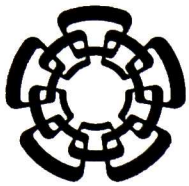
UT-00092-EE1

Tor - 2016



CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS
DEL INSTITUTO POLITÉCNICO NACIONAL

Laboratorio de Tecnologías de Información,
CINVESTAV-Tamaulipas

Método de enriquecimiento de texto a partir de recursos de la Web Semántica

Tesis que presenta:

Dishelt Francisco Torres Paz

Para obtener el grado de:

Maestro en Ciencias en Computación

Director de la Tesis:
Dr. Iván López Arévalo

Cd. Victoria, Tamaulipas, México — Octubre, 2015

226752 - 1001

La tesis presentada por Dishelt Francisco Torres Paz fue aprobada por:

Dr. Hiram Galeana Zapién

Dr. Víctor Jesús Sosa Sosa

Dr. Iván López Arévalo, Director

Cd. Victoria, Tamaulipas, México, 21 de Octubre de 2015

A persona/personas

Agradecimientos

- A *Dios*, por el regalo de la vida, por rodearme de seres extraordinarios, por llenarme de bendiciones y haberme permitido alcanzar otro reto en mi vida profesional.
- Agradezco a mi familia, mis padres, mi hermano y Alejandra. Por acompañarme y apoyarme en cada decisión que he tomado en mi vida. Gracias por su amor, confianza y paciencia que me brindan para enfrentar cada reto que se me presenta. Por que son ustedes mi principal motivación.
- Agradezco al Dr. Iván López Arévalo por su paciencia, guía y cada uno de los consejos que me brindo durante el desarrollo de este trabajo. Mi admiración, respeto y reconocimiento al Dr. Iván.
- Gracias al cuerpo académico y administrativo del CINVESTAV, en especial a mis revisores, Dr. Víctor Jesús Sosa Sosa y Dr. Hiram Galeana Zapién por sus comentarios y críticas para enriquecer este trabajo de tesis.
- Agradezco a mis amigos y compañeros de maestría por brindarme su amistad, por todos esos momentos que compartimos.
- Le doy gracias a CONACyT por el apoyo económico brindado para realizar esta maestría

Índice General

Índice de Figuras

Índice de Tablas

Resumen

Método de enriquecimiento de texto a partir de recursos de la Web Semántica

por

Dishelt Francisco Torres Paz

Laboratorio de Tecnologías de Información, CINVESTAV-Tamaulipas
Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2015
Dr. Iván López Arévalo, Director

Actualmente con el incremento exponencial de la información textual que se genera día a día es cada vez más difícil transformar esa información en un activo útil. El procesamiento de dicha información se dificulta debido a que no tiene una estructura definida y por el poco conocimiento que se tiene sobre el texto. Al encontrarse la información de esa forma, el conocimiento obtenido es poco o inexistente. Una solución a la falta de conocimiento y el poco entendimiento es el enriquecimiento de texto. El enriquecimiento de texto provee conocimiento e información relacionada con el contenido del texto por medio de recursos externos. Existen diferentes maneras para enriquecer un texto, como obtener conocimiento desde definiciones hasta de bases de conocimiento semánticas. En este trabajo se presenta un método de enriquecimiento de texto a partir de recursos de la Web Semántica. El método básicamente sigue tres etapas. La primer etapa obtiene las partes del texto a enriquecer, identifica Entidades Nombradas contenidas en el mismo así como Entidades Nombradas relacionadas con el tema principal. La segunda etapa obtiene información relacionada acerca de las Entidades Nombradas identificadas. Dicha información se obtiene desde la base de conocimiento DBpedia. El tercer paso es la integración de la información obtenida con el texto original, dicha integración se denomina *texto enriquecido*. Para evaluar el desempeño del método de enriquecimiento se realizaron diversos experimentos utilizando los *datasets Reuters* y *20 Newsgroups*. El primer experimento es una evaluación de forma manual en el que se examina la información relacionada mientras que el segundo experimento muestra una comparación entre el texto original y enriquecido. En otro experimento se evaluó la Ganancia de Información (GI) en los *datasets* originales y enriquecidos. Por último se evaluó

el desempeño de tareas de clasificación y *clustering* de texto utilizando los mismos *datasets*. El diseño, implementación y evaluación del método de enriquecimiento muestra que éste identifica relaciones de DBpedia, la cual provee información relacionada al texto obteniendo un mayor conocimiento y entendimiento acerca del mismo.

Abstract

A Text Enrichment Method based in Semantic Web Resources

by

Dishelt Francisco Torres Paz
Information Technology Laboratory, CINVESTAV-Tamaulipas
Research Center for Advanced Study from the National Polytechnic Institute, 2015
Dr. Iván López Arévalo, Advisor

Nowadays the exponential growth of textual information makes difficult transforming that information in a helpful resource. The information processing is a difficult work because the lack of structure and knowledge about it. A solution to process text without knowledge and understanding is text enrichment. The text enrichment provides knowledge and extra information related with the text using external resources. There are many different ways to enrich a text, from definitions to semantic knowledge bases. This work presents a text enrichment method using Semantic Web resources, which basically follows three stages. The first stage gets text sections to identify Named Entities related with the main topic of the text. The second stage gets related information about the Named Entities previously identified. The information is extracted from the semantic knowledge base DBpedia. The third stage is the integration of the retrieved information with the original text, such integration is known as *enriched text*. The evaluation was performed through several experiments, which used the *Reuters* and *20 Newsgroups* datasets. The first experiment evaluates in a manual way the information retrieved by the text enrichment method. In the same way an experiment was performed comparing the original text and the enriched text. Another experiment evaluates the Information Gain (IG) in the original and enriched datasets. Finally, the performance in classification and clustering of text is shown in two different experiments. The design, implementation and evaluation show the text enrichment method identifies relations to DBpedia, which provides related information getting a better knowledge and understanding about the text.

1

Introducción

En este capítulo se presenta de manera general el contexto de la investigación dando una breve introducción a las principales disciplinas en que se sitúa esta propuesta. Además se dan las razones que motivan la realización del trabajo de tesis. Enseguida se menciona el planteamiento del problema que se desea resolver, la hipótesis y la pregunta de investigación. Asimismo, se establece el objetivo general y los objetivos específicos que indican el alcance de la propuesta, responder a la hipótesis formulada así como resolver el problema planteado. Al final del capítulo se menciona la estructura del documento de tesis.

1.1 Descripción del problema

Durante los últimos años se ha incrementado la generación de información almacenada en distintos repositorios. Dicha información se encuentra en constante actualización debido a la necesidad de comunicación y expresión que tienen los seres humanos por naturaleza. La información almacenada se encuentra en una forma difícil de procesar, ya que normalmente es texto plano. Para procesar texto

plano se utilizan diversas técnicas de minería de texto, sin embargo, el desempeño del procesamiento de texto muchas veces no es el esperado, cuando no se tiene suficiente información del texto en cuestión es necesario conocer más acerca del mismo. Cuando se obtiene información (de diferentes repositorios) relacionada con el texto en cuestión éste se puede enriquecer. El enriquecimiento de texto provee conocimiento relacionado a su contenido, por ejemplo: texto relacionado, enlaces a páginas Web, imágenes, mapas, etc. Así, un texto enriquecido con conocimiento relacionado es más sencillo de entender y procesar ya que contiene información y conocimiento de apoyo para aumentar el desempeño de diferentes tareas de minería de texto. El enriquecimiento de texto puede ser implementado de diferentes maneras que afectan directamente a la eficiencia del método de enriquecimiento, por consecuencia, a las tareas de minería de texto y el conocimiento proporcionado al usuario final. Una forma de enriquecer el texto es aprovechar las bases de conocimiento que se encuentren estructuradas con base en la semántica de su contenido. Dichas bases de conocimiento puedan estar estructuradas basándose en los principios de la Web Semántica. La Web Semántica intenta conectar la información (no sólo en la Web, si no a todas las cosas) para proporcionarle significado, con el propósito de mejorar la interacción humano-computadora.

En vista de lo expuesto anteriormente, se plantea la siguiente pregunta de investigación: **¿Cómo se puede, dado un texto plano, obtener conocimiento e información relacionada a partir de bases de conocimiento estructuradas semánticamente?**

1.2 Motivación

A través de los años se ha observado un crecimiento exponencial de la información que diversas entidades pueden generar. Dicha información se encuentra en su mayoría de una manera no estructurada y sin organización alguna, desde su contenido hasta la manera de consultarla. La estructura y organización de la información hace realmente difícil explotar los recursos, principalmente porque está destinada para consumo humano.

Por otro lado, la información que ha surgido a través de los años se procesa día con día para poder estructurarla, obtener un conocimiento de esa información o utilizarla para otro fin. Cuando se trabaja con texto plano es difícil su procesamiento, ya que éste carece de una estructura, de una forma y de un medio para explotar dicha información. Este tipo de limitaciones ha captado la atención de la comunidad científica para construir representaciones de conocimiento a partir de dicha información, con el objetivo de darle una estructura y organización. Las contribuciones y desarrollo en estos campos han sido principalmente en las áreas de Procesamiento del Lenguaje Natural, Minería de Texto, Extracción de Información y Representación de Conocimiento. Estas áreas de la Inteligencia Artificial han desarrollado diversos métodos para confrontar el problema del procesamiento del lenguaje natural en una computadora, además de proporcionar una manera de conectar la información oportunamente.

Con el crecimiento de la información que se genera día con día se ha pensado en procesar el texto con el objetivo de obtener un conocimiento mayor relacionado al mismo. El texto plano es difícil de procesar por lo que es necesario aplicar diversas técnicas de minería de texto, adquirir información y conocimiento adicional. Es decir, el enriquecimiento de texto parte de obtener un conocimiento adicional relacionado con el contenido del texto por medio de recursos o bases de conocimiento externas.

Existen diversas maneras de enriquecer un texto, desde obtener definiciones de su contenido hasta obtener conocimiento desde una ontología. En la literatura existen un número considerable de propuestas que utilizan una base de conocimiento en la que no se toma en cuenta la semántica del texto plano. Lo anterior genera un problema, ya que el enriquecimiento de texto se realiza a partir de términos del contenido del texto, dejando fuera su semántica, significado e información realmente relacionada. La propuesta que se propone en este trabajo de tesis es enriquecer un texto a partir de recursos de la Web Semántica [1], dado que se necesita un procesamiento del texto más complejo debido a la necesidad de obtener la semántica del mismo. Así, un texto enriquecido semánticamente provee la capacidad de obtener conocimiento semánticamente relacionado y con ello aumentar la

eficiencia de tareas de minería de texto, así como proporcionar dicho conocimiento al usuario final. Cuando se obtiene un texto enriquecido semánticamente, el texto puede ser utilizado para muchas aplicaciones, tales como proveer conocimiento al usuario final, almacenar información de acuerdo a su semántica o aplicar búsquedas de acuerdo a su significado.

1.3 Hipótesis

La hipótesis planteada en esta tesis es la siguiente:

Dado un texto, es posible enriquecerlo con conocimiento e información relacionada semánticamente a partir de bases de conocimiento de la Web Semántica.

1.4 Objetivos

Para la verificación de la hipótesis establecida en este trabajo se plantean los siguientes objetivos.

Objetivo general

Obtener un método de enriquecimiento de texto a partir de bases de conocimiento basadas en la Web Semántica.

Objetivos específicos

Con el fin de cumplir con el objetivo principal, a continuación se mencionan distintos objetivos específicos.

- Analizar diferentes técnicas de minería de texto para obtener una representación del texto a enriquecer.
- Determinar una forma de obtener conceptos relacionados para desarrollar el tema principal del texto original.

- Definir un mecanismo de extracción de conocimiento a partir de fuentes de la Web Semántica.
- Definir un mecanismo de integración de información a documentos existentes.

1.5 Estructura de la tesis

Este documento de tesis se compone de 5 capítulos, los cuales se encuentran estructurados de la siguiente manera: En el Capítulo 2 se presenta el marco teórico y el estado del arte relacionado con el trabajo de tesis. Se describe el concepto de Web Semántica, uso y aportaciones en la Web. Asimismo describe el rol que juegan las bases de conocimiento en la mejora de la inteligencia de la Web, haciendo especial énfasis en la base de conocimiento DBpedia. Por otra parte, se mencionan conceptos básicos de la minería de datos y minería de texto, las principales tareas y problemas que se encuentran en estas disciplinas. Finalmente, se mencionan conceptos relacionados con el enriquecimiento de texto, se presentan los diferentes enfoques reportados en la literatura enfatizando las principales aportaciones en el tema. En el Capítulo 3 se presenta el enfoque propuesto para el enriquecimiento de texto. Se describe el diseño y la implementación del método, los principales problemas que se encontraron y la manera de resolverlos. En este capítulo se muestran diversos esquemas y figuras para explicar el método de enriquecimiento: un diagrama general de bloques, esquemas del método y figuras que describen procesos específicos. En el Capítulo 4 se describen diversos escenarios de pruebas así como los diferentes tipos de experimentación utilizados para la evaluación del método de enriquecimiento. También se describen las métricas de evaluación que se utilizaron. A lo largo del capítulo se muestran gráficas correspondientes a cada uno de los resultados obtenidos, así como la interpretación de los mismos. Finalmente en el Capítulo 5 se presentan las conclusiones, aportaciones y limitaciones a las que se han llegado con el trabajo realizado, así como también las ventajas, desventajas y trabajo futuro del mismo.

2

Estado del Arte

En esta sección se presenta el estado del arte, el cual describe conceptos básicos de la Web Semántica y la representación del conocimiento. De la misma manera se mencionan diversas tareas de minería de texto que se han propuesto en la literatura. Además se presentan diferentes enfoques relacionados a enriquecimiento de texto, así como tres trabajos muy similares al que se propone en este trabajo de tesis. Por último se muestra un cuadro que proporciona información acerca de los trabajos relacionados con este protocolo.

2.1 Web Semántica

La Web es un sistema descentralizado que permite la publicación de documentos y vínculos entre estos documentos en Internet [1]. La información contenida en la Web se genera con el objetivo de que sea agradable y comprensible para el ser humano. Actualmente, es casi imposible manejar y procesar esta información debido a que la Web carece de información que ayude a definir el significado de la información contenida y que sea comprensible por computadoras [2].

El *World Wide Web Consortium*, abreviado W3C, es un consorcio internacional que produce recomendaciones para la *World Wide Web*. La Web Semántica es una colaboración de W3C para ser posible el intercambio información entre aplicaciones mediante hipervínculos. La Web Semántica se conoce como una extensión de la Web que permite que el significado de la información se precise en términos bien definidos que puedan ser entendidos por las personas y en el mismo sentido por las computadoras [3]. La Web Semántica es una Web de Datos, es decir, tiene como objetivo principal extender los principios de la Web de documentos a datos, los cuales pueden ser accedidos mediante *URIs* (*Uniform Resource Identifier*); de esta manera los datos pueden ser relacionados unos con otros.

Esencialmente la Web Semántica consiste de un modelo de datos denominado Marco de Descripción de Recursos (*Resource Description Framework*, RDF), una variedad de formatos de intercambio de datos (RDF/XML, N-Triples) y notaciones tales como esquemas RDF (*RDF Schema*, RDFS) y el Lenguaje de Ontologías Web (*Web Ontology Language*, OWL) que facilitan la descripción formal de los conceptos, términos y relaciones dentro de un dominio dado [4]. En la Figura 2.1 se muestra la estructura básica de la Web Semántica.

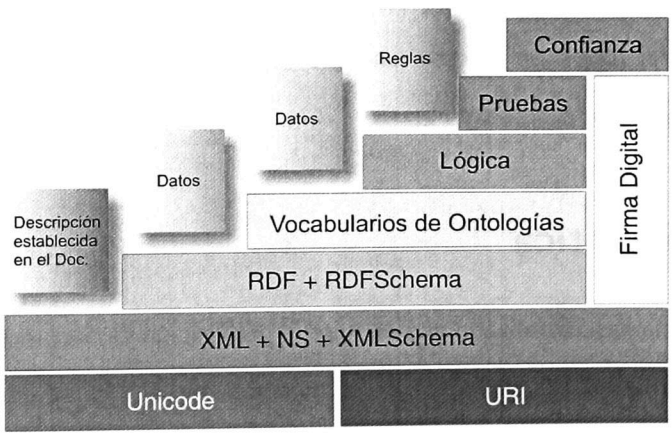


Figura 2.1: Estructura de la Web Semántica.

2.1.1 Linked Data

Conforme la Web se va haciendo más enredada en cuanto a la información, el deseo de acceder a los datos de manera directa es difícil. Debido a que la Web está construida con un conjunto de tecnologías como *URI, HTTP y HTML*, no existe un consenso en los mecanismos de identificación y acceso de objetos, además los documentos recuperados están representados en diferentes formatos. Por lo tanto no es posible implementar aplicaciones utilizando la totalidad de los datos disponibles en la Web como un único espacio global de datos.

Tim Berners-Lee y Tom Bizer [5, 6] propusieron un conjunto de enunciados para conectar datos estructurados en la Web, dichos trabajos se denominan como *Linked Data Principles*. Linked Data proporciona un paradigma de publicación en el que no sólo los documentos, sino también los datos puedan ser candidatos de primera clase en la Web, lo que permitiría extender la Web a un espacio global de datos basada en estándares abiertos, la Web de Datos. Los autores propusieron la idea de que asi como la Web ha revolucionado la forma de conectarse y consumir documentos, también puede revolucionar la forma de descubrir, acceder, integrar y utilizar los datos. La Web es el medio ideal para estos procesos debido a su ubicuidad, madurez, carácter distribuido y escalabilidad.

Para lograr un espacio de datos global y único se establecieron los siguientes principios:

- Utilizar *URIs* como nombre de las entidades nombradas.
- Utilizar *URIs* mediante el Protocolo de Transferencia de Hipertexto (HTTP) para que sea posible que las personas puedan localizar las entidades nombradas.
- Proporcionar información relevante por medio de estándares, por ejemplo RDF, SPARQL cuando se busque determinada *URI*.
- Incluir hipervínculos a otras entidades nombradas de fuentes de datos externas, para que de esta manera puedan ser descubiertas.

El tercer principio de Linked Data define el uso de un modelo de datos único para la publicación de los datos estructurados en la Web. Este modelo es RDF, un simple modelo de datos basado en grafos para su uso en el contexto de la Web. Un documento RDF tiene su estructura de tripletas, es decir, tiene la siguiente forma: *sujeto, predicado, objeto* [7], donde:

- Sujeto y objeto son *URIs* que identifican a una fuente.
- El predicado es también representado por un *URI*, especifica cómo están relacionados el sujeto y objeto.

En Linked Data una tripleta RDF puede tener una *URI* hacia otra fuente de datos. Característica que permite descubrir más información. RDF conecta distintos elementos por medio de enlaces, los cuales tienen un tipo de etiqueta, a diferencia de los enlaces HTML que sólo indican una relación.

Mediante los principios de Linked Data ha sido posible representar de manera semántica información relacionada. Este tipo de información puede estar contenida o provenir de diferentes fuentes, tales como bases de datos, la Web, documentos, correos, etc., con la ventaja de que dicha información que se puede modelar mediante Linked Data no se limita a la Web sino que puede hacerse disponible a cualquier fuente de información [8].

2.1.2 RDF

El *Resource Description Framework* (RDF) fue creado en agosto de 1997 por el *World Wide Web Corsortium* (W3C). El objetivo de dicho marco es la definición de un formato para que los diferentes sistemas de metadatos sean compatibles entre sí, además de definir una arquitectura genérica de *metainformación* [13]. RDF es un modelo que tiene dos fundamentos, la parte de los metadatos y la representación del conocimiento [14]. RDF también facilita la interoperabilidad entre diferentes aplicaciones y repositorios ya que proporciona un mecanismo de intercambio de conocimiento a través de la Web. RDF tiene como principal objetivo establecer un mecanismo que permita describir recursos

que sean independientes de la plataforma y busca la interoperabilidad de los metadatos, en el que sea posible intercambiar y crear descripciones a partir de distintos conjuntos de metadatos. Dicho mecanismo deber ser totalmente ajeno a la aplicación, además de ser tan flexible como se pueda para describir la información que se requiera. Para construir un modelo básico en RDF se utiliza un bloque denominado 'Tripleta', la cual se compone por: *sujeto, predicado y objeto*. Dicha relación se denota como: $P(O, S)$. Se dice que un sujeto S tiene un predicado P con el valor O. En la Figura 2.2 se muestra un ejemplo de la relación anterior.

Figura 2.2: Ejemplo de la relación básica en un modelo RDF.

El modelo permite que cualquier sujeto u objeto pueda ser intercambiado entre sí, haciendo más complejo al grafo y construyendo el mismo como así se desee. En la Figura 2.3 se muestra un ejemplo de un grafo más complejo.

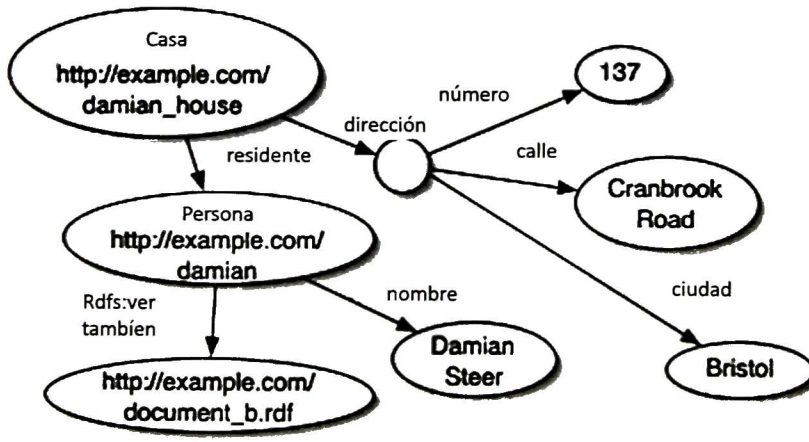


Figura 2.3: Ejemplo de una relación compleja en un modelo RDF.

RDF tiene tres principales características:

- Neutral: Es decir que no está relacionado con ningún sistema ni plataforma.
- Expresivo: Al utilizar etiquetas se dice que son intuitivas.
- Procesable: Puede ser procesado por cualquier sistema puesto que utiliza un formato ASCII, el cual tiene una estructura bien definida.

2.1.3 RDF Schema

RDF Schema está definido como un mecanismo que permite definir un vocabulario particular para la representación de datos en RDF, además de especificar el tipo de objetos que pueden ser aplicados [15]. RDF Schema utiliza una terminología como *Class, subClassOf y Property*. Por ejemplo, *subClassOf* permite la representación jerárquica de clases. Los objetos utilizan *Type* para ser declarados como instancias de estas clases. Además se puede utilizar *domain* y *range* para declarar restricciones para el uso de propiedades.

En la Figura 2.4 se muestra un ejemplo de implementación de RDF Schema, donde la parte de la figura dentro del recuadro representan instancias, mientras que la parte fuera del recuadro define el vocabulario utilizado, los cuales describen a las entidades nombradas.

2.2 Representación de conocimiento

La representación de conocimiento es un área de la Inteligencia Artificial que trata de codificar el conocimiento y razonamiento humano, de manera que sea posible para una computadora procesarlo para obtener un comportamiento inteligente [9]. El formalismo más empleado para representaciones de conocimiento en los últimos años es una ontología.

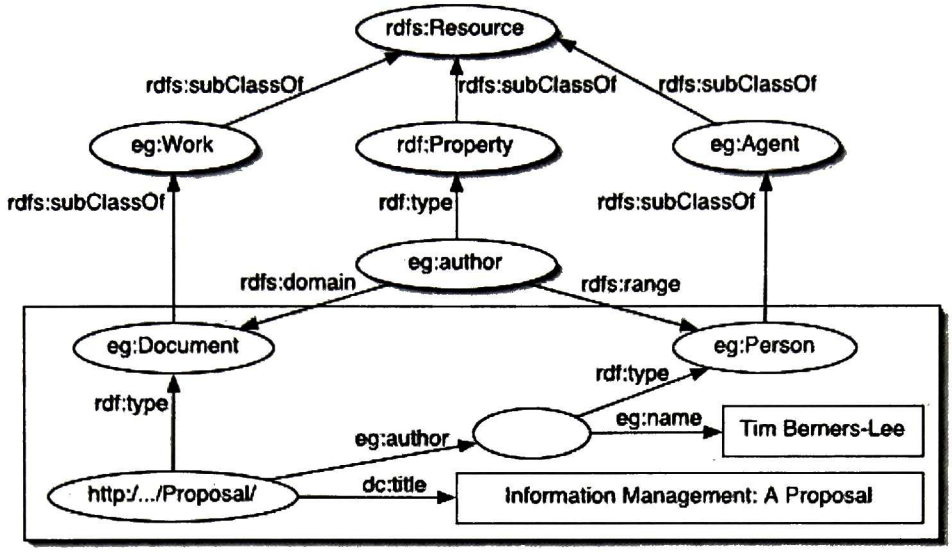


Figura 2.4: Ejemplo de RDF Schema

2.2.1 Ontologías

En Ciencias de la Computación se define a la ontología como una especificación de una conceptualización compartida [12]. En la Web Semántica es primordial la representación del conocimiento, por ello las ontologías son utilizadas para representar dicho conocimiento. Esta representación es necesaria para que se codifique de manera formal. Por tanto, se necesita definir un lenguaje que contenga aspectos importantes que satisfaga los requerimientos anteriores, dichos aspectos se listan a continuación:

- Conceptualización: El lenguaje debe elegir un modelo de referencia adecuado, así también debe proporcionar las primitivas correspondientes para representar conocimiento, tales como la definición de entidades nombradas y relaciones en un dominio.

- Vocabulario: El lenguaje debe comprender la sintaxis, además de una simbología para indicar los conceptos y la gramática para la representación de la conceptualización en una representación explícita.

- Axiomatización: Este aspecto está definido con el objetivo de capturar la semántica para la inferencia de nuevo conocimiento, las reglas y restricciones son necesarias, así como la representación de hechos.

Una característica fundamental de las ontologías es el intercambio de conocimiento entre las mismas para cumplir un grado de colaboración, para ello deben ser considerados tres aspectos cuando se desee desarrollar alguna ontología:

- Extensibilidad: Las ontologías deben de desarrollar una característica incremental, es decir, reutilizar conceptos existentes antes de crear un nuevo concepto.
- Visibilidad: Deben existir acuerdos comunes de sintaxis y semántica entre un productor y un consumidor. Dichos requerimientos son esenciales para el intercambio de conocimiento entre diferentes ontologías.
- Inferencia: Las ontologías además de representar conocimiento, tienen el propósito de permitir la inferencia lógica de hechos a través de la axiomatización. Por ello, las ontologías deben proporcionar estructuras para explotar la inferencia lógica, complejidad y expresividad.

2.2.2 Lenguajes de la Web Semántica

En el contexto de la Web Semántica y la representación del conocimiento se pueden distinguir dos clases de lenguajes: lenguajes para construcción de ontologías y los lenguajes de consulta. A continuación se exponen las dos clases de lenguajes.

2.2.3 Lenguajes para la construcción de ontologías

La Web Semántica es una visión del futuro de la Web donde la información está dando un significado explícito, permitiendo que las máquinas puedan procesar automáticamente e integrar la información disponible en la Web. La Web Semántica se basa en la capacidad de XML para definir

esquemas de etiquetas a medida y en la aproximación flexible de RDF para representar datos. El primer nivel requerido por encima de RDF para la Web Semántica es un lenguaje de ontologías que pueda describir formalmente el significado de la terminología usada en los documentos Web. Si se espera que las máquinas hagan tareas útiles de razonamiento sobre estos documentos, el lenguaje debe ir más allá de las semánticas básicas del RDF Schema.

La construcción de lenguajes para la representación del conocimiento está evolucionando de acuerdo a un enfoque en capas, dado que se establece la capa de razonamiento e inferencia, como se muestra en la Figura 2.1. Estos lenguajes deben cumplir un determinado número de requerimientos para ser útiles para el modelado de sistemas inteligentes, dichos requerimientos se listan a continuación:

- Sintaxis razonable
- Construcción de grandes bases de conocimiento
- Semántica bien definida
- Mecanismo de razonamiento
- Poder de expresividad

Un lenguaje para construir ontologías no sólo necesita tener la habilidad para definir el vocabulario, sino también los medios para definir formalmente la manera en la que trabajará el razonamiento automático. Las ontologías son primordiales en los procesos automatizados, ya que proporcionan un número de características útiles, así como también la representación del conocimiento con el fin de acceder a la información indicada. Específicamente las ontologías ofrecen un vocabulario bien estructurado que describe las relaciones entre diferentes términos, permitiendo a los procesos interpretar su significado de manera flexible pero sin ambigüedades.

Lenguaje de Ontologías Web (OWL)

El Lenguaje de Ontologías Web, OWL (Ontology Web Language), está pensado para ser usado cuando la información contenida en los documentos necesita ser procesada por las aplicaciones, al contrario que en las situaciones donde el contenido sólo necesita ser presentado a los humanos. OWL puede ser usado para representar explícitamente el significado de términos en vocabularios y las relaciones entre esos términos. OWL tiene mayor capacidad para expresar significado y semántica que XML, RDF y RDF-S, de este modo, OWL va más allá de estos lenguajes en su capacidad para representar contenido interpretable por una computadora en la Web [16]. Dada una ontología la semántica formal de OWL especifica cómo derivar sus consecuencias lógicas, es decir, hechos no representados explícitamente en la ontología, pero que implica la semántica.

OWL añade más vocabulario para describir propiedades y clases: entre otros, relaciones entre clases (por ejemplo, desunión), cardinalidad (por ejemplo, 'uno exacto'), igualdad, más tipos de propiedades, características de propiedades (por ejemplo, simetría), y clases enumeradas. La mayor contribución de OWL al extender a RDFS es la habilidad para indicar restricciones sobre cómo se comportan las propiedades que son locales a una clase. Es decir, se pueden definir clases donde una propiedad en particular es restringida para que todos los valores sean de cierta clase (o un tipo de dato) o al menos un valor debe ser de cierta clase.

OWL proporciona tres lenguajes, cada uno con nivel de expresividad mayor que el anterior, diseñados para ser usados por comunidades específicas de desarrolladores y usuarios.

OWL Lite

Está diseñado para aquellos usuarios que necesitan principalmente una clasificación jerárquica y restricciones simples. OWL Lite proporciona una ruta rápida de migración para tesauros y otras taxonomías.

OWL DL

Está diseñado para aquellos usuarios que quieren la mayor expresividad conservando completitud computacional (se garantiza que todas las conclusiones sean computables) y resolubilidad (todos los cálculos se resolverán en un tiempo finito). OWL DL incluye todas las construcciones del lenguaje de OWL, pero sólo pueden ser usados bajo ciertas restricciones. OWL DL es denominado de esta forma debido a su correspondencia con la lógica de descripción (*Description Logics*, en inglés).

OWL Full

Está dirigido a usuarios que quieren máxima expresividad y libertad sintáctica de RDF pero sin garantías computacionales. OWL Full permite una ontología para aumentar el significado del vocabulario preestablecido (RDF o OWL).

2.2.3.1. Lenguajes de consulta

Asi como RDF y OWL son los dos lenguajes dominantes de la Web Semántica, SPARQL es el lenguaje de consulta estándar por defecto para RDF [17]

SPARQL

SPARQL es un lenguaje de consulta para OWL y RDF, OWL puede ser serializado como RDF; SPARQL no entiende de manera nativa OWL. Éste opera sólo en la serialización RDF y no tiene conocimiento de los constructores del lenguaje. A pesar de esto, hay una variedad de lenguajes derivados de SPARQL, SquishQL [18], RQL [20], SPARQL-ST [19], nRQL [21], SPARQL-DL [22], iSPARQL [26], nSPARQL [27].

Por ejemplo, si queremos obtener un listado de nombres de libros a partir de un repositorio de libros podemos ejecutar la siguiente consulta:

```
PREFIX  dc: <http://purl.org/dc/elements/1.1/>
```

```
SELECT  ?title
WHERE   { <http://ejemplo.org/libros> dc:title ?title }
```

SPARQL permite el acceso a información disponible en la Web a través de diversas plataformas como es el caso de DBpedia [28], que provee de acceso a información de Wikipedia. En el siguiente ejemplo se puede observar cómo llevar a cabo una consulta que nos muestre un listado de músicos españoles junto con su nombre, su fecha de nacimiento y de fallecimiento.

```
PREFIX dcterms: <http://purl.org/dc/terms/>
 PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
 PREFIX dbp: <http://dbpedia.org/ontology/>
 SELECT ?musico ?nombreMusico ?fechaNacimiento ?fechaFallecimiento
 WHERE{
      ?musico dcterms:subject
      <http://dbpedia.org/resource/Category:Spanish_musicians>;
      rdfs:label ?nombreMusico ;
      dbp:birthDate ?fechaNacimiento ;
      dbp:deathDate ?fechaFallecimiento .
 FILTER (LANG(?nombreMusico) = "es")
 }
```

Semantic Query-enhanced Web Rule Language

Semantic Query-enhanced Web Rule Language (SQWRL) [29] está basado en el lenguaje SWRL. Por ello, para describir SQWRL es necesario definir a SWRL para conocer el lenguaje de reglas en el cual está basado.

Semantic Web Rule Language

Semantic Web Rule Language (SWRL) es una combinación de los sublenguajes OWL-DL y OWL-Lite de OWL y de los sublenguajes Unary/Binary Datalog RuleML de Rule Markup Language [30]. Aquí una regla axioma consiste de un antecedente (cuerpo) y de un consecuente (cabeza). Un átomo SWRL es de la forma $C(x)$, $P()$, $sameAs(x,y)$, $differentFrom(x,y)$, donde C es una clase, P es una propiedad y x, y pueden ser variables, individuos o valores de datos. La regla SWRL se puede describir de la siguiente forma:

$$C_1(?x) \wedge C_2(?x) \Rightarrow A(?x)$$

Donde C_1 y C_2 y A son clases diferentes y x es la única variable en la regla. En resumen, la semántica de las reglas se pueden describir como: Siempre que las condiciones especificadas en el antecedente se cumplan, entonces las condiciones indicadas en el consecuente también se cumplen.

En la sintaxis de un enunciado de consulta SQWRL se toma una regla SWRL estándar como antecedente y trata a ésta como un patrón específico para una consulta. Usando las funciones de SWRL, éstas definen un conjunto de operadores que pueden ser usados para recuperar datos específicos. Lo más atractivo de este enfoque es que no se requiere extender sintácticamente SWRL. La consultas SQWRL pueden ser almacenadas bajo la sintaxis RDF/XML en una ontología OWL y pueden ser referenciadas mediante una URI.

El núcleo de SQWRL es el operador *sqwrl:select*. Éste puede tomar uno o más argumentos, los cuales son típicamente variables usadas en la especificación del patrón de consulta. Las variables declaradas tanto en SWRL y SQWRL son declaradas con el prefijo *?*. Por ejemplo, la siguiente consulta recupera todas las instancias de la clase o del tipo *Person (persona)* que fueron declaradas explícitamente o inferidas por un motor de razonamiento en una ontología subyacente y muestra el nombre correspondiente:

```
Person(?person) ^ name(?person, ?name) -> sqwrl:select(?person, ?name)
```

La consulta anterior despliega una tabla con dos columnas (de izquierda a derecha), la primera contiene las *URIs* que localiza a cada individuo/instancia y la segunda columna muestra el valor de la propiedad *name* del individuo correspondiente. Además de los operadores SQWRL presentados en las consultas anteriores existen 56 operadores adicionales, todos definidos en la ontología *sqwrl.owl*. Por otra parte, hay operadores para cálculo matemático definidos en *swrlm.owl*, 40 operadores para operaciones con fechas y tiempo en *temporal.owl*, 80 operadores para realizar consultas al esquema del grafo semántico a través de *tbox.owl*, para la consulta de las declaraciones de propiedades de los individuos con *abox.owl*, entre otros.

2.3 Minería de texto

La minería de texto es una aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o corpus textuales [69]. La diferencia entre estas dos aplicaciones está en que con ésta última se pretende extraer conocimiento a partir de los patrones observables en grandes colecciones de datos estructurados que se almacenan en repositorios de datos. En el caso de la minería textual se tomará como punto de partida para la extracción de nuevo conocimiento repositorios documentales o texto. Es decir, información no estructurada.

Si recurrimos a la literatura publicada sobre el tema, se encuentran distintas definiciones. Dan Sullivan [23] proporciona dos definiciones: la primera define minería textual como cualquier operación realizada para extraer y analizar textos procedentes de distintas fuentes externas con el objetivo de obtener inteligencia. La segunda define minería textual como el descubrimiento de información y conocimiento que anteriormente no se conocía a partir de corpus textuales. Asimismo, Sullivan [23] señala cómo la minería textual es el proceso de compilar, organizar y analizar grandes colecciones de documentos para apoyar en la distribución de información a los analistas y a las personas encargadas de tomar decisiones, y para descubrir relaciones entre hechos relacionados que se reparten entre

distintos dominios de investigación.

Existe una clara relación entre minería textual, minería de datos, recuperación de información y lingüística computacional. La minería textual recoge diferentes técnicas y planteamientos desarrollados en otras disciplinas. El objetivo de la minería textual es facilitar el análisis de la información disponible en colecciones de documentos y así la deducción de nuevo conocimiento. Pero las preguntas son: ¿Cómo lograr esto?, ¿Qué nos ofrece una herramienta de minería textual para lograr este propósito? Para dar respuesta a estas preguntas la minería textual ofrece diversas tareas que se pueden utilizar según sea la necesidad, dichas tareas se describen a continuación.

2.3.1 Preprocesamiento

La minería textual adopta una serie de técnicas procedentes de la recuperación de información y de lingüística computacional. El preprocesamiento incluye la eliminación de los signos de puntuación y la extracción de las palabras separadas entre sí por espacios en blanco o signos de puntuación (si éstos no se han eliminado en un paso previo). Entre las subtareas que se pueden realizar están las que se describen a continuación.

Tokenización

Es el proceso de dividir un flujo de texto en palabras, frases, símbolos o elementos, el resultado de dicha división es conocido como *token*. Los *tokens* se convierten en la entrada para un futuro procesamiento en la minería de texto.

Eliminación de palabras vacías

Una tarea habitual en el preprocesamiento de los documentos es la eliminación de palabras vacías, carentes de significado, como son preposiciones, artículos, conjunciones, etc. Sin embargo, no todos los autores coinciden en la conveniencia de eliminar las palabras vacías.

Lematización

Estas tareas intentan mapear las formas verbales a su forma infinitiva y los pronombres a su forma singular. Sin embargo, para alcanzar dicho objetivo, la forma verbal de la palabra deber ser conocida.

Stemming

Como parte del preprocesamiento se suele realizar la normalización de las palabras extraídas del documento. Esta normalización consiste construir la forma básica de las palabras, es decir, a su raíz.

2.3.2 Preprocesamiento lingüístico

Frecuentemente las tareas de minería de texto pueden ser aplicadas sin un preprocesamiento adicional. Sin embargo, algunas tareas de preprocesamiento lingüístico adicionales pueden ser usadas para mejorar la información disponible acerca de las palabras. A continuación se mencionan algunas de ellas.

Part of speech tagging (POS)

Esta tarea también conocida como etiquetado gramatical es el proceso durante el cual cada 'token' del texto es reconocido como una etiqueta, y esta propiedad es colocada junto al 'token'. De esta manera la salida de esta tarea son los 'tokens' o términos del texto y una etiqueta por cada uno de ellos. El término etiqueta se refiere a una secuencia de caracteres tipificando la gramática utilizada en el texto.

Existen diversas maneras de representar la gramática, desde marcar los signos de puntuación hasta palabras específicas. Este tipo de etiquetas están orientadas al reconocimiento estructural y gramático de cada término de la oración, así entonces se puede aplicar un procesamiento futuro.

Este tipo de tareas son usualmente utilizadas como base para un procesamiento futuro más

complejo en el área de análisis de lenguaje natural, como puede ser el análisis estructural, análisis semántico, traducciones, etc.

Extracción de frases nominales

Una frase nominal es la máxima cantidad de información *que un humano* recibe. Una oración usualmente tiene la estructura *Sujeto-Predicado*, en la que el predicado contiene un verbo que describe la acción del sujeto, además de que puede o no apuntar a varios objetos que utiliza el sujeto. Una frase nominal es una frase basada en un pronombre principal, es decir, consiste de un pronombre y varias otras palabras que lo determinan. En algunas ocasiones la base de la frase nominal no es un sujeto, si no un pronombre o alguna otra palabra que puede ser usada independientemente como sujeto de la oración. Por ejemplo: $noun\ \ phrase\ \ extraction$. La frase anterior consiste de tres sujetos, mientras que el sujeto principal es *extraction*. Los sujetos restantes son utilizados para calificar al sujeto principal. Asimismo se encuentran otro tipo de palabras como: Adjetivos, artículos, frases preposicionales, frases relativas subyacentes.

La extracción de frases nominales requiere la detección e identificación de todos los tipos de palabras anteriormente mencionados. Además, esta tarea usualmente se apoya en la tarea de etiquetado gramatical, ya que a partir del etiquetado puede definir reglas de aparición para definir las frases nominales. La dificultad aparece en definir las reglas gramaticales ya que en múltiples ocasiones el sujeto no puede ser específico y por lo tanto no se puede definir como resultado por lo que identificación puede ser un tanto difícil.

Desambiguación del significado de la palabra

La desambiguación del significado de la palabra es un problema abierto de procesamiento de lenguaje natural que incluye el proceso de identificar con qué sentido una palabra está usada en los términos de una oración, cuando la palabra en cuestión tiene polisemia, es decir, pluralidad de significados. La solución de este problema afecta a otras tareas de la lingüística computacional, tales

como el discurso, la mejora de la relevancia en los motores de búsqueda, la resolución de referencia, la coherencia (lingüística), la inferencia, y otros.

Como en todo procesamiento del lenguaje natural, existen dos enfoques principales para la desambiguación del significado de la palabra: enfoque profundo y enfoque superficial. El enfoque profundo supone el acceso a un amplio conjunto de conocimiento del mundo, que permite determinar en qué sentido se utiliza la palabra. Estos enfoques no son muy exitosos en la práctica, principalmente porque tal cuerpo de conocimientos no existe en un formato legible por la computadora, fuera de ámbitos muy limitados. Sin embargo, si ese conocimiento si existe entonces los enfoques profundos serían mucho más precisos que los enfoques superficiales.

2.3.3 Identificación de nombres propios

La extracción de nombres propios relativos a personas, organizaciones, eventos, funciones, así como cantidades monetarias y fechas es una de las principales funciones que debe satisfacer la minería textual. Además, la minería textual también debería permitirnos identificar las relaciones que existen entre estos nombres propios y constatar así 'hechos' descritos en los documentos.

Un tema más complejo en la identificación de nombres propios es la extracción de las relaciones que existen entre los términos. En este sentido, es necesario recurrir a técnicas de *parsing* y análisis sintáctico de las sentencias para identificar los verbos que sirven de nexo entre los nombres propios y tratar de deducir así posibles relaciones.

2.3.4 Identificación de entidades nombradas

La definición de entidad nombrada no es clara, diversos autores han dado diferentes definiciones. Lisa Rau [60] la define como un único identificador en su trabajo de extracción de nombres de un texto. Petasis y colaboradores [61] definen a la entidad nombrada como un pronombre propio, que sirve como nombre para una cosa o alguien, mientras que Nadeau y colaboradores [62] la definen

como nombres para restringir la tarea a sólo aquellas entidades que la definan.

La identificación de entidades nombradas es una tarea que consiste en identificar y clasificar algunos tipos de elementos de la información, es decir, las entidades nombradas [63]. La identificación de entidades nombradas se propone con el objetivo de desarrollar herramientas para buscar y descubrir conocimiento a partir de la identificación de la semántica de textos no estructurados. Esta tarea usualmente es utilizada como base para otras áreas cruciales de administración de información.

A pesar de que existen diversas definiciones de una entidad nombrada, la identificación de Entidades Nombradas se ha propuesto como tarea a resolver en las conferencias CoNLL [65] y ACE [66]. Esta tarea busca localizar elementos del texto con base a su semántica, dichos elementos típicamente corresponden a personas, organizaciones y localizaciones.

Actualmente existen diferentes herramientas para llevar a cabo la identificación de entidades nombradas, hasta el momento la mejor herramienta ha obtenido una precisión y exhaustividad mayor al 90 % [63]. Sin embargo, normalmente las herramientas desarrolladas obtienen valores de entre 60 % y 90 %. Además de las métricas de precisión[1] y exhaustividad[2], el desempeño de una herramienta puede ser evaluado por diferentes métricas, entre ellas se encuentra la validez del contenido, la validez de la generalización de la entidad, la validez de convergencia y la validez de conclusión.

2.3.5 Representación de documentos mediante el modelo vectorial

Una premisa en cualquier aplicación de recuperación y tratamiento documental es la necesidad de representar el contenido de los documentos mediante un modelo. El modelo más utilizado al día de hoy, tanto en los sistemas de indexación como en las aplicaciones de minería textual, es el vectorial. En este modelo un documento se caracteriza mediante el conjunto de términos que representan su contenido. Estos términos podrían ser términos extraídos directamente del texto completo del documento o descriptores asignados al documento por una lista o por una aplicación

[1] Es la proporción de documentos relevantes en el conjunto de documentos recuperados a partir de una consulta.

[2] Es la proporción de documentos relevantes recuperados a partir de una consulta con respecto al total de documentos relevantes.

informática, tomados o no de un lenguaje documental externo. Cualquiera que sea el caso, el documento se representará mediante una secuencia de términos que corresponden con los distintos términos utilizados para describir el contenido del documento.

Un vector es una estructura consistente en un número fijo de elementos, en la cual la posición de cada uno de ellos es significativa. En el modelo vectorial cada documento se considera un vector y cada término que aparece en al menos un documento será un componente del vector.

En este método la recuperación de información se realiza mediante la comparación de la distancia que existe entre los vectores correspondientes a los documentos, y un vector utilizado para representar la ecuación de búsqueda. Entre las ventajas del modelo vectorial frente a otros modelos, como el booleano, se encuentra el hecho de calcular la similitud entre la ecuación de búsqueda y los documentos. Esto permite realizar una ordenación de los documentos recuperados, mostrando al principio de la lista aquellos documentos que son más similares a la ecuación de búsqueda, y al final de la lista los que son menos.

2.3.6 Agrupación de documentos

Se trata de una técnica que permite identificar grupos o clases de objetos similares a partir de un espacio multidimensional. El agrupamiento se define como la organización de una colección de patrones (normalmente representados mediante vectores o como puntos en un espacio multidimensional), en grupos con base a su similitud [70]. Aquellos patrones que pertenezcan a un mismo grupo serán más similares entre sí que con los patrones que pertenecen al resto de los grupos.

El agrupamiento consiste en una clasificación desatendida o no supervisada. Esto diferencia al agrupamiento de las técnicas de clasificación supervisada y que se aplican en la categorización automática. En la clasificación supervisada se debe ordenar un conjunto de objetos en una serie de grupos predefinidos con anterioridad. En el caso del agrupamiento no existirán grupos predefinidos a los que haya que asignar los objetos durante el proceso de clasificación.

2.3.7 Categorización

Se trata de un proceso de clasificación automática con el que se pretende asignar un documento a una clase o tema definido con anterioridad. Un ejemplo de esta aplicación sería la asignación automática de un encabezamiento de materia o una notación de un sistema de clasificación bibliográfico a un documento.

La categorización automática parte de un entrenamiento previo de un programa informático encargado de realizarla. Así, se facilitará al programa informático una serie de documentos a los que ya se ha asignado un tema o clase. De esta forma el programa podrá analizar las características que determinan la asignación de los documentos a una u otra clase. Posteriormente, cuando se procese un nuevo documento, el programa podrá 'deducir' a qué clase pertenece. Esta deducción se basará en la similitud que exista entre el nuevo documento y los utilizados durante la fase de entrenamiento. Tanto el agrupamiento como la categorización parten de una misma base, el cálculo de las similitudes entre documentos; normalmente mediante la identificación de los términos que aparecen de forma conjunta en los documentos.

2.3.8 Relaciones entre términos y conceptos

Entre las técnicas utilizadas por la minería de textos se encuentra la extracción de términos o conceptos y la identificación de relaciones entre estos términos. En apartados anteriores se ha referido a la extracción de términos y a su ponderación para identificar aquellos que resulten más significativos del contenido de los documentos. Otras aproximaciones más complejas, como la 'Latent Semantic Indexing' o el agrupamiento también podrían aplicarse con este propósito.

Tradicionalmente estas asociaciones se han venido usando en proyectos de recuperación de información experimentales. Mediante estas asociaciones entre términos se permitía al usuario recuperar documentos potencialmente relevantes, que no habían sido indexados con los mismos términos que se han utilizado en la ecuación de búsqueda. Esta idea es similar a la propuesta pionera

que Maron y Kuhn [24] hicieron en 1960 con su concepto de recuperación aritmética o asociativa.

En relación con el agrupamiento, de la misma forma que podemos agrupar documentos a partir del número de términos que comparten, sería también posible agrupar términos a partir de los documentos en los que aparecen de forma conjunta. Esta aproximación ha sido descrita por Salton y colaboradores [25].

2.4 Query Expansion

La Expansión de Consultas (*Query Expansion*) es el proceso de reformular una consulta para mejorar el rendimiento de recuperación en las operaciones de recuperación de la información [31]. La Expansión de Consultas es una metodología estudiada en el campo de las Ciencias de la Computación, particularmente en Procesamiento del Lenguaje Natural y la Recuperación de la Información [32]. En el contexto de los motores de búsqueda, la expansión de consultas involucra evaluar una entrada del usuario (las palabras que el usuario ingresa en el área de consulta de búsqueda, y a veces otros tipos de datos) y expandir la consulta de búsqueda para que se ajuste a documentos adicionales.

La expansión de consultas involucra técnicas como:

- Encontrar sinónimos de palabras y buscar también por los sinónimos.
- Encontrar todas las formas morfológicas de las palabras involucradas en la búsqueda, aplicando técnicas de lematización (stemming).
- Corregir errores tipográficos y buscar automáticamente por la forma corregida o sugerirla
- Ponderar los resultados según la relevancia.

En la literatura existen muchos trabajos relacionados con este tipo de enfoques, en los que buscan expandir los términos clave para realizar una búsqueda. Dichos enfoques tienen como objetivo aumentar algunas métricas como la precisión y exhaustividad, cuando se realizan

tareas de recuperación de información. Los trabajos relacionados con la expansión de consultas [33, 34, 35, 36, 37] tienen como principal enfoque enriquecer la consulta con diversos métodos anteriormente descritos. Cuando el enriquecimiento de consultas mejora el desempeño de buscadores y la recuperación de información, surgen diversas propuestas en la literatura para enriquecer documentos y textos no estructurados, es decir, el enriquecimiento de texto (Text Enrichment) o expansión de texto (Text Expansion).

2.5 Enriquecimiento de texto

El enriquecimiento de texto surge con el objetivo de proveer información y recursos sobre entidades nombradas y términos contenidos en el texto. Actualmente en la literatura los autores han enfocado sus esfuerzos utilizando diversas herramientas, diferentes bases de conocimiento, así como las diversas técnicas de enriquecimiento.

Uno de los enfoques que diversos autores han propuesto es el enriquecimiento de texto a partir de sinónimos. Es decir, los términos del texto a enriquecer son procesados apoyándose de herramientas para obtener los sinónimos de dichos términos. Una vez que los términos son procesados y después de haber obtenido sus sinónimos se realizan consultas a partir de los términos originales y los términos procesados. En la Figura 2.5 se muestra la estructura básica del enfoque de enriquecimiento con base en sinónimos.

En el trabajo de XiangHua Fu y colaboradores [38] se propone un enfoque como el anterior explicado. Su propuesta se basa en tres pasos clave, dos de ellos están basados en seleccionar los términos importantes a enriquecer, como es la selección de términos relevantes y a partir de éstos eliminar términos redundantes. El tercer paso clave es el enriquecimiento o expansión de términos, en el que se realizan consultas a WordNet para obtener sus sinónimos y así enriquecer el texto. De la misma manera, Jun Wang [39] propone un enfoque muy similar para enriquecer texto, pero con la diferencia de que parte de una lista de términos preprocesada. A partir de esta lista de términos

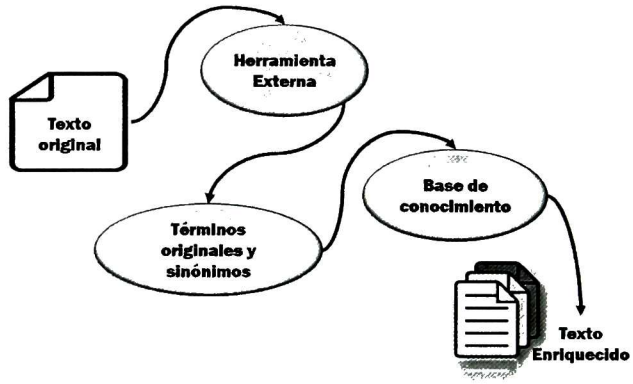


Figura 2.5: Estructura básica del enfoque de enriquecimiento con base en sinónimos

se realizan consultas a WordNet y se obtienen sus sinónimos. La aplicación final que propone Jun Wang es realizar una clasificación, en la cual explica los resultados, mostrando una mejora en la clasificación cuando se enriquece el texto.

Otro tipo de enfoque que se ha propuesto en la literatura es el enriquecimiento de texto a partir de la consulta de hiperónimos e hipónimos. Los hiperónimos son términos que tienen un significado de gran extensión y, por tanto, incluyen a hipónimos más concretos o específicos. Por ejemplo, la palabra flor es un hiperónimo respecto a palabras como clavel, jazmín o margarita.

Para consultar dichos hiperónimos e hipónimos se apoya, al igual que los sinónimos, en una herramienta de consulta para obtener información referente a los mismos. Este enfoque es altamente parecido con los sinónimos, con la única diferencia de que se consultan sus hiperónimos e hipónimos, para obtener otros términos para enriquecer el texto. En la Figura 2.6 se muestra la estructura básica del enfoque de enriquecimiento con base en hiperónimos-hipónimos.

En la literatura actualmente existen pòcas aportaciones que utilizan este enfoque. Uno de ellos es el propuesto por Supakpong Jinarat [41], en el que el objetivo es obtener un enriquecimiento de texto para mejorar el agrupamiento de *snippets* de páginas Web. La idea principal que describe el autor es expandir los *snippets* obtenidos de consultas Web, el contenido del *snippets* es expandido con términos relacionados con hiperónimos-hipónimos y sinónimos. La base de conocimiento que se

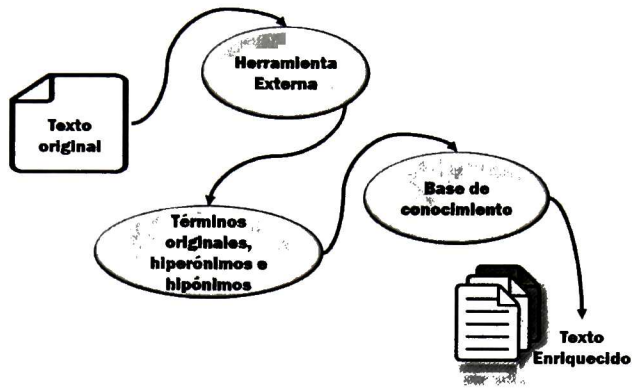


Figura 2.6: Estructura básica del enfoque de enriquecimiento con base en hiperónimos-hipónimos

utiliza es *Open Directory Project* (ODP), es decir, una taxonomía organizada por humanos, con el objetivo de conceptualizar el contenido de la Web.

Otro trabajo con el mismo enfoque fue propuesto por Khaled Abdalgader y colaboradores [42]. La propuesta tiene como objetivo medir la similaridad entre textos cortos apoyándose en la desambiguación del significado de la palabra así como la expansión de términos utilizando también hiperónimos-hipónimos y sinónimos. Los métodos implementados en la propuesta obtienen la información desde WordNet. Así, al obtener términos que tienen referencia hiperónimos-hipónimos y sinónimos es posible desambiguar términos y finalmente medir la similaridad entre textos.

Existe otro enfoque propuesto en la literatura: el enriquecimiento de texto a partir de recursos Web, es decir, el enriquecimiento de texto se realiza a partir de artículos y páginas Web estructuradas. Las propuestas que existen se encuentran diseñadas para diversas aplicaciones, desde clasificación o agrupación hasta expansión de consultas. Una característica en común que se observa en las propuestas es que utilizan artículos de páginas Web, entre ellas predomina Wikipedia como base de conocimiento. Con Wikipedia las propuestas extraen sus enlaces a documentos, categorías y otra información relevante para enriquecer el texto o expandir una consulta. En la Figura 2.7 se muestra la estructura básica de las propuestas que utilizan recursos Web para enriquecer texto.

Como se ha mencionado antes, existen diversos trabajos que utilizan este enfoque, uno de ellos

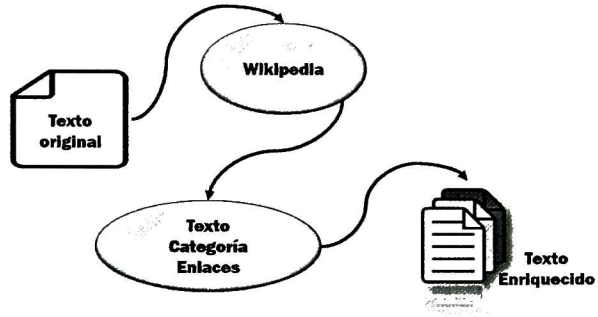


Figura 2.7: Estructura básica del enfoque de enriquecimiento con base recursos Web

es el propuesto por Francisco Bueno y colaboradores [43], en el que presentan una propuesta que describe un modelo de recuperación de información a fin de enriquecer un documento. Los autores definen una arquitectura con el objetivo de experimentar el efecto de combinar diferentes técnicas a fin de enriquecer un documento. Las técnicas principalmente se basan en obtener información a partir de diferentes fuentes de datos Web, es decir, se basa en obtener dicha información desde un ambiente distribuido.

Otra propuesta muy similar es la de Proscovia Olango y colaboradores [44] en el que desarrollaron una herramienta denominada como 'TermPedia' . La herramienta define términos técnicos en textos educacionales. Dichos términos son extraídos desde Wikipedia, los cuales son enlazados a artículos relevantes de Wikipedia, aportando información y explicaciones extra. Por último, plantean una serie de experimentos para estudiar como afecta el enriquecimiento de textos educativos en el estudio cotidiano de los estudiantes.

De la misma forma este tipo de enfoque también es utilizado por Abdullah Bawakid y Mourad Oussalah [45], quienes enriquecen documentos a partir de Wikipedia. El enriquecimiento está definido con base en conceptos, categorías, enlaces y texto de artículos. Los autores proponen extraer los términos importantes a través de valores *TF-IDF* Una vez que se obtienen los términos se realizan consultas a Wikipedia para obtener las características antes mencionadas. Por último, una vez enriquecido el documento realizan clasificaciones basadas en centroides. Esta propuesta la centran en

analizar las mejoras en la clasificación a partir del enriquecimiento, por lo cual, no presentan mucha información detallada cuando se enriquece el texto.

Por otra parte, Christopher Boston y colaboradores [46] presentan un sistema de desambiguación de términos basado en expandir a los mismos. El sistema se enfoca principalmente a realizar consultas en un buscador Web. Cuando se presentan términos ambiguos se expanden dichos términos para desambiguarlos y aumentar la exactitud de la consulta. La base de conocimiento que se utiliza en el sistema son artículos de Wikipedia, específicamente haciendo referencia a enlaces, categorías y definiciones.

Con el mismo enfoque, y además un objetivo muy similar con respecto a los anteriores, Gerasimos Spanakis [47] propone un sistema para adquirir conocimiento desde Wikipedia y mejorar el agrupamiento de documentos. El autor propone que dado un documento se extraen las posibles entidades nombradas y se realizan consultas a Wikipedia. Una vez que se consulta Wikipedia se extraen enlaces, categorías y texto para enriquecer el documento. Cada documento es representado como un vector de conceptos, el cual es utilizado para agrupar los documentos con base en su similaridad.

Los enfoques anteriores han sido propuestos con base en un enriquecimiento en definiciones, diccionarios y artículos Web. Por ello, los últimos aportes se han enfocado en enriquecer al texto de manera semántica, es decir, a partir de bases de conocimiento basadas en la Web Semántica. Con ello se asegura un enriquecimiento mayor gracias a la semántica entre las entidades nombradas de la base de conocimiento. Usualmente este tipo de enfoques utilizan bases de conocimiento como DBpedia, Freebase, Wikitology y ontologías propias ya desarrolladas. En la Figura 2.8 se muestra una estructura básica propuesta por diversos autores para el enriquecimiento de texto con bases de conocimiento semánticas.

En los últimos años se ha observado el esfuerzo por proponer trabajos con este enfoque, ya que experimentalmente ha obtenido mejores resultados, además de que aporta una mayor información, conceptos y entidades nombradas relacionadas debido a su estructura semántica definida. Tal es

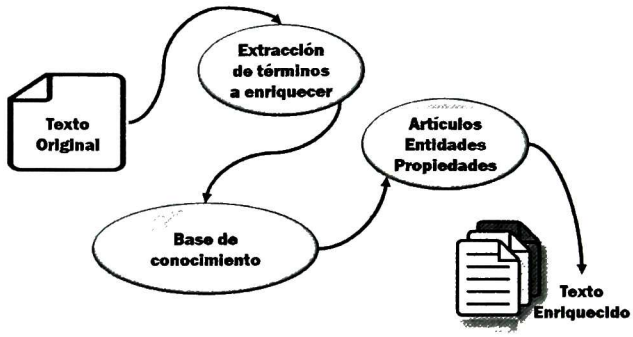


Figura 2.8: Estructura básica del enfoque de enriquecimiento con bases de conocimiento semánticas

el caso de la propuesta Georgios Lioudakis y colaboradores [48], la cual está basada en enriquecer un documento a partir de entidades nombradas de una ontología propia. La propuesta específica es un *framework* para transformar un documento electrónico en un documento enriquecido. El enriquecimiento que proponen es consultar términos conocidos en diferentes ontologías de dominios específicos que están almacenados en varios repositorios en la Web. Otra propuesta muy similar es la de Fraihat Salam [49], quien utiliza una ontología como medio para indexar documentos así como realizar búsquedas de los mismos. Es decir, el enfoque que propone el autor es una alternativa a las búsquedas implementadas actualmente, las cuales están basadas en palabras clave. La propuesta se basa en el diseño de un motor para indexar y buscar documentos con base en su semántica. Al indexar un documento extrae las entidades nombradas clave de un documento y lo indexa con base a este criterio. Para buscar documentos consulta las entidades nombradas identificadas en la consulta, con el objetivo de buscar en las entidades nombradas indexadas.

Otro enfoque muy similar y que utilizan el enriquecimiento semántico para medir la similaridad entre textos es el de Liu Wenyin y colaboradores [50], en el que proponen medir similaridad entre textos, es decir, se basa en saber qué tan parecidos son un par de textos. Para ello se apoya en la base de conocimiento WordNet. Los autores definen un método para calcular la similaridad con base a su contenido léxico. A partir de la base de conocimiento (WordNet) consultan su contenido semántico para verificar la similaridad. El mismo objetivo de Liu Wenyin y colaboradores [50] lo

comparte Aguilar-Lopez y colaboradores [51], es decir, el medir la similaridad, pero Aguilar-Lopez [51] tiene como objetivo medir la similaridad entre páginas Web, utilizando el contenido semántico de las mismas. Aguilar-Lopez propone un enfoque utilizando la semántica de páginas Web. La propuesta parte de realizar consultas a páginas Web, extraer su contenido semántico y determinar su similaridad jerárquica. La base de conocimiento utilizada es WordNet, y a partir de este conocimiento se determina la relevancia de una página.

Por otro lado, existen otros autores que se apoyan en proyectos de años atrás, como es el caso de DBpedia. Un ejemplo es Hiroki Yamakawa y colaboradores [52], en el que presentan un enriquecimiento de texto a partir de conocimiento almacenado en Wikipedia. El enfoque que plantean los autores es mejorar el desempeño de la clasificación de documentos. El objetivo principal de este enfoque es transformar los términos comunes de los documentos en términos enriquecidos semánticamente, es decir, al final el documento es representado por un conjunto de conceptos. Finalmente, los autores presentan una comparación entre diferentes clasificadores y como mejoró un poco el desempeño de cada uno de los clasificadores. El desempeño de los clasificadores puede ser mejor aún, la clave es mejorar la exactitud de los conceptos consultados en Wikipedia. Por su parte Zhaohui Huang y colaboradores [53] proponen un trabajo enfocado al análisis de información de grandes volúmenes. Dado un conjunto de documentos, se extraen sus entidades nombradas y se construye un grafo de relaciones. Por cada documento se obtiene un grafo de relaciones y partir de un algoritmo de descubrimiento se obtiene un solo grafo, es decir, la unión de todos lo grafos. La base de conocimiento que se utiliza es DBpedia, y a través del algoritmo planteado es capaz de observar patrones en el texto de los documentos.

Los trabajos mencionados anteriormente tienen una amplia relación con el trabajo que se plantea en este protocolo, sin embargo, tienen un diferente enfoque, es decir, enriquecen el texto pero utilizando otras técnicas y base de conocimiento. Además, los trabajos anteriores no se enfocan en enriquecer el texto, sino en realizar una tarea final, como puede ser el caso de clasificación o agrupamiento. Por esa razón en algunos trabajos no describen a plenitud el enriquecimiento o en

otros casos el trabajo para enriquecer es mínimo. Por ello, a continuación se presentan tres trabajos que tienen un enfoque muy similar al propuesto, además de que utilizan y describen técnicas de enriquecimiento más complejas y por ello más exactas.

2.6 Trabajo Relacionado

Anteriormente se mencionaron diferentes enfoques en el enriquecimiento de texto que utilizan sinónimos, artículos web, hiperónimos y bases de conocimiento estructuradas semánticamente para proveer un recurso adicional con respecto al contenido del texto. Sin embargo, existen tres trabajos muy similares al que se propone en esta investigación. Dichas propuestas se describen más a detalle a continuación.

2.6.1 Named-Entity Recognition and Text Enrichment using Semantic Web

Chrysoula Zerva y Aliki Kopaneli [54] proponen el desarrollo de una plataforma de enriquecimiento semántico. Específicamente proponen el estudio y desarrollo de dos sistemas independientes que intentan enriquecer texto con base en la Web Semántica (particularmente con DBpedia), en el que dicho texto tiene las siguientes características: Texto plano, texto sin procesamiento alguno y textos de lenguaje natural.

Con el desarrollo de dos sistemas plantean detectar y extraer frases del texto que corresponden a entidades nombradas de DBpedia, es decir, los términos más importantes del texto. A partir de la detección y extracción de dichas entidades nombradas facilitan la adquisición de información extra con el fin de enriquecer el texto inicial. En la Figura 2.9 se muestra la descripción general de la propuesta de Chrysoula Zerva y Aliki Kopaneli [54].

El primer sistema lo denominaron como *'Condensed Representation Extraction and Semantic Text Annotation'* (CRESTA). Dicho sistema está orientado hacia la extracción de un conjunto de

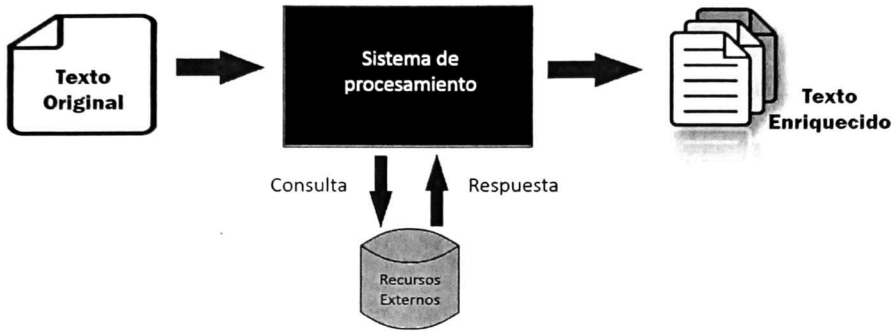


Figura 2.9: Descripción general de la propuesta de Chrysoula Zerva y Aliki Kopaneli [54]

entidades nombradas nominales que pueden ser consideradas como una representación eficiente del texto plano sin procesar. El sistema aprueba una representación cuando los conceptos fundamentales del texto muestran un resumen satisfactorio, así como distinguir su contexto semántico. Además, los autores proponen un sistema de ranking para determinar la potencialidad de las entidades nombradas utilizando como base de conocimiento Wikipedia. En la Figura 2.10 se muestra una descripción específica de CRESTA.

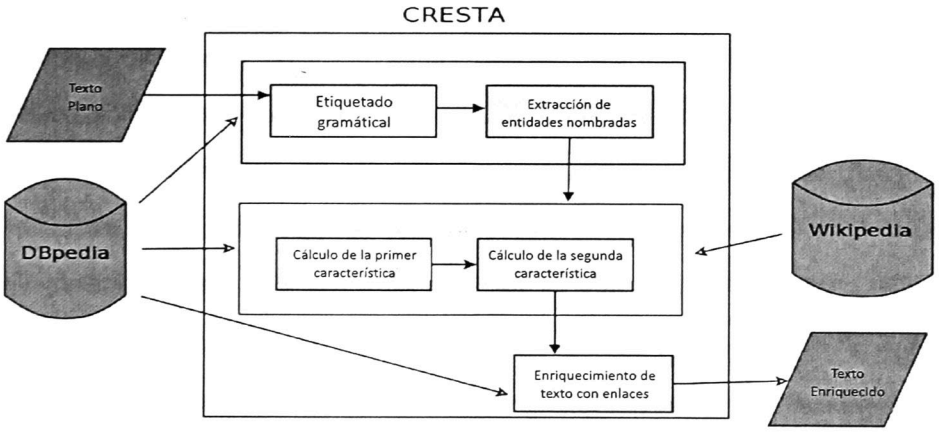


Figura 2.10: Descripción general de CRESTA [54]

El segundo sistema propuesto se denomina como *'Semantic Web based Person Identification'* (SWIPD), el cual está orientado a la detección de referencias del mundo real. Para ello los autores

utilizan conocimiento de DBpedia para identificar las entidades nombradas que cumplen con la condición antes mencionada. En la Figura 2.11 se muestra una descripción de SWIPD.

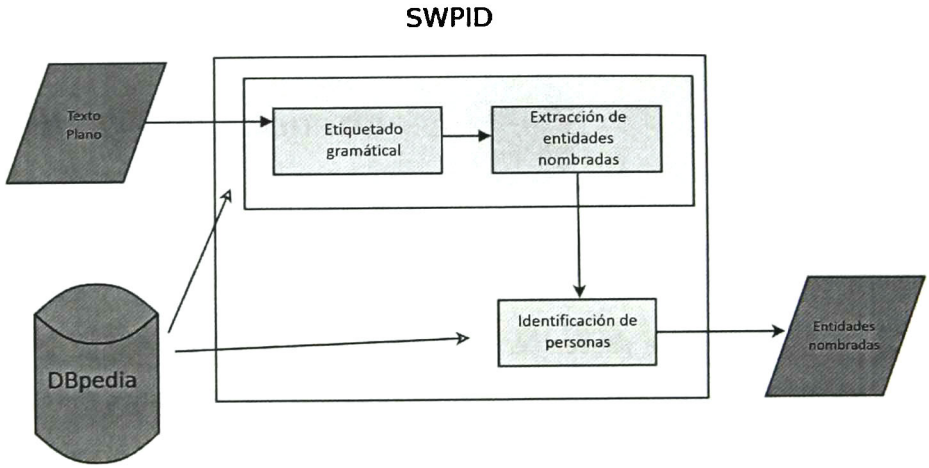


Figura 2.11: Descripción general de SWIPD [54]

2.6.2 Precise Tweet Classification and Sentiment Analysis

Rabia Batool y colaboradores [55] proponen un sistema de análisis y clasificación de mensajes de Twitter mediante el enriquecimiento semántico. El análisis tiene tres objetivos principales: enriquecer el mensaje original, categorizar el mensaje enriquecido y analizar el sentimiento del mismo. Se enriquece el texto del mensaje de Twitter mediante Alchemy API [56]. Su sistema se basa en cinco bloques, los cuales al trabajar como un conjunto son capaces de alcanzar los objetivos antes mencionados.

Preprocesamiento

En este bloque los autores proponen obtener los mensajes de Twitter mediante una API. Dicha API utilizada tiene como salida un documento XML, debido a eso tiene que realizarse un preprocesamiento para traducir el documento XML a texto plano y por último almacenarlo en una base de datos.

Generador de Conocimiento

El objetivo de este bloque es extraer información importante de los Tweets y realizar una clasificación por categorías basadas en el conocimiento obtenido. Para realizar dicha tarea se apoya en Alchemy API. Dicha API recibe como entrada el texto sin estructurar, se aplica procesamiento de lenguaje natural y aprendizaje automático. Como salida se obtienen los términos clave y sentimientos de los usuarios con base a esos términos.

Potenciador de conocimiento

En este bloque se aplica el enriquecimiento al mensaje de Twitter a través de etiquetado gramatical y extracción de entidades nombradas. Una vez más extrayendo conocimiento de Alchemy API.

Extractor de sinónimos

Es un bloque adicional donde se agrega conocimiento con base a consultas de sinónimos desde los términos una vez extraídos. Al obtener términos extra es posible aumentar el conocimiento con otras palabras clave.

Motor de filtrado

En este bloque se aplica una clasificación en diferentes categorías con base al conocimiento extra obtenido, que es la aplicación final.

En esta propuesta los autores se apoyan de Alchemy API, desde preprocesar y enriquecer el mensaje hasta obtener conocimiento, es decir, el objetivo no es desarrollar técnicas de minería de texto, sino enriquecer el texto mismo, los autores basan sus esfuerzos en utilizar herramientas y API externas. De la misma manera, los autores remarcan el incremento de la exactitud de los clasificadores debido al enriquecimiento de texto así como la adquisición de conocimiento extra. En la Figura 2.12 se muestra un diagrama general del sistema antes descrito.

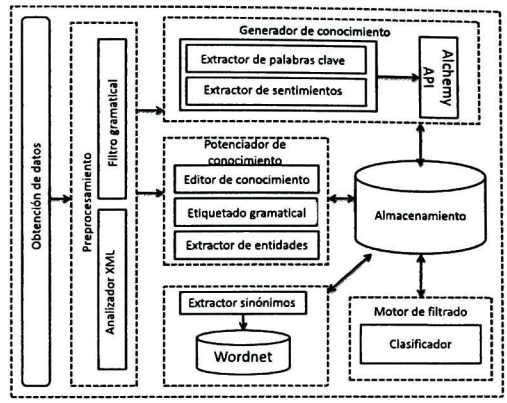


Figura 2.12: Descripción general del sistema desarrollado por Rabia Batool y colaboradores [55]

2.6.3 Clustering of Rough Set Related Documents with Use of Knowledge from DBpedia

Marcin Szczuka y colaboradores [57] proponen un sistema para agrupar artículos científicos con base a su semántica. La propuesta parte de tener documentos agrupados con base a su contenido semántico y con la ayuda de la base de conocimiento de DBpedia enriquece los documentos. Los autores dividen su trabajo en varias etapas, la primera etapa es la aplicación de un preprocesamiento:

- Eliminación de caracteres especiales, números, etc.
- Eliminación de palabras vacías, es decir, *'stopwords'*.
- Aplicar stemming.
- Filtrar únicamente los términos más frecuentes, es decir, los términos que aparezcan con mayor frecuencia.

Así, se obtiene una lista de palabras como representación del texto que se enriquecerá. A partir de la lista de palabras se buscan en DBpedia conceptos relacionados con el término utilizado, es decir, las entidades nombradas. Al tener todas las entidades nombradas se utiliza un valor TF-IDF

para obtener las entidades nombradas más relevantes y menos redundantes. Los autores proponen que las entidades nombradas restantes por documentos forman parte de una categoría.

El último paso que proponen es calcular las distancias de los documentos utilizando las entidades nombradas encontradas como términos; para dicho cálculo se utilizó la distancia del coseno. Para el agrupamiento se utilizó un enfoque jerárquico. En la figura 2.13 se muestra la descripción general del sistema desarrollado.

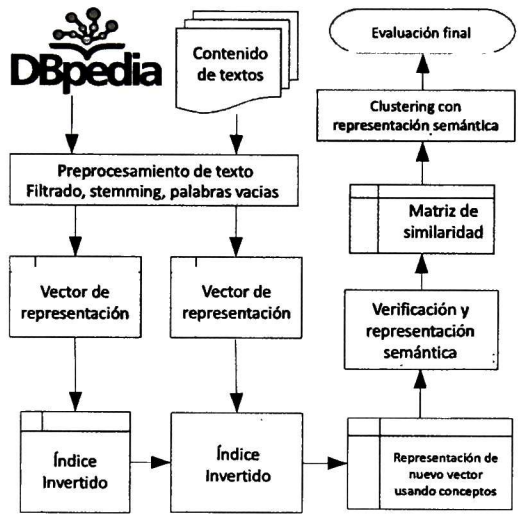


Figura 2.13: Descripción general del sistema desarrollado por Marcin Szczuka y colaboradores [57]

2.7 Resumen

Para situar este trabajo inicialmente se mencionaron diversas disciplinas que contextualizan a la Web Semántica, su diseño y estructura. Asimismo se describieron algunos principios básicos de la Web Semántica. Además, se describieron las bases de conocimiento basadas en la Web Semántica y que pueden ser utilizados como recursos externos para diversas tareas, especialmente la explotación de los recursos para su posterior uso. Por otra parte se definió el concepto de minería de texto y su relación con la minería de datos. En cada uno de los temas abordados se definieron

conceptos, se explicaron algunas de las principales tareas, estrategias, herramientas, etc. De igual forma se mencionaron algunos trabajos realizados por diversos autores que de alguna manera ofrecen resultados, alcances, limitaciones, ventajas y desventajas con respecto a su propuesta. Por último, se describe de una manera más detallada ciertas propuestas que tienen un enfoque, tareas, técnicas y bases de conocimiento muy similar al que se propone en este protocolo, además de describir su experiencia en el desarrollo de la propuesta. Para resumir las propuestas relacionadas, en la Tabla 2.1 se describen las principales características de cada una de las propuestas anteriormente descritas. Asimismo, con los trabajos reportados se puede diferenciar con respecto al trabajo que se propone el extenso procesamiento para obtener las entidades nombradas y palabras relevantes en el texto (las cuales conformaran a la representación del texto), las diferentes fuentes de conocimiento que se utilizaran para obtener un mayor conocimiento de la representación de texto, el método de cuantificación de conocimiento así como la aplicación final del trabajo. Lo anterior, con el objetivo de proveer un mejor desempeño que las propuestas antes descritas. De la misma manera, como se muestra en la Tabla 2.1, las métricas utilizadas en las diferentes propuestas es la precisión y exhaustividad, por tanto, para validar y justificar la propuesta, se pretende aplicar clasificación al texto y conocimiento adquirido de diferentes fuentes de conocimiento.

Autor	Año	Enfoque	Técnica de Enriquecimiento	Base de conocimiento	Método de Evaluación	Elemento Procesado	Precisión %	Exhaustividad %
Chrysoula Zerva	2013	Enriquecimiento de documentos largos	Semántica	Wikipedia DBpedia	Queries	Corpus	30.56	57.96
Jun Wang	2012	Enriquecimiento de documentos largos	Sinónimos	Wordnet	Clustering	Bag of words	24	69
TSO (Servicio Privado)	2014	Enriquecimiento de texto	Semántica	Wikipedia	*	Corpus	*	*
Francisco Bueno	2011	Enriquecimiento de documentos largos	Semántica	Wikipedia	Queries	Corpus	*	*
Hiroki Yamakawa	2011	Enriquecimiento de texto	Semántica	Wikipedia	Clasificación	Corpus	51.98	51.02
Abdullah Bawakid	2011	Enriquecimiento de documentos largos	Semántica	Wikipedia	Clasificación	Bag of words	*	*
Rabia Batool	2013	Enriquecimiento de Tweets	Semántico Sinónimos	Alchemy API	Clasificación	Corpus	*	*
Supakpong Jinarat	2009	Enriquecimiento de Snippets de páginas web	Hiperónimos Hipónimos	ODP (Open Directory Project)	Clustering	Snippet	*	*
Tadej Stajner	2011	Enriquecimiento de Texto	Semántico	Wikipedia	*	Corpus	*	*
Myunggwon Hwang	2011	Desambiguación del sentido de la palabra	Semántico	Wordnet	Queries	Corpus	28.47	64.2
Christopher Boston	2013	Query expansion Desambiguación	Semántico	Wikipedia	Queries	Bag of words	66.82	61.47
Liu Wenyin	2010	Medición de similaridad entre Snippets	Semántico	Wordnet	Correlación Clustering	Bag of words	84.12	*
Dulce Aguilar Lopez	2009	Extraer contenido semántico de páginas web para medir similaridad	Semántico	Wordnet	Queries	Corpus	*	*
Zhaohui Huang	2009	Extraer contenido semántico de páginas web para encontrar patrones	Semántico	DBpedia	*	Corpus	*	*
Khaled Abdalgader	2010	Medición de similaridad para textos cortos	WSD Hiperónimos Sinónimos	Wordnet	Queries	Corpus	75.5	91.5
Marcin Szczuk	2011	Enriquecimiento de documentos largos	Semántico	DBpedia	Clustering	Corpus	*	*

Tabla 2.1: Resumen de los enfoques propuestos para el enriquecimiento de texto

3

Metodología

En este capítulo se presenta el diseño e implementación del método de enriquecimiento de texto a partir de recursos de la Web Semántica.

3.1 Descripción del método

El método de enriquecimiento de texto está diseñado con 5 módulos, los cuales tienen una tarea específica. Dichos módulos son: preprocesamiento, extracción de Entidades Nombradas, extracción de Entidades Nombradas a partir de conceptos clave, estructuración de información y enriquecedor de información. En la Figura 3.1 se muestra un diagrama a bloques del método de enriquecimiento de texto.

Cada bloque de la Figura 3.1 representa a un módulo del método de enriquecimiento, a continuación se describe de manera general cada uno de los bloques:

- **Preprocesamiento**: Realiza operaciones o transformaciones sobre el texto para prepararlo

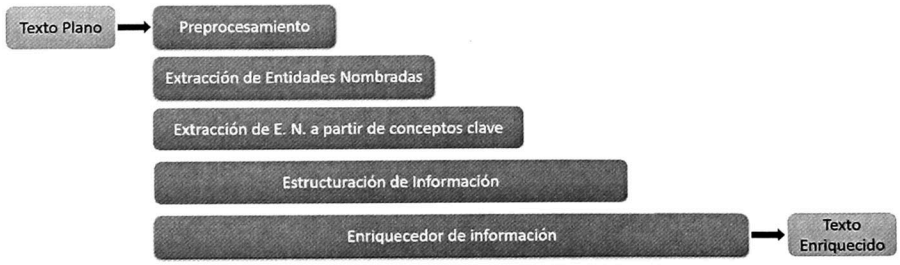


Figura 3.1: Diagrama a bloques del método de enriquecimiento de texto.

para su uso en las siguientes tareas.

- **Extracción de Entidades Nombradas**: Obtiene Entidades Nombradas del texto.
- **Extracción Entidades Nombradas a partir de conceptos clave**: Identifica conceptos clave en el texto y obtiene Entidades Nombradas relacionadas en el texto a partir de la identificación de los conceptos clave.
- **Estructura de Información**: Unifica las entidades encontradas de los módulos anteriores, construye una base de conocimiento, representa el conocimiento.
- **Enriquecedor de entidades**: Es el módulo esencial del método de enriquecimiento, enriquece el texto con información de DBpedia a partir de las Entidades Nombradas identificadas en los módulos anteriores. Asimismo integra la información encontrada al texto original en un nuevo texto denominado *texto enriquecido*.

Para enriquecer un texto a partir de recursos de la Web Semántica el método recibe como entrada el texto al módulo de preprocesamiento, enseguida se identifican y extraen las Entidades Nombradas en el texto, después se identifican los conceptos clave del texto y se obtienen Entidades Nombradas relacionadas con éstos, seguido de la unificación de Entidades Nombradas y de la creación de la base de conocimiento. Por último se enriquece el texto con información de DBpedia y se integra toda esa información con el texto original. De esta manera, como resultado del método, se obtiene un texto

enriquecido a partir de recursos de la Web Semántica. En la Figura 3.2 se muestra un diagrama de tareas del método de enriquecimiento de texto.

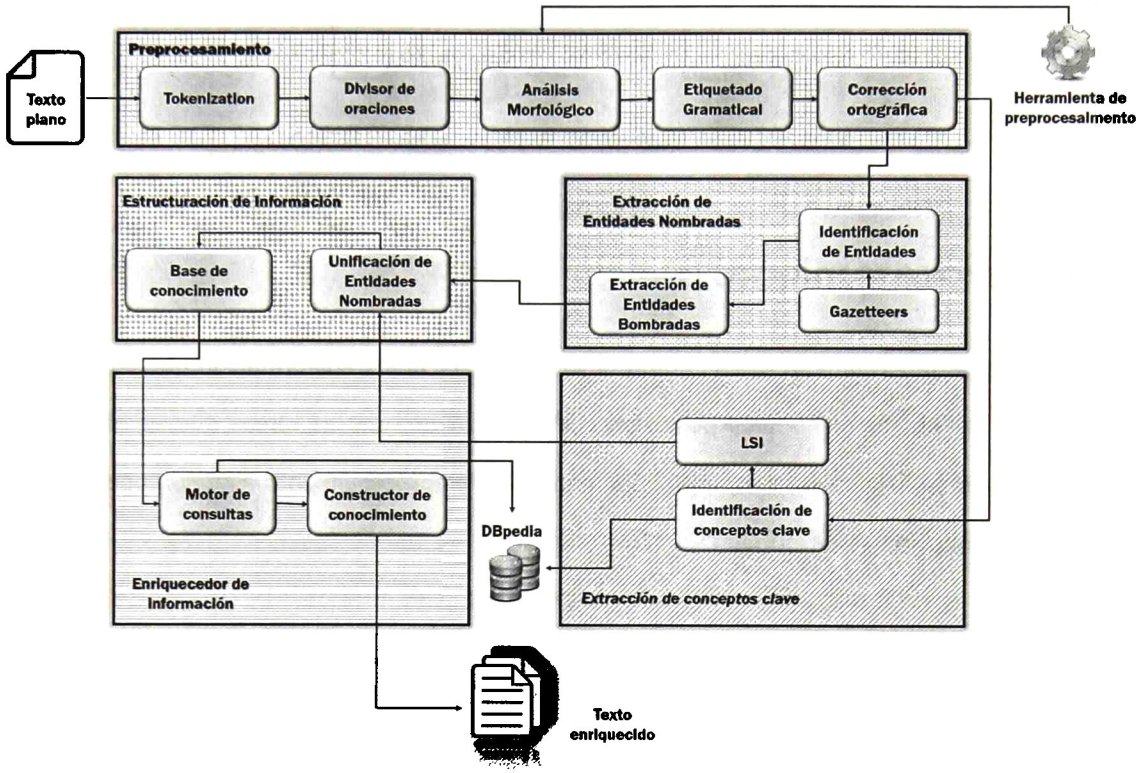


Figura 3.2: Diagrama de tareas del método de enriquecimiento de texto.

Para desarrollar el método de enriquecimiento se optó por aplicar una metodología basada en anotaciones, es decir, una metodología para agregar información a un documento a diferentes niveles, los cuales son: nivel de símbolo, palabra, frase, párrafo, sección y documento. Este tipo de metodología es muy utilizada por procesadores de texto dado que tienen buen desempeño para obtener y agregar información extra al texto. La metodología basada en anotaciones modela el texto como un grafo, donde los nodos son las palabras y las aristas son las conexiones entre una palabra y otra. Cada una de las anotaciones que se generaron en las etapas del método son agregadas como un nodo adicional al texto, el cual representa la anotación de su respectiva palabra. En la Figura 3.3 se muestra un ejemplo de dicha metodología.

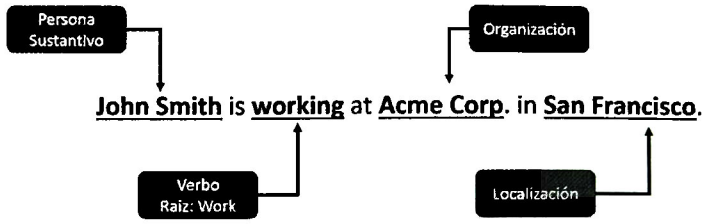


Figura 3.3: Ejemplo de modelado de texto basado en anotaciones.

3.2 Preprocesamiento

El módulo de preprocesamiento tiene como objetivo aplicar una transformación que convierta el texto a un formato útil para módulos posteriores. Para cada una de las tareas que se realizan en esta etapa se utilizó la herramienta Gate [83]. El módulo de preprocesamiento está compuesto por 5 tareas que a continuación se describen.

3.2.1 Divisor de oraciones

El proceso de segmentación divide al texto en párrafos independientes, es decir, realiza una división de texto de acuerdo al número de párrafos que contenga. Después de dividir el texto por párrafos, el divisor de oraciones divide el texto a un nivel más específico, en este caso a nivel de oración. Esto es conocido comúnmente como *'Sentence Splitter'*. La división se realiza a partir de signos de puntuación, secuencias de control o palabras que representan conjunciones.

3.2.2 *Tokenización*

La *tokenización* es la identificación de unidades independientes más pequeñas en el texto. Aunque pareciera que solamente se necesitan palabras, números y símbolos para trabajar, es totalmente lo contrario. Dado que la salida del método debe ser el texto original además de información relacionada, no es posible eliminar signos de puntuación, espacios en blanco o símbolos. Una anotación llamada 'tipo de *token*' es utilizada para modelar los diferentes tipos de *token*. En la Figura 3.4 se muestran

los tipos de *tokens* que se identifican.

Figura 3.4: Tipos de *tokens*.

Por cada tipo de *token* que se identifica se describe el identificador único en el texto basándose en la posición en el texto que representa. Además del identificador único se agrega una anotación llamada como 'tipo de escritura' en el que se describe cómo está escrito el *token*, es decir, si se encuentra escrito en mayúscula, minúscula o alguna combinación de éstas. Esta anotación es importante para la identificación de Entidades Nombradas que se realizará en un módulo posterior.

3.2.3 Análisis morfológico

Una vez que la *tokenización* y el divisor de oraciones es aplicado al texto a enriquecer, se aplican diferentes análisis a cada palabra contenida en el texto. Cada uno de los análisis que se aplican son utilizados en diferentes etapas del método de enriquecimiento de texto.

Dado que en texto se pueden utilizar diferentes formas de una misma palabra, por ejemplo: conjugación verbal o palabras singulares y plurales, es necesario aplicar un análisis morfológico. Este tipo de análisis es aquel proceso que da como resultado las posibles interpretaciones de una palabra. En el análisis morfológico es posible obtener diversa información de las palabras contenidas en un texto, a continuación se describe cada análisis obtenido:

3.2.3.1. *Stemming*

En el proceso conocido como *stemming* se busca la raíz (*stem*) de la palabra para evitar la redundancia de información así como aumentar la exhaustividad [1] en los sistemas de recuperación de información. Un ejemplo de *stemming* se muestra en la Figura 3.5.

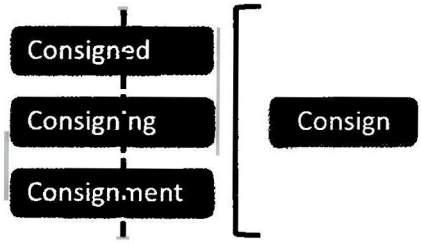


Figura 3.5: Ejemplo de *Stemming*.

3.2.3.2. *Lematización*

La lematización es el proceso de asignación, en forma de etiqueta, de lema (su forma canónica) a una palabra tal como la encontramos en el discurso textual. Es decir, esta tarea agrupa las diferentes formas de una palabra en el que puede aparecer. Pareciera que es la misma tarea que *stemming* pero la lematización se apoya con información o conocimiento extra, específicamente del etiquetado gramatical. Un ejemplo de lematización se muestra en la Figura 3.6.

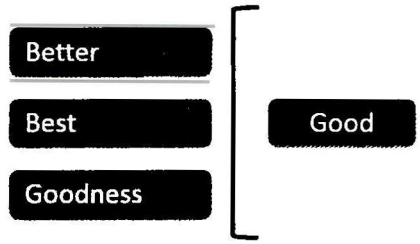


Figura 3.6: Ejemplo de Lematización.

[1] Es la proporción de documentos relevantes recuperados a partir de una consulta con respecto al total de documentos relevantes.

3.2.3.3. Análisis adicionales

Además de *Stemming* y Lematización, otro tipo de análisis fue aplicado al texto como el tipo de *token* o la manera en que se encuentra escrito. Este análisis es aplicado para distinguir los *tokens* en el texto y con ello aplicar análisis posteriores apoyándose en este análisis.

Un *token* puede ser denominado de distintas maneras, entre las cuales se encuentran:

- Palabra: Se refiere a cualquier palabra en el texto.
- Signo de puntuación: Los signos de puntuación utilizados en el texto.
- Símbolo: Son aquellos símbolos que se encuentran en el texto, pero que no forman parte de una palabra, como por ejemplo aquellos en una ecuación matemática.
- Número: Cualquier número en el texto.
- Espacio en blanco: Los espacios en blanco no son descartados, puesto que forman parte del texto de entrada, deben de aparecer en el texto de salida.

Por otra parte, el tipo de escritura también fue analizado. El tipo de escritura ayudará más adelante a determinar Entidades Nombradas. Las diferentes maneras en que un *token* puede ser escrito son:

- Iniciales con mayúscula, por ejemplo: Dishelt.
- Todos los caracteres con mayúscula, por ejemplo: DISHELT.
- Todos los caracteres con minúscula, por ejemplo: dishelt.
- Sin patrón alguno, por ejemplo: DiShELt.

3.2.4 Etiquetado Gramatical

El etiquetado gramatical, también conocido como *Part-Of-Speech tagging, POS tagging o POST*, es el proceso de asignar (o etiquetar) a cada una de las palabras de un texto su categoría gramatical. Existen diversas categorías gramaticales, el número de categorías varían de acuerdo el lenguaje. Un ejemplo de esta tarea se muestra en la Figura 3.7.

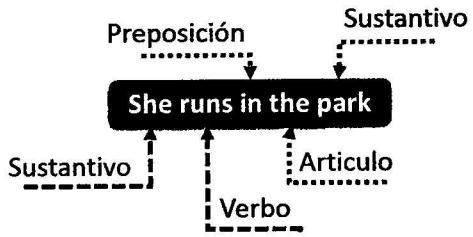


Figura 3.7: Ejemplo del Etiquetado Gramatical.

Esta tarea asigna una etiqueta a cada palabra, describiendo su categoría gramatical. En el ejemplo de la Figura 3.7, las etiquetas asignadas se muestran en la Figura 3.8.

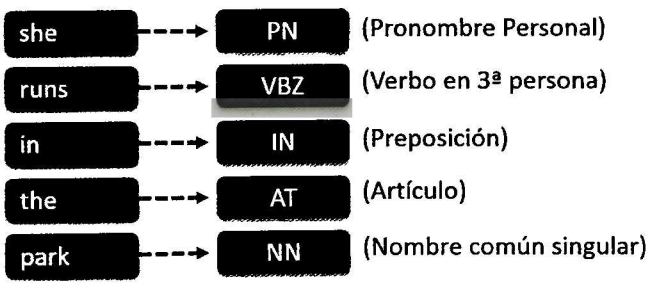


Figura 3.8: Etiquetas asignadas a cada palabra.

3.2.5 Corrección ortográfica

La corrección ortográfica identifica aquellas palabras que se encuentran mal escritas, seguido de buscar posibles sugerencias y corregir la palabra con alguna de ellas.

Para identificar y corregir aquellas palabras que se encuentran mal escritas es necesario acceder a diccionarios. Aspell [82] es un corrector ortográfico de GNU que utiliza diccionarios de diferentes idiomas; el diccionario en inglés es utilizado para identificar palabras mal escritas. GNU Aspell identifica palabras mal escritas y sugiere posibles palabras para corregirla. Para elegir una palabra sugerida como correcta se utiliza la Distancia de Levenshtein. La distancia de Levenshtein es el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra. Se entiende por operación a una inserción, eliminación o la sustitución de un caracter. Por ejemplo:

- casa $\rightarrow$ cala (sustitución de 's' por 'l')
- cala $\rightarrow$ calla (inserción de 'l' entre 'l' y 'a')
- calla $\rightarrow$ calle (sustitución de 'a' por 'e')

Para aplicar la corrección ortográfica como parte del módulo de preprocesamiento se siguen los pasos que en la Figura 3.9 se muestran y que a continuación se describen.

- Se obtienen las palabras del texto.
- Cada palabra es analizada por GNU Aspell para identificar aquellas que están mal escritas.
- Por cada palabra que GNU Aspell identifica como mal escrita, sugiere un conjunto de palabras.
- La palabra correcta es aquella palabra sugerida que tiene como menor distancia de Levenshtein con respecto a la palabra mal escrita.
- Se sustituye la palabra incorrecta con la correcta siempre y cuando la palabra mal escrita no tenga una etiqueta gramatical de pronombre o sustantivo, ya que es muy probable GNU Aspell no conozca nombres de sustantivos y lo identifique como una palabra mal escrita.

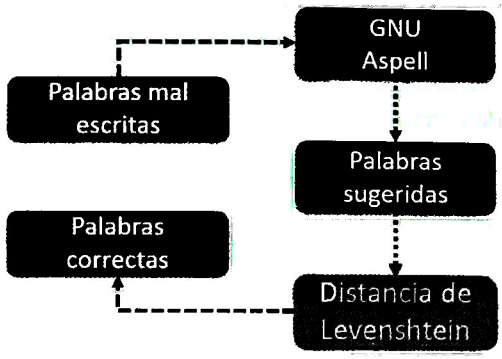


Figura 3.9: Esquema de la corrección ortográfica utilizada.

3.3 Extracción de Entidades Nombradas

El término Entidad Nombrada a una palabra o secuencias de palabras que se identifican como nombre de persona, organización o lugar. Éstas pueden consistir de cualquier tipo de palabra: adverbios, preposiciones, adjetivos, e incluso algunos verbos, pero la mayoría de las Entidades Nombradas están compuestas de sustantivos.

La extracción de Entidades Nombradas está basada en la construcción de Gazetteers y la construcción de un modelo de las Entidades Nombradas a extraer. Dicha extracción es importante en el enriquecimiento de texto debido a que las Entidades Nombradas en un texto son una característica fundamental en la temática que presenta dicho texto. Por ejemplo, en el texto de la Figura 3.10 se muestra un ejemplo de un texto con las Entidades Nombradas identificadas.

The **National Aeronautics and Space Administration** (**NASA**) is the **United States** government agency responsible for the civilian space program as well as aeronautics and aerospace research.
President **Dwight D. Eisenhower** established the **NASA** in 1958 with a distinctly civilian orientation encouraging peaceful applications in space science. Most **US** space exploration efforts have been led by **NASA**, including the **Apollo moon-landing** missions, the Skylab space station and later the Space Shuttle.

Figura 3.10: Ejemplo de un texto con Entidades Nombradas identificadas.

Dicho texto está relacionado con la NASA, por consecuencia es muy probable que encontremos Entidades Nombradas como:

- **NASA**: Nombre de la institución.
- **Dwight D. Eisenhower**: Nombre de quien estableció dicha institución.
- **Apollo program**: Nombre del programa más importante que ha tenido la NASA.

Las Entidades Nombradas encontradas en un texto como el de la Figura 3.10 permiten encontrar información relacionada con dichas Entidades y por consecuencia con el texto. Dado que se utilizó DBpedia como base de conocimiento, al identificar Entidades Nombradas en DBpedia se obtuvieron relaciones y características de las Entidades Nombradas identificadas. La identificación y extracción de Entidades Nombradas es una de las dos formas en el método que determinará qué parte del texto se enriquecerá y qué partes no.

3.3.1 Gazetteers

Un gazetteer consiste en un conjunto de listas que contienen Entidades Nombradas de cierto tipo. Los tipos de Entidades Nombradas comúnmente son de personas, lugares u organizaciones, aunque pueden ser de otro tipo definido por el dominio del problema a resolver, como pueden ser días de la semana o nombres de películas. Estas listas son usadas para buscar ocurrencias de las Entidades Nombradas en un texto, además estas listas están compiladas en una máquina de estados finitos para que la búsqueda sea rápida. En la Figura 3.11 se muestra un ejemplo del modelo de la máquina de estados finitos para tres Entidades Nombradas: Albert Einstein, Albert Finney y Albert Finney Grisar.

La Figura 3.11 muestra que el inicio es 'Albert' y dependiendo de la palabra consecutiva puede llegar a un estado final (representado con los recuadros en negro), lo que significa la identificación de una Entidad Nombrada.

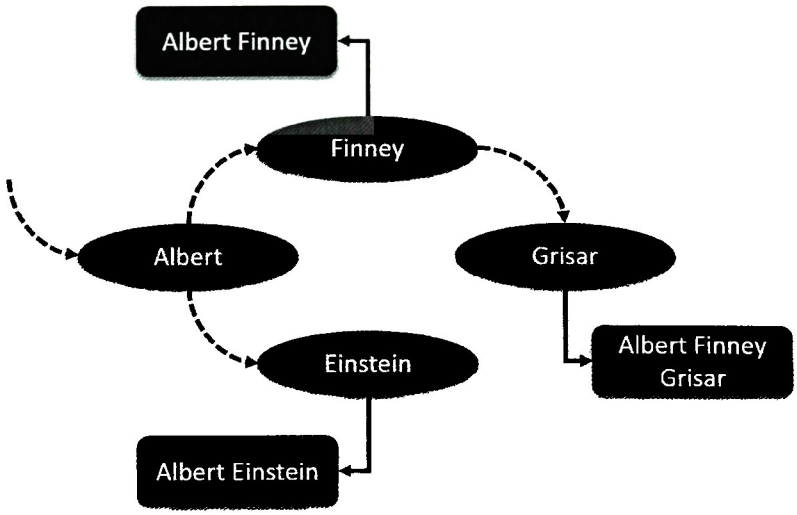


Figura 3.11: Ejemplo de una máquina de estados finitos para la identificación de Entidades Nombradas.

Aunque en un principio las Entidades Nombradas fueron definidas como personas, organizaciones o lugares, las Entidades pueden ser modeladas según sea el dominio de interés. Por ejemplo, si el dominio de interés es el cine, las Entidades Nombradas pueden ser nombres de actores (tipo persona), lugares de filmación (tipo localización) o bien nombres de películas. Éste último no entra en la definición formal de una Entidad Nombrada, sin embargo dado el dominio se puede modelar de esa forma.

Los Gazetteers que se utilizaron se definieron de 4 tipos, los cuales se describen a continuación:

- Personas: Identifica nombres de personas.
- Organización: Para identificar nombres de organizaciones o instituciones.
- Localización: Para identificar nombres de lugares, países, ciudades, etc.
- Misc: Se refiere al tipo de Entidad Nombrada que no pertenece a ninguna de las anteriores, como los son cosas hechas por el hombre, hechos históricos, avances científicos, etc.

Para generar los Gazetteers se realizaron consultas a DBpedia para listar cada artículo de

Wikipedia y modelar la Entidad según sea su naturaleza, es decir, si es una persona, organización, lugar, hechos históricos, etc.

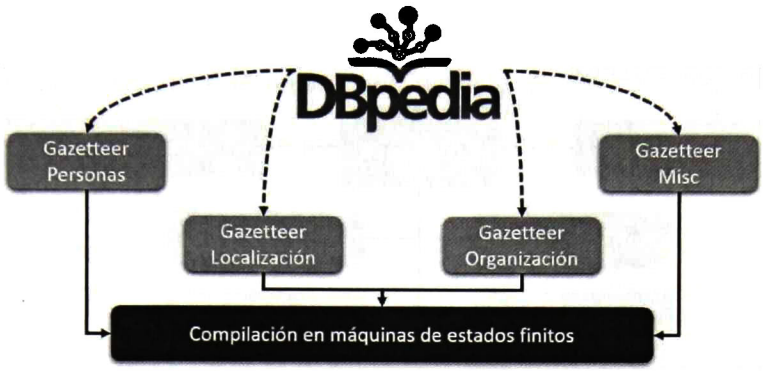


Figura 3.12: Esquema para la generación de Gazetteers.

En la Figura 3.12 se muestra un esquema para la generación de los gazetteers antes mencionados. Por cada tipo de Entidad Nombrada se realiza una consulta a DBpedia para listar todas las Entidades Nombradas de un tipo previamente definido. Cada lista de Entidades Nombradas es compilada en una máquina de estados finitos para utilizarla en el proceso de extracción de Entidades Nombradas.

3.3.2 Modelo de identificación de Entidades Nombradas

Para utilizar cada uno de los Gazetteers que se construyeron a partir de DBpedia, es necesario modelar la extracción de Entidades Nombradas. Con ésto se reutiliza cada una de las Entidades para otro fin, en este caso se utilizarán para buscar información relacionada en DBpedia.

Para modelar la extracción de Entidades Nombradas, se utilizó JAPE (Java Annotation Patterns Engine) [85], el cual provee una transducción sobre las máquinas de estados finitos basados en expresiones regulares. JAPE es una versión de CPSL (Common Pattern Specification Language), desarrollado para el área de Extracción de Información. Básicamente JAPE permite reconocer 'expresiones regulares' sobre documentos, pero con la diferencia que las expresiones regulares trabajan sobre una secuencia de *items*, en este caso es aplicado a una estructura de datos mucho más compleja, como lo es un grafo (estructura definida previamente para modelar el texto). Es decir, JAPE permite

identificar nodos en el grafo que cumplen un conjunto de patrones, dicha identificación constituye una Entidad Nombrada.

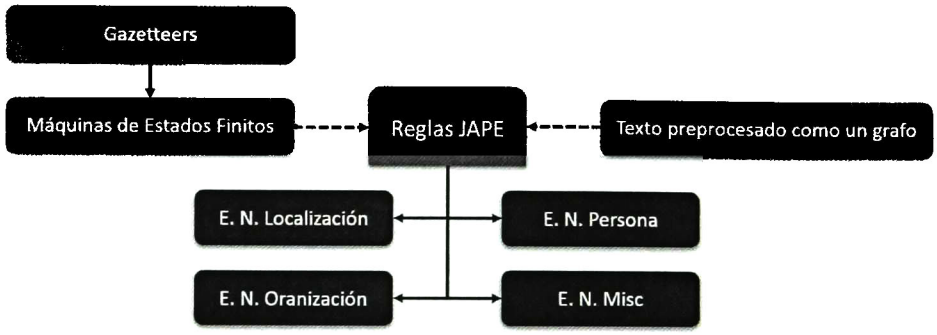


Figura 3.13: Esquema para la generación del modelo de extracción de Entidades Nombradas.

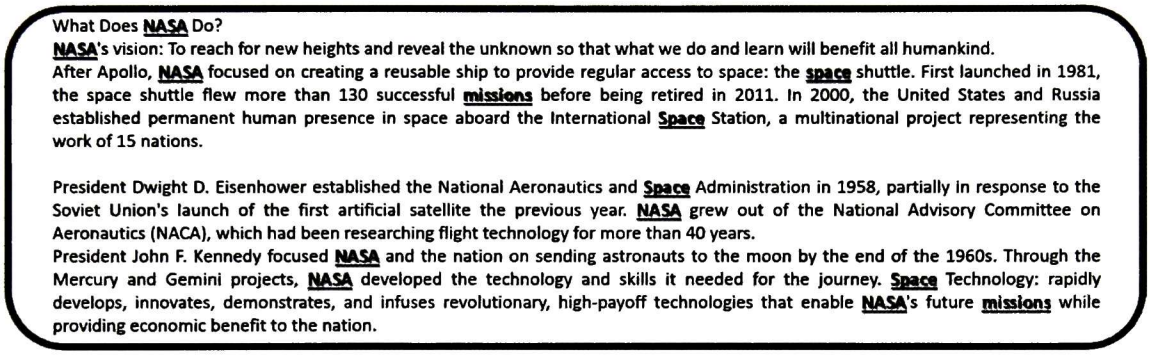
En la Figura 3.13 se muestra un esquema para la identificación y extracción de Entidades Nombradas. Las reglas JAPE modelan los Gazetteers como máquinas de estados finitos y actúan sobre el texto preprocesado anteriormente. Al desarrollar las reglas JAPE se definieron cuatro tipos de salida, los cuales corresponden a los diferentes tipos de Entidades Nombradas definidos previamente: Personas, Localizaciones, Organizaciones y Misc.

3.4 Extracción de Entidades Nombradas relacionadas con conceptos clave en el texto

La extracción de Entidades Nombradas que previamente se describió forma parte esencial en el enriquecimiento de texto, sin embargo es posible que en algunos textos no existan Entidades Nombradas. En esta sección se describe una forma de encontrar Entidades Nombradas que están relacionadas con conceptos clave en el texto. Los conceptos clave son aquellos términos que son fundamentales y los más importantes en el texto pero que pueden o no formar parte de una Entidad Nombrada. Estos conceptos clave por sí mismos no tienen significado alguno, pero un conjunto de conceptos clave representan la idea central del texto y lo más importante del mismo texto.

Básicamente al utilizar conceptos clave para la identificación de Entidades Nombradas se están encontrando relaciones a textos similares de DBpedia con el texto original, dichos textos corresponden a artículos de Wikipedia modelados como Entidades Nombradas.

Por ejemplo, en la Figura 3.14 se muestra un ejemplo de un texto con los conceptos clave identificados, tras la identificación de conceptos clave en dicho texto se determinó que los conceptos clave son: 'NASA', 'space' y 'mission'.


What Does NASA Do?
NASA's vision: To reach for new heights and reveal the unknown so that what we do and learn will benefit all humankind.
After Apollo, NASA focused on creating a reusable ship to provide regular access to space: the space shuttle. First launched in 1981, the space shuttle flew more than 130 successful missions before being retired in 2011. In 2000, the United States and Russia established permanent human presence in space aboard the International Space Station, a multinational project representing the work of 15 nations.

President Dwight D. Eisenhower established the National Aeronautics and Space Administration in 1958, partially in response to the Soviet Union's launch of the first artificial satellite the previous year. NASA grew out of the National Advisory Committee on Aeronautics (NACA), which had been researching flight technology for more than 40 years.
President John F. Kennedy focused NASA and the nation on sending astronauts to the moon by the end of the 1960s. Through the Mercury and Gemini projects, NASA developed the technology and skills it needed for the journey. Space Technology: rapidly develops, innovates, demonstrates, and infuses revolutionary, high-payoff technologies that enable NASA's future missions while providing economic benefit to the nation.

Figura 3.14: Extracto de un texto con conceptos clave identificados.

Siguiendo con el texto de ejemplo de la Figura 3.14, se encontraron Entidades Nombradas de DBpedia relacionadas con los conceptos clave. Por ejemplo, dado los conceptos clave de la Figura 3.14, se encontraron Entidades Nombradas como:

- Apollo program: Artículo referente al programa más importante en la historia de la NASA.
- NASA Earth Observatory: Artículo acerca de los observatorios de la NASA.
- Hubble Space Telescope: Uno de los telescopios más potentes e importantes de la NASA.

Las Entidades Nombradas anteriores no se encuentran en el texto original como tal, pero analizando el tema principal del texto a través de la identificación de conceptos clave se encontró una relación semántica con esas entidades. Lo anterior muestra la principal diferencia entre la identificación de Entidades Nombradas en esta sección y la identificación descrita en la subsección 3.2

A continuación se describen los submódulos de esta etapa del método.

3.4.1 Identificación de conceptos clave

La identificación de conceptos clave obtiene los términos más importantes en el texto, los cuales mantienen la idea central y más importante del texto completo. Para identificar los conceptos clave se utilizó el algoritmo de *Meaning Circulation* [72], el cual identifica los términos más importantes en un texto utilizando una metodología basada en grafo. Básicamente el algoritmo realiza los siguientes pasos:

- Preprocesamiento
- Modelado y construcción del grafo con base en una ventana de 2 y una ventana de 5.
- Cálculo de nodos (términos) más importantes basado en la medida denominada 'Betweenness centrality measure'.

3.4.1.1. Preprocesamiento

El preprocesamiento aplicado en el algoritmo de Dmitriv [72] es la remoción de artículos, conjunciones, preposiciones y aquellas palabras que no aportan ningún significado pero que aparecen frecuentemente, es decir, la remoción de 'Stopwords' o 'palabras vacías'. Después de eliminar palabras vacías, solamente se toman en cuenta la palabras del texto, dejando fuera los signos de puntuación, símbolos y números. Por último, se aplica un análisis morfológico a cada una de las palabras para obtener la raíz de cada palabra. A continuación se resumen los pasos del preprocesamiento:

- Eliminación de 'stopwords'
- Eliminación de signos de puntuación, símbolos y números
- Stemming

3.4.1.2. Construcción del grafo

Una vez que el texto ha sido preprocesado es posible construir un grafo para analizar el texto e identificar los términos más importantes. Como ya se mencionó anteriormente, en el módulo de preprocesamiento del método de enriquecimiento de texto las palabras se encuentran modeladas como un grafo, conectadas unas con otras según el orden en el texto y cada una de ellas tienen un conjunto de anotaciones. El modelo de grafo que propone Dmitriv Paranyushkin [72] es un grafo en el que se denoten la conexión entre palabras en una determinada ventana, con el objetivo de identificar aquellas palabras que aparecen muy frecuentemente junto a otras.

Para construir el grafo se realizó lo siguiente: se analizó el texto con una ventana de tamaño 2 y después con una de tamaño 5. Por ejemplo, en la Figura 3.15 se muestra un extracto de un texto para describir la construcción del grafo.

What Does NASA do in space?
NASA's vision: To reach for new heights like Apollo program so that what we do and space will benefit all humankind.
After Apollo, NASA focused on creating a reusable ship like apollo to provide regular access to space: the interstellar space shuttle.

Figura 3.15: Extracto de un texto para la construcción de un grafo.

El primer paso para la construcción del grafo es realizar el análisis con una ventana de tamaño 2. Es decir, se crea un nodo por cada palabra y se crea una arista entre cada 2 palabras. En la Figura 3.16 se muestra el texto de la Figura 3.15 una vez aplicado el preprocesamiento que anteriormente se describió: la eliminación de palabras vacías, tomar sólo palabras, su transformación a su raíz y la asociación de palabras con una ventana de tamaño 2.

|nasa space|nasa vision|reach height|apollo program|space benefit|humankind apollo|nasa focus|create reuse|
|ship apollo|access space|interstellar space|shuttle|

Figura 3.16: Texto analizado con un 'gap' de tamaño 2.

Cada ventana representa 2 nodos y un enlace entre ellos, si un nodo ya se encuentra representado

sólo se agrega el enlace y el nodo al cual está conectado. En la Figura 3.17 se muestra el mismo texto representado como un grafo tras asociar palabras en una ventana, dicho grafo es no conexo. Se observa que existen 4 grupos de nodos interconectados pero no así entre los grupos. Este tipo de grafo afecta la extracción de términos importantes del texto ya que podrían quedar nodos importantes sin conectar, por tal motivo se aplica la asociación del texto con una ventana de tamaño 5.

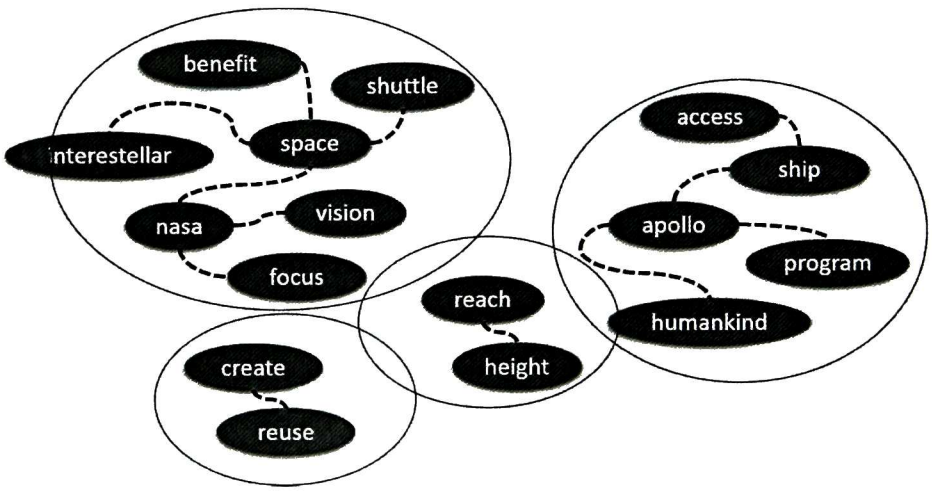


Figura 3.17: Construcción del grafo con un 'gap' de tamaño 2.

En la Figura 3.18 se muestra el texto asociado con un ventana de tamaño 5. Cuando se tiene una ventana de tamaño 5 se conecta la primera palabra y la última palabra de la ventana. En la Figura 3.18 las palabras que se conectan están representadas en negrita dentro de cada ventana.

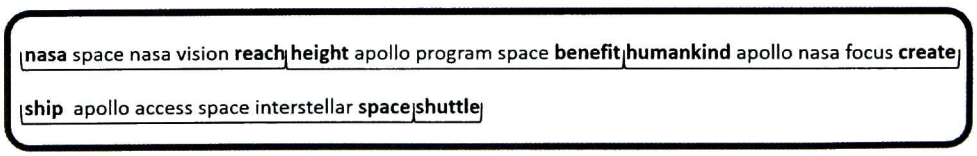


Figura 3.18: Texto analizado con una ventana de tamaño 5.

Tras la asociación con una ventana de tamaño 5 el grafo conecta grupos de nodos que con una ventana de tamaño 2 no estaban conectados. En la Figura 3.19 se muestra el grafo construido con una ventana de tamaño 2 seguido de una asociación utilizando una ventana de tamaño 5. En dicha

Figura se muestra con línea continua los enlaces creados a partir de la segunda asociación. Estos enlaces permiten generar un grafo conexo además de que se obtendrá un mejor desempeño cuando se analice el grafo para obtener los nodos (términos) más importantes del grafo.

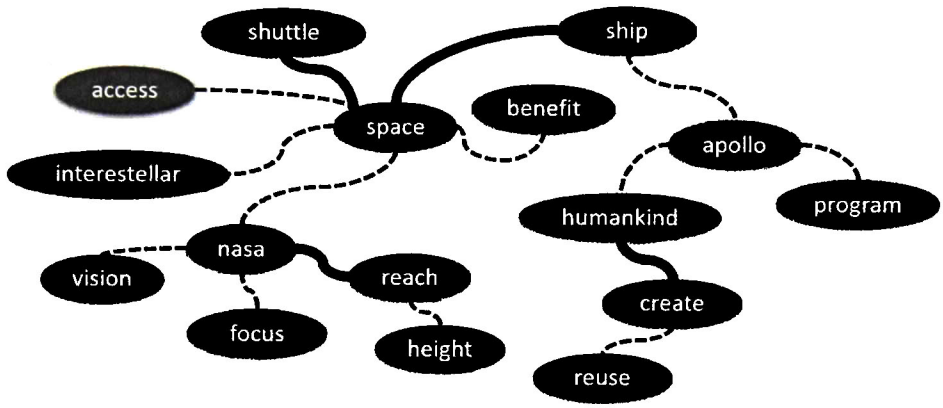


Figura 3.19: Grafo construido con una ventana de tamaño 2 y 5.

El análisis del texto con una ventana de tamaño 5 evita que nodos queden sin conectar a otros nodos importantes. Sin embargo, en ocasiones quedarán nodos sin conectar dado que la ventana de tamaño 5 puede no conectar a todos los grupos sin conectar. Dmitriv Paranyushkin [72] realizó una experimentación en el que los resultados aseguran que los grupos que pueden quedar sin conectar no afectan en la identificación de nodos o términos importantes.

Es necesario mencionar una última consideración dado que en el ejemplo anterior no se presentó el caso. Es el caso en el que una pareja de nodos aparece repetidas veces en el texto, en el grafo es representado con un peso mayor, es decir, en primera instancia las parejas de nodos tienen un peso de 1, si se repite esa aparición su peso es de 2. Al realizar lo anterior se asegura que se le de más importancia a que las parejas de nodos que aparezcan con mayor frecuencia en el texto.

3.4.1.3. *Cálculo de nodos (términos) más importantes basado en la medida 'Betweenness centrality measure'*

Una vez que el texto fue modelado como grafo, los nodos o términos más importantes se obtienen a partir de la medida 'Betweenness centrality measure'. A cada nodo se aplica dicha medida, la cual indica que los nodos más centrales tienen una mayor influencia en el grafo ya que son parte esencial en la comunicación entre los nodos del grafo. Es decir, esta medida es igual al número de rutas más cortas desde ese vértice a todos los vértices restantes. Actualmente existen diversas herramientas para el análisis de grafos, las cuales son usadas para calcular este tipo de medidas. En este caso se utilizó **GraphStrem** [74] para calcular la 'Betweenness centrality measure'. En el ejemplo del extracto de texto que muestra la Figura 3.15 y que la Figura 3.19 muestra el grafo construido, los nodos con un mejor *score* fueron: **nasa**, **space** y **apollo**, los cuales son los nodos centrales en el grafo, representando los términos más importantes del texto.

3.4.2 Obtención de Entidades Nombradas candidatas de DBpedia a partir de conceptos clave

Este submódulo identifica Entidades Nombradas de DBpedia a partir de conceptos clave que anteriormente se identificaron. El proceso para identificar estas Entidades Nombradas utiliza consultas a DBpedia para analizar texto relacionado con estas Entidades Nombradas con los conceptos clave.

Para realizar búsquedas de texto en Entidades Nombradas se utiliza la utilidad de DBpedia conocida como 'Full Text Search'. Esta utilidad permite hacer búsquedas de texto en propiedades de una Entidad Nombrada. Por ejemplo, encontrar Entidades Nombradas que contengan la palabra 'Albert' como nombre de una persona o encontrar todas aquellas Entidades Nombradas que contengan la palabra 'México' en su lugar de nacimiento. Esta tipo de búsquedas de texto en propiedades de una Entidad Nombrada son muy útiles ya que en muchas ocasiones no se tiene el valor de determinada propiedad de la Entidad Nombrada, por lo tanto no se puede realizar el

'*matching*' correspondiente. DBpedia permite realizar este tipo de búsquedas a partir del prefijo ***bif:contains***. Este prefijo se utiliza en una consulta SPARQL especificando a la propiedad(es) en las que se aplica. Asimismo, otra característica importante de esta utilidad es que los resultados pueden ser ordenados de acuerdo a su frecuencia de aparición en una Entidad Nombrada, en orden ascendente o descendente, según sea la cantidad de *hits* que encuentre la utilidad. Recordando que las Entidades Nombradas que se encuentran en DBpedia son el resultado de modelar los artículos de Wikipedia como una Entidad, asignando diferentes propiedades y características. Una propiedad fundamental que se utiliza es el resumen, el cual corresponde al resumen de un artículo de Wikipedia y presenta la información más relevante e importante de cada artículo. Aunque el resumen es muy importante, también se toman en cuenta otras propiedades que contengan texto plano, como son el nombre, lugar de nacimiento, etc.

Cada uno de los conceptos clave del texto original son utilizados para realizar búsquedas en las propiedades que contengan texto plano de cada Entidad Nombrada. Es decir, una Entidad Nombrada es tomada en cuenta siempre y cuando contengan los conceptos clave más importantes del texto original. El número de conceptos clave para realizar la búsqueda es determinado en la sección 4.9 (Experimentación). En la Figura 3.20 se muestra un esquema para la identificación de Entidades Nombradas a partir de los conceptos clave previamente identificados en el texto original.

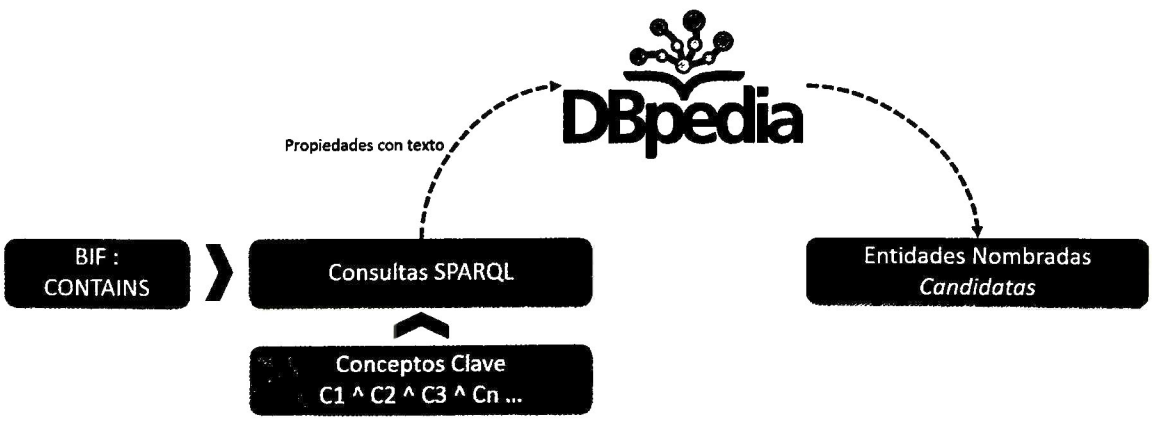


Figura 3.20: Esquema para la identificación de Entidades Nombradas a partir de conceptos clave.

Siguiendo el ejemplo del extracto de texto que muestra la Figura 3.14, los conceptos clave que se determinaron para los ejemplos dados fueron **nasa**, **space**, **mission**. Algunas de las Entidades Nombradas resultantes fueron:

- Apollo program: Artículo referente al programa más importante en la historia de la NASA.
- NASA Earth Observatory: Articulo acerca de los observatorios de la NASA.
- Hubble Space Telescope: Uno de los telescopios más potentes e importantes de la NASA.

Las Entidades Nombradas resultantes son denominadas *candidatas* por que son el resultado de consultas basadas en la aparición de los conceptos clave y no así por un análisis en el significado y temática de la Entidad Nombrada, es decir, no se obtuvieron a partir de una identificación propia de Entidades Nombradas en el texto original. Éstas son denominadas de esa manera porque es necesario realizar un mayor análisis para verificar si esa Entidad Nombrada realmente tiene una relación con el texto original. Esta tarea se aborda y se soluciona en el submódulo siguiente.

3.4.3 Latent Semantic Indexing

Las Entidades Nombradas que se identifican con base en conceptos clave no han sido analizadas para verificar si éstas realmente tienen relación con el texto en original. Es decir, aunque los conceptos clave aparezcan en la Entidad Nombrada no es indicio confiable de que la Entidad Nombrada tenga relación con el texto original. Por ejemplo, en la Figura 3.21 se muestra un ejemplo que caracteriza el comportamiento antes mencionado. A partir del texto de la Figura 3.14 en el que el tema principal es la NASA se extraen los conceptos clave: **nasa** , **space**, **earth** y **technology**. Las Entidades Nombradas resultantes son 4, tras una verificación manual se llegó a la conclusión de que todas tienen relación con el texto, excepto una Entidad Nombrada denominada como 'The Phenomenauts'. Dicha Entidad Nombrada contiene los conceptos clave antes mencionados, pero no tiene una relación directa con el tema principal del texto debido a que es un grupo musical. Este tipo de resultados erróneos se deben

de evitar, una solución a este tipo de resultados es aplicar el algoritmo *Latent Semantic Indexing* (LSI). LSI es un algoritmo que permite medir la similitud de documentos dadas unas palabras, pero con la diferencia de que no sólo toma en cuenta las apariciones de las palabras en los documentos, sino que ve el conjunto de documentos como un todo para determinar aquellos documentos que contienen esas palabras y aquellos documentos que no las contienen pero contienen otras palabras que están semánticamente cerca, descartando aquellos documentos que están semánticamente distantes. En otras palabras, LSI analiza todo el contenido del documento y recibe un conjunto de palabras, tras analizar cada uno de los documentos con base al conjunto de palabras determina aquellos documentos semánticamente similares.

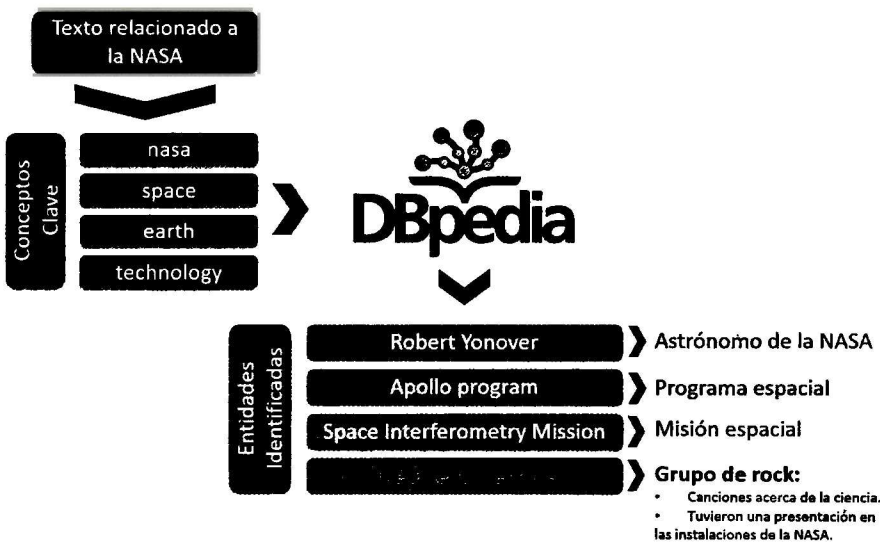


Figura 3.21: Ejemplo que caracteriza un resultado erróneo en la identificación de Entidades Nombradas a partir de conceptos clave.

Adaptando el algoritmo LSI al problema antes mencionado, las palabras son los conceptos clave identificados previamente, con la diferencia de que el número de conceptos clave a considerar será mayor por que es necesario dar más información al algoritmo acerca del documento original. El conjunto de documentos a analizar son las Entidades Nombradas candidatas (el contenido en texto plano como los resúmenes). Las Entidades Nombradas resultantes del algoritmo LSI cuyo contenido

en texto plano tiene relación semántica con los conceptos clave se utilizan para enriquecer el texto en un proceso posterior. En la Figura 3.22 se muestra un esquema del algoritmo LSI en el método de enriquecimiento de texto.

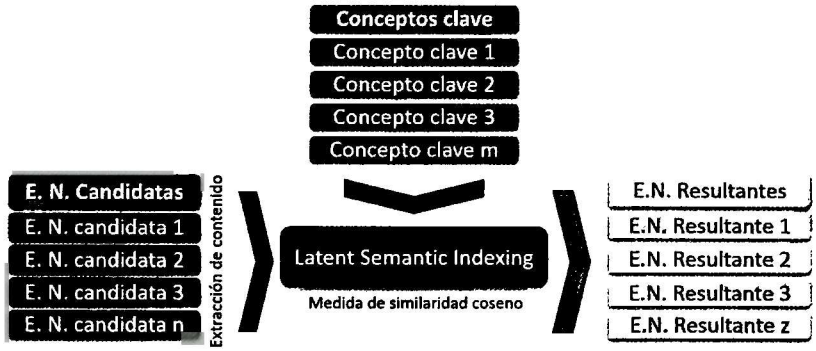


Figura 3.22: Esquema del algoritmo LSI en el método de enriquecimiento de texto.

Por último, hay dos criterios a considerar, el primero es determinar el número de conceptos clave con los que se trabajó en el algoritmo LSI. Para obtener las Entidades Nombradas candidatas se consideraron **4 conceptos clave**, este número fue determinado tras la experimentación descrita en la sección 4.9 (Experimentación). El otro criterio es la medida de similaridad utilizada para determinar la semejanza entre los documentos y los conceptos clave utilizados, dicha medida es la similitud de coseno. Dado que tras comparar los documentos con los conceptos clave se obtuvo un 'score', fue necesario determinar cuál es el umbral para determinar los documentos relevantes, excluyendo a aquellos irrelevantes. El umbral que se obtuvo fue de **0.865**, este umbral fue determinado tras la experimentación descrita en la sección 4.9 (Experimentación).

3.5 Estructuración de Información

El método de enriquecimiento de texto está orientado a proveer al usuario final información adicional relacionada a un texto original. La información proveída debe ser clara, precisa y organizada con respecto al texto original. Existen diversos factores que impulsaron a diseñar una Estructuración

de Información, los cuales son descritos a continuación:

- **Redundancia**: Se desea tener la menor redundancia posible, es decir, utilizar los recursos lo menos posible, un ejemplo es no repetir las Entidades Nombradas encontradas. Puesto que cuando se obtenga información se tendrá redundancia de información.

- **Consultas múltiples**: Debido a que se utilizará DBpedia, el tiempo de respuesta es un factor fundamental. Por esa razón se desea consultar a DBpedia el menor número posible de veces y no así consultas múltiples sobre una misma Entidad Nombrada.

- **Procesamiento**: El método de enriquecimiento de texto utiliza, en diversas ocasiones, recursos según se requiera. El proceso de búsqueda de *tokens*, Entidades Nombradas o información relacionada debe ser lo más eficiente posible. Los resultados obtenidos de cada procesamiento deben estar organizados con respecto al *token* procesado para su posterior utilidad. Asimismo, la información obtenida en el módulo denominado como 'Enriquecedor de Información' debe estar organizada adecuadamente para su posterior consumo.

- **Presentación al usuario final**: Dado que el texto enriquecido debe presentarse al usuario final de una manera clara, es necesario presentar el texto original tal y como es. Haciendo énfasis en la información obtenida sobre las Entidades Nombradas y conceptos clave identificados.

En la Figura 3.23 se muestra el diseño de la Estructuración de Información, enseguida se describe cada una de las tareas.

3.5.1 Índice de *tokens*

El índice de *tokens* organiza las palabras, números, símbolos y espacios del texto original. Asigna una posición con respecto al texto y los resultados del preprocesamiento aplicado se organizan de acuerdo al *token*, por ejemplo el tipo, la raíz o la categoría del *token*. Este índice permite tener un acceso rápido y ordenado al conjunto de *tokens* que corresponde al texto original a procesar.

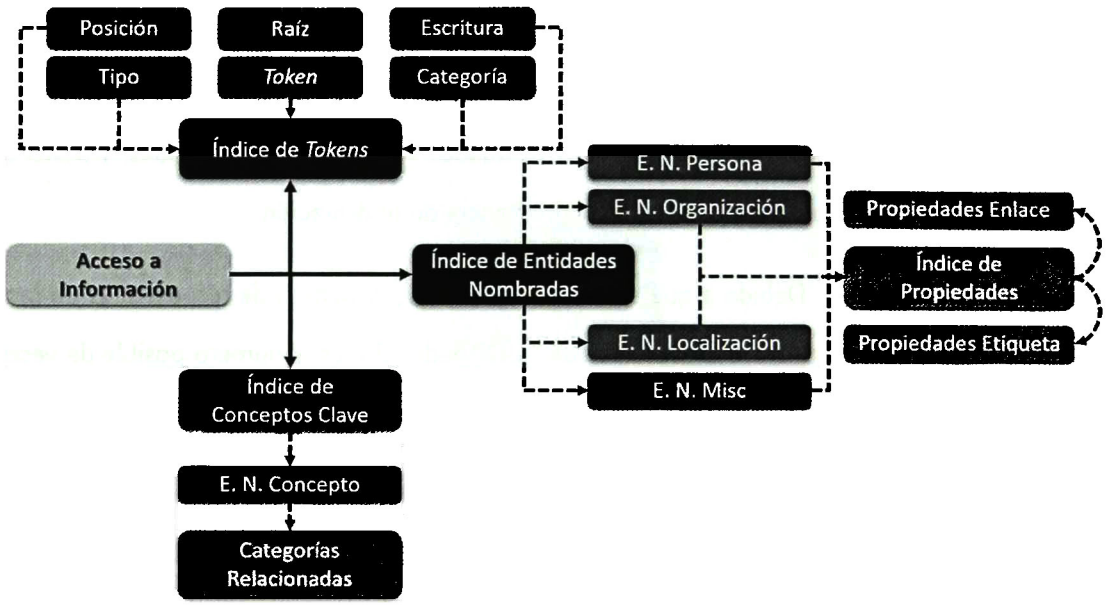


Figura 3.23: Diseño de la Estructuración de Información.

3.5.2 Índice de Entidades Nombradas

Permite la organización de las Entidades Nombradas encontradas en el bloque de 'Extracción de Entidades Nombradas' y la 'Extracción de Entidades Nombradas a partir de conceptos clave'. Dado que existen 4 tipos de Entidades Nombradas: Persona, Organización, Localización y Misc, se organizaron para evitar la duplicación de información, evitando así las múltiples consultas y el almacenamiento de información innecesaria. Cada Entidad Nombrada tiene un conjunto de propiedades, las cuales se dividen en propiedades enlace y propiedades etiqueta. Dichas propiedades se describen en la sección 3.5. Cada Entidad Nombrada contiene un conjunto de *tokens*, el cual hace referencia a la organización en el Índice de *tokens*.

3.5.3 Índice de Conceptos Clave

Permite la organización de los conceptos clave identificados, las Entidades Nombradas relacionadas con los conceptos clave y las categorías relacionadas. Las categorías relacionadas se

describen en la sección 3.5. Los conceptos clave identificados hacen referencia a *tokens* del Índice de *tokens*. Las Entidades Nombradas identificadas a partir de conceptos clave hacen referencia a un conjunto de *tokens* al igual que el Índice de Entidades Nombradas. Por cada Entidad Nombrada identificada se obtiene la categoría a la que pertenece dicha Entidad Nombrada. Estas categorías son organizadas de la misma forma.

3.5.4 Acceso a Información

Este acceso es el que se utiliza en todo el método de enriquecimiento de texto, permite acceder a los *tokens*, a información relacionada con las Entidades Nombradas identificadas almacenando información resultante del procesamiento así como información relacionada con DBpedia. De la misma forma permite acceder a las Entidades Nombradas identificadas a partir de los conceptos clave así como las categorías a las que pertenecen. Durante la ejecución del método la Estructuración de Información es utilizada para almacenar información resultante del procesamiento así como información relacionada como DBpedia.

3.6 Enriquecedor de Información

El enriquecedor de Información es el módulo encargado de obtener información relacionada de DBpedia con el texto a partir de las Entidades Nombradas encontradas en el módulo de 'Extracción de Entidades Nombradas' y 'Extracción de Entidades Nombradas' a partir de conceptos clave. Además el Enriquecedor de Información obtiene información relacionada, es el módulo encargado de crear el texto enriquecido. Es decir, genera el texto original con las Entidades Nombradas identificadas y los enlaces a información relacionada, enlaces a Wikipedia y enlaces a páginas externas. El Enriquecedor contiene 2 sub-módulos: motor de consultas y constructor de conocimiento, los cuales se describen a continuación.

3.6.1 Motor de consultas

La base de conocimiento que el Enriquecedor de Información utiliza es DBpedia. El motor de consultas toma una Entidad Nombrada identificada y obtiene información relacionada de DBpedia. El enriquecedor obtiene 2 propiedades fundamentales de una Entidad Nombrada de DBpedia:

- Propiedades Etiqueta: Estas propiedades describen características de una Entidad, por ejemplo la fecha de nacimiento o el nombre completo.
- Propiedades Enlace: Estas propiedades describen las relaciones de la Entidad Nombrada en relación a otras Entidades Nombradas, por ejemplo padres de una persona, obteniendo la relación de padre y la etiqueta de nombre.

Además de las propiedades mencionadas anteriormente se obtuvieron los enlaces a los artículos de Wikipedia y enlaces a otras páginas con información relacionada. En la Figura 3.24 se muestra un ejemplo de las relaciones y características que se mencionaron anteriormente. Por ejemplo, si la Entidad Nombrada identificada es Albert Einstein, se obtiene sus Propiedades Etiquetas, por ejemplo el año de nacimiento y el *abstract*. De la misma manera se obtienen las Propiedades Enlace, las cuales representan relaciones con otras Entidades, por ejemplo el nombre de su esposa y tutor doctoral. Por último los enlaces a Wikipedia y externos también son obtenidos.

Básicamente el método de enriquecimiento de texto encuentra relaciones a artículos o documentos modelados como Entidades Nombradas en DBpedia. Estas relaciones pueden ser divididas en 2 tipos de relaciones, aquellas relaciones que se encuentran a partir de la Extracción de Entidades Nombradas y aquellas relaciones que se encuentran a partir de Conceptos Clave. Un ejemplo que muestra las relaciones se encuentra en la Figura 3.25. En dicha Figura se muestra un texto enriquecido y un extracto del grafo de DBpedia. En la Figura 3.25 se muestran los dos tipos de relaciones que se mencionaron anteriormente. Las relaciones a partir de la Extracción de Entidades Nombradas se denotan con una línea continua, mientras que las relaciones a partir de Conceptos Clave se denotan

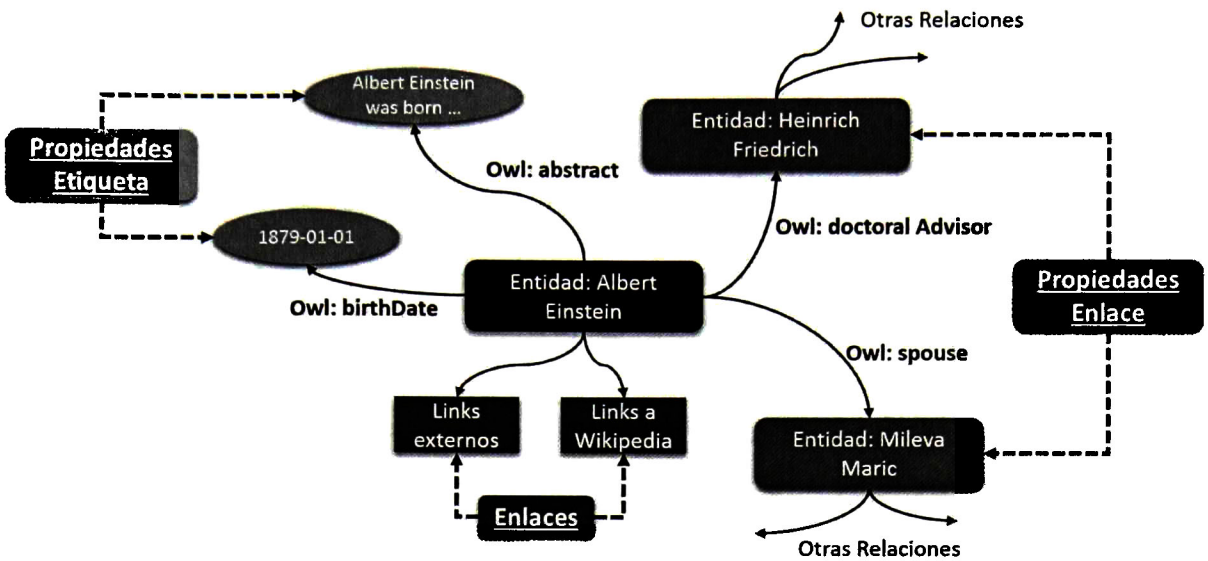


Figura 3.24: Ejemplo de las propiedades y características obtenidas de una Entidad Nombrada.

con una línea punteada. Como se muestra en dicha Figura las relaciones a partir de la Extracción de Entidades Nombradas tienen una relación directa en puntos distantes del grafo de DBpedia pero que aparecen en un mismo texto. Por otro lado, las Entidades Nombradas a partir de conceptos clave tienen una relación estrecha entre ellas. Dado que la identificación de Entidades Nombradas de este tipo es a partir de una representación resumida del texto (conceptos clave identificados), las Entidades identificadas representan el tema principal o son muy cercanas.

DBpedia tiene organizadas sus Entidades Nombradas en una jerarquía de conceptos. Los conceptos superiores en la jerarquía son conceptos que se denominan como categorías, ya que estos conceptos son más generales que los conceptos inferiores. Esta categoría puede ser vista como un tópico general a los conceptos inferiores. Haciendo uso de la organización de las Entidades Nombradas en una jerarquía de conceptos, al identificar un concepto (Entidad Nombrada a partir de conceptos clave) se obtiene la categoría a la que pertenece dicho concepto. Esta categoría se utiliza para realizar una asignación de un tema al texto original. En la Figura 3.26 se muestra un ejemplo de la identificación de conceptos (Entidades Nombradas a partir de conceptos clave) y la obtención de una

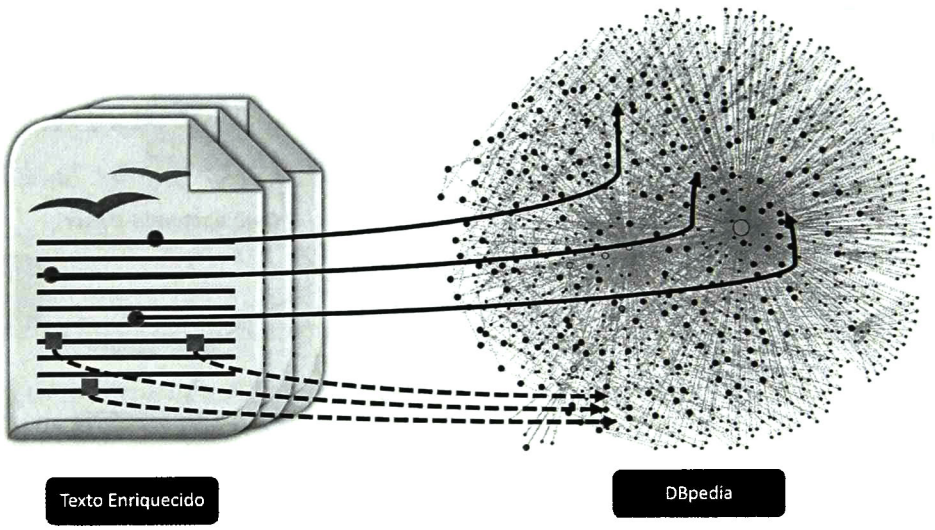


Figura 3.25: Relaciones en el texto a DBpedia.

categoría. Las figuras representadas con un círculo representan a los conceptos, cuando se identifica el concepto en DBpedia se obtiene la categoría a la que pertenece. Dicha categoría corresponde a otro concepto más general, el cual describe a los conceptos inferiores en la jerarquía.

3.6.2 Constructor de conocimiento

Este submódulo es el último del módulo de Enriquecedor de Información y la última parte del Método de Enriquecimiento de texto. El constructor de conocimiento construye el texto enriquecido a partir de la Estructuración de Información descrita en la sección 3.4. En la Figura 3.27 se muestra un esquema de la construcción del texto enriquecido.

El constructor de conocimiento genera un texto más amplio a partir de 2 documentos, al conjunto de estos dos documentos se le denomina como *texto enriquecido*. El primero es un documento que es equivalente al texto original, pero además hace énfasis en las Entidades Nombradas identificadas y los conceptos clave identificados. El segundo documento generado es el que contiene la Información relacionada con el texto. Este documento se construye a partir de la información relacionada con las Entidades Nombradas y los Conceptos Clave. Finalmente, en cada Entidad Nombrada y Concepto

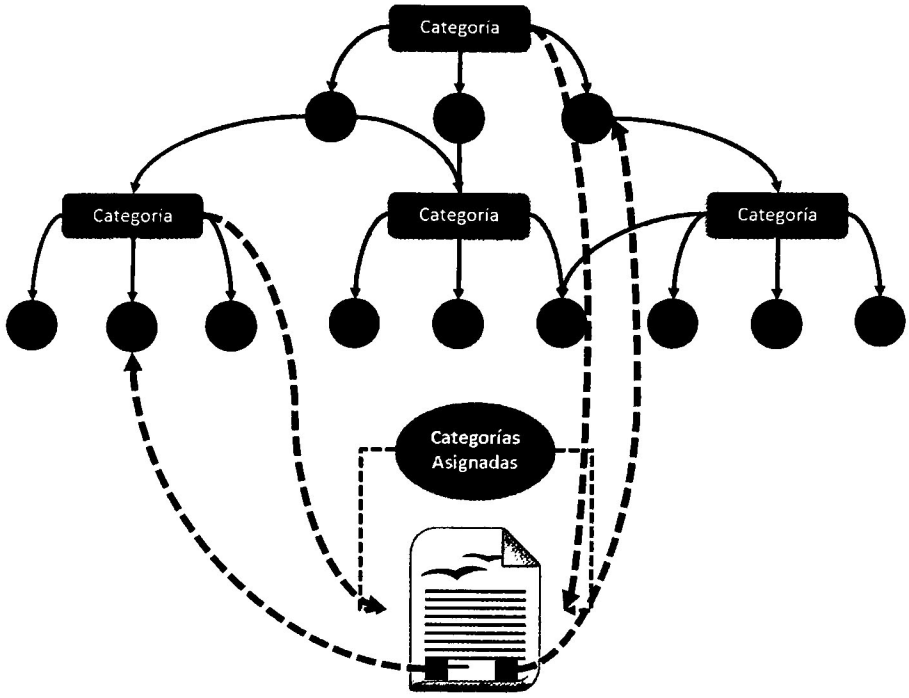


Figura 3.26: Ejemplo de la obtención de categorías a partir de conceptos clave.

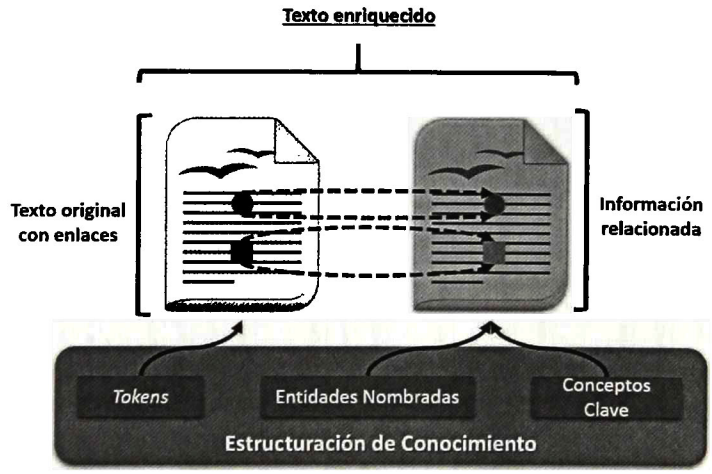


Figura 3.27: Esquema de la construcción del texto enriquecido.

Clave identificados en el primer documento se construye un enlace a la posición exacta de la información relacionada en el segundo documento.

3.7 Resumen

El método de enriquecimiento de texto está compuesto por 5 módulos: Preprocesamiento, Extracción de Entidades Nombradas, Extracción de Entidades Nombradas a partir de conceptos clave, Estructuración de Información y Enriquecedor de Información. El método consta de 3 pasos fundamentales, identificación de las secciones del texto a enriquecer, la búsqueda de información relacionada con estas secciones y la integración de esa información con el texto original. Por otra parte, el funcionamiento de las distintas tareas en cada etapa se apoyan en otras, lo cual fue descrito en las secciones anteriores. Asimismo en la Figura 3.28 se muestra un diagrama de la secuencia del Método de Enriquecimiento de texto, denotando las tareas más importantes y el orden de ejecución de cada una de ellas.

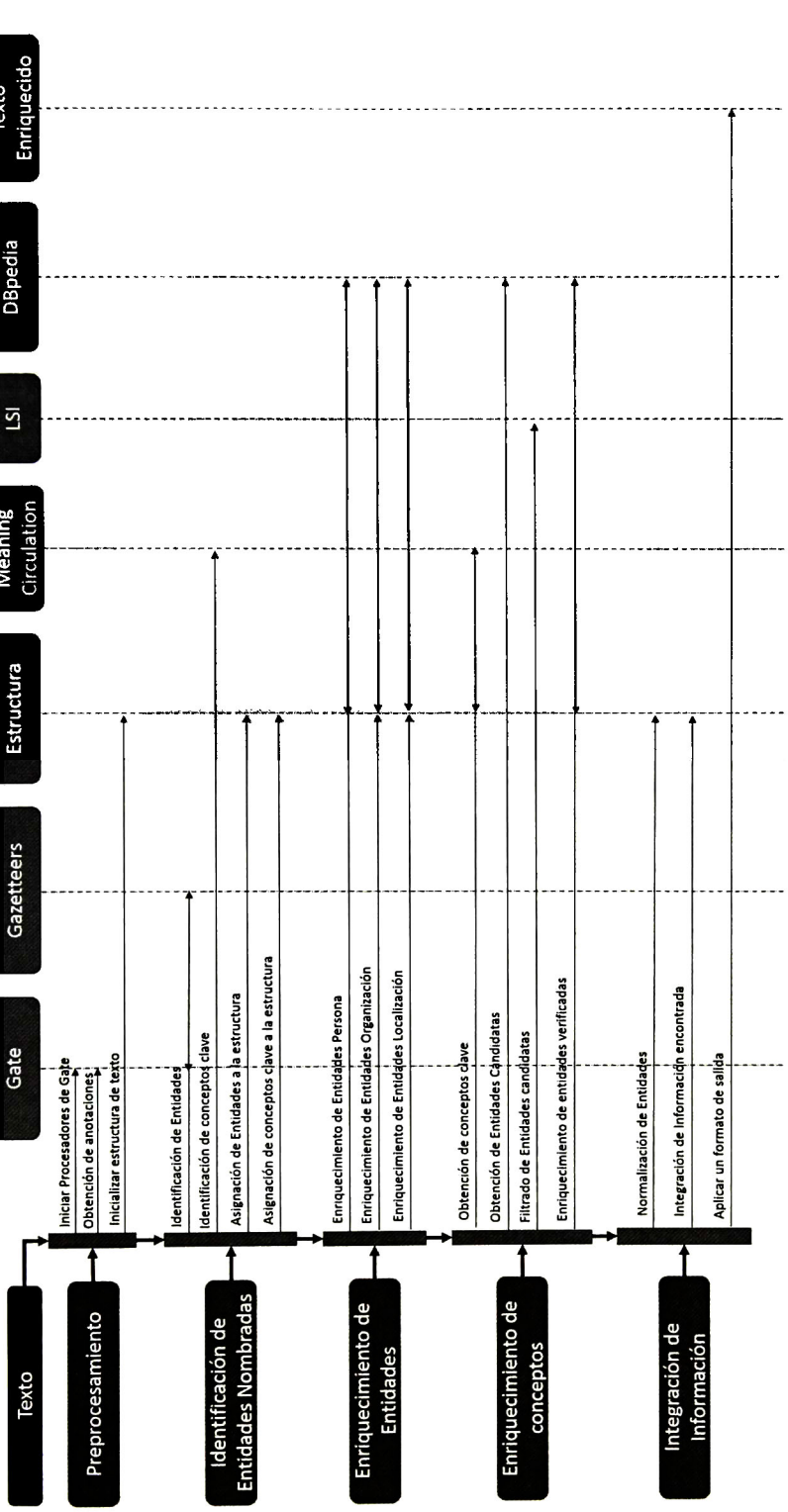


Figura 3.28: Diagrama de tareas del Método de Enriquecimiento de Texto.

4

Resultados

En este capítulo se muestran los experimentos y resultados obtenidos de la implementación del método de enriquecimiento de texto a partir de recursos de la Web Semántica.

4.1 Introducción

La evaluación del método de enriquecimiento de texto está dividida en 5 experimentos. Cada experimento evalúa el desempeño del método de enriquecimiento de texto con un diferente enfoque. El primer experimento es el más simple, el cual sirvió para ajustar el método a lo largo de su desarrollo. Esta evaluación es de forma manual, es decir, se determinó mediante una inspección manual si la información encontrada tenía alguna relación con el texto original. Dicho experimento se describe en la sección 4.4. El segundo experimento está basado en la comparación del texto original y el texto enriquecido, es decir, comparar qué tan diferente es el texto enriquecido con respecto al original, este experimento se describe en la sección 4.5. Para los experimentos 3, 4 y 5 se utilizaron diversos *datasets*, los cuales se denominaron como *datasets originales*. Cada *dataset* fue enriquecido con

información determinada por el método y se les denominó *datasets enriquecidos*. En la sección 4.6 se describe un experimento de clasificación de texto. Los *dataset* original y enriquecido son evaluados en tareas de clasificación utilizando diferentes algoritmos. En el cuarto experimento se determina la Ganancia de Información (GI), en este experimento se evalúa la GI con los mismos *datasets*, dicho experimento es descrito en la sección 4.7. Por último, en el quinto experimento se realiza el *clustering* utilizando los *datasets* originales y enriquecidos, el experimento es descrito en la sección 4.8.

4.2 Infraestructura

Las características del hardware empleado para la realización de los experimentos es la siguiente:

- Procesador Intel® I7 a 2.93 GHz
- Memoria RAM de 8 GB
- Disco Duro de 2 TB
- Sistema Operativo: Ubuntu 14.04

Para enriquecer cada uno de los *datasets* que se utilizaron en cada experimento se desarrolló una infraestructura virtual debido a que se utilizó DBpedia. DBpedia limita el número de peticiones en un determinado tiempo para una IP específica. En la Figura 4.1 se muestra un esquema que representa la infraestructura desarrollada.

En la infraestructura desarrollada se utilizaron 4 máquinas virtuales, las cuales fueron utilizadas para realizar las tareas del motor de consultas. Cada máquina virtual se conecta a una VPN (Virtual Private Network) para obtener una IP diferente y así evitar que DBpedia limite la conexión. Los equipos virtuales tienen las siguientes características:

- 1 Procesador

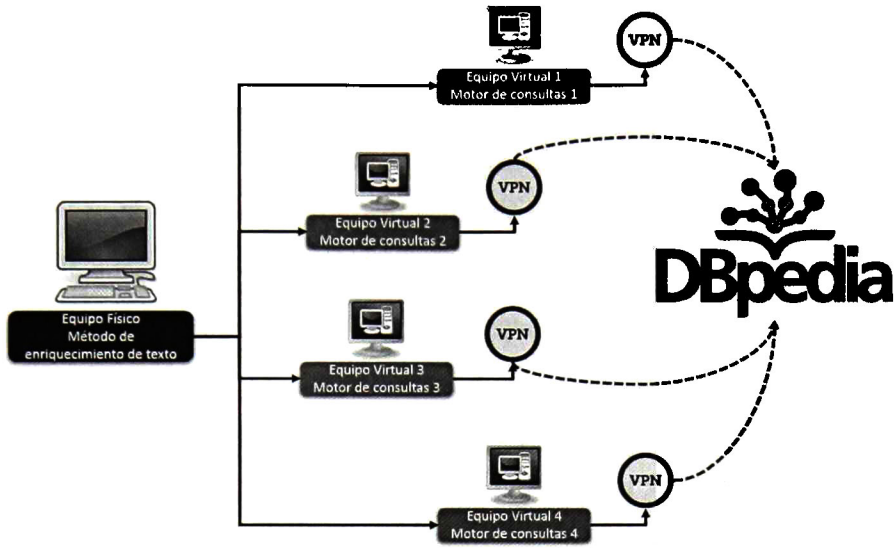


Figura 4.1: Esquema que representa la infraestructura desarrollada para enriquecer los *datasets* utilizados

- Memoria RAM de 1GB

- Disco Duro de 50 GB

- Sistema Operativo: Ubuntu 14.04

El procesamiento completo del método de enriquecimiento de texto se realizó en el equipo físico a excepción de la parte del motor de consultas, el cual se conecta con DBpedia para obtener información. Los resultados de cada equipo virtual son enviados al equipo físico para que el método de enriquecimiento de texto obtenga los resultados de DBpedia y complete el proceso.

4.3 Representación del texto enriquecido para la experimentación

El método de enriquecimiento de texto está enfocado a proveer al usuario final información relacionada a un texto original. La información relacionada se entrega al usuario en un conjunto de documentos para su fácil entendimiento. Como se mencionará en secciones posteriores, el enriquecimiento de texto ayuda a mejorar tareas de clasificación y *clustering*, por lo tanto se han desarrollado experimentos relacionados a estas tareas. Debido a ésto, se necesita representar el texto enriquecido como un solo documento y no como un conjunto de documentos.

Para representar un documento enriquecido se utilizó el algoritmo de *Meaning Circulation* [72], el cual representa un documento como grafo. Cuando se identifica una Entidad Nombrada en el texto se obtienen los términos más importantes de la información obtenida utilizando este algoritmo. El conjunto de términos obtenidos se concatenan enseguida de cada Entidad Nombrada identificada. Por ejemplo en la Figura 4.2 se muestra un extracto de un texto. En ese extracto se identificaron tres Entidades Nombradas: *Albert Einstein, General Theory of Relativity y Gravity*.

In 1915, **Albert Einstein** developed the **General Theory of Relativity**, one of the two pillars of modern physics and **gravity**.

Figura 4.2: Ejemplo de un extracto de un texto con Entidades Nombradas identificadas.

Tras el enriquecimiento de texto se obtuvo información relacionada con las Entidades Nombradas. Con el algoritmo de *Meaning Circulation* [72] se obtuvieron los términos más importantes a partir de la información relacionada. El texto enriquecido a partir del texto original en la Figura 4.2 se muestra en la Figura 4.3. El texto presentado en las Figuras 4.2 y 4.3 fueron utilizados para realizar los experimentos siguientes.

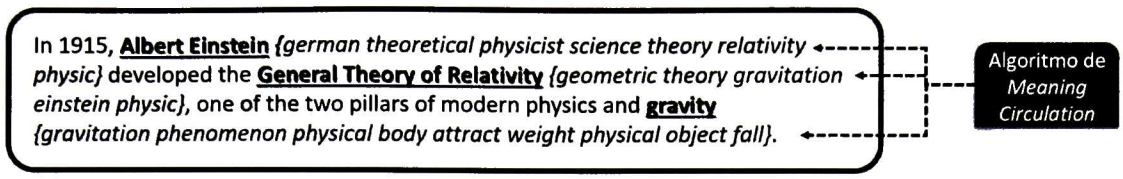


Figura 4.3: Texto enriquecido a partir del texto en la Figura 4.2.

4.4 Primer experimento

El primer escenario de prueba está basado en evaluar las Entidades Nombradas que el método encuentra para enriquecer el texto. Para realizar dicha evaluación se descargaron documentos desde *Springer Link*, específicamente de la sección en que los documentos están divididos por disciplina.

En la Tabla 4.1 se muestran las diferentes disciplinas a las que pertenecen los documentos así como la cantidad de documentos por disciplina. Se utilizaron un total de 100 documentos.

Disciplina	Cantidad de documentos
Arquitectura	10
Astronomía	10
Química	10
Computación	10
Geología	10
Economía	10
Nutrición	10
Derecho	10
Medicina	10
Física	10

Tabla 4.1: Cantidad de documentos por disciplina.

Por cada documento se evaluó de forma manual cada una de las Entidades Nombradas que el método determinó como relevantes en el contexto del texto original. A continuación se muestra la forma en que se calculó la precisión por documento.

$$P = \frac{Número\ de\ Entidades\ relevantes}{Número\ de\ Entidades\ identificadas}$$

Para determinar la precisión por disciplina se promedió la precisión de cada documento que pertenece a una disciplina determinada. En la Tabla 4.2 se muestra la precisión promedio de las Entidades Nombradas obtenidas por cada disciplina.

Disciplina	Precisión
Arquitectura	0.960
Astronomía	1.000
Química	0.927
Computación	0.955
Geología	0.920
Economía	0.904
Nutrición	0.940
Derecho	0.975
Medicina	0.925
Física	1.000

Tabla 4.2: Precisión por disciplina.

En general, las Entidades Nombradas que el método obtiene son bastante precisas y tienen relación semántica con el texto original. Así, determinando con alta precisión las Entidades Nombradas con las que se apoyará el método se asegura un enriquecimiento correcto y relacionado con el texto. Por lo contrario, si el enriquecimiento no fuera el correcto no ayudaría, sino por el contrario sería contraproducente para el enriquecimiento debido a que se introduciría información no relacionada con el texto original. Al realizar este experimento se obtuvo una precisión promedio de **0.9508**, tomando en cuenta la precisión por disciplina.

4.5 Segundo experimento

El segundo experimento evalúa qué tan diferente es el texto enriquecido con respecto del original. Para realizar la comparación entre el texto original y el enriquecido se aplica el algoritmo de *Meaning Circulation* [72] al texto original y enriquecido, obteniendo una lista de términos más importantes del texto original y otra lista de términos más importantes del texto enriquecido. Cada lista es muestra

representativa del texto original y enriquecido, respectivamente. En la Figura 4.4 se muestra un diagrama que ilustra la forma en que se diseñó el segundo experimento.

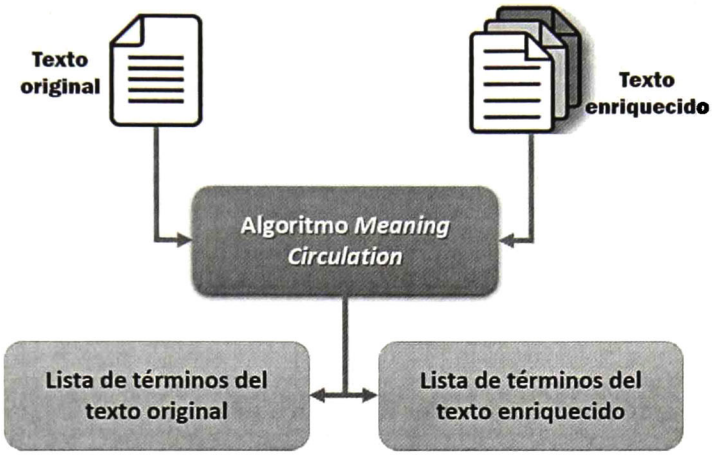


Figura 4.4: Esquema desarrollado para el segundo experimento.

Para realizar este experimento se utilizaron 100 diferentes documentos, los mismos utilizados en el experimento 1. Cada documento fue enriquecido a través del método de enriquecimiento de texto. Es decir, se obtuvieron 100 diferentes documentos con sus respectivos 100 documentos enriquecidos. Por cada documento original se obtuvo una lista de términos, la cual se denomina como *lista original*. De la misma manera se obtuvo una lista de términos por cada documento enriquecido, la cual se denomina como *lista enriquecida*.

Para comparar el texto original y enriquecido se realizaron diversos análisis; el primero evalúa la aparición de un término en la lista correspondiente al texto original (lista original) y en la lista correspondiente al texto enriquecido (lista enriquecida). Es decir, toma un término de la lista original y se verifica si se encuentra en la lista enriquecida. El proceso anterior se repite por cada uno de los términos de la lista original.

La Figura 4.5 muestra en cuántas ocasiones un término de la lista original sigue apareciendo en la lista enriquecida sin importar su posición. Los términos más importantes de la lista original siguen apareciendo, mientras que los términos menos significativos en pocas ocasiones siguen apareciendo

en la lista enriquecida. El análisis anterior indica que los términos más importantes del texto siguen siendo parte fundamental del texto enriquecido, mientras no así con términos menos significativos en muchas ocasiones.

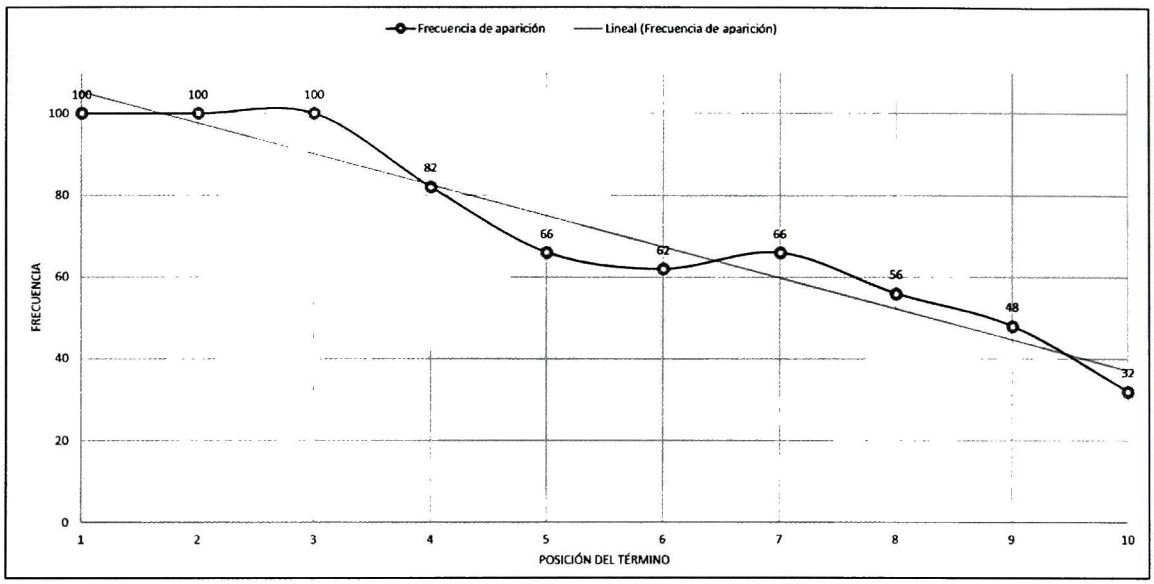


Figura 4.5: Frecuencia de aparición de términos de la lista original en la lista enriquecida.

Un segundo análisis muy similar al anterior se realizó con las listas del texto original y enriquecido. Este análisis parte de observar no solamente si un término de la lista original aparece en la lista enriquecida, sino que toma en cuenta la posición en que se encuentra en su respectiva lista. Es decir, si el primer término de la lista original aparece en la misma posición en la lista enriquecida, se toma como un resultado favorable. En la Figura 4.6 se muestra la frecuencia de aparición en la misma posición entre la lista original y la lista enriquecida. La Figura muestra un comportamiento esperado siempre y cuando el enriquecimiento se esté realizando de manera correcta, ya que en 68 ocasiones el término más importante prevalece tanto en la lista original como en la enriquecida. En 40 ocasiones prevalece el segundo y tercer término en la misma posición. Por tanto, el tema principal del texto original prevalece en el texto enriquecido. Por el contrario, los términos menos significativos, es decir los términos en la posición 8, 9 y 10 de la lista original, se encuentran en menos de 20 ocasiones en

la lista enriquecida.

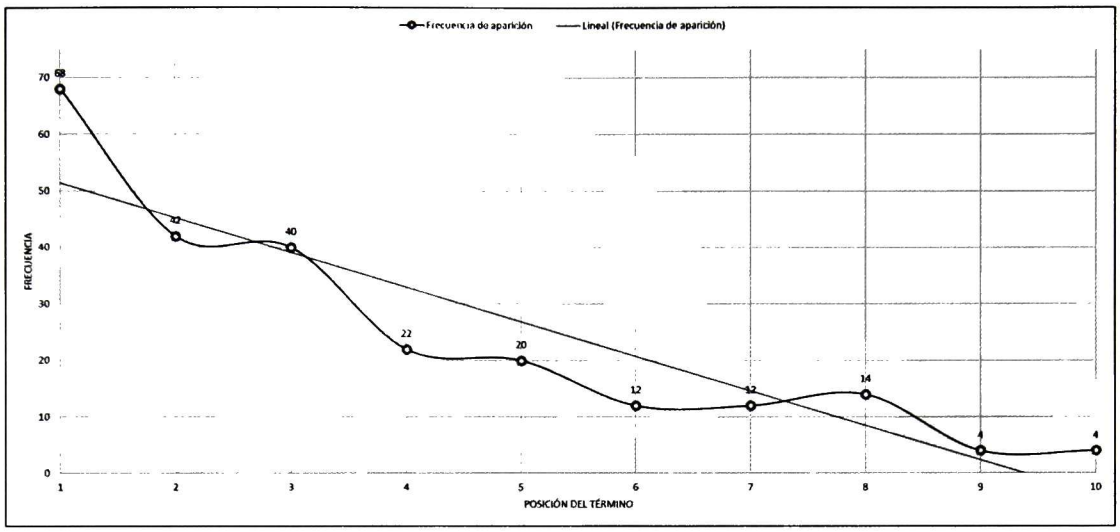


Figura 4.6: Frecuencia de aparición en la misma posición de términos entre la lista original y enriquecida.

Además de los dos análisis anteriores se realizó un tercer análisis con las listas generadas del texto original y el texto enriquecido respectivamente. El tercer análisis no sólo observa la frecuencia de aparición en la misma posición entre los términos de las listas, sino que además toma en cuenta la posición a la que se ha movido el término. Es decir, toma en cuenta a qué posición se ha movido el término de la lista original en la lista enriquecida. En la Figura 4.7 se muestra el intercambio de posición entre los términos de la lista original y la lista enriquecida. Cada sub-barra muestra la frecuencia en la que ha cambiado a determinada posición. Además de mostrar el intercambio de posiciones, la frecuencia de cada barra muestra en cuántas ocasiones sigue apareciendo el término de la lista original en la lista enriquecida. Es decir, la Figura, además de proporcionar el intercambio de posiciones, también aporta un resumen de los análisis anteriores.

El análisis muestra un comportamiento en el que los términos más importantes intercambian su posición entre ellos con una mayor frecuencia, es decir, el término en la primera posición de la lista original se mantiene en 68 ocasiones en la misma posición de la lista enriquecida, en 12 lo intercambia a la segunda posición y en 8 en la tercera posición. A las demás posiciones sólo se intercambia en

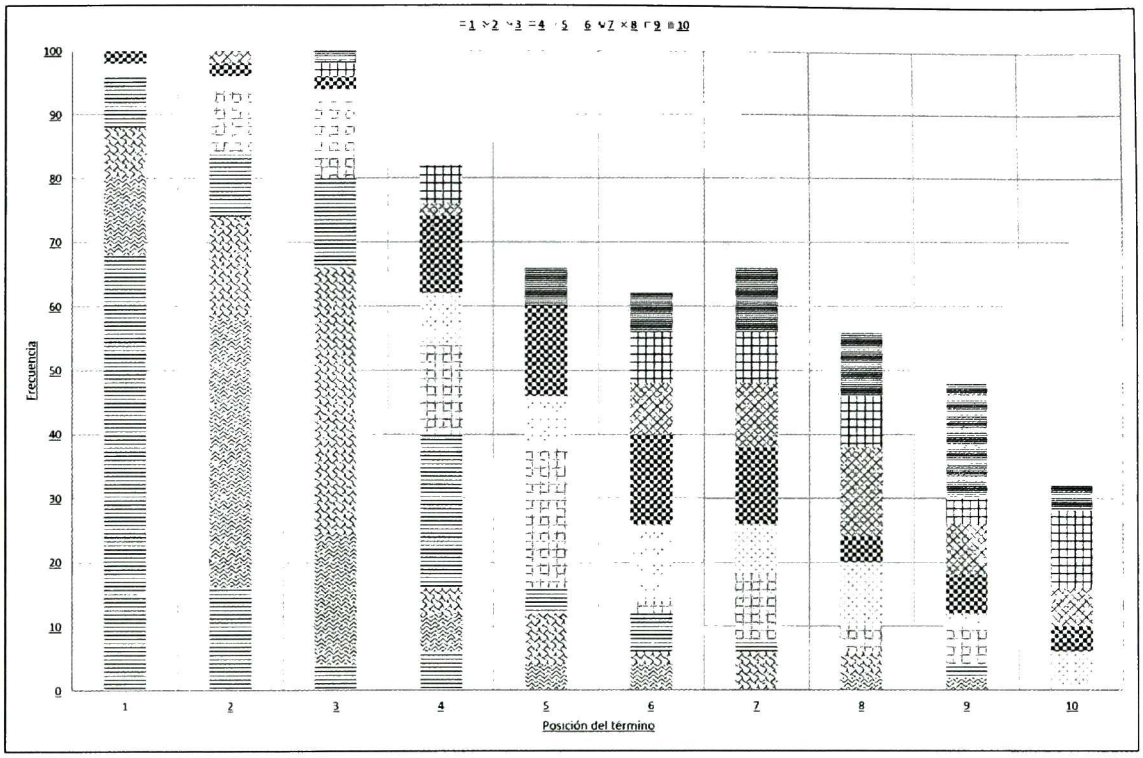


Figura 4.7: Intercambio de posición entre los términos originales y términos enriquecidos.

menos de 20 ocasiones. Lo anterior muestra un comportamiento favorable, ya que el tema principal se preserva en el texto enriquecido, por ello se asume que el enriquecimiento se realizó de una manera correcta.

Un último análisis se realizó con base en los tres anteriores, es decir, se calificó el enriquecimiento de texto dependiendo de la posición de aparición de los términos entre las listas generadas. Es decir, con base al intercambio de los términos más importantes (1 posición, 2 posición, 3 posición) se calificó como bueno, regular o malo el enriquecimiento aplicado. En la Tabla 4.3 se muestran los criterios de calificación utilizados para cada uno de los documentos de texto que se enriquecieron. Cada criterio está basado en los términos más importantes (1 posición 3 posición) de la lista original y en qué posición se encuentran en la lista enriquecida. Por ejemplo, si los términos de la lista original se encuentran en las posiciones 1-3 de la lista enriquecida, el enriquecimiento se considera como bueno,

si se encuentran en las posiciones 4-7 se considera como regular y finalmente si se encuentran en las posiciones 8-10 se considera como malo.

Criterio de posición	Calificación
1-3	Bueno
4-7	Regular
8-10	Malo

Tabla 4.3: Criterios de calificación para el enriquecimiento de texto aplicado.

En la Figura 4.8 se muestra la calificación para el enriquecimiento realizado, mostrando que en 78 ocasiones se considera como un enriquecimiento bueno, en 20 ocasiones se considera como un enriquecimiento regular y solo en 2 ocasiones un enriquecimiento malo.

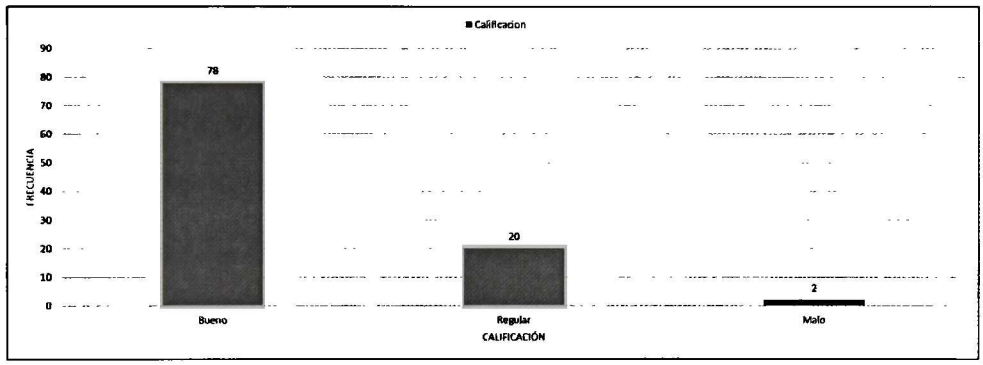


Figura 4.8: Calificaciones obtenidas.

4.6 Tercer experimento

El tercer experimento evalúa el texto original y el texto enriquecido con diferentes algoritmos de clasificación. Para evaluar el desempeño en la clasificación se optó por utilizar *datasets* que son utilizados comúnmente en la clasificación de texto. Un *dataset* utilizado en este experimento es conocido como *Reuters* y el segundo *dataset* es 20 *Newsgroups*. El *dataset* Reuters es un conjunto de noticias de una agencia con sede en el Reino Unido, conocida por suministrar información a medios de comunicación y mercados financieros. El *dataset* 20 *Newsgroups* es un conjunto de documentos

de diferentes tópicos (20) obtenidos de noticias diversas. Para el *dataset* Reuters se utilizaron dos versiones diferentes, la versión que tiene 8 diferentes clases de documentos y la versión de 52 clases. Mientras que para el *dataset* 20 *Newsgroups* se utilizó la versión normal de 20 clases. En la Tabla 4.4 se muestra la cantidad de documentos utilizados en cada *dataset*.

Dataset	Cantidad de documentos
Reuters 8	2500
Reuters 52	3000
20 Newsgroups	4000

Tabla 4.4: Cantidad de documentos utilizados por *dataset*.

Para obtener un panorama más claro sobre el rendimiento del método enriquecimiento de texto se realizó una comparativa de los resultados del texto enriquecido con el texto original utilizando diferentes algoritmos de clasificación. Los algoritmos con los que se realizó la comparativa se listan a continuación:

- K-vecinos más cercanos (K-Nearest Neighbor, K-NN).
- Máquinas de vectores de soporte (Support Vector Machine, SVM).
- Red neuronal de alimentación hacia adelante (Feed-Forward Neural Network, FFNN).
- Red neuronal de función de base radial (Neural Network Radial Basis Function, NNRBF).
- Red Bayesiana (Bayesian Network, BN).

Las implementaciones que se utilizaron por cada uno de los algoritmos fueron de la herramienta WEKA [71]. Por cada algoritmo se llevó a cabo una serie de combinaciones de parámetros de entrada, de los cuales fueron tomados los que mejor rendimiento obtuvieron en cada caso.

	Predicción positiva	Predicción negativa
Clase positiva	verdaderos positivos (TP)	falsos negativos (FN)
Clase negativa	falsos positivos (FP)	verdaderos negativos (TN)

Tabla 4.5: Matriz de confusión para clasificación binaria.

4.6.1 Medidas de Evaluación

Para evaluar un sistema de clasificación de texto se utilizan comúnmente la precisión, exhaustividad y *F-measure*, las cuales son medidas comunes en el área de recuperación de información [86].

La precisión es la medida de correspondencia entre las etiquetas verdaderas y las etiquetas positivas dadas por el clasificador, mientras que la exhaustividad es la medida de efectividad del clasificador para identificar etiquetas positivas. *F-measure* es la media armónica entre la precisión y la exhaustividad para obtener un resumen de la eficacia. La matriz de confusión contiene información acerca del aprendizaje y predicción de un modelo de clasificación. En la Tabla 4.5 se muestra la matriz de confusión para clasificación binaria.

A partir de la matriz de confusión se pueden calcular las medidas de rendimiento (precisión, exhaustividad y *F-measure*) de clasificación binaria como sigue:

- **Precisión**:

$$P = \frac{TP}{TP + FP}$$

- **Exhaustividad**:

$$E = \frac{TP}{TP + FN}$$

- ***F-measure***:

$$F = \frac{(1 + \beta^2) * P * E}{\beta^2 * P + E}$$

La matriz de confusión para un problema de clasificación multi-clase es una generalización del

	C_1	C_2	C_3	C_N
C_1	TP_1	FN_{12}	FN_{13}	FN_{1N}
C_2	FN_{21}	TP_2	FN_{23}	FN_{2N}
C_3	FN_{31}	FN_{32}	TP_3	FN_{3N}
C_N	FN_{N1}	FN_{N2}	FN_{N3}	TP_N

Tabla 4.6: Matriz de confusión para problemas de clasificación multi-clase.

caso binario. En la Tabla 4.6 muestra dicha generalización para problemas de clasificación multi-clase.

Para un problema multi-clase de N clases, la matriz de confusión tendrá una dimensión de NxN, como se observa en la Tabla 4.6. Sin embargo, este tipo de problemas se puede reducir a problemas de clasificación binaria si cada clase se considera de forma separada frente a la unión del resto de clases, obteniendo por tanto N matrices de confusión. A partir de lo anterior, a continuación se presentan las medidas de evaluación para clasificación multi-clase antes mencionadas.

- **Precisión (Multiclase)**:

$$P_M = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i}$$

- **Exhaustividad (Multiclase)**:

$$E_M = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FN_i}$$

- ***F-measure*** **(Multiclase)**:

$$F_M = \frac{\sum_{i=1}^{N} P * E}{\sum_{i=1}^{N} \beta * P + E}$$

4.6.2 Evaluación de *datasets*

Cada uno de los *datasets* fueron evaluados con los diferentes modelos de clasificación. Los resultados obtenidos en el *dataset Reuters* en su versión de 8 clases se muestran en la Tabla 4.7.

El mismo procedimiento fue aplicado para el *dataset Reuters* en la versión de 52 clases. En la Tabla 4.8 se muestran los resultados correspondiente al *dataset* antes mencionado.

	Reuters 8							
	Original				Enriquecido			
	Correctos	Precisión	Exhaustividad	F-Measure	% Correctos	Precisión	Exhaustividad	F-Measure
KNN	71.410	0.787	0.714	0.684	76.810	0.779	0.724	0.741
SVM	93.47	0.938	0.931	0.933	94.020	0.939	0.936	0.940
NNRBF	75.200	0.782	0.753	0.728	78.300	0.814	0.771	0.772
BN	55.230	0.616	0.552	0.442	59.230	0.618	0.572	0.462
FFNN	46.412	0.441	0.457	0.459	48.780	0.523	0.477	0.481

Tabla 4.7: Resultados obtenidos en clasificación con el *dataset Reuters* (8 clases).

	Reuters 52							
	Original				Enriquecido			
	% Correctos	Precisión	Exhaustividad	F-Measure	Correctos	Precisión	Exhaustividad	F-Measure
KNN	64.199	0.736	0.642	0.606	71.140	0.768	0.699	0.705
SVM	85.460	0.835	0.855	0.832	86.100	0.849	0.877	0.861
NNRBF	67.800	0.687	0.678	0.634	69.910	0.699	0.681	0.656
BN	46.720	0.457	0.467	0.347	47.820	0.437	0.497	0.367
FFNN	41.120	0.462	0.378	0.389	43.200	0.466	0.401	0.481

Tabla 4.8: Resultados obtenidos en clasificación con el *dataset Reuters* (52 clases).

La misma evaluación fue aplicada para el *dataset* 20 *Newsgroups*. En la Tabla 4.9 se muestran los resultados correspondientes.

	20 Newsgroups							
	Original				Enriquecido			
	% Correctos	Precisión	Exhaustividad	F-Measure	Correctos	Precisión	Exhaustividad	F-Measure
KNN	74.090	0.774	0.741	0.745	75.190	0.781	0.759	0.757
SVM	86.120	0.851	0.854	0.857	88.410	0.886	0.881	0.882
NNRBF	73.460	0.739	0.735	0.735	76.460	0.759	0.757	0.750
BN	40.646	0.424	0.409	0.401	44.740	0.456	0.440	0.450
FFNN	40.090	0.442	0.362	0.391	43.660	0.468	0.411	0.441

Tabla 4.9: Resultados obtenidos en clasificación con el *dataset* 20 *Newsgroups*.

Los resultados en cada una de las tablas anteriores tienen los valores de precisión, exhaustividad, *F-Measure* y porcentaje de documentos correctamente clasificados. Como se observa en cada uno de los resultados de los *datasets* utilizados, el desempeño de la clasificación de cada uno de los algoritmos utilizando el *dataset* enriquecido es mejor que el desempeño de la clasificación con los *datasets* originales. Lo anterior muestra que al enriquecer un texto se obtiene un mejor entendimiento aportando mayor información y por consecuencia el clasificador tiene un mejor desempeño cuando se utiliza un *dataset* enriquecido que uno original.

4.7 Cuarto experimento

En este experimento se evalúa la Ganancia de Información (GI) en los *datasets* utilizados anteriormente, es decir, Reuters y 20 Newsgroups.

La Ganancia de Información se define como la impureza que presenta un *dataset*, es decir, la entropía que presenta un conjunto de datos. El concepto de entropía viene de la Teoría de Información, se dice que a más entropía más información, por el contrario a menos entropía se tiene menos información. Por esa razón, un *dataset* que se utilizará para tareas de clasificación, es deseable que tenga una alta entropía, es decir, un grado de impureza alto. Una alta entropía en un *dataset* aporta más información para que el algoritmo de clasificación tenga la mayor información posible para clasificar un determinado documento de ese *dataset*. Específicamente se puede decir que la Ganancia de Información determina qué atributo de un conjunto de datos es más útil para discriminar entre las clases que componen ese conjunto de datos.

Dataset	Cantidad de documentos
Reuters 8	2500
Reuters 52	3000
20 Newsgroups	4000

Tabla 4.10: Cantidad de documentos utilizados por *dataset*.

La cantidad de documentos que se utilizó en cada *dataset* se muestra en la tabla 4.10. El primer análisis que se realizó determinó qué tan parecidos son los términos que aportan mayor información al *dataset* original, con respecto a los términos que aportan mayor información al *dataset* enriquecido. Por cada *dataset* se realizó el mismo análisis, en la Figura 4.9 se muestra la comparación entre la aparición de los términos del *dataset* Reuters 8 original y el Reuters 8 enriquecido. De la misma manera en la Figura 4.10 se muestra la comparación entre la aparición de los términos del *dataset* Reuters 52 original con respecto a los términos del *dataset* Reuters 52 enriquecido. Asimismo, en la Figura 4.11 la comparación entre el *dataset* 20 *Newsgroups* original y enriquecido.

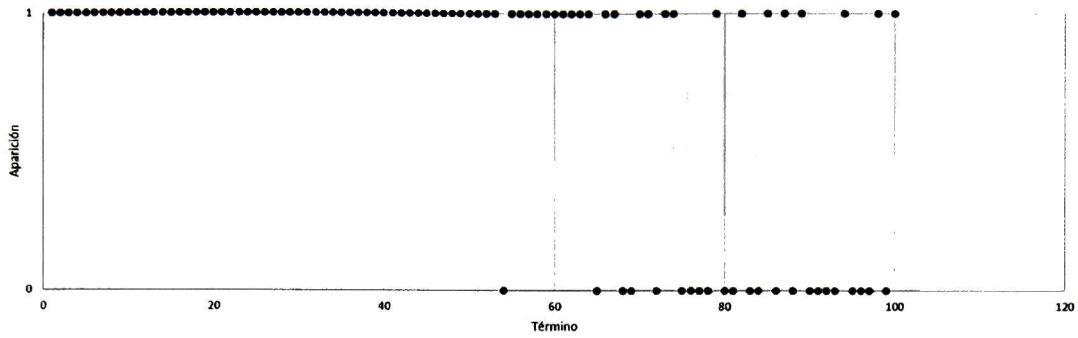


Figura 4.9: Comparación de los términos que aportan mayor información entre el *dataset* Reuters 8 original y enriquecido.

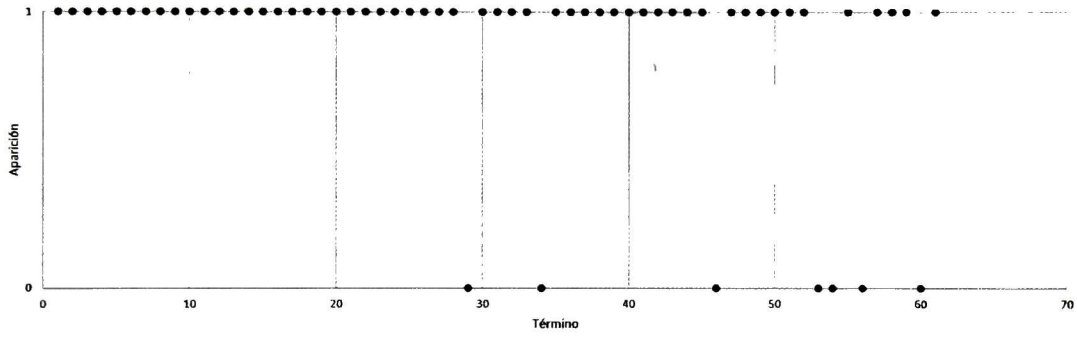


Figura 4.10: Comparación de los términos que aportan mayor información entre el *dataset* Reuters 52 original y enriquecido.

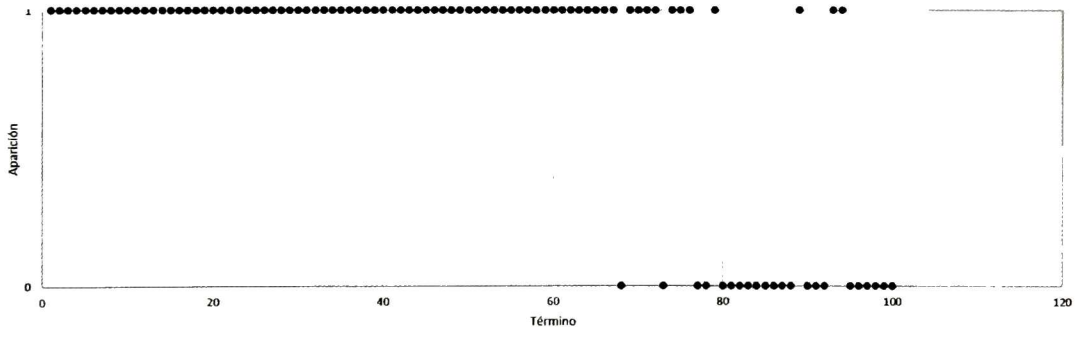


Figura 4.11: Comparación de los términos que aportan mayor información entre el *dataset* 20 Newsgroups original y enriquecido.

Como se muestra en la Figura 4.9, los primeros 67 términos aparecen tanto en el *dataset Reuters* 8 original como en el enriquecido. A partir del término 67 los términos empiezan a desaparecer en el *dataset* enriquecido. Asimismo, en el *dataset Reuters* 52 original, los primeros 30 términos del *dataset* original siguen apareciendo en el enriquecido. Por último, en el *dataset 20 Newsgroups*, los primeros 75 términos siguen apareciendo tanto en el *dataset* original como en el enriquecido. Lo anterior muestra que el *dataset* enriquecido no presentó una modificación abrupta dado el enriquecimiento que se aplicó.

Un segundo análisis se realizó con el objetivo de observar qué tanto cambian de posición los términos relevantes en los *datasets*. Lo que se busca en este análisis es observar qué tanto cambia la relevancia en un *dataset* y otro, es decir, si el término más relevante sigue estando entre las posiciones más relevantes del *dataset* enriquecido, si no es así, se desea saber a dónde se ha movido. Por cada *dataset* se realizó el mismo análisis, las Figuras 4.12, 4.13 y 4.14 se muestran el análisis de los *datasets Reuters* 8, *Reuters* 52 y 20 *Newsgroups*, respectivamente.

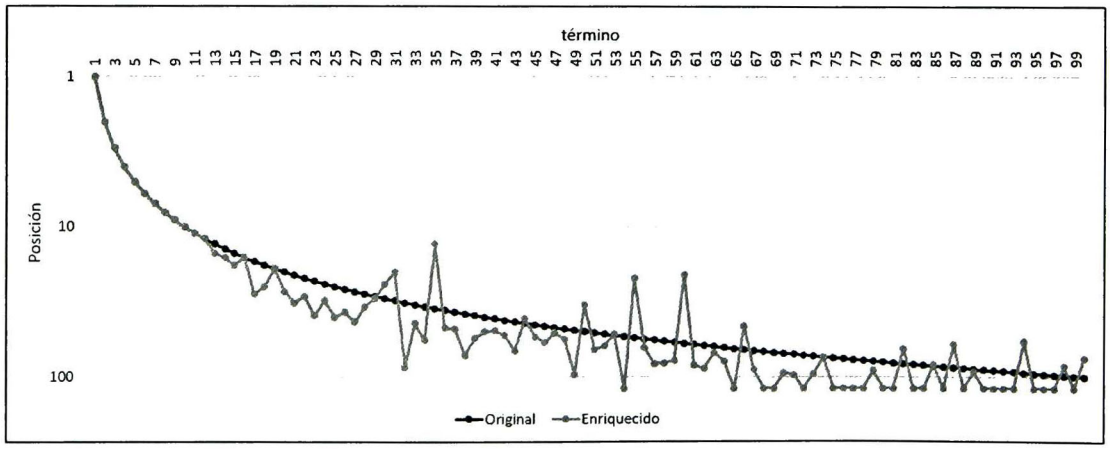


Figura 4.12: Posición de los términos que aportan mayor información al *dataset* original y enriquecido utilizando el *dataset* Reuters 8.

Como se observa en la Figura 4.12, los primeros 12 términos del *dataset* original siguen apareciendo en el mismo orden y posición del *dataset* enriquecido. La Figura 4.13, muestra que los primeros 14 términos del *dataset* original siguen apareciendo en el mismo orden y posición que

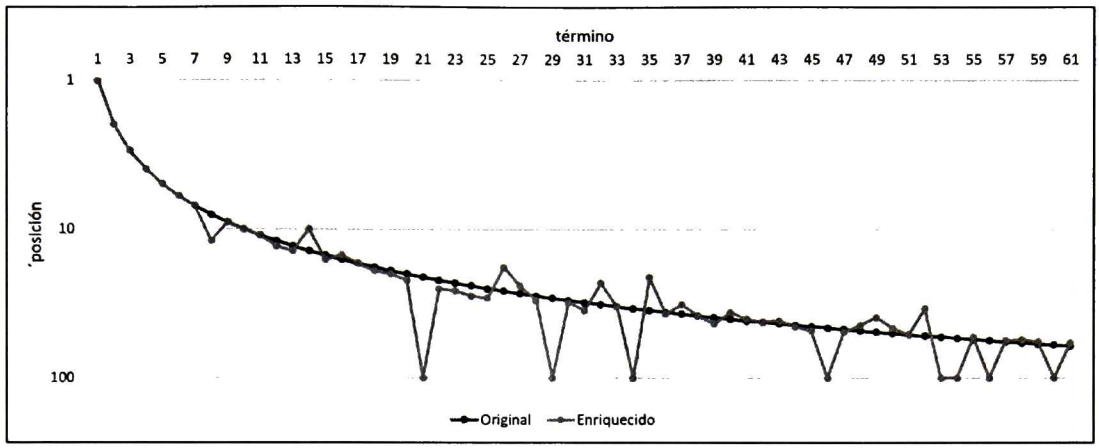


Figura 4.13: Posición de los términos que aportan mayor información al *dataset* original y el enriquecido utilizando el *dataset* Reuters 52.

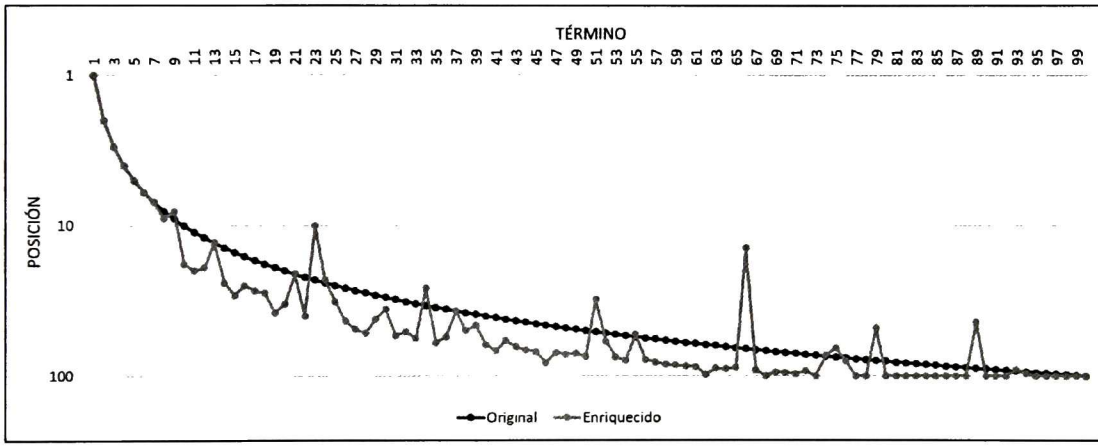


Figura 4.14: Posición de los términos que aportan mayor información al *dataset* original y el enriquecido utilizando el *dataset* 20 Newsgroups.

en el *dataset* enriquecido. Asimismo en la Figura 4.14 se muestra que los primeros 8 términos del *dataset* original siguen apareciendo en el mismo orden y posición del *dataset* enriquecido. Lo anterior indica que los primeros términos y más relevantes de los *datasets* con que se experimentó no han cambiado de forma drástica en los *datasets* enriquecidos, por lo que muestra que dichos términos siguen siendo importantes y que no han sufrido una modificación abrupta dado el enriquecimiento de texto.

Por último, se obtuvo la Ganancia de Información entre el *dataset* original y el enriquecido utilizando cada uno de los *dataset*. Las Figuras 4.15, 4.16 y 4.17 muestran el análisis de los *datasets* *Reuters* 8, *Reuters* 52 y 20 *Newsgroups* respectivamente.

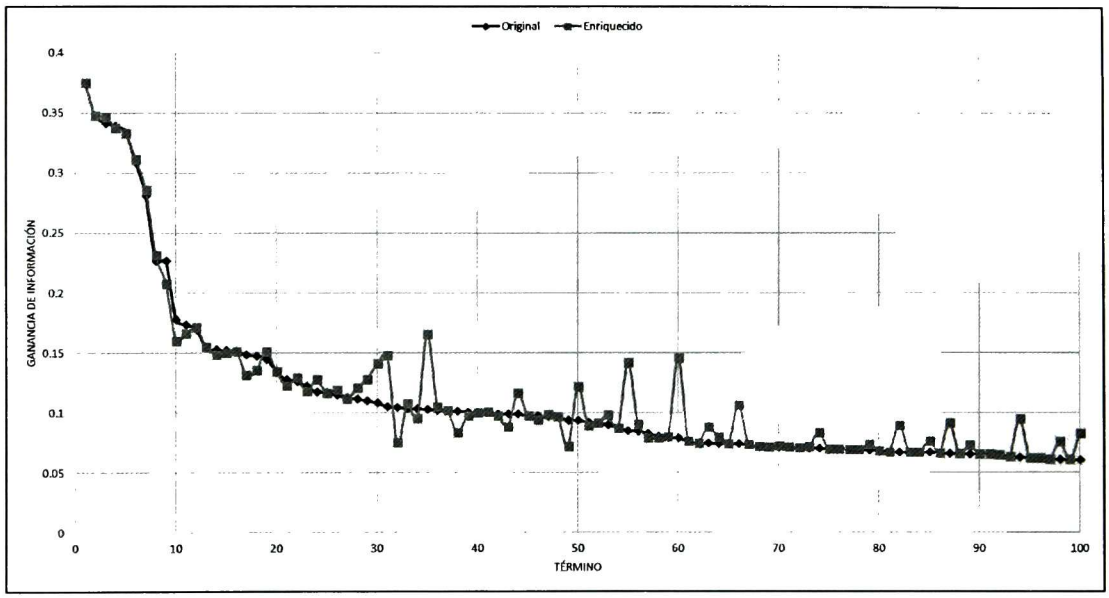


Figura 4.15: Comparación entre la Ganancia de información del *dataset* *Reuters* 8 original y el enriquecido.

Cada una de las gráficas anteriores muestran que los términos que aportan mayor información en el *dataset* enriquecido tienen una mayor Ganancia de Información que en el *dataset* original. De esta manera se comprueba que al enriquecer un texto se obtiene más información acerca de términos fundamentales en el *dataset* para proveer al algoritmo de clasificación. Por esa razón, en el experimento 3, relacionado con clasificación, cada uno de los algoritmos se comporta de una mejor manera cuando se utiliza un *dataset* enriquecido.

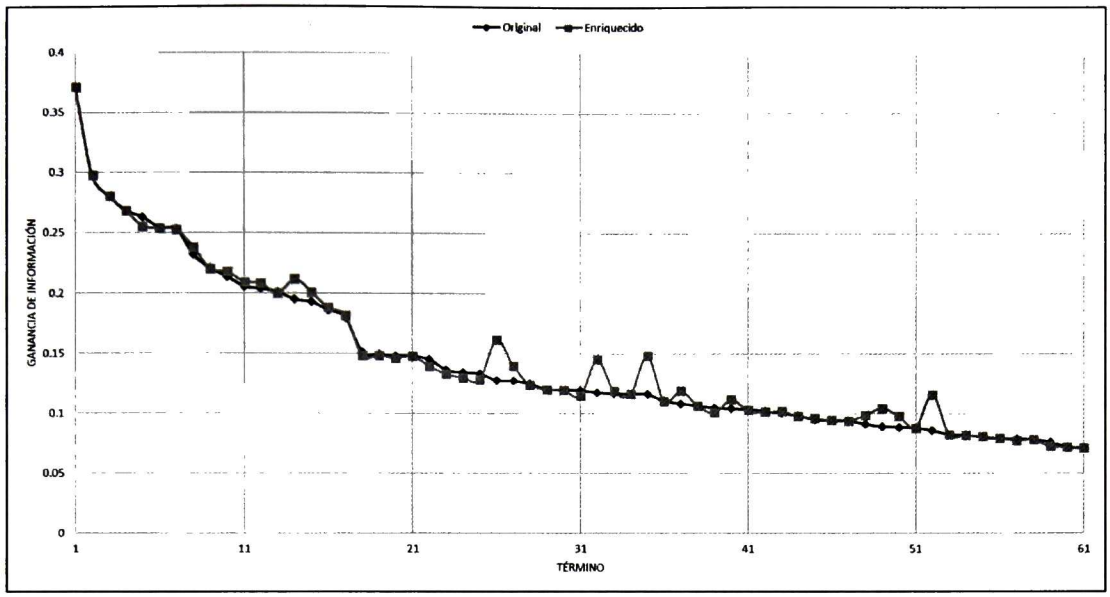


Figura 4.16: Comparación entre la Ganancia de Información del *dataset Reuters* 52 original y el enriquecido.

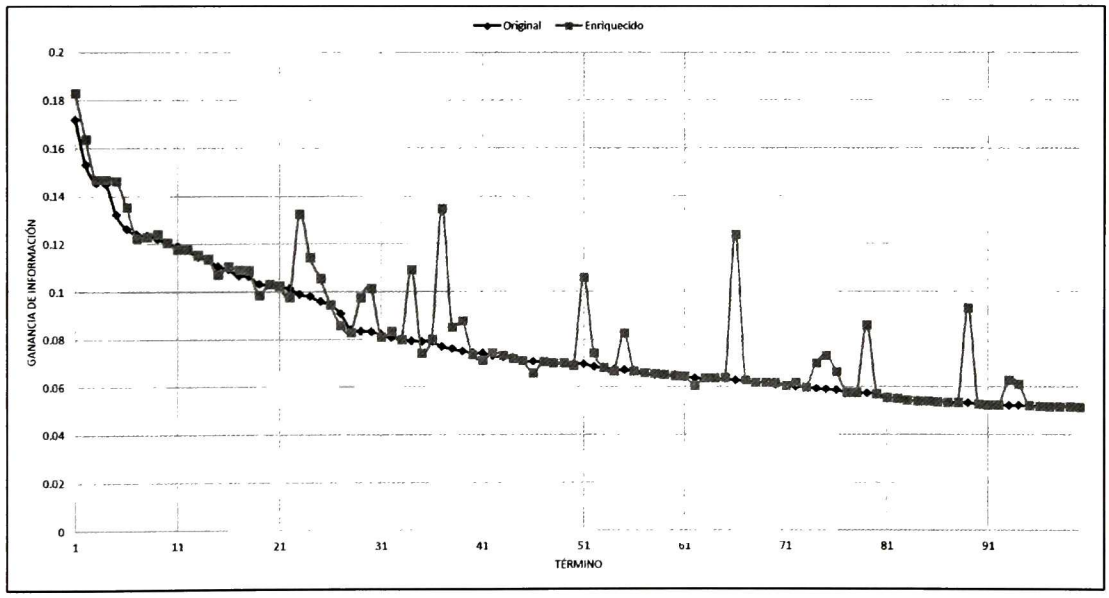


Figura 4.17: Comparación entre la Ganancia de Información del *dataset 20 newsgroups* original y el enriquecido.

4.8 Quinto experimento

El quinto experimento está relacionado con la agrupación de documentos (clustering). Para realizar este experimento se utilizaron los mismos *datasets* con los que se realizaron los experimentos anteriores.

Este experimento está basado en el algoritmo de *clustering* llamado CBC (*Clustering by Committee*), este algoritmo de agrupamiento de términos permite descubrir temas contenidos en los documentos de un *dataset*, es decir, es un algoritmo enfocado a la agrupación de documentos utilizando el contexto de las palabras como característica para agrupar aquellas que tienden a aparecer en contextos similares. Dicho algoritmo fue propuesto por Pantel [75], se utilizó una implementación desarrollada por Ana Ríos [79]. En general el algoritmo consta de cinco etapas:

- Construcción de la matriz de pesos: Esta matriz de pesos se construye a partir de la matriz de palabra-contexto, la cual se describe posteriormente.
- Reducción de la dimensionalidad: Los vectores más representativos de la matriz se *mapean* al vector que mejor los representa usando un mapa auto-organizativo.
- Obtener los K-centroides: Se construyen los K centroides con la información de L términos similares.
- Validación de centroides candidatos: Se busca obtener un conjunto de centroides que sean más disimilares entre sí para que representen de manera más eficiente los grupos finales.
- Agrupamiento de términos: Los x términos se agregan al centroide del grupo con el cual tienen mayor similaridad.

Para evaluar el desempeño de *datasets* enriquecidos y originales en algoritmo de *clustering* se desarrollaron una serie de pasos. En la Figura 4.18 se muestra un esquema que se implemetó para

evaluar el desempeño de *datasets* originales y enriquecidos aplicado a tareas de *clustering*. Dicho esquema fue diseñado por Jesús Cervantes [80].

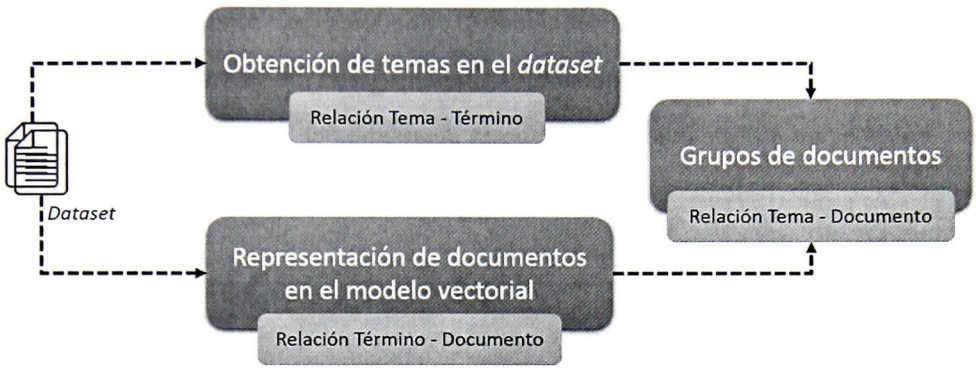


Figura 4.18: Esquema para el experimento de *clustering* de documentos.

La Figura 4.18 contiene tres módulos diferentes, la obtención de temas en el *dataset*, la representación de documentos en el modelo vectorial y los grupos de documentos. El módulo para la obtención de temas identifica y extrae información potencialmente relevante, agrupando documentos de acuerdo a características similares entre términos. Este módulo obtiene una relación entre los temas identificados y los términos que pertenecen a un tema en específico, dicha relación es denominada como **Relación Tema - Término**. El módulo para la representación de documentos utiliza el modelo espacio vectorial ponderado y modela a los documentos como un vector, en el que cada vector contiene los términos del documento con una ponderación en el documento así como la ponderación en el *dataset* completo. Este módulo produce una relación entre los términos del *dataset* y los documentos, dicha relación se denomina **Relación Término - Documento**. Por último, el módulo de grupos de documentos es el resultado de relacionar el módulo de obtención de temas en el *dataset* y el módulo de representación de documentos. Por lo tanto, a esta relación se le denomina como **Relación Tema - Documento**. Con los grupos de documentos resultantes se evaluó y comparó el desempeño del *clustering* utilizando el *dataset* normal y enriquecido.

4.8.1 Representación de documentos en el espacio vectorial

Para la representación de los documentos del corpus se utilizó el modelo espacio vectorial ponderado. Este modelo representa los documentos como vectores, donde los elementos que lo componen son los términos que describen mediante un grado de relevancia (ponderación) al documento. En la Figura 4.19 se muestra el esquema del módulo de representación de documentos.

Figura 4.19: Esquema del módulo de representación de documentos.

Para obtener una representación de los documentos es necesario realizar una serie de tareas para la obtención de los términos relevantes, las cuales se describen a continuación:

4.8.1.1. Preprocesamiento

En esta etapa se preprocesa el texto para su posterior manipulación, esta etapa consta de tres sub-etapas las cuales se describen a continuación:

- Normalización: Se convierte a minúscula el texto, se eliminan los signos de puntuación y caracteres especiales.
- Eliminación de *stopwords*: Se eliminan las palabras vacías del texto.
- *Stemming* Los términos resultantes deben pasar por un proceso de reducción de palabras con variantes morfológicas, para ello se aplica *stemming*.

4.8.1.2. TF-IDF

Uno de los métodos de ponderación de texto más representativos es el denominado TF-IDF (del inglés, Term Frecuency, Inverse Document Frecuency) [78], el cual establece la relación entre la frecuencia del término en el documento y la frecuencia de aparición del término en el resto de los documentos. El método de ponderación es expresado de la siguiente manera:

$$TF - IDF_{ij} = TF_{ij}. \log_2 \frac{N}{n}$$

donde TF_{ij} es la frecuencia del término t_i en el documento d_j, N es el tamaño del corpus y n es la cantidad de documentos en donde el término t_i aparece.

De esta manera cada término en el vector (documento) tiene una ponderación basada en la ocurrencia respecto al documento y al *dataset*.

4.8.1.3. Relación Término-Documento

A partir de la obtención de los vectores se obtiene una relación y representación denominada como Término-Documento. Un ejemplo de dicha representación se muestra en la Figura 4.11. La matriz representa un conjunto de vectores (filas), los cuales están compuestos por un conjunto de términos (columnas) cuyos valores reales representan el grado de relevancia del elemento con respecto al vector.

Término \Doc	**doc 1**	**doc 2**	**doc 3**
término 1	0.110	0.090	0.000
término 2	0.190	0.540	0.100
término 3	0.000	0.000	0.440

Tabla 4.11: Relación Término-Documento

4.8.2 Obtención de temas en el *dataset*

En el *dataset* el contenido de los documentos es variable y éste no tiene una estructura definida por lo tanto es difícil tener un conocimiento previo de la información contenida en los documentos. El uso de patrones léxico-sintácticos y categorías gramaticales permiten identificar términos candidatos para la identificación de temas. Por lo tanto, este módulo establece un modelo para la representación de los datos a partir de las ocurrencias de términos, identificando sustantivos y verbos dentro de oraciones en el texto y utilizando el algoritmo *CBC* para la obtención de grupos que representen temas de los documentos de entrada. En la Figura 4.20 se muestra el esquema del módulo de obtención de temas en el *dataset*.

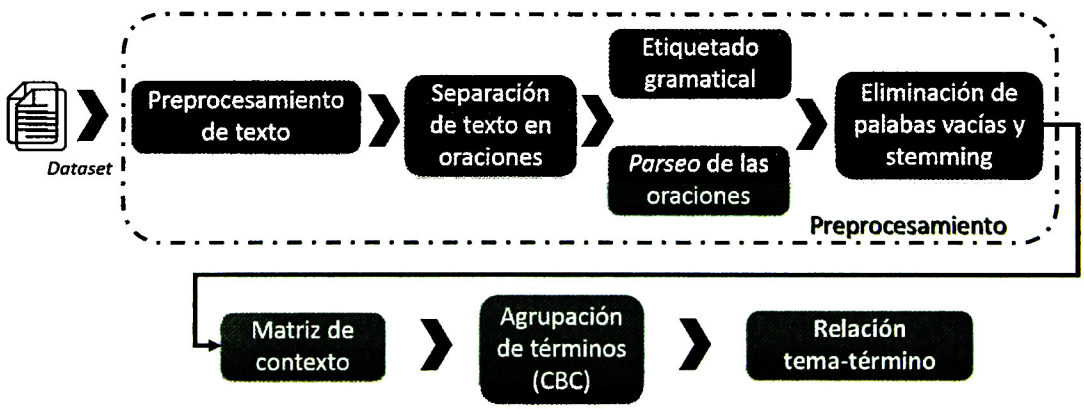


Figura 4.20: Esquema del módulo de obtención de temas en el *dataset*.

4.8.2.1. Preprocesamiento

Para aplicar el algoritmo de *clustering* (CBC) se deben obtener las características (términos más importantes) a partir de los documentos, por consiguiente, éstos requieren pasar por un preprocesamiento que permita extraer sus características. El proceso aplicado en este trabajo incluye la ejecución de varias tareas, tales como:

- **Preprocesamiento del documento:** Dado que en este módulo es necesaria la integridad de

las palabras para su categorización gramatical, el preprocesamiento en esta etapa se limita a convertir las palabras a minúsculas y a la eliminación de caracteres especiales.

- **Separación del texto en oraciones:** Una vez que se tienen los archivos de texto plano es conveniente aplicar un proceso de separación del texto en oraciones que permita manipular el texto con base en elementos funcionales, como una oración que contiene una estructura gramatical.

- **Etiquetado de oraciones y *parseo* gramatical de las oraciones:** Dado que es necesario identificar las características de las oraciones, tal como la relación gramatical entre verbos y sustantivos, se aplica un etiquetado y *parseo* gramatical de las oraciones. Con base en herramientas de procesamiento de lenguaje natural se pueden extraer las características del tipo Sustantivo y Verbo sobre los enunciados del texto para entonces construir una matriz del tipo *palabra-contexto*. Tanto para el etiquetado de oraciones y *parseo* gramatical se utilizó el *parser de Stanford* [76], el cual identifica las relaciones verbo-sustantivo.

- **Eliminación de palabras vacías (*stopwords*) y *stemming*.** Las palabras que no tienen un significado relevante respecto al contenido del texto como son artículos, preposiciones, entre otras, no son tomadas en cuenta como características.

4.8.2.2. Matriz de contexto

En esta etapa se usa un *parser* gramatical para extraer el contexto gramatical donde cada palabra ocurre. Para ello, en el submódulo de etiquetado gramatical y parseo de las oraciones se extrajeron las relaciones y se obtuvo el contexto donde cada palabra ocurre. Una vez procesado y etiquetado gramaticalmente el *dataset* se procede a la obtención de relaciones verbo-sustantivo, ya que estas categorías son esenciales para formar oraciones. De esta manera se puede encontrar qué términos semánticamente similares (sustantivos) aparecen en un mismo contexto (verbo). Para identificar y hacer un conteo de las ocurrencias de las relaciones en el corpus se creó una matriz de contexto,

donde se estableció la obtención de relaciones entre los verbos y sustantivos del *dataset*. Al final de este procedimiento, aquellas relaciones que tengan al menos una aparición en el *dataset* son seleccionadas. En la Tabla 4.12 se muestra un ejemplo de la matriz de contexto donde cada columna representa los sustantivos del corpus, mientras que cada fila representa la relación entre un verbo y los sustantivos del corpus. Cada relación verbo-sustantivo tiene asignado un valor numérico que representa el número de apariciones de esta relación en el corpus. Así, se pueden identificar las relaciones verbo-sustantivo que aparecen al menos una vez dentro del *dataset* de documentos.

Verbo \Sustantivo	**computer**	**John**	**Peter**
wait	3	0	1
run	0	1	5
buy	0	2	2

Tabla 4.12: Matriz de contexto.

4.8.2.3. Agrupación de términos (CBC)

La siguiente etapa de este módulo es la agrupación de términos con características similares. Este módulo utiliza una implementación del algoritmo *CBC* [75]. El algoritmo CBC genera como resultado las agrupaciones de términos con características comunes (contextos), donde cada grupo es expresado como un tema. Dicho resultado es adaptado para crear una matriz de relaciones Tema-Término. El resultado del algoritmo CBC se muestra en la Tabla 4.13.

Tema \Término	**término 1**	**término 2**	**término 3**
Tema 1	0.210	0.000	0.040
Tema 2	0.120	0.005	0.600
Tema 3	0.000	0.230	0.000

Tabla 4.13: Relación Tema-Término

4.8.3 Relación Tema-Documento (Grupos de documentos)

Dado que el resultado del algoritmo CBC son grupos de términos y lo que se necesita para evaluar el desempeño del enriquecimiento de texto en *clustering* es determinar qué tan bien formados se encuentran los grupos resultantes, se obtiene una relación entre las relaciones *Tema-Término* y *Término-Documento*. Es decir, para integrar los módulos de obtención de temas y representación de documentos del método es necesario buscar la manera de establecer una relación entre las matrices Tema-Término (Tabla 4.13) y Término-Documento (Tabla 4.11). Esta integración se puede lograr apoyándose del álgebra de matrices utilizando el producto de matrices, donde el producto de la matriz $A = (A_{ij}) \in M_{n\times m}$ por la matriz $B = (B_{jk}) \in M_{m\times p}$ es la matriz $C = (C_{ik} \in M_{n\times p})$, sus elementos son expresados de la siguiente forma [81]:

$$c_{ik} = a_{i1}b_{1k} + a_{i2}b_{2k} + ... + a_{im}b_{mk} = \sum_{i=1}^{m} a_{ij}b_{jk}$$

De manera gráfica se puede expresar de la siguiente forma:

$$A_{i,j}\begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix} \times B_{i,j}\begin{bmatrix} b_{1,1} & b_{1,2} \\ b_{2,1} & b_{2,2} \\ b_{3,1} & b_{3,2} \end{bmatrix} =$$

$$C_{i,j}\begin{bmatrix} a_{1,1}\cdot b_{1,1} + a_{1,2}\cdot b_{2,1} + a_{1,3}\cdot b_{3,1} & a_{1,1}\cdot b_{1,2} + a_{1,2}\cdot b_{2,2} + a_{1,3}\cdot b_{3,2} \\ a_{2,1}\cdot b_{1,1} + a_{2,2}\cdot b_{2,1} + a_{2,3}\cdot b_{3,1} & a_{2,1}\cdot b_{1,2} + a_{2,2}\cdot b_{2,2} + a_{2,3}\cdot b_{3,2} \end{bmatrix}$$

Así, con el producto de estas matrices se puede obtener una relación con las dimensiones *Tema y Documento* donde el enlace de éstas es la dimensión *Término* de ambas. En la Tabla 4.14 se puede observar una representación de la matriz *Tema-Documento*. Esta matriz se puede considerar como un conjunto de temas o vectores compuestos por valores reales que representan el grado de pertenencia de los documentos hacia los temas.

Tema \ Documento	Documento 1	Documento 2	Documento 3
Tema 1	0.410	0.100	0.010
Tema 2	0.810	0.000	0.000
Tema 3	0.100	0.180	0.000

Tabla 4.14: Relación Tema-Documento

Una vez que la relación **Tema-Documento** es construida a partir de las relaciones **Tema-Término** y **Término-Documento** se realiza la agrupación de documentos por tema, según la ponderación asignada. Es decir, se examina cada una de las ponderaciones de los documentos a los distintos temas con el objetivo de asignar el documento al tema con mayor grado de pertenencia. Dado que un documento puede tener un grado de pertenencia a diferentes temas, dicho problema se resuelve eligiendo a la mayor ponderación. En la Figura 4.21 se muestra un posible escenario en la asignación de documentos en el que un documento pertenece a dos temas, pero la asignación a un tema se realiza con base a su ponderación.

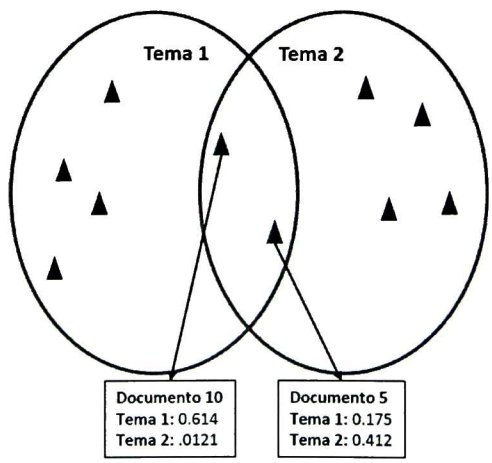


Figura 4.21: Asignación de documentos a los temas correspondientes.

4.8.4 Análisis de la agrupación de documentos con los *datasets* de prueba.

A cada uno de los *datasets Reuters 8, Reuters 52 y 20 Newsgroups* se aplicó el proceso anteriormente descrito. Como resultado final se obtuvieron los grupos de documentos que se formaron por cada *dataset*.

En la Figura 4.22 se muestran los grupos formados con el *dataset Reuters 8* original, mientras que en la Figura 4.23 se muestran los grupos formados con el *dataset Reuters 8* enriquecido. Las Figuras muestran la relación entre el número de grupos y la cantidad de documentos por cada grupo. En el *dataset* original se muestra que se formaron 20 grupos, mientras que en el enriquecido solo se formaron 14 grupos.

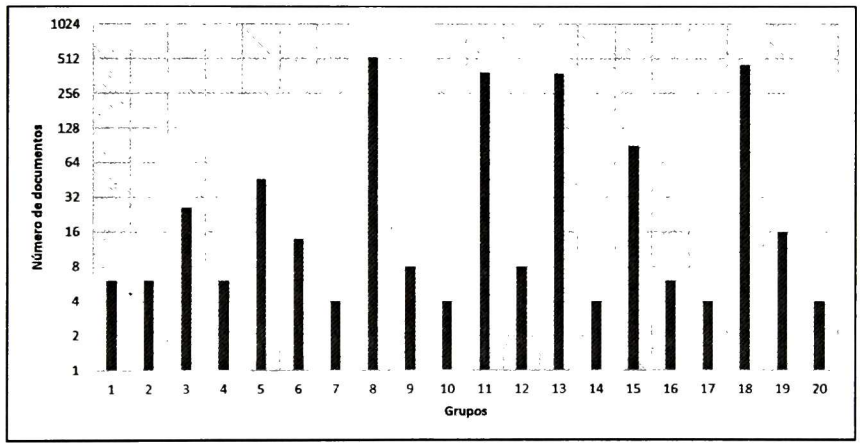


Figura 4.22: Grupos resultantes del *dataset Reuters 8* Normal.

En la Figura 4.24 se muestran los grupos formados con el *dataset Reuters* 52 original y en la Figura 4.25 se muestran los grupos formados con el *dataset Reuters* 52 enriquecido. Los grupos formados con el *dataset Reuters* 52 original fueron 16 y con el *dataset Reuters 52* enriquecido con la misma cantidad. Ambos *datasets* tienen la misma cantidad de grupos, sin embargo, los grupos formados tienen diferente cantidad de documentos.

De la misma manera, en la Figura 4.26 se muestran los grupos formados con el *dataset 20*

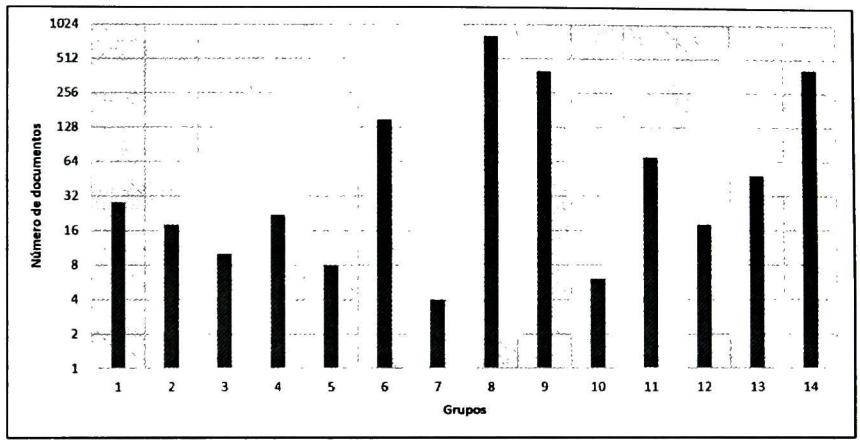


Figura 4.23: Grupos resultantes del *dataset Reuters 8* Enriquecido.

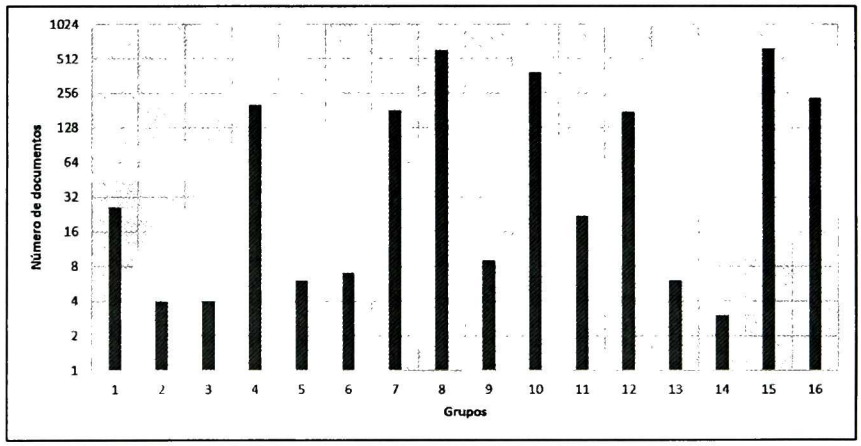


Figura 4.24: Grupos resultantes del *dataset Reuters 52* Normal.

Newsgroups original y en la Figura 4.27 se muestran los grupos formados con el *dataset 20 Newsgroups* enriquecido. Con el *dataset 20 News Groups* original se formaron 22 grupos, al igual que para el *dataset 20 News Groups* enriquecido. De la misma manera que en los *datasets Reuters* 52, se formaron la misma cantidad de grupos entre los *datasets*, pero los grupos del *dataset* original tienen diferente distribución con respecto a los grupos del *dataset* enriquecido.

En cada uno de los *datasets* originales y enriquecidos se muestran ciertas diferencias. Por ejemplo, en el *dataset Reuters 8* enriquecido disminuye el número de grupos, lo cual implica el cambio en el número de documentos por grupo. Por otro lado, en los *datasets Reuters 52 y 20 Newsgroups* no

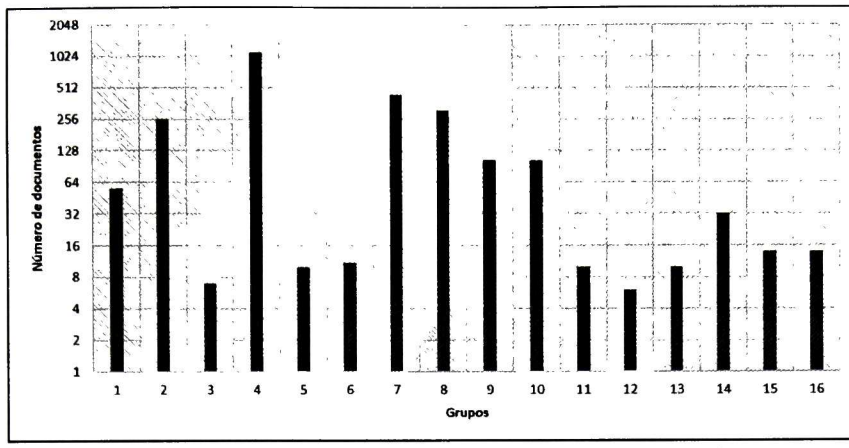


Figura 4.25: Grupos resultantes del *dataset Reuters 52* Enriquecido.

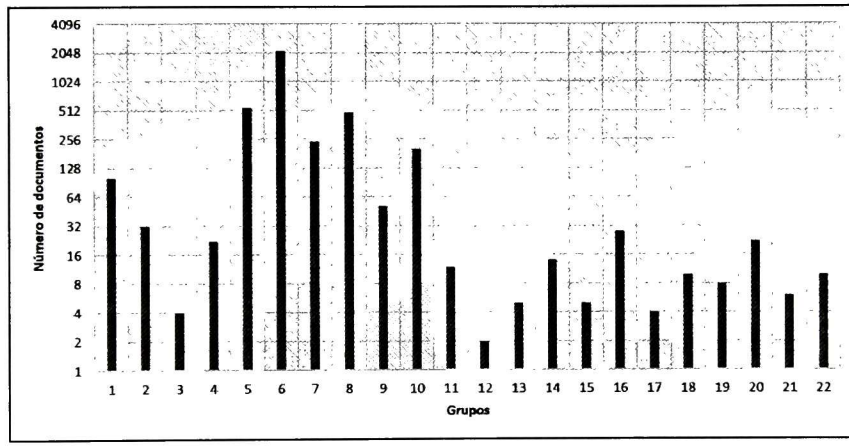


Figura 4.26: Grupos resultantes del *dataset 20 Newsgroups* Normal.

muestran un cambio en el número de grupos, pero si muestran una re-agrupación de documentos, es decir, la distribución de documentos es diferente. Lo resultados y Figuras anteriores muestran que en los *datasets* originales y enriquecido existe una diferencia visible en la distribución de los documentos y en el número de grupos formados, lo que muestra que el enriquecimiento aplicado está beneficiando al desempeño del *clustering*. De lo contrario, si se observara un comportamiento muy similar tanto para los *datasets* originales y enriquecidos, el enriquecimiento no estaría beneficiando al desempeño del *clustering*, lo cual atendería a la causa de que no se está enriqueciendo el texto de manera correcta.

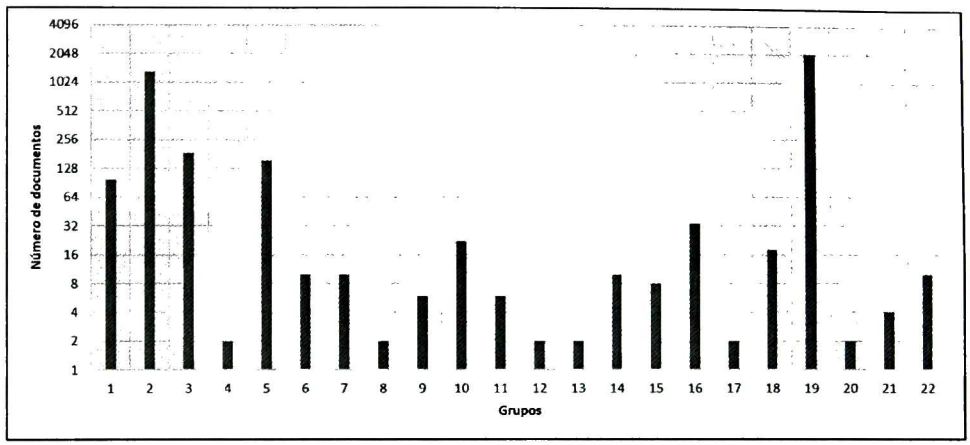


Figura 4.27: Grupos resultantes del *dataset 20 Newsgroups* Enriquecido.

4.8.5 Evaluación del desempeño del *clustering* (CBC)

Después de aplicar el *clustering* a los *datasets Reuters y 20 News Groups* se evaluó la calidad de los grupos formados por el algoritmo de *clustering* CBC. Es muy difícil evaluar la calidad de los resultados de *clustering* dado que no hay intervención humana en la creación de grupos y no se sabe si los grupos están formados correctamente dado que no existe un *dataset* de comparación. Generalmente para evaluar el desempeño del *clustering* se utilizan métodos de inspección manual, dichos métodos tratan de examinar los *clusters* a través del estudio de la amplitud de los centroides o leer los documentos de cada *cluster* generado. Otra forma de evaluar el desempeño del *clustering* es la evaluación denominada como *Ground truth*. Este tipo de evaluación prueba el *dataset* sometido al algoritmo de *clustering* sin clases originales, es decir, reemplaza las etiquetas de clases originales y considera como nuevas etiquetas de clases a cada *cluster* generado. El número de *clusters* generados es el nuevo número de etiquetas de clases para el *dataset*. En la Figura 4.28 se muestra la asignación de nuevas etiquetas de clase.

Una vez que se asignó la nueva etiqueta de clase a cada *cluster* generado se aplicó una evaluación indirecta. Este tipo de evaluación aplica otra tarea de minería de datos para evaluar la calidad de los *clusters* generados; la tarea que se eligió fue la clasificación. Para aplicar la evaluación indirecta

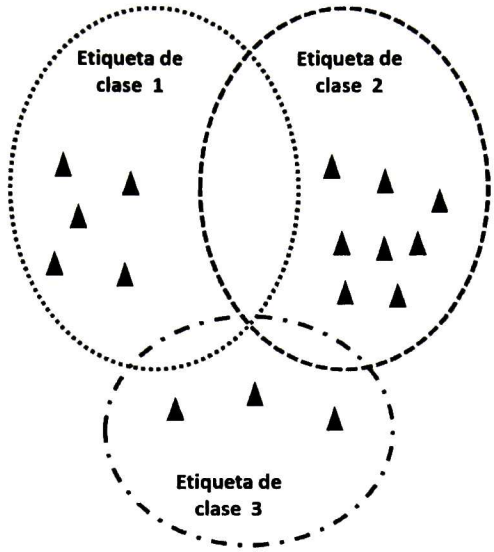


Figura 4.28: Asignación de nuevas etiquetas de clase.

se realizó una división del número de documentos generados en una proporción 60 %-40 %. El 60 % de cada *cluster* es utilizado para entrenar el clasificador y el 40 % de cada *cluster* es para probar el modelo generado por el clasificador. En la Figura 4.29 se muestra la división del número de documentos generados.

Después de realizar la agrupación de documentos y dividir cada grupo de documentos en una proporción 60 %-40 % se generó un modelo con el *dataset* de entrenamiento, posteriormente se sometió a prueba el modelo generado con el *dataset* de prueba. Para realizar dicha evaluación se utilizaron tres algoritmos de clasificación. Los algoritmos que se utilizaron para probar los *datasets* generados fueron:

- K-vecinos más cercanos (K-Nearest Neighbor, K-NN).
- Máquinas de vectores de soporte (Support Vector Machine, SVM).
- Red Bayesiana (Bayesian Network, BN).

En cada uno de los modelos de clasificación generados se modificaron los parámetros y se tomó el mejor resultado obtenido. En la Tabla 4.15 se muestran los resultados obtenidos entre los *datasets*

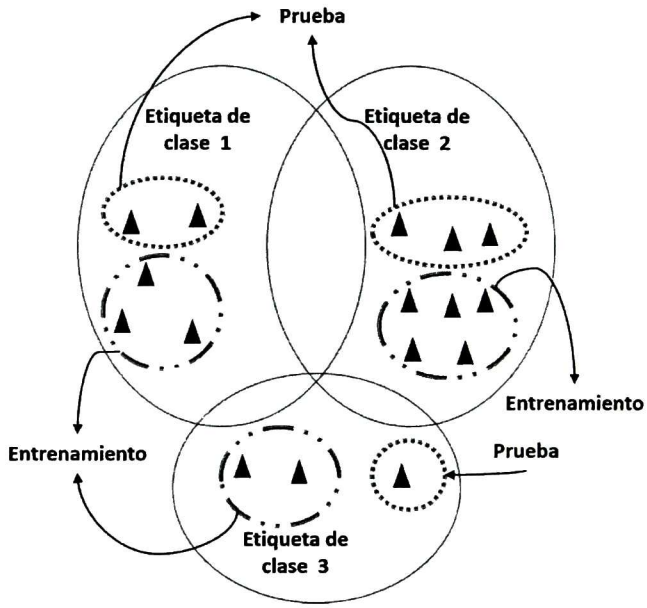


Figura 4.29: División del número de documentos generados en una proporción de 60 %-40 %.

Reuters 8 original y enriquecido. De la misma forma en la Tabla 4.16 se muestran los resultados obtenidos entre los *datasets* Reuters 52 original y enriquecido. Por último en la Tabla 4.16 se muestran los resultados obtenidos entre los *datasets* 20 Newsgroups original y enriquecido.

En las Tablas 4.15, 4.16, 4.17 se muestran los resultados por cada uno de los clasificadores, por cada clasificador se muestra el porcentaje de documentos correctamente clasificados, la Precisión, Exhaustividad y Medida F. Dichas tablas muestran una comparación entre los *datasets* originales y enriquecidos. El *dataset Reuters 8* muestra un comportamiento en el que el *dataset* enriquecido tiene un mejor desempeño que el *dataset* original, dicho comportamiento se muestra en cada uno de los algoritmos de clasificación que se utilizaron. Sin embargo, las máquinas de vectores de soporte tuvieron un mejor desempeño en ambos *datasets*, pero tuvo aún mejor desempeño en el *dataset* enriquecido. De la misma manera, el *dataset Reuters 52* tiene un comportamiento muy similar, es decir, el *dataset* enriquecido tuvo un mejor desempeño. Incluso los resultados obtenidos son muy similares a los obtenidos con el *dataset* Reuters 8, lo anterior se debe a que forman parte de un mismo *dataset* original, con la diferencia que tiene más clases. Finalmente, en el *dataset 20 Newsgroups*

se comportó de la misma manera que los *dataset* anteriores, es decir, el desempeño del *dataset* es mejor que el del *dataset* original. Los resultados obtenidos con este *dataset* son mejores, tanto para el *dataset* enriquecido como para el original, dado que los documentos que contiene este *dataset* son más largos y por lo tanto contienen más información acerca de el tema. Al contener más información, el enriquecimiento es mayor, es por ello que aunque el desempeño del *dataset* original es bueno, es aún mayor el desempeño del *dataset* enriquecido.

Con lo anterior se muestra que el enriquecimiento de texto aporta una mayor información acerca del documento, incrementando y mejorando el desempeño en el algoritmo de *clustering*.

	Reuters 8							
	Normal				Enriquecido			
	% Correctos	Precisión	Exhaustividad	Medida F	% Correctos	Precisión	Exhaustividad	Medida F
KNN	29.280	0.332	0.293	0.289	52.780	0.538	0.528	0.516
SVM	46.890	0.479	0.469	0.464	61.250	0.611	0.613	0.591
BN	28.280	0.355	0.283	0.187	54.23	0.507	0.542	0.487

Tabla 4.15: Resultados de la clasificación de los *clusters* generados por los *datasets* Reuters 8 original y enriquecido

	Reuters 52							
	Normal				Enriquecido			
	% Correctos	Precisión	Exhaustividad	Medida F	% Correctos	Precisión	Exhaustividad	Medida F
KNN	27.380	0.354	0.184	0.188	42.780	0.516	0.428	0.441
SVM	45.49	0.451	0.455	0.443	57.260	0.558	0.573	0.557
BN	33.260	0.252	0.333	0.232	48.960	0.475	0.491	0.444

Tabla 4.16: Resultados de la clasificación de los *clusters* generados por los *datasets* Reuters 52 original y enriquecido

	20 Newsgroups							
	Normal				Enriquecido			
	% Correctos	Precisión	Exhaustividad	Medida F	% Correctos	Precisión	Exhaustividad	Medida F
KNN	55.580	0.414	0.556	0.407	58.24	0.536	0.582	0.552
SVM	60.590	0.551	0.606	0.567	70.11	0.661	0.701	0.671
BN	55.760	0.561	0.578	0.551	64.81	0.604	0.648	0.617

Tabla 4.17: Resultados de la clasificación de los *clusters* generados por los *datasets* 20 Newsgroups original y enriquecido

4.9 Ajuste de parámetros utilizados por el método de enriquecimiento de texto

Para que el funcionamiento del método de enriquecimiento de texto sea el correcto es necesario ajustar diversos parámetros. La definición de dichos parámetros se realizó una experimentación apoyándose en la Teoría del Límite Central [84], el cual define que tras 31 experimentos se obtiene una validez estadística para determinar un conjunto de parámetros. Los parámetros que se ajustaron fueron:

- Número de conceptos clave para identificar Entidades Nombradas.
- Umbral de discriminación de Entidades Nombradas relevantes.

4.9.1 Definición del número de conceptos clave para identificar Entidades Nombradas

El primer parámetro a definir es el número de conceptos clave para identificar Entidades Nombradas. La búsqueda de estas Entidades esta relacionada con DBpedia, cuando se realiza una búsqueda relacionada con los conceptos clave previamente identificados.

Para definir dicho parámetro se realizaron 31 experimentos. En cada uno de los experimentos se enriqueció un documento variando el número de conceptos clave para determinar el número de Entidades Nombradas obtenidas. El número de conceptos clave se modificó desde 2 hasta 6 conceptos clave iniciales. Los 2 conceptos clave utilizados indican una búsqueda muy amplia mientras que la búsqueda con 6 conceptos clave restringe los resultados.

En la Figura 4.30 se muestra los resultados obtenidos en cada uno de los experimentos. En cada experimento se muestran los resultados obtenidos variando el número de conceptos clave utilizados. En el eje *'x'* se encuentra cada uno de los experimentos realizados, mientras que en el eje *'y'*

se encuentra el número de Entidades Nombradas encontradas. Dicha Figura muestra que cuanto menor sea el número de conceptos clave en la búsqueda se obtiene un mayor número de Entidades Nombradas y mientras se utilice una mayor cantidad de conceptos claves las Entidades Nombradas obtenidas es menor.

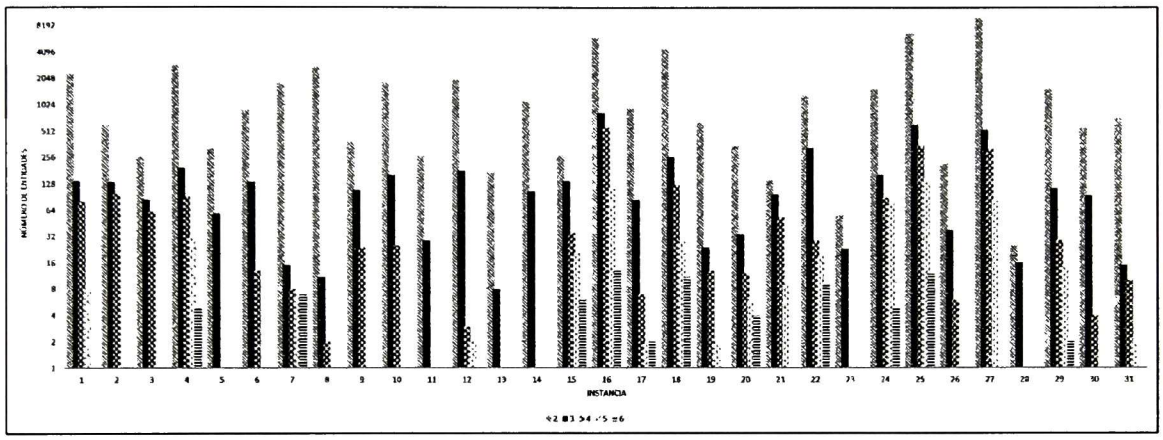


Figura 4.30: Determinación del número de conceptos clave para la identificación de Entidades Nombradas candidatas.

El número de Entidades Nombradas encontradas fue promediado por cada número de conceptos clave utilizados. Es decir, se obtuvo un promedio del número de Entidades Nombradas encontradas utilizando 2, 3, 4, 5, y 6 conceptos clave. En la Figura 4.31 se muestra el promedio de Entidades Nombradas obtenidas utilizando un diferente número de conceptos clave.

Tras la anterior experimentación se concluyó que el número de conceptos clave que se utilizan es fundamental para encontrar Entidades Nombradas candidatas, debido a que reduce el número de documentos (contenido de las Entidades Nombradas) a analizar. No es deseable analizar un alto número de Entidades Nombradas por que el tiempo de ejecución se incrementará. De la misma forma no es deseable realizar la búsqueda con pocos conceptos clave por que es posible que en muchas ocasiones no se encuentre nada o muy pocas Entidades. Debido al análisis anterior, se determinó configurar la búsqueda con 4 y 3 conceptos clave, iniciando con 4 conceptos clave y en caso de que no se encuentre nada realizar la búsqueda con 3.

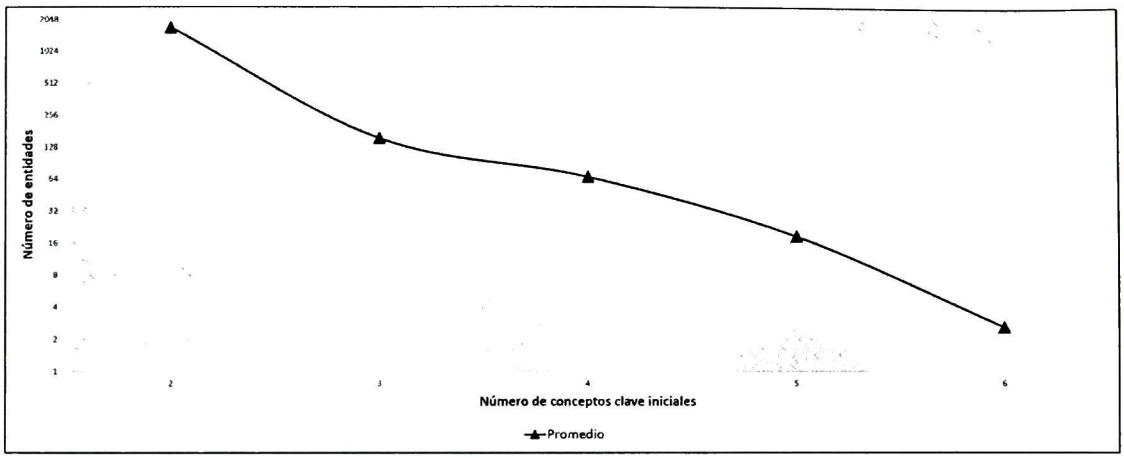


Figura 4.31: Número de Entidades Nombradas promedio utilizando un número de conceptos clave diferente.

4.9.2 Definición del umbral de Entidades Nombradas relevantes

Este parámetro está relacionado con el algoritmo Latent Semantic Indexing (LSI), el cual es utilizado para filtrar las Entidades Nombradas candidatas y considerar solamente aquellas Entidades Nombradas relevantes al texto original. Este parámetro es calculado para 2 casos, el primero utilizando 3 conceptos iniciales y el segundo utilizando 4 conceptos iniciales.

Como se ha mencionado antes, tras analizar los documentos (contenido de Entidades Nombrada) y obtener un *score* que representa la similaridad de cada documento con respecto a los conceptos clave identificados previamente. El parámetro representa el umbral para discriminar aquellos documentos relevantes del conjunto de documentos analizados. Para definir dicho umbral se realizó una prueba similar a la anterior, es decir, se realizaron 31 experimentos. En cada experimento se obtuvo un conjunto de *scores*, los cuales representan la similaridad de cada documento. Cada documento analizado fue evaluado manualmente para determinar su relevancia con respecto al texto original. Se tomaron en cuenta solamente los *scores* de los documentos relevantes. Al final de cada experimento se obtiene un conjunto de umbrales correspondiente a los documentos relevantes. Para obtener el umbral por cada experimento se obtiene la mediana de ese conjunto. Finalmente el umbral final es

el promedio de los umbrales por cada experimento.

En la Figura 4.32 se muestran los umbrales obtenidos por cada experimento. El promedio de las medianas utilizando los *scores* de los documentos relevantes con 3 conceptos clave es **0.865**.

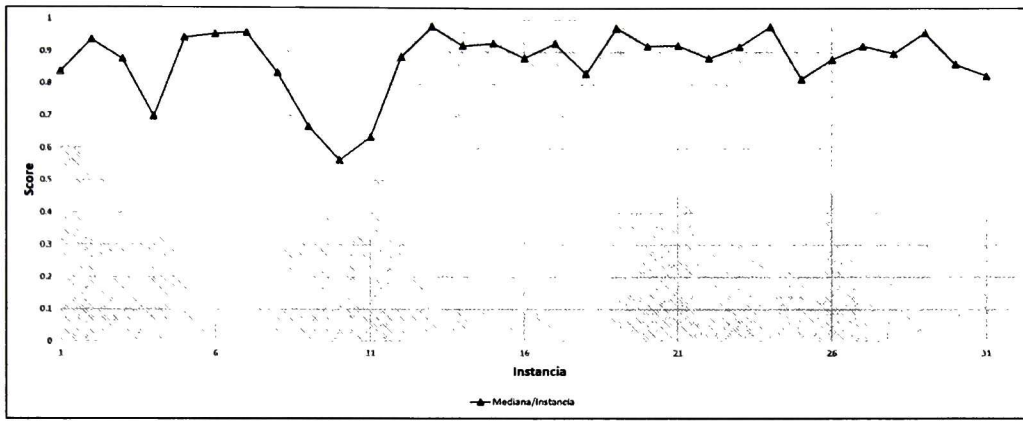


Figura 4.32: Umbral obtenido en cada uno de los experimentos utilizando 3 conceptos clave.

De la misma forma se determinó el umbral utilizando 4 conceptos clave. El promedio de las medianas utilizando los *scores* de los documentos relevantes con 4 conceptos clave es **0.671**. En la Figura 4.33 se muestran los umbrales obtenidos por cada experimento. En muchos experimentos no se obtuvieron resultados dado que la búsqueda es muy restringida, pero los resultados son más acertados, comportamiento mostrado en dicha Figura. Por esa razón se optó por realizar la búsqueda primero con 4 conceptos clave y después con 3.

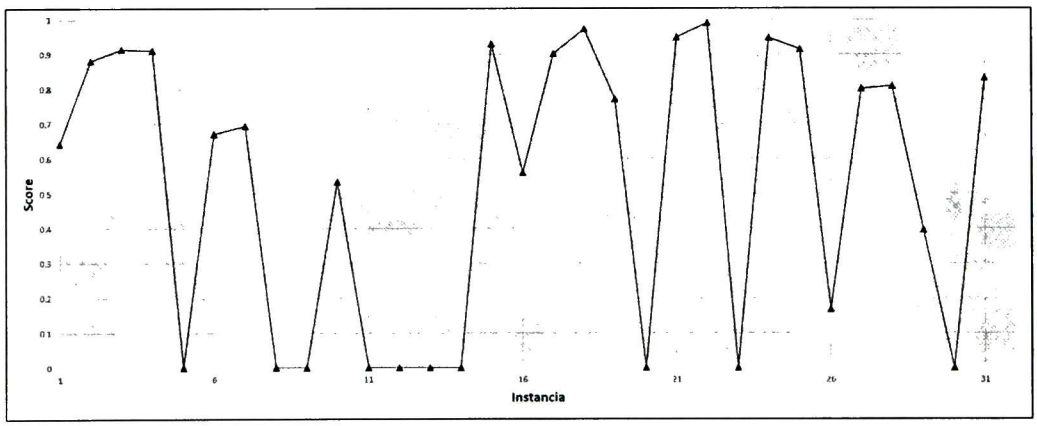


Figura 4.33: Umbral obtenido en cada uno de los experimentos utilizando 4 conceptos clave.

5

Conclusiones y Trabajo Futuro

En este capítulo se muestran las conclusiones de este trabajo de tesis. Asimismo se presentan las aportaciones, dificultades y trabajo a futuro.

5.1 Conclusiones

En este trabajo de tesis se presentó un método de enriquecimiento de texto a partir de recursos de la Web Semántica, el cual consistió en 5 módulos: Preprocesamiento, Extracción de Entidades Nombradas, Extracción de Entidades Nombradas a partir de conceptos clave, Estructuración de Información y Enriquecedor de Información. Tras la implementación del método de enriquecimiento, éste fue evaluado en 5 diferentes experimentos. Dichos experimentos demostraron que el enriquecimiento aplicado se realiza de forma adecuada. Durante el proceso de desarrollo de la metodología y de la validación experimental del enfoque propuesto se obtuvieron diversas conclusiones, las cuales se mencionan continuación.

La utilización de las bases de conocimiento para obtener información relacionada es fundamental

para el enriquecimiento de texto. En el Estado del Arte se mencionaron diferentes tipos de bases de conocimiento, desde aquellas definidas como diccionarios o consultas Web hasta las que tienen una estructura definida como las bases de conocimiento con un estructura semántica. En este trabajo de tesis se utilizó una base de conocimiento con estructura semántica, DBpedia. La organización de DBpedia permite acceder a recursos para proveer información adicional relacionada semánticamente, éstos recursos son obtenidos por el método de enriquecimiento texto mediante relaciones en el texto, tras el análisis del texto y la identificación de Entidades Nombradas. Por lo tanto, la minería de texto tiene un rol primario en el método de enriquecimiento dado que identifica las secciones en el texto para relacionar con la base de conocimiento.

Para evaluar el texto enriquecido se realizó una comparación con el texto original. Dicha evaluación mostró que el texto enriquecido tiene una alta similitud con el original a pesar de la información adicional agregada. Lo cual demuestra que no existe una gran diferencia entre el significado de la información adicional y el texto original. El enriquecimiento de texto puede afectar distintas partes del texto, las cuales unas son más importantes que otras. El experimento 4 muestra que el enriquecimiento de texto afecta a las partes más importantes del texto, lo cual denota que aporta una mayor información a partes importantes del texto, es decir el texto enriquecido denota un aumento de Ganancia de Información. Una evaluación más completa se realizó utilizando los *datasets* 20 *NewsGroup* y *Reuters*, los cuales fueron enriquecidos para utilizarlos en tareas de clasificación y *clustering*. En la clasificación de documentos se utilizaron diversos modelos de clasificación, teniendo un mejor desempeño cuando se utilizó un *dataset* enriquecido con respecto al utilizar uno original. De la misma forma se utilizaron los *datasets* antes mencionados para hacer *clustering* de documentos. El algoritmo utilizado fue CBC, el cual tuvo un mejor desempeño utilizando un *dataset* enriquecido que el original. Las anteriores evaluaciones muestran que el enriquecimiento de texto identifica, obtiene e integra información afectando al texto original, pero no solamente afecta al texto, sino que afecta al texto original de forma positiva.

Por último, el método de enriquecimiento consta de tres pasos fundamentales. El primer paso es

la identificación de partes del texto a enriquecer, es decir la Identificación de Entidades Nombradas y la Identificación de Entidades Nombradas a partir de conceptos clave. Estas dos tareas son fundamentales en el método debido a que encuentran las partes más importantes del texto a enriquecer. Mientras la Extracción de Entidades Nombradas identifican partes específicas del texto, la Identificación de Entidades Nombradas a partir de conceptos clave identifica el tema principal y lo más importante del texto original. El segundo paso es la obtención de información relacionada a partir de las partes del texto a enriquecer. La manera en cómo enriquecer el texto es importante dado que la información de DBpedia tiene una estructura semántica. El último paso es la integración de esa información, determinando la manera en cómo proporcionar esa información al usuario final y cómo integrar esa información para tareas de clasificación y *clustering*. Si bien el texto original carece de una estructura y forma, con estos tres pasos se analizó y modeló el texto para encontrar una forma de encontrar vínculos y relaciones con otros artículos en DBpedia, incrementando así el conocimiento e información con respecto al tema del texto original.

5.2 Aportaciones

Las principales aportaciones que se obtuvieron a partir de este trabajo de tesis son:

- El diseño de una estructura para representar y explotar información de recursos como DBpedia.
- Un método para identificar Entidades Nombradas a partir de conceptos clave.
- Un método de enriquecimiento de texto a partir de recursos de la Web Semántica.
- La implementación del método de enriquecimiento de texto.
- Un documento de tesis que reporta el trabajo de investigación realizado.

5.3 Dificultades y Limitaciones

Durante el desarrollo de la metodología propuesta se presentaron distintas dificultades, las cuales se listan a continuación:

- El principal reto en el método de enriquecimiento es la identificación de Entidades Nombradas a partir de conceptos clave. En muchas ocasiones en un texto no se encuentran Entidades Nombradas, aún así el método obtiene Entidades Nombradas relacionadas con el tema del texto. La manera en que encuentra estas Entidades es identificando Entidades que estén relacionadas con el tema del texto. Para realizar esta tarea lo ideal sería tener un análisis completo de las Entidades Nombradas en DBpedia, lo cual es una tarea muy compleja debido a que existen millones de Entidades, por consecuencia encontrar relaciones en el texto es una tarea difícil dado que analizar un alto número de Entidades Nombradas es computacionalmente costoso. El método que se desarrolló aborda esa problemática y trata de obtener estas relaciones a partir del menor número posible de Entidades Nombradas.

- Los *datasets* utilizados tienen un formato poco manejable, es decir, se tuvo que aplicar un difícil procesamiento para poder utilizar esos *datasets*. Especialmente el *dataset* Reuters, el cual no se encontró en texto plano sino en formato SGML, el cual es un formato que no es popular y por lo tanto no hay herramientas de procesamiento.

- DBpedia limita el número de consultas que se pueden realizar desde una misma IP Para ello se construyó la infraestructura mencionada en la sección 4, la cual permite realizar diferentes consultas a través de una VPN, obteniendo una IP diferente para realizar un mayor número de consultas.

En cuanto las limitaciones en el desarrollo del método a continuación se describen:

- El método de enriquecimiento de texto depende de DBpedia, la cual se encuentra *online*. Aunque es posible instalar una versión local es necesario actualizarla constantemente para obtener la última versión. DBpedia es una base de conocimiento construida a partir de hechos históricos, personajes, temas en general y temas científicos, sin embargo, aunque DBpedia consta de temas científicos, existen otros temas muy específicos que no contiene, lo cual limita el enriquecimiento de texto muy específico.

- Dado que el método de enriquecimiento es un proceso que consta de diversas tareas, realiza consultas a la base de conocimiento, maneja mucha información y genera documentos adicionales, el método consume un tiempo considerable pero admisible para una aplicación dirigida al usuario final.

- El método utiliza diversas tareas de minería de texto que son sensibles a problemas en el texto original, es decir si el texto original tiene muchos problemas de redacción y escritura, el método de enriquecimiento de texto no funcionará correctamente. Este tipo de problemas son abordados a lo largo del método de enriquecimiento, por ejemplo la corrección ortográfica, pero aún así el método lo tolera.

5.4 Trabajo a Futuro

A continuación se listan algunas ideas consideradas como trabajo futuro para la mejora de este trabajo de tesis:

- Incorporar al método la tarea *Named Entity Resolution*. Esta tarea busca averiguar a qué Entidad Nombrada se refiere en el caso de que existiera ambigüedad con otra(s). Aunque la herramienta GATE provee un módulo que solventa esta tarea, éste es muy básico. *Named Entity Resolution* es un problema que actualmente se encuentra en investigación y por lo tanto

no está del todo resuelto. Añadir esta tarea al método permitiría tener un enriquecimiento más robusto dado que tendría un módulo de desambiguación propio.

- Evaluación del método de enriquecimiento de texto en dominios muy específicos. El método fue evaluado con texto relacionado con noticias, el cual contiene temas de un dominio semi-general. Una tarea interesante sería evaluar el comportamiento del método con texto acerca de temas muy específicos.

- Diseño y construcción de una base de conocimiento propia. El método utiliza a DBpedia como base de conocimiento, por lo tanto el tipo de textos que puede enriquecer está limitado a la información que se encuentra en DBpedia. El enfoque que se podría seguir sería construir una base de conocimiento propia a partir de texto no estructurado. Dicho enfoque es un tema de estudio actualmente, pero daría la libertad de enriquecer texto de cualquier dominio dado que el método no estaría limitado.

- Ejecución del método en un ambiente a gran escala. El método está constituido por diversas tareas, las cuales consumen bastante tiempo de procesamiento. Una posible solución es dividir el texto en partes para que cada una de esas partes se procesen por separado, teniendo un tiempo menor de ejecución. Este tipo de tareas podría realizarse en un ambiente paralelizado, es decir, en un *cluster* de procesamiento.

Bibliografía

[1] J. Cardoso, 'The semantic web vision: Where are we?', Intelligent Systems, IEEE, vol. 22, no. 5, pp. 84?88, 2007.

[2] Jakob Nielsen. Designing Web Usability: The Practice of Simplicity. New Riders Publishing, Thousand Oaks, CA, USA, 1999. ISBN 156205810X.

[3] W3C (2001). W3c semantic web. www.w3.org/2001/sw/. Última consulta: 22 de septiembre de 2014.

[4] M. C. Schraefel, 'What is an analogue for the semantic web and why is having one important in Proceedings of the eighteenth conference on Hypertext and hypermedia?', HT 07, (New York, NY, USA), pp. 123-132, ACM, 2007.

[5] Tim Berners-Lee. Linked data design issues, sep 2009. URL http://www.w3.org/DesignIssues/LinkedData.html. Última consulta: 22 de septiembre de 2014.

[6] Tom y Berners-Lee Tim Bizer, Christian y Heath. Linked Data - The Story So Far. International Journal on Semantic Web and Information Systems (IJSWIS), 5(3):1?22, Mar 2009. ISSN 1552-6283. doi: 10.4018/jswis.2009081901.

[7] Y. Kim, B. Kim, and H. Lim, 'The index organizations for rdf and rdf schema', in Advanced Communication Technology, 2006. ICACT 2006. The 8th International Conference, vol. 3, pp. 4 pp. ?1874, feb. 2006.

[8] Heath, T. and Bizer, C. (2011). Linked Data: Evolving the Web into a Global Data Space. Morgan y Claypool, San Rafael, CA, First edition.

[9] J. Durkin. Expert systems: design and development. Macmillan, 1994. ISBN 9780023309700. URL http://books.google.es/books?id=9-BQAAAAMAAJ. Última consulta: 22 de septiembre de 2014.

[10] Hofweber, Thomas, 'Logic and Ontology', en Edward N. Zalta (en inglés), Stanford Encyclopedia of Philosophy (Spring 2009 Edition edición).

[11] Luciano Floridi, editor. The Blackwell guide to the philosophy of computing and information, volume 14 of Blackwell philosophy guides. Blackwell, Malden, Mass. [u.a.], 2004. URL http://www.gbv.de/dms/goettingen/355339978.pdf. Última consulta: 22 de septiembre de 2014.

[12] Thomas R. Gruber. A translation approach to portable ontology specifications. Knowl. Acquis., 5(2):199?220, jun 1993. ISSN 1042-8143. URL http://dx. doi.org/10.1006/knac.1993.1008. Última consulta: 22 de septiembre de 2014.

[13] RDF Vocabulary Description Language 1.0: RDF Schema. World Wide Web Consortium, Septiembre 2014. URL http://www.w3.org/TR/2004/REC-rdf-schema-20040210/. Última consulta: 22 de septiembre de 2014.

[14] Johan Hjelm. Creating the Semantic Web with RDF: Professional Developer's Guide. John Wiley y Sons, Inc., New York, USA, 2001. ISBN 0471402591.

[15] RDF Schema 1.1. World Wide Web Consortium. http://www.w3.org/TR/rdf-schema . Ultima Consulta 27 de septiembre de 2014.

[16] Tim Berners-Lee and Mark Fischetti. Weaving the web: The original design and ultimate destiny of the world wide web by its inventor. Harper, San Francisco, 1999.

[17] Eric Prud?hommeaux and Andy Seaborne. SPARQL query language for rdf. W3C Recommendation, 4:1?106, 2008. URL http://www.w3.org/TR/rdf-sparql-query/. Última consulta: 22 de septiembre de 2014.

[18] Libby Miller, Andy Seaborne, and Alberto Reggiori. Three implementations of squishql, a simple rdf query language. In International Semantic Web Conference 02, pages 423?435, 2002.

[19] Matthew Perry, Prateek Jain, and Amit P. Sheth. SPARQL-ST: Extending SPARQL to Support Spatiotemporal Queries Geospatial Semantics and the Semantic Web. volume 12 of Semantic Web and Beyond, chapter 3, pages 61?86. Springer US, Boston, MA, 2011. ISBN 978-1- 4419-9445-5.

[20] Gregory Karvounarakis, Sofia Alexaki, Vassilis Christophides, Dimitris Plexousakis, and Michel Scholl. Rql: a declarative query language for rdf. In WWW?02, pages 592-603, 2002.

[21] Volker Haarslev, Ralf Möller, and Michael Wessel. Querying the semantic web with racer. In In Proceedings of the KI-2004 International Workshop on Applications of Description Logics (ADL 04, 2004).

[22] Evren Sirin and Bijan Parsia. SPARQL-DL: SPARQL Query for OWL-DL. In Proceedings of the Third International Workshop on OWL: Experiences and Directions (OWLED 07, Innsbruck, Austria, 2007. URL http://ceur-ws.org/Vol-258/paper14.pdf. Última consulta: 29 de septiembre de 2014.

[23] Document warehousing and text mining, Dan Sullivan, p. 324, 2001.

[24] Maron, M. E.; Kuhns, J. L. 'On relevance, probabilistic indexing and information retrieval' En: Journal of the ACM, 1960, v. 7, n. 3, pp. 216-244.

[25] Salton, Gerard; McGill, Michael J. Introduction to modern information retrieval. New York. McGraw Hill, 1983.

[26] Christoph Kiefer, Abraham Bernstein, Hong Lee, Mark Klein, and Markus Stocker. Semantic Process Retrieval with iSPARQL. The Semantic Web: Research and Applications, 4519:609- 623, 2007. ISSN 0302-9743.

[27] Christoph Kiefer, Abraham Bernstein, Hong Lee, Mark Klein, and Markus Stocker. Semantic Process Retrieval with iSPARQL. The Semantic Web: Research and Applications, 609- 623, 2007.

[28] Jens y Kobilarov-Georgi y Auer Sören Bizer, Christian y Lehmann, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia a crystallization point for the web of data. Web Semantica, 154-165, September 2009.

[29] Martin J. O'Connor and Amar K. Das. Sqwrl: A query language for owl. In Rinke Hoekstra and Peter F. Patel-Schneider, editors, OWLED, volume 529 of CEUR Workshop Proceedings. CEUR-WS.org, 2008. URL http://dblp.uni-trier.de/db/conf/semweb/owled2009.htmlOConnorD08. Última Consulta 3 de octubre de 2014.

[30] Ian Horrocks, Peter F. Patel-Schneider, Harold Boley, Said Tabet, Benjamin Grosof, and Mike Dean. SWRL: A Semantic Web Rule Language Combining OWL and RuleML. W3C Member Submission, 2004.

[31] Efthimis N. Efthimiadis. Query Expansion. In: Martha E. Williams (ed.), Annual Review of Information Systems and Technology (ARIST), v31, pp 121-187, 1996

[32] Y. Qiu and H.P. Frei. Concept Based Query Expansion. In Proceedings of SIGIR-93, 16th ACM International Conference on Research and Development in Information Retrieval, Pittsburgh, SIGIR Forum, ACM Press, June 1993.

[33] Zhi-Yun Zheng, Query Expansion for Answer Document Retrieval in Chinese Question Answering System, Machine Learning and Cybernetics, Proceedings of 2005 International Conference, 2005.

[34] Hang Cui, Query Expansion by Mining User Logs, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 15, NO. 4, JULY/AUGUST 2003.

[35] Rashmi Chauhan, Domain Ontology based Semantic Search for Efficient Information Retrieval

through Automatic Query Expansion, 2013 International Conference on Intelligent Systems and Signal Processing (ISSP).

[36] Wen Syan Li, Supporting web query expansion eficiently using multigranularity indexing and query processing, Data and Knowledge Engineering 35 239-257, (2000).

[37] Olga Vechtomova, Query expansion with terms selected using lexical cohesion analysis of documents, Information Processing and Management.

[38] D. Liu, Improving Text Classification with Concept Index Terms and Expansion Terms, ISNN 2011, Part III, LNCS 6677, pp. 485-492, 2011.

[39] Jun Wang, Improving Short Text Clustering Performance with Keyword Expansion The Sixth ISNN, AISC 56, pp. 291-298, 2009.

[40] Richard Goodwin, Data Enrichment Service, TSO, 2012, http://openup.tso.co.uk/des/enriched (Última consulta 21 de octubre de 2014).

[41] Supakpong Jinarat, Web Snippet Clustering Based on Text Enrichment with Concept Hierarchy, ICONIP 2009, Part II, LNCS 5864, pp. 309-317, 2009.

[42] Khaled Abdalgader, Short-Text Similarity Measurement Using Word Sense Disambiguation and Hyperonim Expansion, AI 2010, LNAI 6464, pp. 435-444, 2010.

[43] Francisco Bueno, Enrichment of text documents using information retrieval techniques in a distributed environment, Expert Systems with Applications 37 (2010) 8348-8358, 2010.

[44] Proscovia Olango, Effect of document enrichment on e-learning, The 2nd International Conference on Integrated Information, 2013.

[45] Abdullah Bawakid and Mourad Oussalah, Centroid-based Classification Enhanced with Wikipedia, 2010 Ninth International Conference on Machine Learning and Applications, 2010.

[46] Christopher Boston, Wikimantic: Toward effective disambiguation and expansion of queries, Data and Knowledge Engineering 90, 2014.

[47] Gerasimos Spanakis, Exploiting Wikipedia Knowledge for Conceptual Hierarchical Clustering of Documents, The Computer Journal, Vol. 55 No. 3, 2012.

[48] Georgios V. Lioudakis, eDocuments Intelligent Enrichment from Distributed Knowledge Resources, IBM, NBIC (Nanotechnology, Biotechnology, Information technology, and Cognitive science) Conference, 2005.

[49] Fraihat Salam, New Semantic Indexing and Search System based on Ontology, 2013 Fourth International Conference on Emerging Intelligent Data and Web Technologies, 2013.

[50] Liu Wenyin, A short text modeling method combining semantic and statistical information, Information Sciences 180, 2010.

[51] Dulce Aguilar-Lopez, Toward the semantic search by using ontologies, 2008 5th International Conference on Electrical Engineering, Computing Science and Automatic Control (CCE 2008), 2008.

[52] Hiroki Yamakawa, Semantic Enrichment of Text Representation with Wikipedia for Text Classification, Institute of Electrical and Electronics Engineers (IEEE) Conference, 2010.

[53] Zhaohui Huang, Semantic Text Mining with Linked Data, 2009 Fifth International Joint Conference on INC, IMS and IDC, 2009.

[54] Chrysoula Zerva and Aliki Kopaneli, Named-Entity Recognition and Text Enrichment using Semantic Web, Graduate Thesis, 2013.

[55] Rabia Batool, Precise Tweet Classification and Sentiment Analysis, Institute of Electrical and Electronics Engineers (IEEE)Conference, 2013.

[56] Alchemy api, (Última consulta octubre de 2014). Disponible en: www.alchemyapi.com.

[57] Marcin Szczuka, Clustering of Rough Set Related Documents with Use of Knowledge from DBpedia, RSKT 2011, LNCS 6954, pp. 394?403, 2011.

[58] Tadej Stajner, Enricher, Service oriented to text enrichment, 2008.

[59] Myunggwon Hwang, Automatic Enrichment of Semantic Relation Network and Its Application to Word Sense Disambiguation, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, vol. 23, no. 6, june 2011.

[60] N. Chinchor, P. Robinson, MUC-7 named entity task definition, in: 7th Conference on Message Understanding, 1997

[61] N. Chinchor, P. Robinson, MUC-7 named entity task definition, in: 7th Conference on Message Understanding, 1997.

[62] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Linguisticae Investigationes 30 (7) (2007).

[63] R. Grishman, B. Sundheim, Message understanding conference: a brief history, in: 16th Conference on Computational Linguistics, 1996, pp. 466-471.

[64] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proceedings of HLT/EMNLP, vol. 5, pp. 339-346, 2005.

[65] A.M. Popescu, O. Etzioni, Extracting product features and opinions from reviews, in: Proceedings of HLT/EMNLP, vol. 5, 2005, pp. 339-346.

[66] NIST, Automatic Content Extraction Evaluation (ACE08). Official Results, 2008.

[67] Treetagger, http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger, última consulta: octubre de 2014.

[68] Wordnet, http://wordnet.princeton.edu, última consulta: octubre de 2014.

[69] Hearst (1999), Untangling Text Data Mining, Proc. of ACL?99: The 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999.

[70] Rousseeuw, P.J.; Kaufman, L. , Finding Groups in Data: An Introduction to Clúster Analysis, (1990).

[71] Weka, página oficial: http://www.cs.waikato.ac.nz/ml/weka. Última consulta: Abril de 2015.

[72] Paranyushkin, D. (2011). Identifying the pathways for meaning circulation using text network analysis. Nodus Labs, Berlin.

[73] Gate: An general architecture for text engineering. Página oficial: https://gate.ac.uk/ Última consulta: abril de 2015.

[74] GraphStream, Página oficial: http://graphstream-project.org Última consulta: octubre de 2015.

[75] Patric Andr´e Pantel. 'Clustering By Committee'. Tesis Doctoral. University of Alberta. 2003.

[76] Marie-Catherine de Marneffe, Bill MacCartney and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In LREC 2006.

[77] Harris, Z. 'Distributional Structure'. The Philosophy of Linguistics. New York: Oxford University Press. 1985. pp. 26-47.

[78] Ramos, J. (1999). Using TF-IDF to Determine Word Relevance in Do cument Queries. Technical rep ort, Department of Computer Science, Rutgers University, 23515 BPO Way, Piscataway, NJ, 08855e

[79] Obtención de axiomas en el aprendizaje de ontologías, Ana Bertha Ríos Alvarado, Tesis Doctoral, Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2013.

[80] Organización de documentos en Español, digitalizados y semánticamente relacionados, Jesús Ángel Cervantes de la Fuente, tesis de maestría,Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional, 2014.

[81] Steegmann-Pascual, C., Ro dríguez-Velázquez, J. A., and Pérez, J. (2002). Álgebra de Matrices. Technical report, Universitat Oberta de Catalunya.

[82] Aspell, http://aspell.net/ . Última consulta: Octubre de 2015.

[83] Gate: https://gate.ac.uk/ Última consulta octubre de 2015.

[84] Grinstead, Charles M.; Snell, J. Laurie (1997). Central Limit Theorem.

[85] Dhaval Thakker, Taha Osman, Phil Lakin, JAPE Grammar, http://gate.ac.uk/sale/thakker-jape-tutorial/GATE%20JAPE%20manual.pdf. Última consulta octubre de 2015.

[86] Christopher D. Manning, Foundations of Statistical Natural Language Processing, 1999, Massachusetts Institute of Technology.

CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS AVANZADOS DEL IPN

Parque Científico y Tecnológico TECNOTAM
Carretera a Soto la Marina Km. 5.5 C.P. 87130
Ciudad Victoria, Tamaulipas, México
Tel: +52 (834) 107 0220 Fax: +52 (834) 107 0224
E-mail: admin@tamps.cinvestav.mx

www.tamps.cinvestav.mx

Cinvestav
Tamaulipas

Los abajo firmantes, integrantes del jurado para el examen de grado que sustentará el C. Dishelt Francisco Torres Paz, declaramos que hemos revisado la tesis titulada:

"Método de enriquecimiento de texto a partir de recursos de la Web Semántica"

Y consideramos que cumple con los requisitos para obtener el grado de Maestro en Ciencias en Computación.

Atentamente,

Dr. Víctor Jesús Sosa Sosa

Dr. Hiram Galeana Zapién

Dr. Iván López Arévalo