos contienen datos de muy diverso tipo (texto, imágenes, audio, etc) que son, por lo tanto, no estructurados o semi-estructurados, a diferencia de las bases de datos. Además, los documentos son hipertextos, al hacer referencia a otros documentos a través del hipervínculo. Estos hipervínculos pueden ser recorridos o no por distintos usuarios, según las secuencias de navegación por la WEB. Esta diversidad permite minar la WEB basándose en tres conceptos: el contenido, la estructura y el uso [Hernández.2007] [Chakrabarti.2003].

La *Minería WEB* se define como el uso de técnicas de minería de datos para descubrir y extraer información desde la World Wide WEB [Hernández.2007].

La minería WEB se clasifica en tres áreas de interés en función de la parte de la WEB que se mina: minería de contenido, minería de la estructura y minería del uso [Hernández.2007] [Chakrabarti.2003].

- *Minería del contenido de la WEB*: Describe el descubrimiento de información útil desde los contenidos textuales y gráficos de los documentos de la WEB, y tiene sus orígenes en el procesamiento del lenguaje natural y en la recuperación de información.
- *Minería de la estructura de la WEB*: Trata de descubrir el modelo subyacente a la estructura de enlaces de la WEB y analiza, fundamentalmente, la topología de los hipervínculos. Este modelo se utiliza para categorizar páginas WEB y es útil para generar información como la similitud y relación entre diferentes sitios WEB, así como para detectar páginas autoridades y páginas concentradoras.
- *Minería del uso de la WEB*: Es el proceso de analizar la información sobre los accesos WEB disponibles en los servidores WEB minando datos secundarios derivados de la interacción de los usuarios mientras interactúan con la WEB. Estos datos incluyen los archivos logs de acceso al servidor, logs del navegador, logs de los servidores proxy, perfiles de usuario, sesiones y transacciones, etc.

2.2.2 Secuencias Frecuentes Maximales

Uno de los problemas que ha surgido en la minería de datos es el análisis de la información que mantiene un orden en sus registros simbólicos a través del tiempo, es decir, de la información descrita de manera secuencial. Este problema ha sido abordado por el área de investigación denominada *Minería de Patrones Secuenciales*, en donde el objetivo principal

es extraer datos que se presentan frecuentemente, pero preservando su orden secuencial dentro de la información, los cuales se llaman *patrones secuenciales*. Como el texto mantiene un orden secuencial de las palabras entonces también es posible analizar y descubrir patrones válidos, novedosos, potencialmente útiles y entendibles por el humano a partir de información textual [García.2007].

En la Minería de Patrones Secuenciales, una *secuencia de palabras se considera frecuente* si ésta se encuentra en al menos un cierto número de documentos (umbral mínimo de frecuencia).

Una *secuencia S*, denotada por $\langle s_1,s_2,...,s_k\rangle$, es una lista ordenada de *k* elementos. Una secuencia de *longitud k* es denominada *k*-secuencia.

Sean $P=\langle p_1,p_2,\ldots,p_t\rangle$ y $S=\langle s_1,s_2,\ldots,s_m\rangle$ secuencias, *P* es una *subsecuencia* de *S con GAP=0*, denotado como $P \subseteq S$ si existe un entero $i \geq 1$ tal que $p_1=s_i$, $p_2=s_{i+1}$, $p_3=s_{i+2}$, $\ldots,p_t=s_{i+(t-1)}$.

Un documento *W* se puede considerar como una secuencia de palabras, denotado también como $\langle w_1,w_2,\ldots,w_n\rangle$.

La *frecuencia* de una secuencia *S* en una colección de documentos $\{W_1, W_2, \ldots, W_j\}$ considerados como secuencias, denotada por S_f o $\langle s_1,s_2,...,s_t\rangle_f$, es el número de documentos en los cuales *S* aparece por lo menos una vez, esto es, $S_f = \{ W_i \mid S \subseteq W_i \}$.

Dado un umbral definido por el usuario (β), una secuencia *S* es *frecuente* si $S_f \geq \beta$.

Una secuencia frecuente *S* es *maximal* si *S* no es subsecuencia de alguna otra secuencia frecuente [García.2007].

Para ejemplificar estos conceptos consideremos un conjunto de oraciones extraídas de una colección de documentos, mostradas en la figura 2.3, al analizar cuáles son las secuencias frecuentes maximales con un umbral de 2, esto es, que al menos se repita 2 veces; tenemos que hay 13 secuencias frecuentes maximales con una palabra (tamaño 1), 3 secuencias frecuentes maximales con dos palabras (tamaño 2) y 2 secuencias frecuentes maximales con tres palabras (tamaño 3). Este resultado se visualiza en la figura 2.4. Cabe mencionar que los signos de puntuación y los números aunque no son una palabra, para fines de este trabajo de investigación se van a considerar como tal.

1. DE LAS DELEGACIONES DEL MUNICIPIO DE LANDA DE MATAMOROS , QUERETARO , MEXICO . ACATITLAN PROVIENE DE LOS
2. EN LA COSTA SUR DE INGLATERRA , REINO UNIDO . EL EDIFICIO DE LA ACADEMIA
3. ES UN EQUIPO DE FUTBOL DE LA CIUDAD DE GUADALAJARA , MEXICO . ACTUALMENTE MILITA EN LA PRIMERA DIVISION
4. MUNICIPIO DE LA REGION SIERRA OCCIDENTAL DEL ESTADO DE JALISCO , MEXICO . AYUTLA SIGNIFICA
5. DE HELICOPTEROS , CON SEDE EN ITALIA Y EL REINO UNIDO . FUE CREADA EN JULIO DEL
6. TIENE SERVICIO DE DRENAJE Y AGUA POTABLE . SU CONTRUCCION ES GENERALMENTE A BASE DE CEMENTO , TEJA Y / O TABIQUE
7. MATERIAL DE CONSTRUCCION EN EL QUE SE UTILIZA CEMENTO COMO CONGLOMERANTE . LOS MORTEROS POBRES O ASPEROS
8. COMO SUCEDIO CON LOS GOBIERNOS DE @LINK , REINO UNIDO , POLONIA O ITALIA A FAVOR DE LA POLITICA
9. ES DE TODOS LOS SCOUTS DE MEXICO . EL NOMBRE DE MEZTITLA , ES UNA PALABRA
10. LA ETNIA MIXTECA , Y SE TRATA DEL SEGUNDO MUNICIPIO MAS POBRE DE LA REPUBLICA MEXICANA . METLATONOC ES UN

Figura 2. 3 .Conjunto de oraciones extraídas de una colección de documentos.

Al obtener las secuencias frecuentes maximales con umbral 2, el resultado es:

13 SFM's de tamaño 1----------
[3] DEL
[4] LOS
[4] EL
[5] ES
[2] UN
[2] CON
[2] ITALIA
[4] Y
[2] A
[2] CEMENTO
[3] O
[2] SE
[2] COMO
3 SFM's de tamaño 2----------
[2] MUNICIPIO DE
[2] EN LA
[5] DE LA
2 SFM's de tamaño 3----------
[3] , MEXICO .
[2] REINO UNIDO .

Figura 2. 4 SFM obtenidas.

Como se puede visualizar en la figura 2.4 la secuencia que consta de sólo la palabra MEXICO no es una SFM, ya que está contenida en otra secuencia de mayor longitud.
El algoritmo empleado para obtener SFM, es el desarrollado por el Dr. René A. García Hernández [García.2007]

Los *patrones léxicos* son aquellos patrones que trabajan en un nivel léxico sin tomar en cuenta elementos sintácticos o semánticos, y estos pueden ser obtenidos a partir de secuencias frecuentes maximales.

2.2.3 Métricas de evaluación

El proceso de evaluación en la tarea de extracción de información consiste en comparar un conjunto de patrones realizados por el humano contra el conjunto de patrones generados por un sistema de extracción a ser evaluado. Las métricas básicas para medir el desempeño de los sistemas de extracción de información son la precisión y el recuerdo [Feldman.2007] [Orta.2008] [Ortega.2007].

La *precisión* está definida como la fracción de casos recuperados que son relevantes:
Precisión = |{casos relevantes} ∩ {casos recuperados}| / {casos recuperados}

El *recuerdo* está definido como la fracción de casos recuperados con la consulta del sistema:
Precisión = |{casos relevantes} ∩ {casos recuperados}| / { casos relevantes}

F-measure es la métrica armónica basada en la precisión y el recuerdo definida como:

$$\text{F-measure} = 2PR/(P+R)$$

Para este trabajo de investigación quedan definidas como:

$$\text{Precisión} = \frac{|\{\text{hipervínculos relevantes}\} \cap \{\text{fragmentos de texto recuperados}\}|}{|\{\text{ fragmentos de texto recuperados }\}|}$$

$$\text{Recuerdo} = \frac{|\{\text{hipervínculos relevantes}\} \cap \{\text{fragmentos de texto recuperados}\}|}{|\{\text{ hipervínculos relevantes}\}|}$$

$$\text{F-measure} = 2PR/(P+R)$$

2.4 Resumen

En este capítulo se han revisado los conceptos necesarios para abordar el problema planteado en este trabajo de tesis.

Recordemos que el problema de investigación de este trabajo de tesis consiste en identificar si hay patrones en el texto que el humano sigue para la construcción de hipervínculos y qué tan valiosos serían para su aplicación en la construcción de hipervínculos en una colección de documentos dada y bajo un dominio, considerando que es posible que a partir de la detección de patrones léxicos en los hipervínculos creados por el humano, se puedan detectar fragmentos de texto candidato a hipervínculo en texto plano, replicando los patrones léxicos encontrados; concluimos que para ello necesitamos hacer minería del contenido de la WEB con la tarea de reglas de asociación secuencial para obtener patrones léxicos que nos permitan identificar texto candidato a hipervínculo. Además es posible la evaluación de los patrones léxicos a través de las medidas de Precisión, Recuerdo y F-measure.

Capítulo 3

Estado del arte

A la fecha existen diferentes trabajos de investigación dedicado a la construcción de hipervínculos de manera automática. En el 2007 el foro INEX convocó a investigadores al foro denominado Enlaces Wikipedia (The Link the Wiki – LTW), cuyo objetivo fue proporcionar un foro para el descubrimiento de enlaces en Wikipedia y para evaluar objetivamente el desempeño de este tipo de algoritmos. En este capítulo se describen los trabajos más representativos del foro de Enlaces Wikipedia (LTW). Además, se describirán diferentes trabajos que hacen descubrimiento de patrones de extracción con el objetivo de generar un conjunto de patrones léxicos, a través de la herramienta de secuencias frecuentes maximales, a ser utilizados en la tarea de extracción de información.

3.1 FORO INEX 2007

INEX es una iniciativa para la evaluación de la recuperación de XML establecido en el 2002 con más de 90 organizaciones participantes en todo el mundo. En el 2007 organizó un foro de enlaces de Wikipedia (LTW) destinado a la discusión de la evaluación de descubrimiento de enlaces en Wikipedia así como métodos automáticos en el descubrimiento de enlaces, esto en una colección de Wikipedia en Inglés de 2006 que contiene 660,000 documentos en 4 GB de tamaño.

Recordemos que los autores de hipertexto deben identificar tanto las palabras que contienen el hipervínculo como el documento destino del hipervínculo con el fin de construir una red de conocimiento.

Un sistema de descubrimiento de enlaces en Wikipedia selecciona automáticamente una serie de fragmentos de textos candidatos a contener un hipervínculo, y varios destinos del hipervínculo para cada fragmento de texto. Esto es llamado, descubrimiento de enlaces salientes. El foro LTW está dirigido sólo a vínculos de documento a documento.

Uno de los trabajos presentados en el foro LTW es el de Delip Rao [Rao.2007] llamado "Vinculando entidades: Extracción de entidades en una base de conocimiento", en este trabajo se describen métodos para la tarea de enlazar entidades incluyendo variaciones de manipulación en entidades nombradas, ambigüedad de entidades y entidades que no están incluidas en la base de conocimientos, y se muestra como pueden ser tratados cada uno de ellos. Este trabajo se centra en la vinculación de organizaciones, entidades geopolíticas y de las personas en Wikipedia, derivadas de una base de conocimientos en el idioma inglés, formada con los títulos de las páginas de Wikipedia.

Otro trabajo es el de Oskar Gross, "Filtrado de entidades nombradas basado en asociación-concepto de grafos" [Gross.2007] que propone un método centrado en el análisis de redes de asociación de palabras mediante grafos, que relacionan los documentos. El método se

basa en la idea de que un documento está relacionado con una entidad nombrada cuando ambos están relacionados con los mismos conceptos. Un punto importante del modelo es el proceso de eliminar las aristas y nodos que no son necesarios, dejando sólo las asociaciones que están directamente relacionadas con las entidades nombradas.

La vinculación entre los recursos digitales se está convirtiendo en una forma cada vez más importante para encontrar la información. A través de la navegación, los usuarios pueden fácilmente entender el contexto y darse cuenta de las relaciones de información relacionada. "un método basado en el Perfil de Entidades en una búsqueda de conocimiento Automático" desarrollado por Xitong Liu [Liu.2007] trata de la construcción de perfiles de entidades basándose en entidades de búsqueda en Wikipedia. Esto es, dada una entidad, lo primero es recuperar su página en Wikipedia a través del URL asociado, así se construye un perfil de la entidad, formado por el nombre del URL y por el texto de ancla que aparece en la página HTML. Encontrando así que, este perfil de entidad es importante en la vinculación de entidades con documentos.

Por último, el trabajo llamado: "Vinculación de documentos a conocimiento enciclopédico desarrollado por Rada Milhacea [Milhacea.2007], describe un sistema capaz de enriquecer automáticamente un texto con hipervínculos al conocimiento enciclopédico, llamado Wikify!. Este sistema se basa en dos tareas de recuperación de información, como son, la identificación de palabras clave obtenidas a partir de los títulos de los documentos, para construir un catálogo, por lo que, el hipervínculo rojo que trabaja Wikipedia no existiría con este sistema; y la desambiguación del sentido de la palabra. Siendo la primera tarea para identificar que palabras deben tener hipervínculo (ancla) y la segunda es para identificar hacia donde (página destino) se dirige el hipervínculo.

Como se puede ver, los trabajos descritos anteriormente se basan en Entidades nombradas como ancla del hipervínculo, esto es, se toman entidades nombradas como fragmento de texto candidato a hipervínculo. En cambio, nuestro problema es determinar qué fragmentos

de texto son candidatos a tener hipervínculo sin la premisa de que sólo las entidades nombradas pueden tener hipervínculo.

3.2 Descubrimiento de conocimiento en texto

Dentro del procesamiento del lenguaje natural existen diferentes tareas que abordan el problema de la búsqueda de información. La tarea de extracción de información, que consiste en identificar descripciones de eventos en texto en lenguaje natural y por consiguiente, extrae la información relacionada a dichos eventos.
Existen diferentes trabajos relacionados con la extracción de información, entre ellos:

- Respondiendo a preguntas de definición mediante el descubrimiento de patrones léxicos [Denicia.2007].
- Descubrimiento automático de hipónimos a partir de texto no estructurado [Ortega.2007].
- Métodos basados en patones léxicos para la extracción de información [Orta.2008].

Estos trabajos de investigación anteriormente mencionados llevan a cabo la tarea de extracción de información con el descubrimiento y aplicación de patrones de extracción. Además los tres se pueden ver como un proceso de descubrimiento de conocimiento en texto, por lo que se describirán a través de las fases del KDD.

3.2.1 Fase de integración y recopilación:

En esta fase se determinan las fuentes de información que pueden ser útiles y dónde conseguirlas. En la tabla 3.1 se identifican los tipos de colecciones utilizadas en los diferentes trabajos investigados; así como el tamaño de la colección. Como se puede ver en la tala 3.1, las colecciones de documentos utilizadas mientras más grandes mejor. También se puede observar que los textos utilizados vienen en textos comunes o cotidianos.

Tabla 3. 1 Fase de integración y recopilación.

	[Denicia.2007]	[Ortega.2007]	[Orta.2008]
Tamaño de la colección	454,045	WEB	550
Tipo de colección	Noticias en español de la agencia EFE de 1994 y 1995	WEB	Noticias de desastres naturales

3.2.2 Fase de selección, limpieza y transformación:

En esta fase se seleccionan, limpian y transforman los textos, para normalizar los textos de acuerdo al propósito de cada trabajo de investigación, dejando así una colección de textos lista a ser minada por la herramienta seleccionada para tal efecto. En la tabla 3.2 se visualizan los tres trabajos de investigación que se analizaron, mostrando que la transformación consistió en que los autores de dichos trabajos de investigación colocaron, de forma manual, etiquetas en el texto para normalizar los datos.

Tabla 3. 2 Fase de selección, limpieza y transformación.

	Denicia.2007	Ortega.2007	Orta.2008
Normalizar datos por medio de etiquetas	Si	Si	Si
Quien etiqueta la colección	Autor	Autor	Autor

3.2.3 Fase de minería de datos

En esta fase se pretende extraer información útil a partir de una colección de documentos, por medio de la identificación y exploración de patrones interesante que se encuentran en los datos de texto no estructurados o semi-estructurado en los documentos de estas colecciones. En la tabla 3.3 se muestra que la técnica de minería de texto para la obtención de dichos patrones en los trabajos de investigación que se están revisando, es la técnica de Secuencias Frecuentes Maximales.

Tabla 3. 3 Fase de minería de texto.

	Denicia.2007	Ortega.2007	Orta.2008
Técnica de minería de texto para la obtención de patrones	SFM	SFM	SFM

3.2.4 Fase de evaluación e interpretación

Para medir la calidad de los patrones descubiertos por el algoritmo de minería de datos se transforman estos patrones en patrones de búsqueda, construyendo plantillas, como se ve en la tabla 3.4. Además, para evaluar su método propuesto se partió el conjunto de datos en dos, uno de entrenamiento y otro de prueba. Se utilizaron las métricas de precisión y recuerdo para evaluar los patrones candidatos, filtrándolos por su calidad y dejando patrones definitorios, para así generar un conocimiento en cada uno de los trabajos analizados de la tabla 3.4.

Tabla 3. 4 Fase de evaluación e interpretación.

	Denicia.2007	Ortega.2007	Orta.2008
Como se transforma a un patrón de búsqueda	Plantilla	Plantilla	Plantilla
Métrica de evaluación de patrones candidatos	Precisión y recuerdo	Precisión y recuerdo	Precisión y recuerdo
Se filtran los patrones candidatos según su calidad y se quedan patrones definitivos	Si	Si	Si
Conocimiento	Se encontró un conjunto de patrones léxicos de la tupla Definición-concepto	Se encontró un conjunto de patrones léxicos de la tupla Hipónimo-hiperónimo	Los métodos son útiles pero la cantidad de documentos utilizados provoco que los patrones léxicos recuperados fueran pocos.

3.2.5 Fase de difusión y uso

Además del conocimiento obtenido, descrito en la sección anterior, de los trabajos Denicia, Ortega y Orta; se han identificado como conocimiento que la técnica de secuencias frecuentes maximales es una buena técnica para la extracción de patrones en texto y que las medidas de precisión y recuerdo son útiles para la evaluación de los patrones obtenidos. En la tabla 3.5 se resume esto.

Tabla 3. 5 Fase de difusión y uso.

	Denicia.2007	**Ortega.2007**	**Orta.2008**
Se identificó que la técnica de SFM es buena para la identificación de patrones.	Si	Si	Si
Se identificó que las medidas precisión recuerdo ayudan en la evaluación de patrones recuperados con algún modelo	Si	Si	Si

3.3 Resumen

Como se vio en este capítulo, la construcción de hipervínculos en una colección de documentos no es un problema reciente. Sin embargo, ninguno de estos trabajos se basa en el proceso de descubrimiento de conocimiento en texto, menos aún, utilizan la herramienta de SFM para la extracción de patrones léxicos que ayuden en la tarea de extracción de información. Más bien, parten de los títulos de los documentos o de entidades nombradas.

Capítulo 4

Método propuesto

Dado que el propósito de esta investigación es descubrir conocimiento novedoso y útil en la construcción de hipervínculos elaborados por el humano, identificando patrones de búsqueda, para así localizar fragmentos de texto candidato a tener hipervínculo en una colección de documentos en texto plano, se considera que el proceso KDT es de gran utilidad para tal fin, así mismo la utilización de secuencias frecuentes maximales en la minería del contenido de la WEB. Por lo que en este capítulo propone un método que nos lleve a conocimiento útil y novedoso y que ayude a cumplir el objetivo de este trabajo de investigación.

La colección de donde se propone hacer la extracción de patrones de búsqueda es un subconjunto de documentos de Wikipedia en Español 2008 y la colección donde se propone

hacer la búsqueda de los patrones léxicos es un subconjunto de documentos en texto plano de Wikipedia en español 2008, cabe mencionar que serían colecciones diferentes.

4.1 Arquitectura general

El método propuesto está basado en el descubrimiento de conocimiento en bases de datos (KDD), así como, en las fases que lo componen. En particular se profundizará la fase de minería, haciendo minería de contenido de la WEB (en Wikipedia en español 2008), ya que es la más complicada. Después se continuará con la fase de evaluación e interpretación para saber qué características tienen las palabras que pudieran ser candidatas a tener hipervínculo, por medio de las reglas de asociación secuencial (patrones léxicos); y obtener así un conocimiento útil y novedoso, que pueda ser difundido y usado en la construcción automática de hipervínculos. Por ello, se propone el esquema de la figura 4.1, como el proceso de descubrimiento de conocimiento en la WEB, para el desarrollo de nuestro método y la detección de fragmentos de texto como candidato a hipervínculo. El cual cuenta con las mismas fases del proceso KDD sólo que ahora los datos iniciales son documento recuperados de la WEB y el proceso de minería es del contenido de la WEB; ya que se va a utilizar la colección de Wikipedia en español 2008.

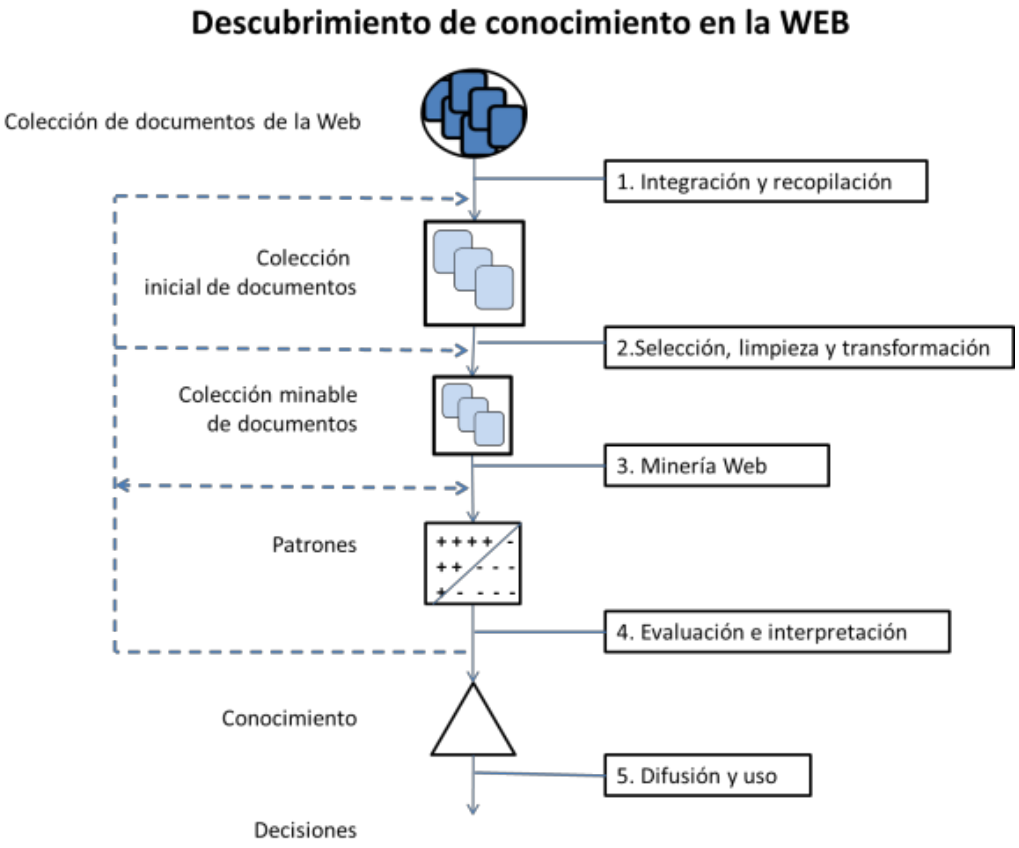

Figura 4. 1 Descubrimiento de conocimiento en la WEB.

4.1.1 Integración y recopilación

En esta fase, como se ve en la figura 4.2, se realizó la recopilación de la colección de documentos de Wikipedia en español 2008 ya que fue la colección más completa en español con hipervínculos que fue posible descargar de manera libre y está integrada por 1,362,467 archivos, todos ellos elaborados en formato HTML.

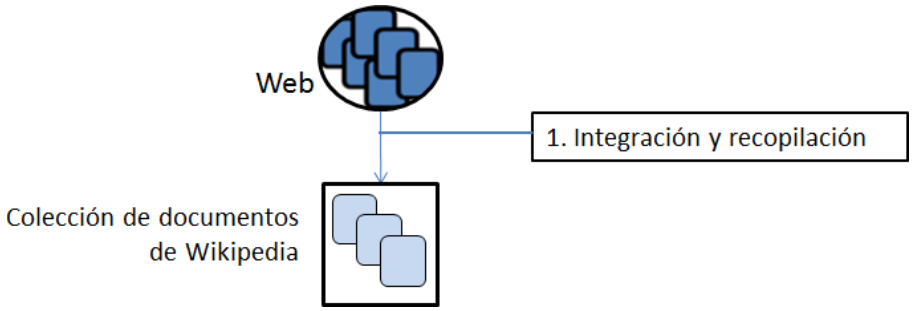

Figura 4. 2 Fase de integración y recopilación.

4.1.2 Selección, limpieza y transformación

En esta fase la colección inicial de documentos que es Wikipedia en español 2008, es transformada en una colección minable de documentos a través de un proceso de selección, limpieza y transformación, como se muestra en la figura 4.3. Los pasos detallados se ven a continuación.

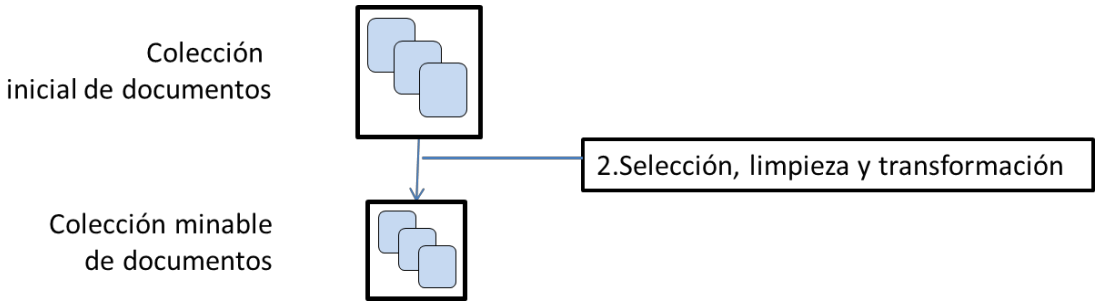

Figura 4. 3 Fase de selección limpieza y transformación.

Existen diferentes tipos de documentos en Wikipedia, algunos de ellos son documentos de foros de discusión, discusión de usuarios, imágenes, sonidos, videos, portales, marcos de presentación, tablas de categorías, etc. Todos ellos enriquecen el sitio de Wikipedia, este trabajo sólo se enfoca en documentos textuales que son los posiblemente tengan textos con

hipervínculo. Haciendo una selección de sólo documentos textuales, se obtuvieron alrededor de 699 mil documentos. Este proceso se representa en la figura 4.4.

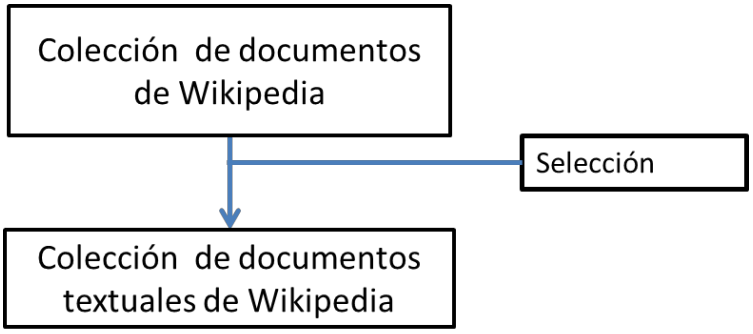

Figura 4. 4 Proceso de selección.

A continuación se realizan tres colecciones diferentes de documentos que corresponden a tres tipos de limpieza y transformación:

1. Para obtener una colección de documentos con sólo los hipervínculos que existen en ellos.
2. Para obtener una colección de documento en texto plano (sin formato alguno, únicamente caracteres alfanuméricos y signos de puntuación).
3. Para obtener una colección de documentos listos para ser minados por un algoritmo de minería de textos.

En estos tres procesos, la limpieza y transformación se lleva a cabo con Expresiones Regulares (ver apéndice A) [Friedl.2006]. Los procesos de limpieza y transformación se representan en la figura 4.5, y se detallan a continuación.

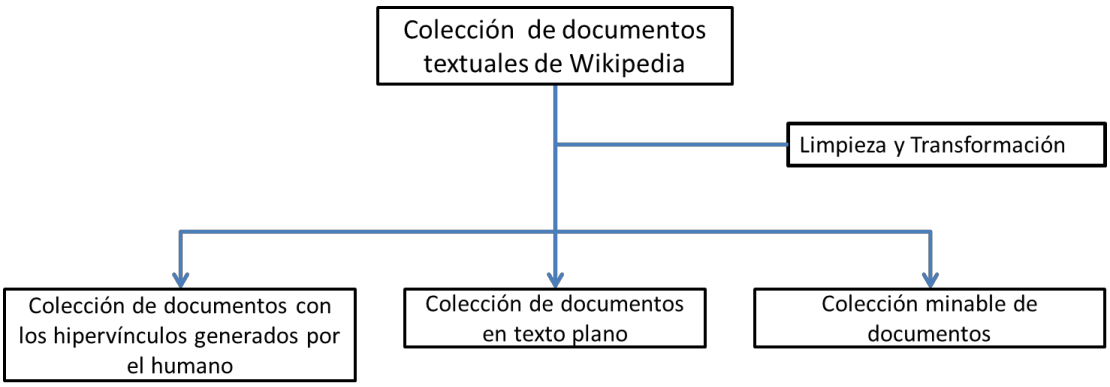

Figura 4. 5 Procesos de limpieza y transformación.

El primer proceso consiste en obtener colecciones de documentos que sólo contengan las palabras que tienen hipervínculo generado por el humano, separadas por el número de palabras que tiene el hipervínculo, para poder evaluar los fragmentos de texto candidatos a hipervínculo que se obtenga que en la fase de evaluación e interpretación. En la figura 4.6 se muestra un ejemplo de un documento en HTML que se limpia y transforma en uno que sólo tiene las palabras que son hipervínculo y que después se separa en colecciones por número de palabras.

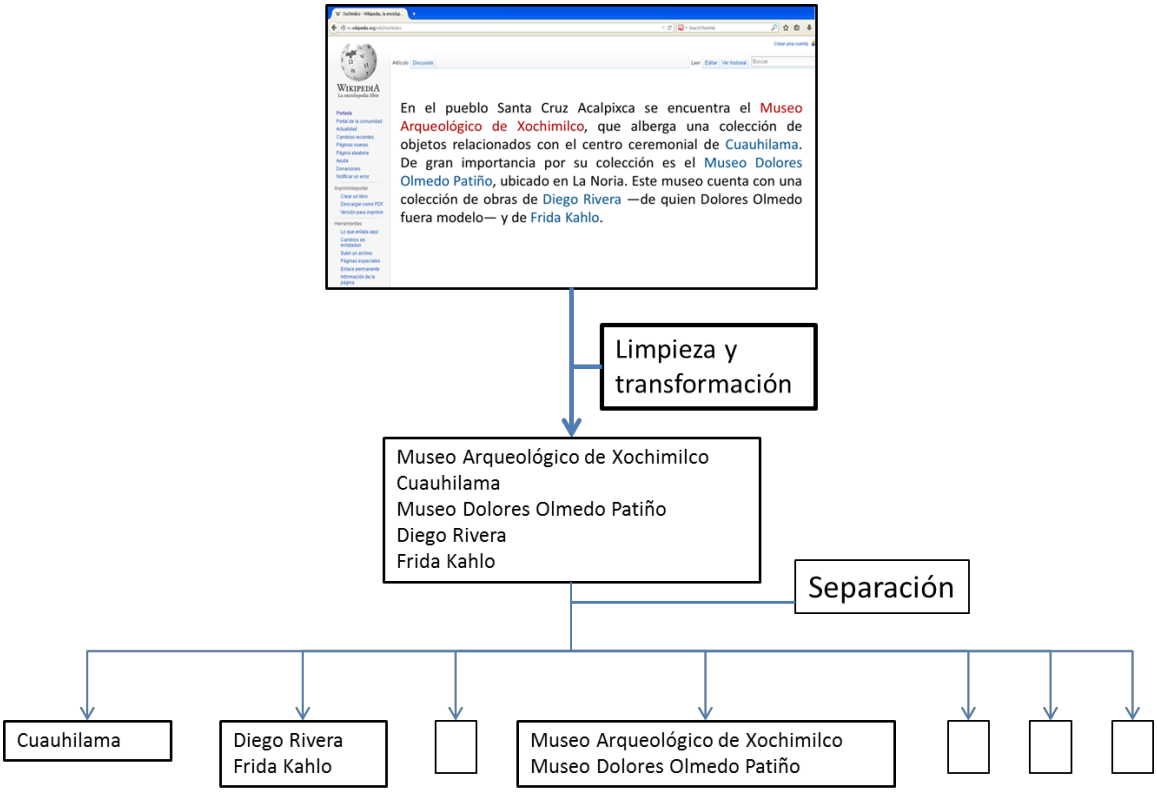

Figura 4. 6 Primer proceso de limpieza y transformación.

El segundo proceso consiste en obtener cada documento en texto plano (sin formato alguno, sólo caracteres alfanuméricos y signos de puntuación), esto con el fin de que una vez detectados los patrones léxicos de los hipervínculos construidos por los humanos (fase de minería), se puedan aplicar estos patrones al texto plano para identificar fragmentos de texto candidatos a hipervínculo, y a su vez sean posible su comparación con la colección del proceso anterior (fase de evaluación e interpretación). La figura 4.7 muestra un ejemplo de un documento en HTML que se limpia y transforma en un documento en texto plano.

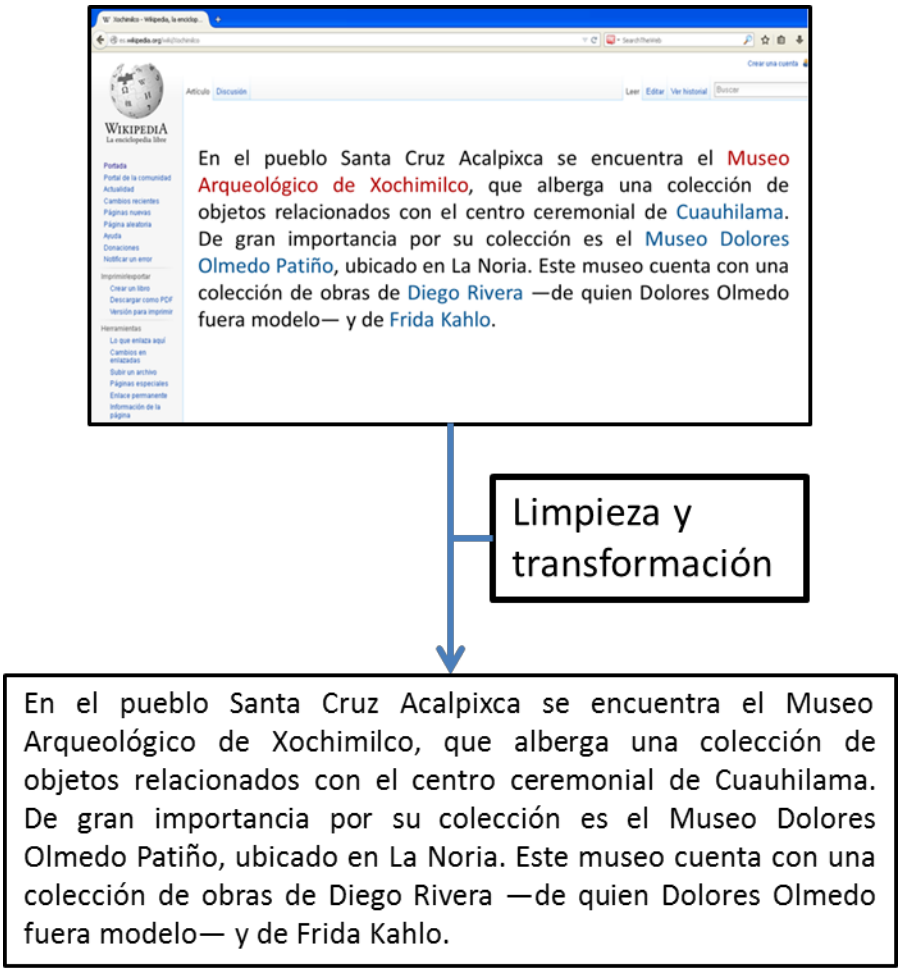

Figura 4. 7 Segundo proceso de limpieza y transformación.

El tercer proceso de limpieza y transformación consiste en obtener una colección de documentos listos para ser minados por el algoritmo DIMASP [García.2007] con el objetivo de obtener patrones léxicos que pudieran detectar fragmentos de texto candidato a tener hipervínculo. Este tercer proceso de limpieza y transformación para cada documento, se lleva a cabo a través de 4 pasos, todos ellos mediante la utilización se expresiones reglares y se describen a continuación:

1. Limpiar la colección seleccionada de Wikipedia en español 2008 de etiquetas HTML, excepto las etiquetas que enmarcan los hipervínculos. En la figura 4.8 se observa

como un documento HTML se limpia y transforma en un documento cuyas únicas etiquetas HTML son la de hipervínculos.

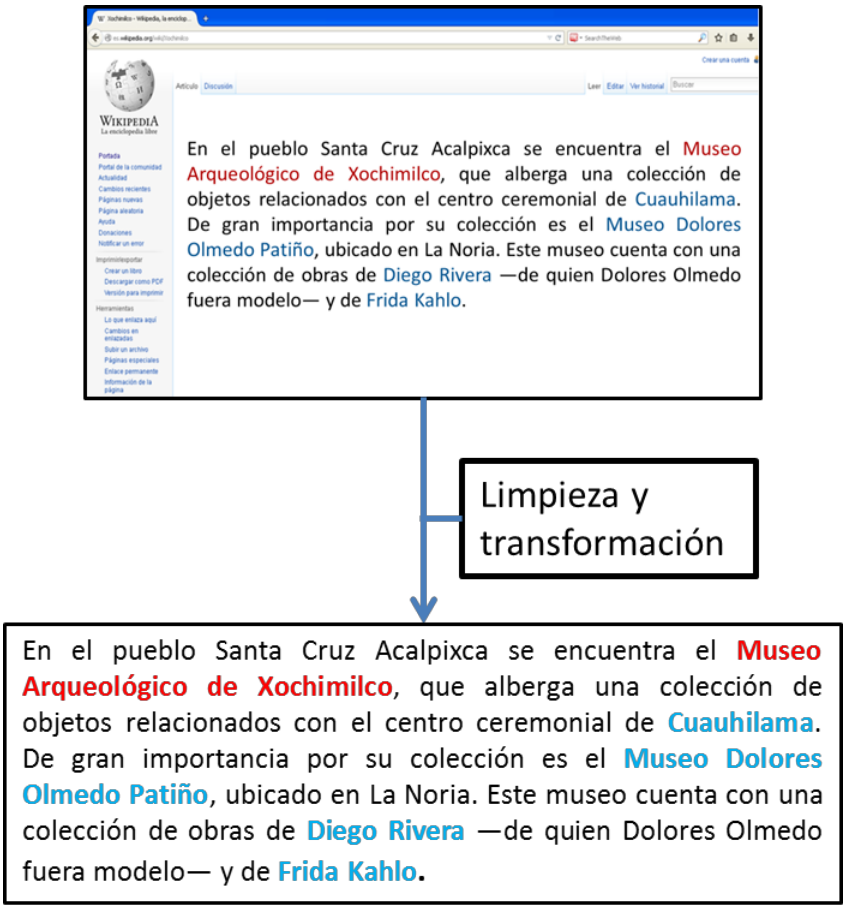

Figura 4. 8 Paso 1/4 en la preparación de documentos a ser minados.

2. En trabajos previos como el de Ortega [ortega2007] y el de Denicia [Denicia.2007] el contexto de búsqueda está delimitado por etiquetas llamadas <frontera izquierda><frontera centro> y <frontera derecha>. Normalizando así los patrones de búsqueda para Hiperónimo/Hipónimo [Ortega.2007] y para Concepto/Definición [Denicia.2007], como por ejemplo:

<frontera izquierda> Hiperónimo <frontera centro> Hipónimo <frontera derecha>
<frontera izquierda> Concepto <frontera centro> Definición <frontera derecha>

En el presente trabajo el contexto de búsqueda se define de la siguiente manera: por cada hipervínculo se seleccionan 20 palabras antes y 20 palabras después de él, sin que se traslapen las etiquetas de hipervínculo, ya que para este estudio se considera que 20 palabras son suficientes para definir el contexto izquierdo y derecho de cada hipervínculo, normalizando de esta manera los patrones de búsqueda, dejando una línea por cada hipervínculo de la siguiente manera:

<frontera izquierda> Hipervínculo <frontera derecha>

Cabe mencionar que los signos de puntuación y cualquier otro símbolo se respetan como una palabra, ya que se consideran importantes en el contexto de los hipervínculos. En la figura 4.9 se muestra un ejemplo de como un documento de Wikipedia en español 2008 se transforma en un documento que contienen una línea por hipervínculo con su respectivo contexto izquierdo y derecho.

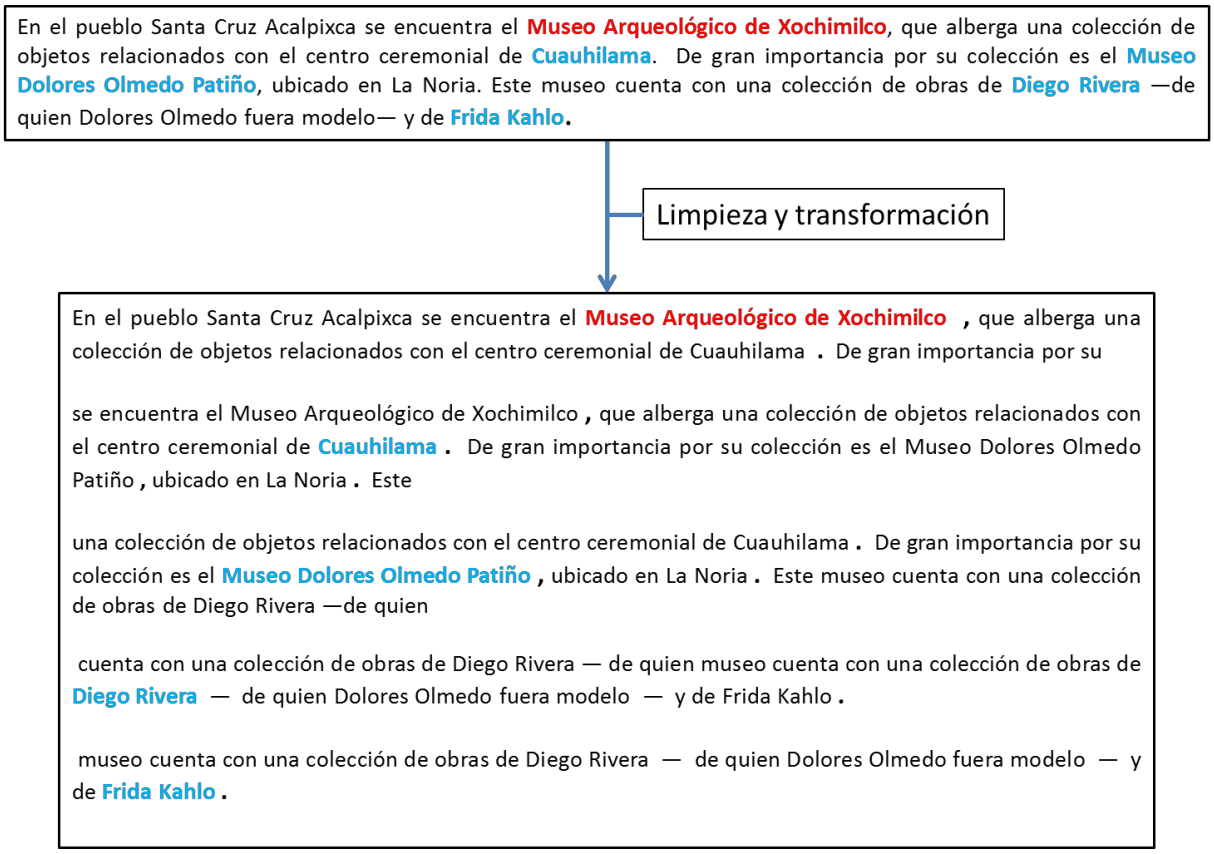

Figura 4. 9 Paso 2/4 en la preparación de documentos a ser minados.

3. Continuando con la normalización, se identificó que para poder delimitar el número de palabras que intervienen en la construcción de un hipervínculo, es necesario trabajar los hipervínculos construidos por el humano de acuerdo al número de palabras que lo componen. Por lo cual, cada palabra que forma el hipervínculo en los documentos de las colecciones a minar, se transforma en una etiqueta <@LINK>. En la figura 4.10 se muestra un ejemplo de la transformación de un documento que tiene hipervínculos en otro que contiene la etiqueta <@LINK> por cada palabra que forma el hipervínculo. Esta normalización hace más eficiente el análisis, la búsqueda y aplicación de patrones en el texto.

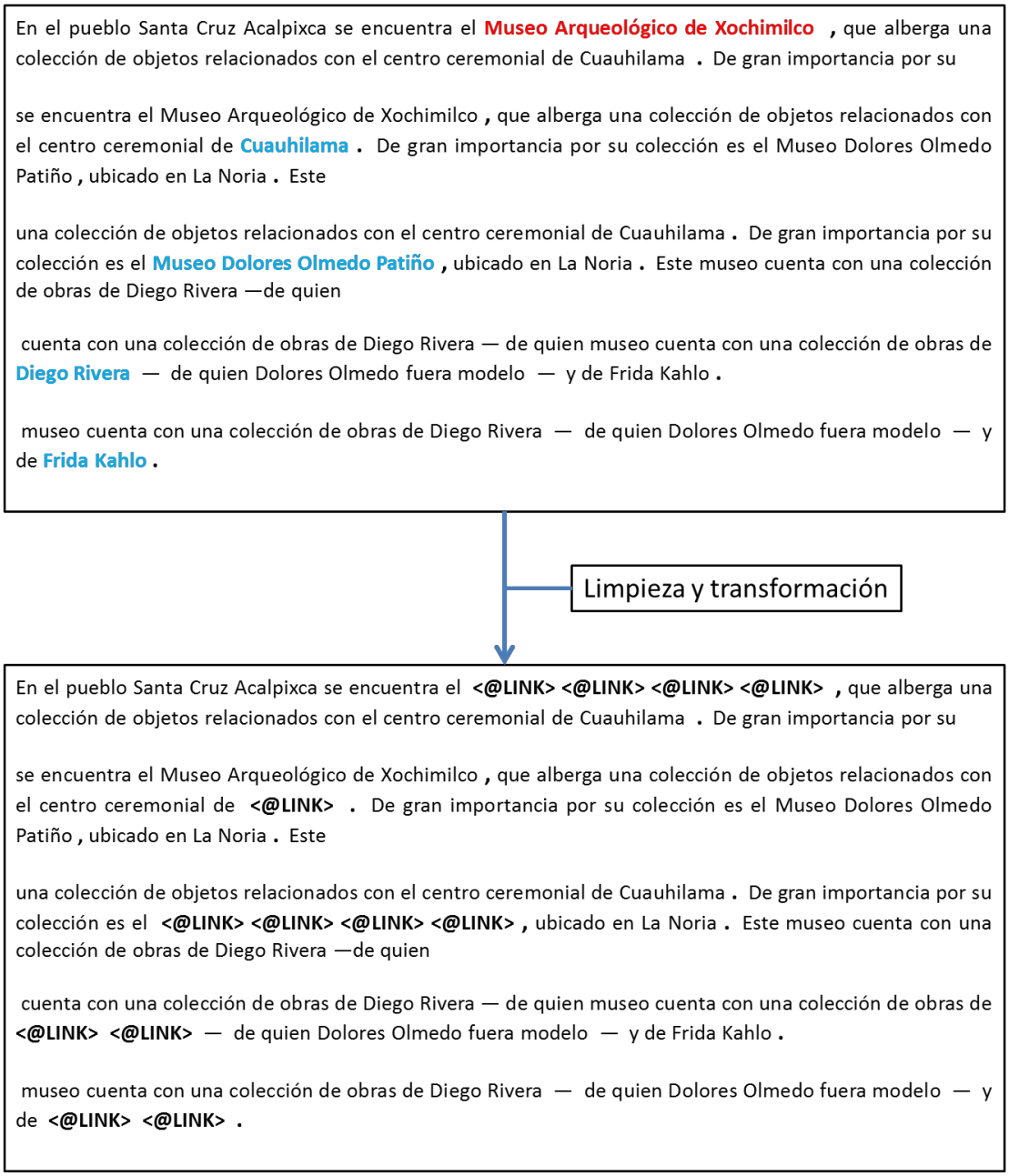

Figura 4. 10 Paso 3/4 en la preparación de documentos a ser minados.

4. Todas las líneas de hipervínculos de una palabra forman una colección, las líneas de hipervínculos de dos palabras forman otra colección, de tres palabras forman otra colección, así sucesivamente, formando colecciones donde cada una contiene hipervínculos con el mismo número de etiquetas <@LINK> en el hipervínculo.

   En la figura 4.11 se muestra un ejemplo para la colección de líneas con hipervínculos de 2 palabras. En el primer documento se tienen 5 líneas con diferentes cantidades de etiquetas <@LINK>, al hacer la transformación se genera una colección de líneas que sólo tienen dos etiquetas <@LINK>.

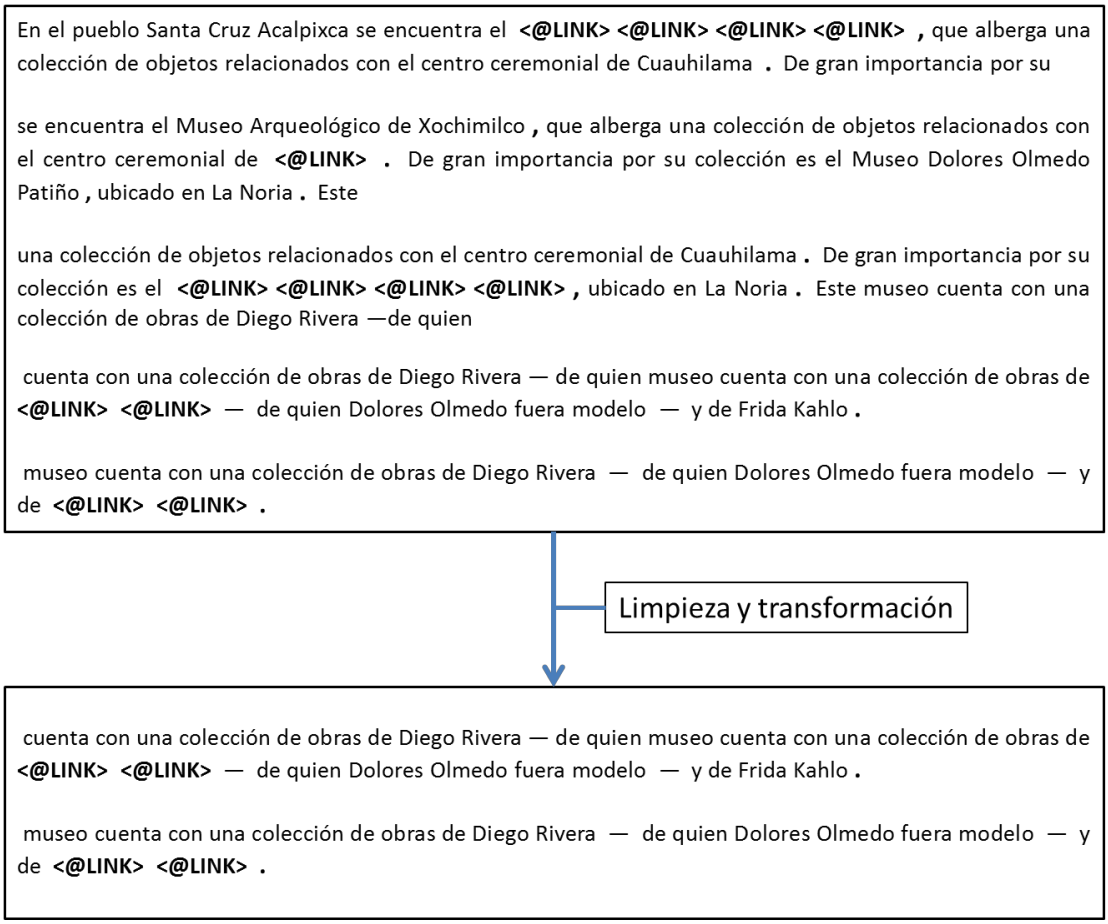

Figura 4. 11 Paso 4/4 en la preparación de documentos a ser minados

4.1.3 Minería de contenido de la WEB

En esta fase se obtienen una colección de patrones léxicos al minar los archivos generado en la fase de selección, limpieza y transformación, a través de la técnica de minería de texto referente a reglas de asociación secuencial. En la figura 4.12 se ve el esquema de esta fase, en la cual dada una colección minable de documentos, aplicamos la herramienta de minería de texto DIMASP [García.2007] y obtenemos una colección de secuencias frecuentes maximales, las cuales contienen los patrones léxicos que se desean identificar.

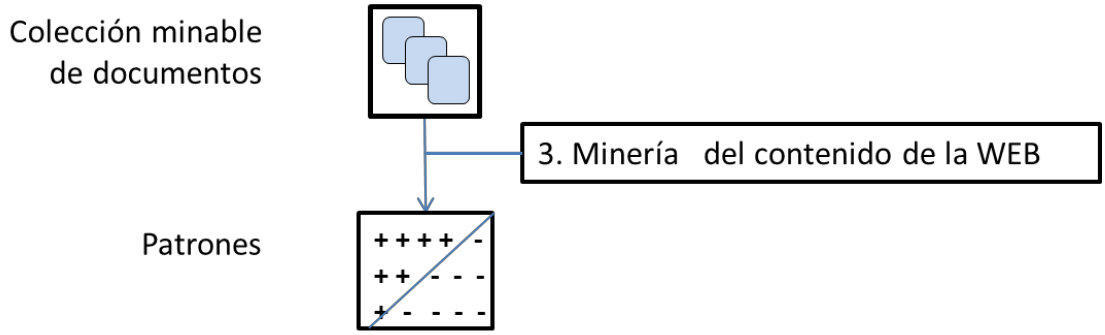

Figura 4. 12 Fase de minería WEB.

La utilización del algoritmo DIMASP fue debido a que en los trabajos de Orta [Orta.2008], Denicia [Denicia.2007] y Ortega [Ortega.2007], fue una excelente herramienta para la obtención de secuencias frecuentes maximales

En la figura 4.13 se ve un ejemplo para la colección formada con líneas que contienen cuatro palabras en el hipervínculo y se muestra como sería la salida después de aplicarle el algoritmo de secuencias frecuentes maximales llamado DIMASP.

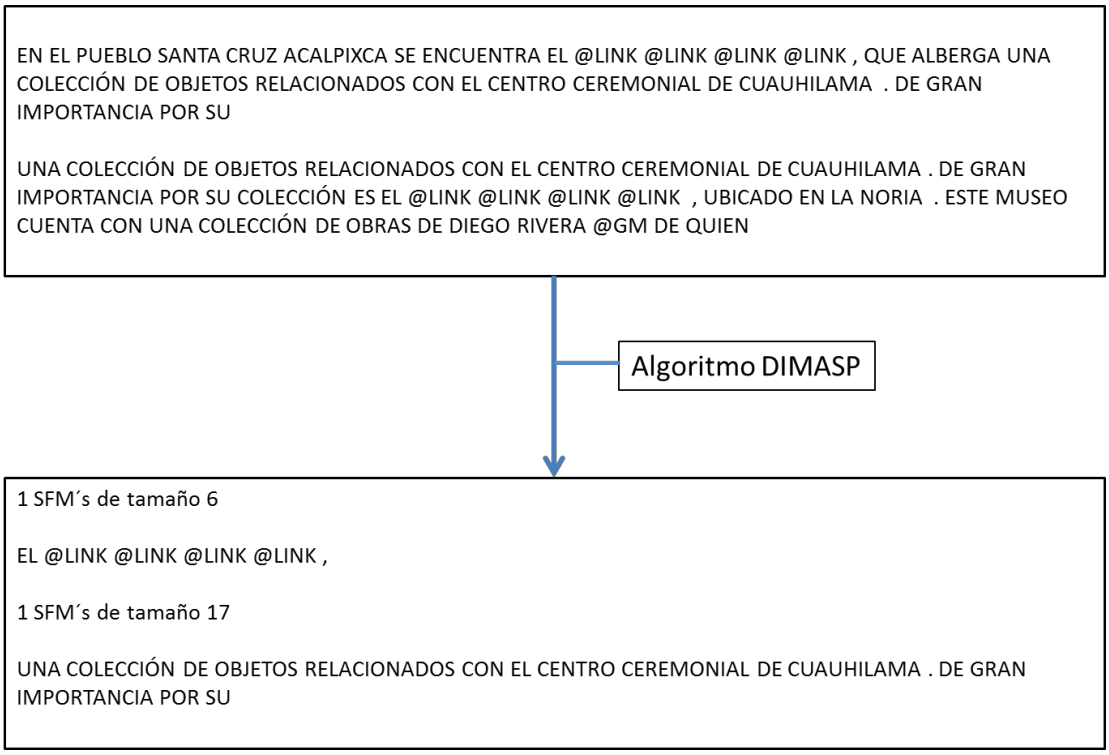

Figura 4. 13 Ejemplo de la aplicación del algoritmo de SFM.

Como podemos visualizar existen secuencias frecuentes maximales que contienen la etiqueta <@LINK> y otras que no contienen tal etiqueta. Pero el objetivo de este trabajo de investigación es identificar fragmentos de texto candidato a hipervínculo por medio de patrones léxicos y para ello deben tener las siguientes dos características:

- Que contengan etiquetas <@LINK>
- Que contenga texto antes y después de las etiquetas <@LINK>, esto es el contexto izquierdo y derecho de las etiquetas.

Haciendo una depuración para el ejemplo de la figura 4.14, sólo es de utilidad el patrón léxico:

El <@LINK> <@LINK> <@LINK> <@LINK> ,

Hasta aquí, se han obtenido los patrones léxicos que contienen las etiquetas <@LINK> y su contexto izquierdo y derecho.

4.1.4 Evaluación e interpretación

En esta fase de describen los procesos para evaluar e interpretar los resultados obtenidos. En la figura 4.14 se visualiza la entrada que es un conjunto de patrones léxicos obtenidos después de hacer la minería del contenido de la WEB para obtener un conocimiento útil y novedoso.

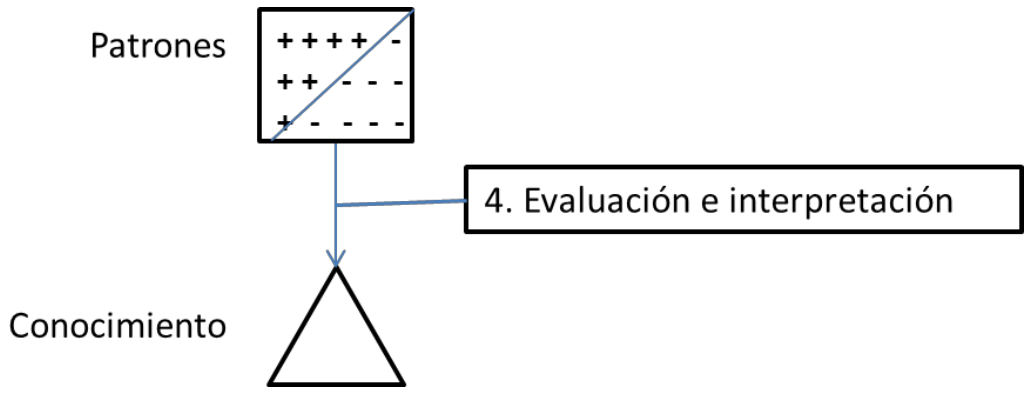

Figura 4. 14 Fase de evaluación e interpretación.

Esta fase consta de varios procesos:

1. Transformar los patrones léxicos obtenidos en la minería, a patrones de búsqueda.
2. Aplicar los patrones de búsqueda a la colección de texto plano, obtenida en el paso dos de la fase de selección limpieza y transformación, para obtener colecciones de fragmentos de texto candidato a tener hipervínculo, una por cada número de palabras.
3. Comparar los fragmentos de texto candidatos a tener hipervínculo con la colección de documentos con hipervínculos, obtenida en paso uno de la fase de selección, limpieza y transformación
4. Por último, pero el proceso más importante, interpretar los resultados obtenidos.

En el primer proceso, ilustrado en la figura 4.15, se tienen la colección de patrones léxicos obtenida de la fase de minería del contenido de la WEB y es transformada a una colección de patrones de búsqueda por medio de expresiones regulares. Cabe mencionar que sólo se muestra una línea como ejemplo, pero se trata de colecciones de patrones.

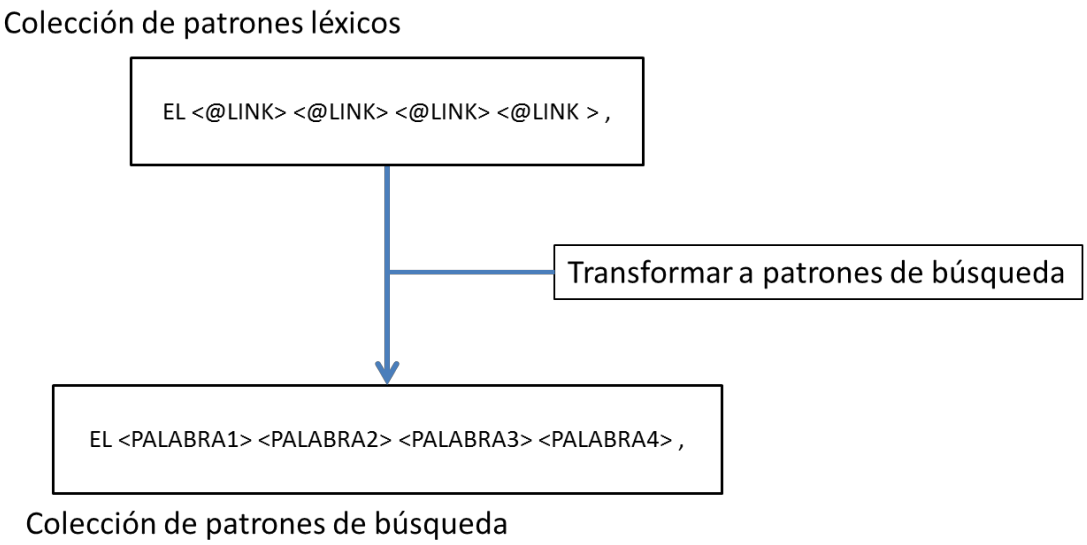

Figura 4. 15 Proceso 1/4 de la fase de evaluación e interpretación.

En el segundo proceso, ilustrado en la figura 4.16, se tiene la colección de patrones de búsqueda que se aplican a la colección de texto plano, que se obtuvo en el proceso dos de la fase de selección, limpieza y transformación, para obtener la colección de fragmentos de texto candidatos a tener hipervínculo. Cabe mencionar que únicamente se muestra una línea como ejemplo, pero se trata de colecciones de patrones.

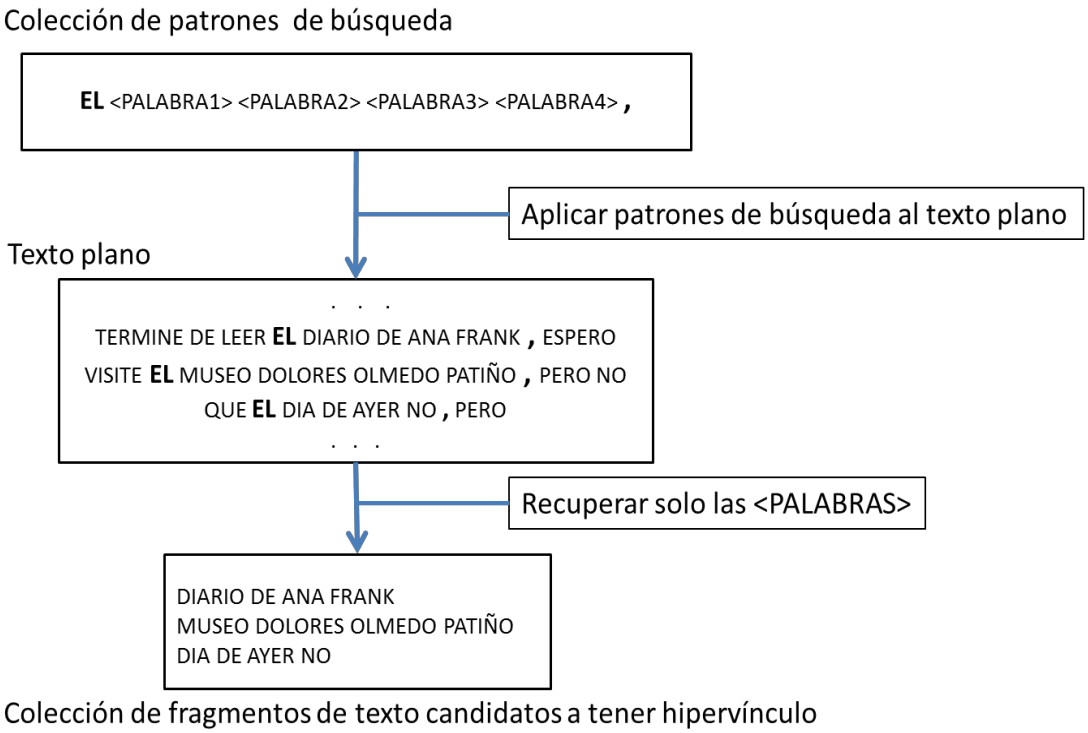

Figura 4. 16 Proceso 2/4 de la fase de evaluación e interpretación.

En el tercer proceso se compara la colección de fragmento de texto obtenidos de cada colección (de una palabra, dos palabras, tres palabras,…) del proceso anterior, con cada colección obtenida en el proceso uno de la fase de selección limpieza y transformación, separadas por número de palabras. En la figura 4.17 se hace un ejemplo de este proceso y se visualiza que ambos conjuntos coinciden en un fragmento de texto y se la medida F-mesure se describe a continuación.

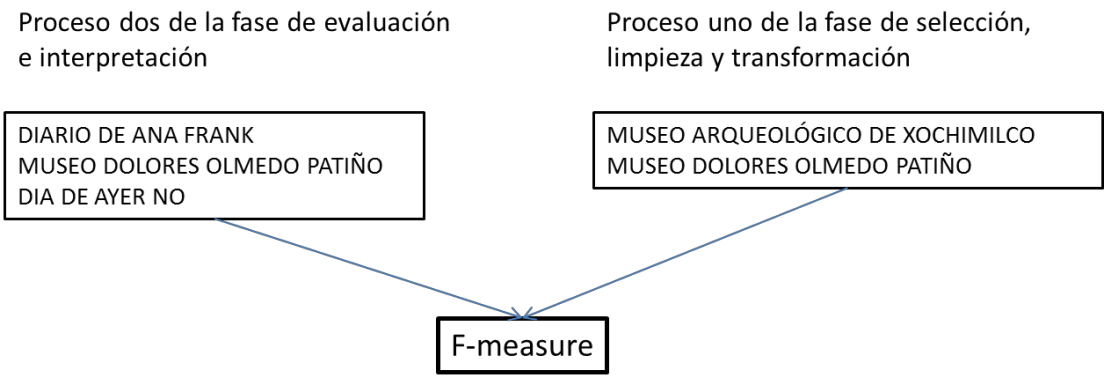

Figura 4. 17 F-measure

Las métricas de evaluación serían las siguientes:

$$\text{Precisión} = \frac{|\{\text{hipervínculos relevantes}\} \cap \{\text{fragmentos de texto recuperados}\}|}{|\{\text{ fragmentos de texto recuperados }\}|}$$

$$\text{Recuerdo} = \frac{|\{\text{hipervínculos relevantes}\} \cap \{\text{fragmentos de texto recuperados}\}|}{|\{\text{ hipervínculos relevantes}\}|}$$

Hipervínculos relevantes = 2

Fragmentos de texto recuperados = 3

Intersección entre relevantes y recuperados = 1

Precisión = 1 / 3 = 0.33

Recuerdo = 1 / 2 = 0.5

F-measure = 2 P R / (P + R) = 2 (0.33)(0.5) / (0.33 + 0.5) = 0.33 / 0.83 = 0.39

En el cuarto proceso de la fase de evaluación e interpretación se interpretan los valores obtenidos:

En la colección de 4 palabras en el hipervínculo se ve que el recuerdo es bueno (0.5) ya que recordó 1 fragmento de texto, de 2 posibles hipervínculos y la precisión en baja (0.33) ya

que recupero más fragmentos de texto de los 2 posible hipervínculos, lo que hace que la medida F-measure, no se buena (0.39); recordemos que la medida F-measure tiene un intervalo de 0 a 1, siendo 0 muy malo y 1 muy bueno.

Con esto se finalizan los cuatro procesos de la fase de evaluación e interpretación, cabe recordar que los ejemplos vistos en esta sección sólo son ilustrativos, en la fase de experimentación se verán colecciones reales de Wikipedia en español 2008

El descubrimiento de conocimiento se obtiene al refinar lo patrones léxicos en grandes colecciones, para que estos identifiquen fragmentos de texto candidato a tener hipervínculo con una buena exactitud, en base a la medida de F-measure. En el capítulo siguiente se verán varios experimentos que nos aportaran conocimiento útil y relevante referente a los hipervínculos y a la colección de Wikipedia en español 2008.

4.1.5 Difusión y uso

En esta fase se estaría dando a conocer los mejores patrones léxicos para su uso en la construcción automatizada de hipervínculos, una vez que se tengan los fragmentos de texto candidato a hipervínculo, recordemos que en eso consiste este trabajo de investigación. La fase de difusión y uso se queda como trabajo futuro.

4.3 Resumen

Una vez detallado el método propuesto con cada fase y proceso, recordemos que el objetivos de este trabajo de tesis en la detección de fragmentos de texto como candidatos a hipervínculos en una colección de documentos –Wikipedia 2008 en español – y que se debe de identificar si los autores de hipertexto siguen algún patrón para detectar fragmentos de textos que puedan tener hipervínculo, y mostrar si el contexto ayuda a determinar que fragmento de texto puede ser un hipervínculo.

Capítulo 5

Experimentación

Este capítulo se basa en el método propuesto para descubrir conocimiento descrito en el capítulo anterior. De esta manera se seguirán cada una de las fases del método propuesto con el fin de ver si es posible que a partir de la detección de patrones léxicos en los hipervínculos creados por el humano se pueden detectar fragmentos de texto candidato a hipervínculo; replicando los patrones léxicos encontrados.

Para realizar la experimentación se construyó una colección aleatoria de 10 mil documentos, la cual le llamó C10MIL. Además se construyeron dos colecciones aleatorias, una de 500 documentos (C500A) y otra de 500 documentos (C500B). Todas las colecciones fueron extraídas aleatoriamente de Wikipedia 2008 en español. Además, la experimentación se realizó con base en la cantidad de palabras que forman el hipervínculo

La fase de integración y recopilación quedó descrita en su totalidad en el capítulo anterior, por lo que iniciaremos este capítulo desde la fase de selección, limpieza y transformación, continuando con la fase de minería del contenido de la WEB y finalizando con la fase de evaluación e interpretación; para así realizar el descubrimiento de conocimiento del contenido de la WEB referente a los hipervínculos.

5.1 Fase de selección, limpieza y transformación.

Esta fase se llevó a cabo como se describe y ejemplifica en el capítulo anterior; generando siete colecciones de documentos normalizados. Cada una de las 7 colecciones se construyó de acuerdo a la cantidad de palabras que hay en cada hipervínculo, en la colección de 10 mil documentos de Wikipedia en español 2008. Considerando las 7 colecciones se encontraron 493,078 hipervínculos en total. La colección de hipervínculos de una palabra tiene 293,159 hipervínculos. Cabe señalar que aunque se construyeron 7 colecciones solo se experimentará con la más grande por razones del tiempo de procesamiento.

En la figura 5.1 se muestra un análisis de los hipervínculos de hasta siete palabras de la colección C10MIL y se puede ver que:

- Los hipervínculos de una palabra son los más representativos con el 60%.
- Los hipervínculos de 2 palabras son el 20%.
- Los de tres palabras son el 13%.
- Los de 4, 5, 6 y 7 palabras son muy poco representativos en la colección, entre todos son menos del 7%.
- En la colección C10MIL hay 493,078 hipervínculos.

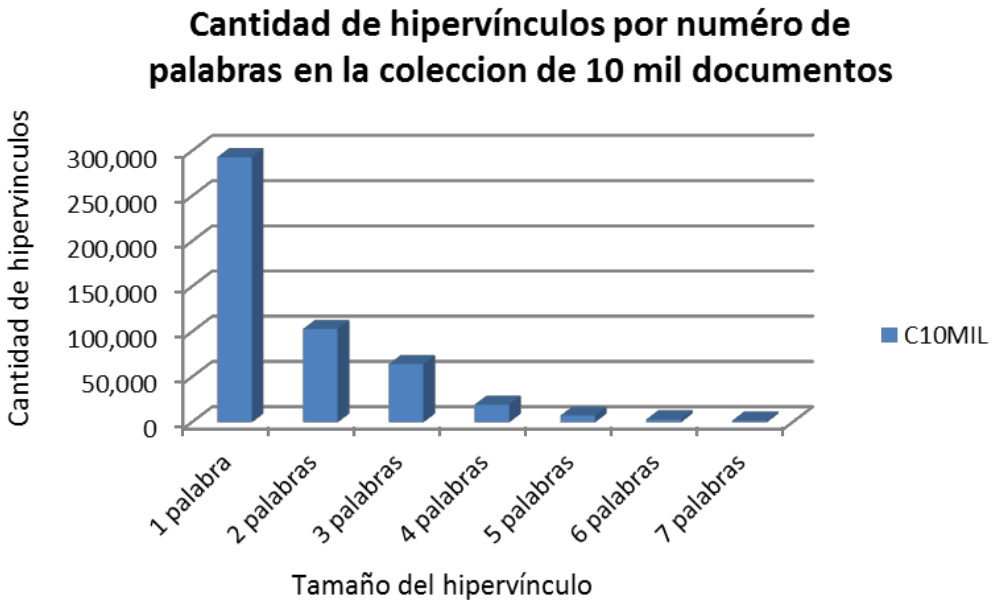

Figura 5. 1 Cantidad de hipervínculos por palabra en la colección C10MIL.

5.2 Minería del contenido de la WEB.

En esta fase se aplicó la herramienta de SFM [García.2007] a la colección minable donde los hipervínculos son de una palabra (la que se generó en la fase anterior) ya que es la más representativa de la colección de documentos C10MIL con 493,078 hipervínculos. Esto se realizó con dos umbrales de frecuencia mínima, uno del 0.1% (293) y otro del 1.0% (2,931) de acuerdo a la cantidad de documentos en la colección para minar.

En la figura 5. 2 se visualiza la cantidad de patrones léxicos recuperados con umbral del 0.1% y del 1.0%, donde a menor umbral se recupera una mayor cantidad de patrones léxicos. En el apéndice B se pueden visualizar los patrones léxicos de está colección con umbral 1.0%.

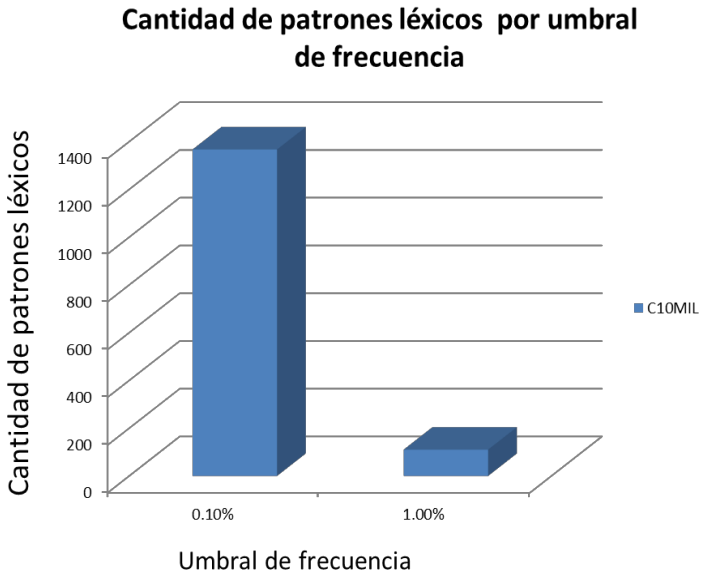

Figura 5. 2 Cantidad de patrones léxicos con dos umbrales frecuencia mínima

5.2 Fase de evaluación e interpretación

En esta fase se realizaron experimentos con la colecciones C10MIL, haciendo evaluaciones con las medidas de precisión, recuerdo y F-measure, a los patrones obtenidos con C500A y C500B.

La organización de los experimentos es la siguiente: En un primer acercamiento a la experimentación de forma cualitativa se buscaron los patrones léxicos de una palabra. En un segundo experimento se muesta una evaluación de manera individual de los patrones. Después se realizó una evaluación de los patrones léxicos en su conjunto sin repetición, siendo este el tercer experimento. Por último en un cuarto experimento se hizó la evaluación de los patrones léxicos en su conjunto con repetición.

5.2.1. Primer acercamiento a la experimentación de forma cualitativa.

Con la colección de hipervínculos de una palabra y utilizando el umbral de 0.1% se encontraron en el método propuesto 163 hipervínculos. En la tabla 5.1 se presentan los 24 mejores patrones según su F-measure.

Un analisis de esto es:

- Por ejemplo, el patrón "(<@LINK>)" significa que todo aquello que está entre paréntesis de una palabra debe ser un hipervínculo.
- Por ejemplo, el patrón de búsqueda " DE <@LINK> ," recuperaría fragmentos de texto de una palabra donde esa palabra tiene la palabra "DE" como contexto derecho y la palabra "," como contexto izquierdo.
- En general se puede ver que los patrones dependen del contexto, pero este no depende del dominio ya que solo intervienen preposiciones y signos de puntuación.

Tabla 5. 1 Patrones léxicos de una palabra en el hipervínculo.

(<@LINK>)	DE <@LINK> ,	DE LA <@LINK> ,
, <@LINK> ,	DE <@LINK> , EL	DEL <@LINK> ,
, <@LINK> .	DE <@LINK> .	EN <@LINK> ,
, <@LINK> DE	DE <@LINK> A	EN <@LINK> .
, <@LINK> Y	DE <@LINK> DE	EN <@LINK> Y
A <@LINK> ,	DE <@LINK> EN	LA <@LINK> DE
A <@LINK> .	DE <@LINK> SE	Y <@LINK> ,
DE <@LINK> (	DE <@LINK> Y	Y <@LINK> .

En la figura 5.3 se visualiza un ejemplo de la aplicación de los patrones léxicos encontrados en un parrafo de Wikipedia, donde en **negritas** se identifican los hipervinculos hechos por el humano. En ***negritas y cursivas*** se identifican los hipervinculos hechos por el humano pero que no tienen destino (conocidos como hipevínculos rojos). Los fragmentos de textos obtenidos por el método propuesto son identificados con color gris de fondo.

El **10 de octubre** de **1637** la ciudad de **Breda**, tras diez meses de asedio, fue tomada de nuevo por el príncipe de Orange, **Federico Enrique de Nassau**, tras permanecer bajo control español durante doce años. Pese a los numerosos intentos del Cardenal-Infante fue imposible volver a adueñarse de esta fortificación estratégica. Fernando de Austria también perdió, frente a los franceses, ***Chapelles***, ***Landrey*** y ***Damvilliers***. No pudo conquistar **Maubeuge**, y este proceso supuso pérdidas territoriales frente a Francia. Fernando sí fue capaz de tomar **Amberes**, ***Chastillon*** y **Geldern**, pero en cambio **perdió Arras** en **1640**.

Figura 5. 3 Ejemplo de hipervínculos hechos por el humano y fragmentos de texto candidato a hipervínculo.

Como se puede ver en la figura 5.3 se muestra como el método propone hipervinculos que no había como ciudad, asedio y Orange. También propone hipervínculos que coinciden con el humano como **Octubre**, **Breda**, **Nassau**, **Geldern** y **1940**. Aún más interesante es ver como el método permite encontrar hipervínculos rojos al igual que el humano como ***Chapelles***, ***Landrey***, ***Damvilliers*** y ***Chastillon***; lo cual no sucedería con los métodos del estado del arte basados en el título.

Cabe mencionar que los casos de ciudad, asedio y Orange son considerados en la evaluación como un error por no parecerse al humano. Sin embargo desde nuestro punto de vista el caso de Orange parece no ser un error.

La pregunta que surge es ¿si la calidad de los patrones depende del umbral de frecuencia?

5.2.2 Segundo experimento para evaluar de manera individual los patrones

Para ver qué tan buenos o malos son los patrones léxicos recuperados con la técnica de SFM se hace la evaluación de los patrones léxicos recuperados de la colección C10MIL con una

palabra en el hipervínculo, variando el umbral de frecuencia, aplicándolos a una colección de 500 documentos (C500A).

En la figura 5.4 se ven los 24 mejores patrones léxicos recuperados de una palabra en el hipervínculo, de un total de 163 y en la gráfica se ve que los 24 patrones son robustos porque aunque se varía el umbral los patrones persisten.

Otras características que se ven en la figura 5.4 de los 24 mejores patrones según su F-measure son:

- A menor umbral (175) aparecen mayor cantidad de patrones léxicos, conforme se aumenta el umbral aparecen o desaparecen patrones y al tener un umbral muy alto (12,000) es menor la cantidad de patrones léxicos.
- Los patrones léxicos con mejor F-measure y que se mantienen hasta un umbral de 12,000 apariciones como frecuencia minima son: “ DE <@LINK> . ” , “ , <@LINK> , ” y “ EN <@LINK> , “, es decir son patrones muy persistentes.
- En el umbral 1,400 se obtienen la mayor cantidad de patrones léxicos con la mejor medida de F-measure.
- Hay patrones como el de “, <@LINK> y “ que aparece en el umbral 125 y desaparece en el umbral 6000 sin cambiar su contexto.

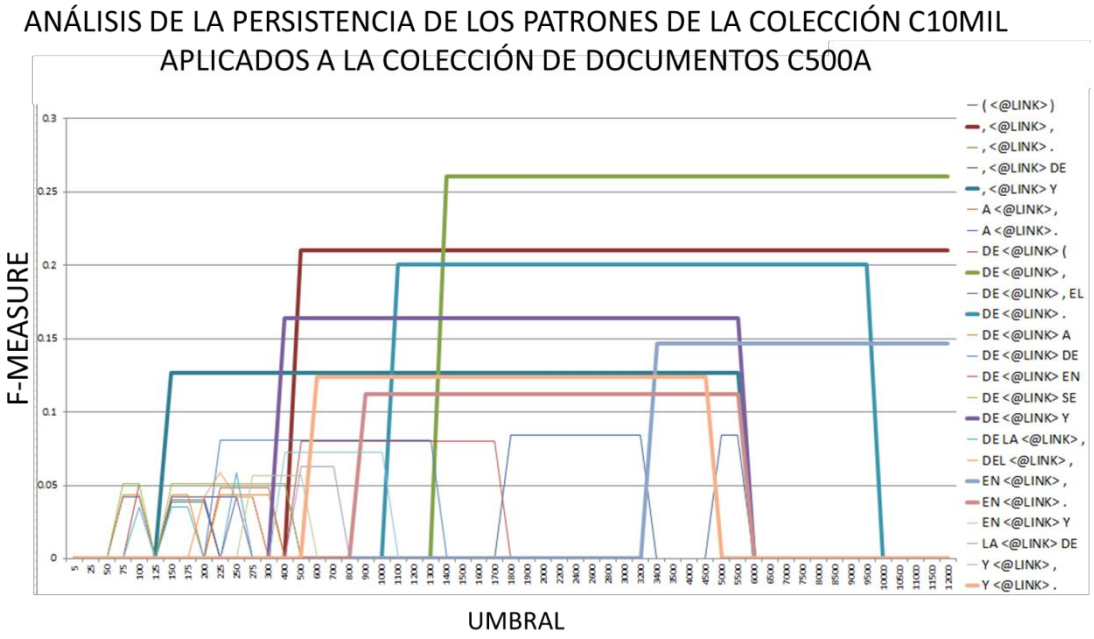

Figura 5. 4 Patrones extraídos de la colección C10mil y evaluados individualmente en C500A variando el umbral.

Con el experimento anterior fue posible saber la calidad de cada patrón. Sin embargo, para saber de manera general qué tan bien funciona el método propuesto para una colección de documentos se utiliza la colección C500B a la cual se le aplicaron los mejores patrones.

5.2.3 Tercer experimento para evaluar los patrones léxicos en su conjunto sin repetición.

En esta evaluación, tanto la colección de hipervínculos hechos por el humano como los hipervínculo automáticos, no pueden contener elementos repetidos.

En la figura 5.5 se graficaron las medidas de precisión, recuerdo y F-measure, que se obtuvieron al aplicar los patrones léxicos obtenidos de la colección C10MIL en la colección C500B; variando el umbral. La mejor medida F-measure se encuentra en el umbral 1800 siendo de **0.41.** Esta gráfica muestra como con el umbral 1400 se recupera muchos hipervínculos que posiblemente algunos sean error, ya que el recuerdo está cerca de 0.58 y la precisión está debajo de 0.31, decrementando la evaluación final del F-measure. La gráfica 5.5 permite conocer que si se quiere mejores fragmentos de texto hay que moverse al umbral 1800 donde el F-measure tiene el valor de 0.41. Recordemos que esta gráfica es para el total de patrones léxicos de la colección C10MIL aplicados a la colección C500B.

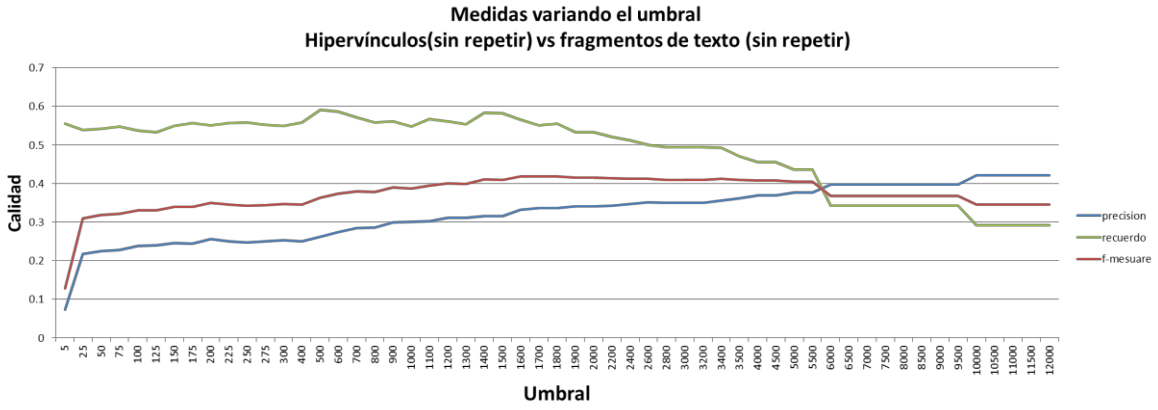

Figura 5. 5 Medidas variando el umbral en la colección C500B.

En esta segunda evaluación se van a tomar los mejores patrones de la colección C10MIL con el umbral de 500 y de 1400, porque es al parecer hay recuerdo más alto de acuerdo al F-measure.

En la figura 5.6 se graficaron las medidas de precisión, recuerdo y F-measure, analizando en el umbral 500, variando la cantidad de mejores patrones léxicos; según su recuerdo y según

su precisión. Con el umbral 500 no se mejoró el F-measure que fue de 0.39 seleccionando los mejores 70 patrones.

En la siguiente gráfica se usa la notación de 500-P25MP para especificar que en el umbral 500 se seleccionaron los mejores 25 patrones de acuerdo a la precisión o la notación de 500-R70MP para especificar que en el umbral 500 se seleccionaron los mejores 70 patrones de acuerdo al recuerdo.

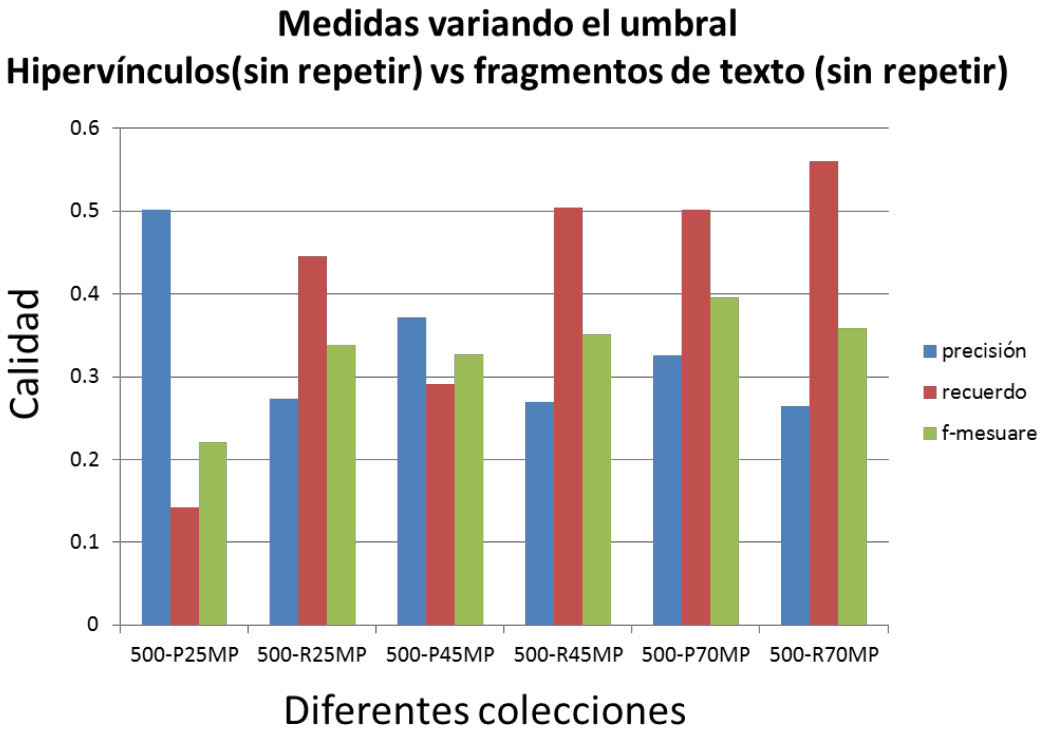

Figura 5. 6 Medidas variando la cantidad de patrones con umbral 500

En la figura 5.7 se graficaron la medidas de precisión, recuerdo y F-measure con umbral 1400 variando la cantidad de mejores patrones según su recuerdo y según su precisión de la colección C500A. Obteniendo el mejor F-measure (0.41) en el umbral 1400 con los mejores 20 patrones.

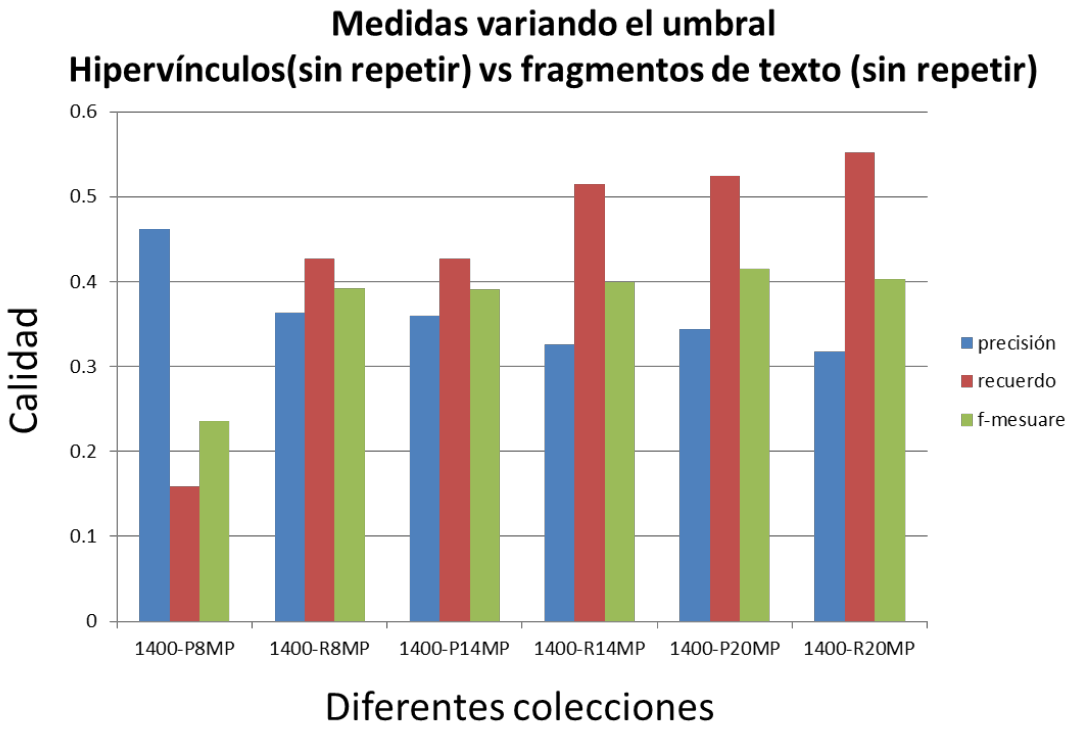

Figura 5. 7 Medidas variando el umbral y la cantidad de patrones.

La mejor F-measure en esta evaluación con repetición se encuentra en el umbral 1400 con todos los patrones y la medida **0.41**, al igual que con sólo los 20 mejores patrones de la colección C500A con umbral de 1400.

5.2.4 Cuarto experimento para evaluar los patrones léxicos en su conjunto con repetición.

En esta evaluación, tanto la colección de hipervínculos hechos por el humano como los hipervínculo automáticos, pueden contener elementos repetidos.

En la figura 5.8 se graficaron las medidas de precisión, recuerdo y f-measure, variando el umbral. La mejor medida F-measure se encuentra en el umbral 1400 siendo de **0.48**. Esta gráfica muestra como con el umbral 50 se recupera muchos hipervínculos, que posiblemente algunos sean error ya que el recuerdo está cerca de 0.9 y la precisión está debajo de 0.3, decrementando la evaluación final de F-measure. La gráfica 5.8 permite reconocer que si se quiere mejores fragmentos de texto hay que moverse al umbral 1400. Recordemos que esta gráfica es para el total de patrones léxicos de una palabra de la colección C10MIL.

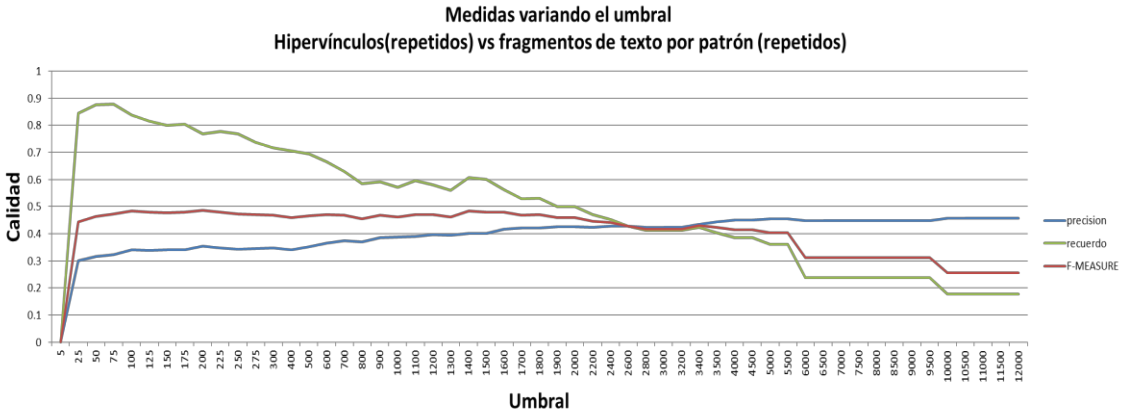

Figura 5. 8 Medidas variando el umbral en colección C500B.

Otro experimento consiste en no tomar todos los patrones de la colección C10MIL, sino seleccionar los mejores patrones léxicos con el umbral 75 por que es donde al parecer hay recuerdo más alto de acuerdo la medida F-measure.

En la figura 5.9 se graficaron la medidas de precisión, recuerdo y F-measure, analizando en el umbral 75 de aplicar los patrones léxicos de la colección C10MIL en la colección C500B, variando la cantidad de mejores patrones léxicos; según su recuerdo y según su precisión en

la colección C500A. Con el umbral 75 se mejoró un poco el F-measure con 0.50 seleccionando los mejores 500 patrones.

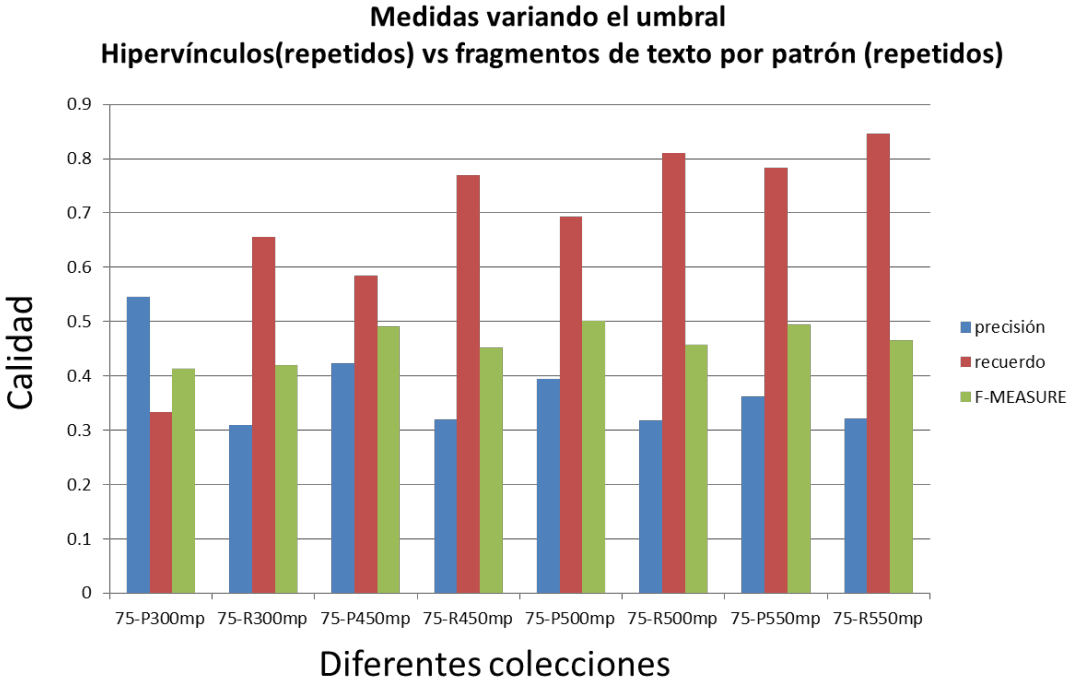

Figura 5. 9 Medidas variando la cantidad de patrones con umbral 75.

En la figura 5.10 se graficaron las medidas de precisión, recuerdo y F-measure, analizando en el umbral 200, variando la cantidad de mejores patrones léxicos; según su recuerdo y según su precisión. Con el umbral de 200 se mejoró un poco el F-measure con 0.51 seleccionando los mejores 210 patrones.

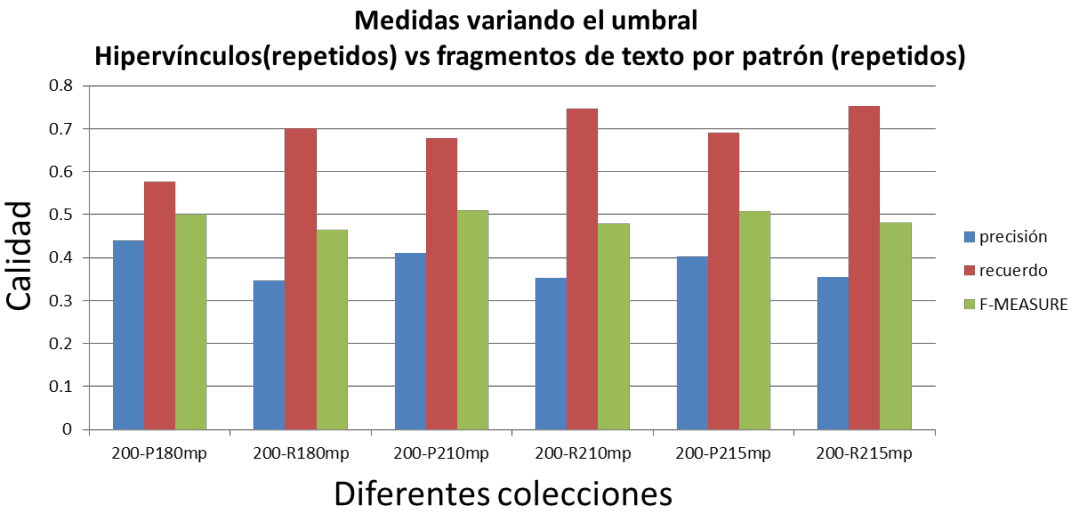

Figura 5. 10 Medidas variando la cantidad de patrones con umbral 200.

En la figura 5.11 se graficaron la medidas de precisión, recuerdo y F-measure, analizando en el umbral 1400, variando la cantidad de mejores patrones léxicos; según su recuerdo y según su precisión. Con el umbral de 1400 no se mejoró F-measure.

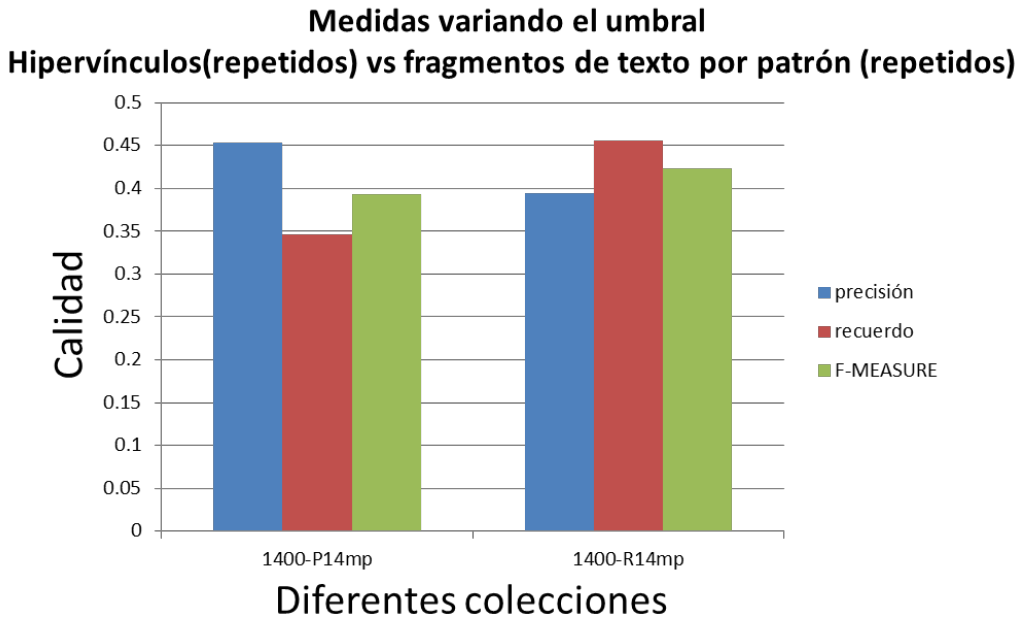

Figura 5. 11 Medidas variando la cantidad de patrones con umbral 1400.

Después de estas 4 graficas se resume que en esta tercera evaluación la mejor F-measure se encuentra en el umbral 200 con los mejores 210 patrones y F-measure es igual a **0.51**.

5.6 Resumen

La métrica F-measure ha sido de gran utilidad en la experimentación, ya que de acuerdo a los resultados es posible detectar fragmentos de texto candidato a hipervínculo a partir de patrones léxicos. Con base a la primera evaluación se considera posible hacer la evaluación para hipervínculos de 2 o más palabras como se realizó esta.

Capítulo 6

Conclusiones

Dado que el propósito de esta investigación era descubrir conocimiento novedoso y útil en la construcción de hipervínculos elaborados por el humano, identificando patrones de búsqueda, para así localizar fragmentos de texto candidato a tener hipervínculo en una colección de documentos en texto plano, se propuso un método basado en el proceso KDT que consiste de 5 fases; donde una vez hecha la integración y recopilación de documentos de Wikipedia 2008 en español, se realizó la selección, limpieza y transformación de los documentos para normalizar los datos que serían minados en la fase de Minería del contenido de la WEB. El proceso de minería se llevó acabo por medio de la herramienta de SFM y la evaluación e interpretación de los datos se realizó con las métricas de precisión, recuerdo y F-measure. Descubriendo así conocimiento que consistió en identificar algunos patrones léxicos en los hipervínculos creados por el humano y que al transformarlos en

patrones de búsqueda, fue posible detectar fragmentos de texto candidato a hipervínculo en texto plano, lo cual era el objetivo de esta investigación y donde la fase de difusión y uso se queda como trabajo futuro y que consistirá de utilizar los fragmentos de texto en la construcción automática de hipervínculos.

- El contexto ayuda a identificar texto candidato a hipervínculo.
- Las palabras vacías y los signos de puntuación son de utilidad en la identificación de texto candidato a hipervínculo.
- Al etiquetar los hipervínculos por número de palabras, los patrones léxicos son clasificados por número de palabras también, y la identificación de palabras candidatas a hipervínculos es más confiable.
- Al trabajar con patrones léxicos, es posible trabajar el método en otro dominio y en un lenguaje diferente al español.
- Los patrones léxicos que son menos óptimos en la identificación de texto candidato a hipervínculos son los que contienen más de tres palabras como candidatas a hipervincular.

Trabajo futuro

- Identificar patrones léxicos en colecciones mayores de documentos como por ejemplo 100 mil, o en los 669 mil documentos de la colección de Wikipedia 2008 en español, para tener mayor certidumbre de que los patrones de búsqueda identificados son los mejores en la detección de fragmentos de texto candidato a hipervínculo.
- Hacer mejoras en el código de los procesos para optimizar tiempos y así poder procesar una gran cantidad de documentos máquina.
- Realizar la minería del contenido de la WEB variando el gap en el método ya que sólo se trabajó con gap=0 en la técnica de secuencias frecuentes maximales, para así ver si es posible identificar mejores patrones de búsqueda de fragmentos de texto candidato a hipervínculo.

- Realizar la minería del contenido de la WEB con STEAMING, para ver si con esto se mejoran los patrones de búsqueda.
- Probar el método en colecciones de otros dominios para ver si los patrones recuperados son similares a los ya encontrados.
- Probar el método en colecciones de otros lenguajes para ver si los patrones recuperados son similares a los ya encontrados o si es posible sólo hacer la traducción de palabras del español a otro lenguaje.
- Hacer agrupación de patrones con contexto parecido (agrupamiento de instancias) como por ejemplo para los patrones “, <@LINK> , “, “ , <@LINK> <@LINK> , “, “, <@LINK> <@LINK> <@LINK> ,” podrían generalizarse en una expresión regular como “, <@LINK> {1,3} , “.

Apéndices

A. Expresiones regulares

Las expresiones regulares constituyen un mecanismo bastante potente para realizar manipulaciones de cadenas de texto [Friedl.2006].

En la tabla que sigue se muestran los caracteres comodín usado para crear los patrones y su significado, junto a un pequeño ejemplo de su utilización.

	Significado	**Ejemplo**	**Resultado**
****	Marca de carácter especial	/\$ftp/	Busca la palabra $ftp
^	Comienzo de una línea	/^-/	Líneas que comienzan por -
$	Final de una línea	/s$/	Líneas que terminan por s
.	Cualquier carácter	/\b.\b/	Palabras de una sola letra

	(menos salto de línea)		
|	Indica opciones	/(L|l|f|)ocal/	Busca Local, local, focal
( )	Agrupar caracteres	/(vocal)/	Busca vocal
[ ]	Conjunto de caracteres opcionales	/escrib[aoe]/	Vale escriba, escribo, escribe

La tabla que sigue describe los modificadores que pueden usarse con los caracteres que forman el patrón. Cada modificador actúa sobre el carácter o el paréntesis inmediatamente anterior.

	Descripción	**Ejemplo**	**Resultado**
*****	Repetir 0 o más veces	/l*234/	Valen 234, 1234, 11234...
+	Repetir 1 o más veces	/a*mar/	Valen amar, aamar, aaamar...
?	1 o 0 veces	/a?mar/	Valen amar, mar.
{n}	Exactamente n veces	/p{2}sado/	Vale ppsado
{n,}	Al menos n veces	/(m){2}ala/	Vale mmala, mmmala....
{m,n}	entre m y n veces	/tal{1,3}a/	Vale tala, talla, tallla

Los siguientes son caracteres especiales para indicar caracteres de texto no imprimibles, como puedan ser el fin de línea o un tabulador, o grupos predefinidos de caracteres (alfabéticos, numéricos, etc...)

	Significado	**Ejemplos**	**Resultado**
\b	Principio o final de palabra	/\bver\b/	Encuentra ver en "*ver de*", pero no en "*verde*"

\B	Frontera entre no-palabras	/\Bver\B/	Empareja ver con *"Valverde"* pero no con *"verde"*
\d	Un dígito	/[A-Z]\d/	No falla en "A4"
\D	Alfabético (no dígito)	/[A-Z]\D/	Fallaría en "A4"
\0	Carácter nulo		
\t	Caracter ASCII 9 (tabulador)		
\f	Salto de página		
\n	Salto de línea		
\w	Cualquier alfanumérico, [a-zA-Z0-9_]	/\w+/	Encuentra *frase* en "*frase*.", pero no el . (punto).
\W	Opuesto a \w ([^a-zA-Z0-9_])	/\W/	Hallaría sólo el punto (.)
\s	Carácter tipo espacio (como tab)	/\sSi\s/	Encuentra *Si* en *"Digo Si ",* pero no en *"Digo Sientate"*
\S	Opuesto a \s		
\cX	Carácter de control X	\c9	El tabulador
\oNN	Carácter octal NN		
\xhh	El hexadecimal hh	/\x41/	Encuentra la *A* (ASCII Hex41) en *"letra A"*

B. Patrones léxicos en la colección C10MIL de Wikipedia 2008 en español con umbral de frecuencia mínima del 1.0%.

DE <@LINK>,

LA <@LINK>,

EL <@LINK> ,

. EN <@LINK> ,

, <@LINK> ,

DE <@LINK> .

EN <@LINK> .

(<@LINK>)

, <@LINK> Y

DE <@LINK> Y

Y <@LINK> .

Y <@LINK> ,

, <@LINK> <@LINK> ,

Y <@LINK> <@LINK> .

DE <@LINK> <@LINK> Y

(<@LINK> <@LINK>)

Y <@LINK> <@LINK> ,

DE LA <@LINK> <@LINK> ,

DE <@LINK> <@LINK> ,

DE <@LINK> <@LINK> .

EL <@LINK> <@LINK> ,

LA <@LINK> <@LINK> .

DEL <@LINK> <@LINK> ,

, <@LINK> <@LINK> Y

DEL <@LINK> <@LINK> .

EL <@LINK> <@LINK> .

. EL <@LINK> <@LINK> <@LINK> DE

DE <@LINK> <@LINK> <@LINK> .

DE LA <@LINK> <@LINK> <@LINK> .

LA <@LINK> <@LINK> <@LINK> Y

EL <@LINK> <@LINK> <@LINK> Y

EN LA <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> (

DEL <@LINK> <@LINK> <@LINK> .

EL <@LINK> <@LINK> <@LINK> ,

, <@LINK> <@LINK> <@LINK> ,

DE <@LINK> <@LINK> <@LINK> ,

EL <@LINK> <@LINK> <@LINK> .

DE LA <@LINK> <@LINK> <@LINK> ,

, <@LINK> <@LINK> <@LINK> DE

, EL <@LINK> <@LINK> <@LINK> DE

DEL <@LINK> <@LINK> <@LINK> ,

, <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> Y

EN LA <@LINK> <@LINK> <@LINK> <@LINK> ,

EN <@LINK> <@LINK> <@LINK> <@LINK> .

LA <@LINK> <@LINK> <@LINK> <@LINK> EN

DEL <@LINK> <@LINK> <@LINK> <@LINK> .

EN LA <@LINK> <@LINK> <@LINK> <@LINK> .

, <@LINK> <@LINK> <@LINK> <@LINK> Y

Y <@LINK> <@LINK> <@LINK> <@LINK> .

(<@LINK> <@LINK> <@LINK> <@LINK>)

DE <@LINK> <@LINK> <@LINK> <@LINK> ,

EL <@LINK> <@LINK> <@LINK> <@LINK> ,

DE LA <@LINK> <@LINK> <@LINK> <@LINK> ,

EL <@LINK> <@LINK> <@LINK> <@LINK> .

LA <@LINK> <@LINK> <@LINK> <@LINK> (

DE LA <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> .

DEL <@LINK> <@LINK> <@LINK> <@LINK> ,

, <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> Y

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> Y

EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> (

EN LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

EN LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

EN EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> (

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> EN

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

EN EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

EN LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

(<@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK>)

EN EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

EN LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> DE

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> (

, <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK>.

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> Y

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> EN

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

EN EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

<@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

, EN EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> Y CANTON DE

EN LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> DE

EL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DEL <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

, <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

DE LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> .

DE <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> ,

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> (

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> , LA

LA <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> <@LINK> Y

Referencias

[Chakrabarti.2003] Chakrabarti Soumen. “Mining the WEB”. Morgan Kaufmann Publishers. 2003.

[Denicia.2006] Denicia Claudia, Montes Manuel, Villaseñor Luis, García Rene. “A Text Mining Approach for Definiting Question Answering”. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México. 2006.

[Denicia.2007] Denicia Carral María Claudia. “Respondiendo Preguntas de Definición Mediante el Descubrimiento de Patrones Léxicos. Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México.2007.

[Fayyad.1996] Fayyad Usuma, Piatetsky-shapiro Gregory, and Smyth Padhraic. “From data minning to knowledge discovery in databases”.

American Association of Artificial Intelligence. 1996.

[Feldman.2007] Feldman Ronen. “The Text Mining Handbook”. Cambridge Univerisy Press. 2007.

[Fiedl.2006] Friedl Jeffrey E. F. “Mastering Regular Expressions”. O’REILLY. 2006.

[García.2003] García Hernández René Arnulfo. “Modelado Gramatical de un Subconjunto del Lenguaje Español”. Tesis de Maestría. CENIDET. Morelos. México. 2003.

[García.2007] García Hernández René Arnulfo. “Desarrollo de Algoritmos para el Descubrimiento de Patrones Secuenciales Maximales. INAOE. Puebla, México. 2007.

[Gelbukh.2010] Gelbukh Alexander, Sidorov Grigori. “Procesamiento Automático del Español con Enfoque en Recursos Léxicos Grandes”. IPN. 2010.

[Gross.2007] Gross Oskar, Doucet Antoine, and Toivonen Hannu, “Named Entity Filtering Based on Concept Association Graphs”. Department of Computer Science P. O. Box 68, University of Helsinki Finland. CICLing 2007

[Hernández.2007] Hernández Orallo José. “Introducción a la Minería De Datos”. PRENTICE HALL. 2007.

[Liu.2007] Liu Xitong and Fang Hui. “Entity Profile based Approach in Automatic Knowledge Finding”. University of Delaware Newark, DE, USA. TREC

2007

[Merca20.2013] Mercadotecnia publicidad medios. Merca2.0. 2013. http://www.merca20.com/los-10-periodicos-mas-influyentes-en-mexico/

[Milhacea.2007] Milhacea Rada and Csomai Andras. “Wikify! Linking documents to Encyclopedic Knowledge”. Department of Computer Science; University of North Texas, Texas; EUA. Association for Computing Machinery (ACM) Conference on Information and Knowledge Management (CIKM), 2007, Lisbon, Portugal.

[Netcraft.2015] Netcraft,. January 2015 Web Server Survey, Reino Unido, 2015, http://news.netcraft.com/archives/web_server_survey.html

[Orta.2008] Orta Palacios Claudia Patricia. “Métodos Basados en Patrones Léxicos para la Extracción de Información”, Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México. 2008.

[Ortega.2007] Ortega Mendoza Rosa María. “Descubrimiento automático de hipónimos a partir de texto no estructurado”, Tesis de Maestría. Instituto Nacional de Astrofísica, Óptica y Electrónica. Puebla, México. 2007.

[Osorio.2013] Osorio Juan Carlos. “Guía práctica de XHTML, JAVASCRIPT Y CSS”. Alfaomega. 2013.

[Rao.2007] Rao Delip, McNamee Paul, and Dredze Mark. “Entity Linking: Finding Extracted Entities in a Knowledge Base”. Department of Computer Science, Johns Hopkins University. Springer-Verlag Berlin Heidelberg. 2007

[Rfc.1896] IETF. “The text/enriched MIME content-type”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Rfc1341] IETF. “MIME (Multipurpose Internet Mail Extensions)”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Rfc1945] IETF. “Hypertext Transfer Protocol 1.0”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Rfc1983] IETF. “Internet Users Glossary”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Rfc2616] IETF. “Hypertext Transfer Protocol 1.1”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Rfc2774] IETF. “An HTTP Extensión Framework”. The Internet Engineering Task Force. 2014. http://www.ietf.org

[Scott.2013] Scott Peña Patricia. “Internet”. Anaya. 2013.

[Wei.2007] Wei Che Huang Darren, Xu Yue, Trotman Andrew, and Geva Shlomo. “Overview of INEX 2007 Link the Wiki Track”. Faculty of Information Technology, Queensland University of Technology, Brisbane

Queensland Australia. Springer-Verlag Berlin Heidelberg. 2007.

[WikiEn.2015] Wikipedia Organization. USA. 2015.
http://en.wikipedia.org/wiki/Wikipedia

[WikiEs.2013] Wikipedia Organization. USA. 2013.
http://es.wikipedia.org/wiki/Hiperenlace

[Witten.2011] Witten Ian H. “Data Minning”. MORGAN KAUFMANN. 2011.