Resumen Tema 4: Minería de contenido / Minería de texto

José Alberto Benítez Andrades

Enero 2011

En este trabajo se resumen las conclusiones obtenidas después de haber realizado la lectura de los artículos propuestos Marti A. Hearst *"Untangling Text Data Mining"* , Jordi Turmo, *Information Extraction, Multilinguality and Portability,* MAnselmo Peñas, F. Verdejo y J.Gonzalo *"Terminology Retrieval: towards a synergy between thesaurus and free-text searching".*

1. Creación de corpus.

1.1. ¿Qué es un corpus?

La definición de **corpus** teniendo en cuenta el uso que se le da actualmente, en el ámbito de la lingüística o lexicografía de corpus, o en la lingüística computacional en general, no es tan fácil como parece. En principio, se llama corpus a cualquier colección de textos (corpus como cuerpo textual). Sin embargo, cuando se utiliza este término en la lingüística actual, tiene implicaciones que van más allá del análisis de un cuerpo textual (como por ejemplo pueden ser dos novelas de un autor o un artículo de algún periódico).

Estas implicaciones son patentes en algunas de las definiciones de corpus que se han propuesto en los últimos años. Por ejemplo, *Leech* introduce el concepto de corpus de la siguiente manera:

On the face of it, a computer corpus is an unexciting phenomenon: a helluva lot of text, stored on a computer. Donde se refleja que, aunque sea de un modo bastante simplista, podemos considerar que un corpus no es más que una colección de texto en formato magnético, aunque Leech completa su definición recalcando que la habilidad que poseen los ordenadores para buscar, recuperar, ordenar y hacer cálculos sobre cantidades masivas de texto nos ha brindado la oportunidad de comprender y de explicar el contenido de esos córpora de formas que no eran imaginables en la era que él denomina "pre-computacional". De hecho, dado que los avances tecnológicos van tan unidos al desarrollo de la lingüística de corpus tal y como hoy en día la conocemos, Leech argumenta que debe denominarse Computer Corpus Linguistics, ya que el término "lingüística de corpus" se usaba antes del advenimiento de los ordenadores digitales (Leech ibid).

Hay un consenso en la comunidad científica relativo al hecho de que un corpus no ofrece únicamente información sobre sí mismo, es decir, sobre su contenido, sino que representa una sección más amplia de la lengua seleccionada de acuerdo a una tipología específica:

[...] a corpus is a collection of texts assumed to be representative of a given language, dialect, or other subset of a language to be used for linguistic analysis. Francis (1982: 17)

Tognini-Bonelli hace hincapié sobre la característica de la representatividad del corpus:

[...] a corpus is a collection of naturally-occurring language text, chosen to characterize a state or variety of a language. Sinclair (1991: 171)

La definición que ofrecen *Atkins, Clear y Ostler* añade otro aspecto esencial en la creación de un corpus: el corpus debe ser construido de acuerdo a una serie de criterios explícitos:

[a corpus is] a subset of an ETL (Electronic Text Library) built according to explicit design criteria for a specific purpose.

La definición más estandarizada la ofrece el grupo de trabajo dedicado a los córpora textuales de *EAGLES (Expert Advisory Group on Language Engineering Standards)* :

Corpus: A collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language.

En esta definición se recogen tres aspectos fundamentales que se deben tener en cuenta a la hora de definir los corpus: un corpus debe estar **compuesto por textos producidos en situaciones reales ("pieces of language") y la inclusión de los textos que componen el corpus debe estar guiada por una serie de criterios lingüísticos explícitos para asegurar que pueda usarse como muestra representativa de una lengua**. Todos los estudiosos dedicados al corpus están de acuerdo en que éstos son aspectos fundamentales en la creación y definición de los córpora, aunque no por ello dejan de ser aspectos controvertidos y que en ocasiones han dado lugar a posturas diferentes.

1.2. Posibles usos y utilidad de un corpus

La información léxica es usada para diferentes tipos de etiquetado, las taxonomías existentes se usan para etiquetado semántico y se usa la información léxica para la desambiguación de textos escritos en lenguaje natural, se pueden construir diccionarios por medio de análisis de corpus, construcción semiautomática de léxico específico de un dominio, y adquisición automática de información léxica semántica.

1.3. Creación de corpus a partir de la web

El corpus debe estar en formato electrónico, mediante adaptación de material ya en forma electrónica, escaneado o tecleando. Lógicamente, se deberán tener en cuenta los copyrights pertinentes como salvaguarda contra la explotación y la piratería. Se deberá tener en cuenta el diseño(Lenguaje escrito y hablado, de formal a informal, de literario a ordinario, etc.) Las características que debe tener nuestro corpus al crearle deben ser: Cantidad grande, con calidad auténtica, simple, mejor texto sin enriquecimiento y documentado.

Es frecuente encontrar también muchos corpus construidos en base a noticias de prensa, estos recursos son fáciles de obtener y tienen múltiples aplicaciones especialmente en el desarrollo de herramientas de tratamiento de corpus.

Pero a día de hoy la web es una fuente prácticamente inagotable de documentos de todo tipo y en muchos idiomas que pueden utilizarse para compilar un corpus. Hay que tener en cuenta, de todas maneras, que algunos sitios de la web sólo indexan documentos, es decir que estos sitios sólo cumplen una función de portal de acceso. El acceso al documento real requiere algún otro tipo de iniciativa.

Los corpus obtenidos a partir de la web se suelen denominar “oportunísticos” en oposición a los corpus planificados (sólo incorporan documentos previamente seleccionados) y a pesar de sus inconvenientes (falta de control, documentación, falta de representatividad, etc.) suelen utilizarse para di-

ferentes propósitos (para complementar corpus de referencia con material actualizado y crear corpus exploratorios de lenguajes de especialidad entre otros usos).

Para compilar un corpus de este tipo es suficiente con la utilización sistemática de cualquiera de los motores de búsqueda comerciales existentes (tales como Yahoo, Google y AltaVista entre otros). Esta aproximación tiene el inconveniente de que la inversión de tiempo puede ser importante y no siempre justificable. Para solventar este inconveniente existen algunas herramientas que automatizan el proceso de búsqueda de documentos. Tanto si la exploración se hace manualmente o a través de programas especializados, la idea básica es utilizar una o más palabras semilla que permitan recuperar (combinándolas de alguna manera) algunos documentos. Estos documentos una vez explorados convenientemente, permiten obtener nuevos términos que se utilizan en combinación (o no) con los anteriores para obtener nuevos documentos y así se repite el ciclo hasta completar un corpus del tamaño deseado.

Los corpus así obtenidos son necesariamente sobre lenguajes de especialidad. Obtener un corpus general es más difícil aunque no imposible. Existen varias técnicas para explorar sistemáticamente la web o bien utilizar la iniciativa ODP (Open Directory Project, http://www.dmoz.org/) como se sugiere en Liu et al. (2006).

1.4. Ejemplos de algunos corpus y su finalidad

Los corpus más importantes son:

- **Brown Corpus:** En 1961/1963, Kucera y Francis publicó su obra clásica Computational Analysis of Present-Day American English (1967), que proporcionan las estadísticas básicas en lo que hoy se conoce simplemente como el Brown Corpus. El Brown Corpus fue una selección cuidadosamente compilada del actual Inglés Americano, por un total de alrededor de un millón de palabras extraídas de una amplia variedad de fuentes. Kucera y Francis realizaron una variedad de análisis computacionales, gracias a los cuales consiguieron compilar una obra rica y variada, que combina elementos de la lingüística, la psicología, la estadística y la sociología. Ha sido muy ampliamente utilizado en la lingüística computacional , y fue durante muchos años uno de los recursos más citados en el campo. Poco después de la publicación del primer análisis léxico estadístico, la editorial Bostón Houghton-Mifflin Kucera tuvo una oferta millonaria para crear un nuevo diccionario American Heritage Dictionary . El Brown Corpus ha sido la base de otros corpus posteriores a él como el Lancaster-Oslo-Bergen Corpus o el SUSANNE . El corpus original (1961) contenía 1.014.312 palabras muestra de 15 categorías de texto (de prensa: reportaje, editorial y comentarios, religión, habilidad y hobbies, bellas artes, temas varios, aprendidas, ficción: general, misterior y detectives, ciencia, aventura, romance, y humor). El Brown Corpus es un corpus de tipo "POS Tagging" (Part-Of-Speech tagging, de análisis léxico) y posee **82 etiquetas distintas**.

- **SUSANNE Corpus:** Susanne es la abrevación de Surface and Underlying Structural ANalysis of Natural English. Es un corpus procedente del Brown Corpus, que originariamente procedía de 64 de las 500 muestras del Brown Corpus y su versión inicial está formado por unas 130.000 palabras. Al igual que el Brown Corpus, es de tipo "POS Tagging" (Análisis Léxico), y está formado por 353 etiquetas, y las temáticas que contiene son las A,G,J y N del Brown Corpus (Prensa, Bellas Letras, Aprendidas y Ficción: Aventura y Occidental). El Corpus SUSANNE se creó, con el patrocinio del Comité Económico y Social Research Council (Reino Unido), como parte del

proceso de elaboración de una taxonomía completa del lenguaje de la ingeniería orientada y el esquema de anotación de la gramática (lógica y de la superficie) de Inglés. El Corpus SUSANNE, a pesar de haberse creado con el patrocinio de TEI, no cumple la normativa. El esquema

analítico SUSANNE ha sido desarrollado sobre la base de las muestras de ambos Inglés británico y americano. Fue inicialmente orientado hacia la lengua escrita solamente, y de hecho contiene muestras exclusivamente del lenguaje escrito. Sin embargo, en trabajos posteriores patrocinado por primera vez por el Royal Signals and Radar Establishment, se produjeron extensiones al sistema para anotar los fenómenos distintivos estructurales del lenguaje hablado, y ha aplicado a estas muestras de los últimos Inglés hablado espontáneo (modificación mostrada en el Corpus CHRISTINE). La primera etapa del Corpus CHRISTINE, que incluye análisis de una equilibrada sección del Inglés hablado en todas partes del Reino Unido en la última década, fue lanzado en agosto de 1999 y es uno de los corpus orales para poder analizar el lenguaje hablado. El Corpus

SUSANNE abarca un subconjunto de aproximadamente 130.000 palabras del Brown Corpus de Inglés Americano, anotado, de acuerdo con el esquema de Susanne. Los motivos originales para la producción de esta base de datos incluye el de proporcionar mejores estadísticas para el análisis probabilístico, pero en este sentido, el Proyecto SUSANNE fue alcanzado después de su creación por los proyectos (en particular, Mitchell Marcus Pennsylvania proyecto Treebank) que han utilizado métodos cuasi-industrial para generar cuerpos mucho más grandes de material a analizar gramaticalmente. Sin embargo, el Corpus Susanne sí que mejora notablemente el análisis probabilístico en comparación con el Brown Corpus.

- **Penn Treebank:** El Treebank Penn, es un corpus de más de 4,5 millones de palabras de Inglés Americano. Durante la primera fase de tres años del Proyecto Penn Treebank (1989 - 1992), este corpus posee 2 tipos de etiquetado de, de tipo léxico y de tipo sintáctico. Sus orígenes están en la Universidad de Pennsylvania. El conjunto de muestras del que está compuesto, procede de

  diferentes corpus distintos, concretamente de los siguientes: Dept. of Energy abstract, Dow Jones

  Newswire stories, Dept. of Agriculture bulletins, Library of America texts, MUC-3 messages, IBM Manual sentences, WBUR radio transcripts, ATIS sentences, Brown Corpus, retagged.

2. Extracción de Información textual (Automatic Information Extraction):

2.1. Definición y objetivos

La extracción de información es la tarea que se encarga de identificar descripciones de eventos en textos en lenguaje natural y por consiguiente, extraer la información relacionada a dichos eventos [Patward S. & Riloff E.,2006]. En otras palabras, un sistema de extracción de información 32 encuentra y enlaza la información relevante, mientras ignora la extraña e irrelevante [Cowie J. & Lehnert W., 1996].

La creciente disponibilidad de fuentes on-line en formato texto y el número potencial de enfoques de adquisición del conocimiento de datos textuales, tales como la Extracción de Información ha llevado a incrementos en la investigación de extracción de la información, como generar bases de datos de los documentos, así como también la adquisición de conocimiento útil para tecnologías emergentes como responder a preguntas e integración de la información, entre otras relacionadas con la minería de texto.

La tecnología de extracción de la información responde a retos más difíciles que los de recuperación de la información, ya que mientras en IR la respuesta a una consulta es simplemente una lista de documentos potencialmente relevantes, en la IE el contenido relevante de esos documentos tiene que ser localizado y extraído del texto. Este contenido relevante es decidido a priori, lo que hace que haya una clara dependencia del dominio de la tecnología IE. Cuando se traten nuevos dominios, se necesitará nuevo conocimiento específico y tiene que ser adquirido por tales sistemas.

Los inicios de la extracción de información se ubican a mediados de los años 60´s. Sin embargo, es a finales de los 80´s cuando esta tecnología comienza a tener auge. Esto se debió a tres factores: el poder computacional, el exceso de información textual existente de forma electrónica y la intervención de la Agencia de Defensa de los Estados Unidos (DARPA).

DARPA patrocinó durante los años de 1987 a 1998 las siete conferencias sobre entendimiento de mensajes (MUC). Asimismo, durante los años de 1990 a 1998 DARPA activó el TIPSTER (Programa de Investigación sobre Recuperación y Extracción de Información), donde las MUC fueron incluidas.

Las MUC fueron las que inicialmente fomentaron las competencias entre distintos grupos de investigación. Las cuales se llevaron a cabo con el objetivo de desarrollar sistemas de extracción de información. Es por ello que también definieron sus propios métodos de evaluación. En cada una de las MUC se han empleado diferentes dominios. En MUC-1 y MUC-2 se utilizaron noticias sobre operaciones navales, posteriormente, en MUC-3 y MUC-4 se empleó el dominio sobre atentados terroristas en América Latina. Después, en MUC-5 [Chinchor N. & Sundheim B., 1993] se hizo uso de noticias sobre fusiones de empresas y anuncios de productos microelectrónicos. De igual forma, en MUC-6 [Sundheim B., 1993] se utilizaron noticias sobre sucesión de directivos. Asimismo, en MUC-7 [Chinchor N., 1998] hicieron uso de dos dominios, uno sobre noticias de accidentes de avión y otro sobre lanzamiento de misiles y artefactos (para un estudio más completo véase [Grishman R., 1993]).

A continuación, se muestra un ejemplo de cómo sería el funcionamiento de un sistema de extracción de información. El siguiente texto es una parte de un documento que pertenece al dominio de sucesión de directivos extraído de un texto libre [Turmo J. et al., 2006].

- *A.C.Nielsen Co. dijo que George Garrick, de 40 años, presidente de los recursos de información de Londres que se basa en la operación de servicios de información europea, se convertirá en presidente de Nielsen Marketing Research, una unidad de la corporación Dun&Bradstreet. Él será el sucesor de John I.Costello quién renunció en marzo.*

La salida de un sistema de extracción de información es un conjunto de registros por noticia de entrada. En la tabla siguiente, se muestra el registro extraído del fragmento de texto mostrado en esta sección. Cabe mencionar que cada registro está compuesto por campos. Dichos campos se establecen desde las primeras etapas del sistema de extracción y se agrupan en lo que se denomina plantilla de extracción. Es importante señalar que cada campo representa información relevante de acuerdo al dominio, la cual será útil para el análisis del conjunto de documentos textuales de entrada.

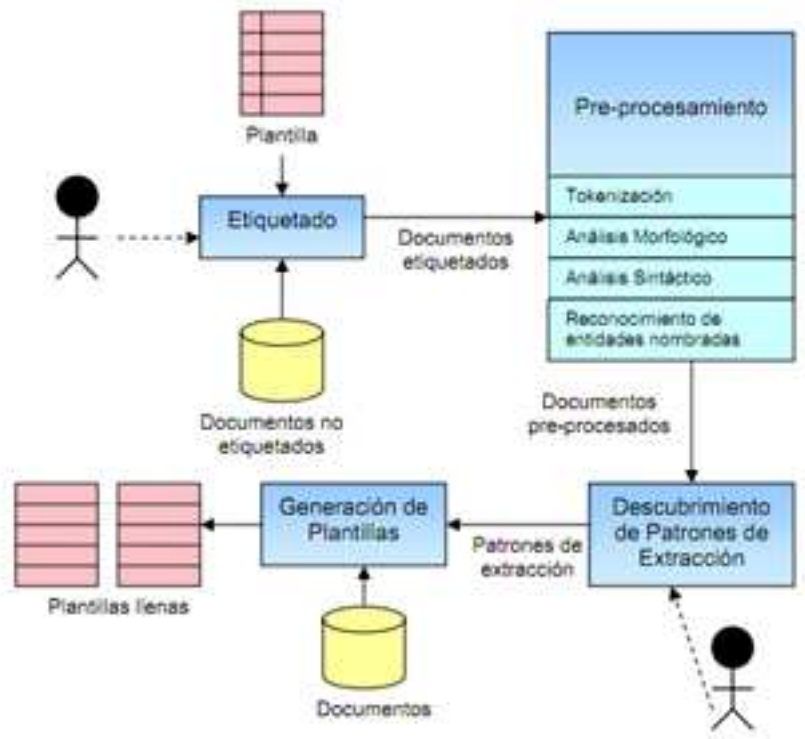
INFORMACIÓN	DE DIRECTIVOS
PERSONA ENTRANTE	George Garrick
PERSONA SALIENT	John I.Costello
PUESTO	Presidente
ORGANIZACIÓN	Nielsen Marketing Research

Se han construido diversos métodos de extracción de información hasta la fecha. No obstante, los trabajos que presentan dichos métodos se caracterizan por emplear dos tipos de enfoques: el supervisado y el nosupervisado.

2.2. Arquitectura de un sistema de EI

Para poder explicar la arquitectura de un sistema de EI, me ha sido de gran ayuda la lectura del artículo de Jordi Turmo , *Information Extraction, Multilinguality and Portability*, que viene perfectamente explicado en la sección 4.

En general, la combinación de los módulos en cascada, permite la realización de las siguientes funciones, en mayor o menor medida :



- **Preprocesado de documento**: El preprocesamiento de los documentos puede alcanzarse gracias a una variedad de módulos como por ejemplo: los text *zoners*[1], *segmenters* [2](también llamados splitters), *filters* [3], *tokenizers* [4], *lexical analyzers* [5], *disambiguators* [6], *stemmers* y *lemmatizers*, entre otros. Muchos sistemas tienen ventaja sobre recursos y propuestas generales disponibles. Especialmente interesantes para el EI son los módulos de reconocimiento NE. El proceso de reconocimiento de NE puede ser rápido utilizando *finite-state transducers* y diccionarios de búsqueda. Los resultados dependen de las fuentes de información que se utilizan. Por ejemplo, Grishman, utilizó las siguientes fuentes: un pequeño diccionario geográfico, que contenía los nombres de todos los países y las ciudades más importantes; un diccionario de empresas

[1] convierten un texto en un conjunto de zonas de texto
[2] a cargo de las zonas de la segmentación en unidades apropiadas, por lo general frases
[3] seleccionan los segmentos relevantes
[4] obtienen unidades léxicas
[5] realizan análisis morfológicos y clasificación y reconocimiento NE
[6] POS taggers, semantic taggers, etc.

derivado de Fortune 500; un diccionario de Agencia Gubernamental; un diccionario de nombres comunes; y un diccionario de términos especificos.

- **Análisis del discurso e interpretación semántica** , para enlazar interpretaciones relacionadas entre sentencias. Esto se hace por medio de coreferencia y resolución de anáfora y otros tipos de inferencias semánticas

  - Un análisis completo implica un gran y relativamente ilimitado espacio de búsqueda y en consecuencia es caro.
  - Un análisis completo no es un proceso robusto porque el árbol sintáctico global no se alcanza siempre. Para suplir tal falta, se intenta que el análisis cubra la mayor subcadena de la frase. Algunas veces, sin embargo, la meta global llevó a opciones de análisis localmente incorrectas
  - Un análisis completo puede producir resultados ambiguos. Se consigue usualmente más de una interpretación sintáctica. En esta situación se debe seleccionar la interpretación más correcta
  - Las gramáticas de amplio espectro, necesarias para hacer una análisis completo son difíciles de afinar. Tratando con los nuevos dominios, nuevas construcciones sintácticas podrían ocurrir en los textos especializados y no ser reconocidos
  - Un análisis del vocabulario no puede manejar situaciones fuera del vocabulario
  - Una vez que los constituyentes han sido analizados, los sistemas resuelven dependencias específicas del dominio entre ellas, generalmente usando las restricciones semánticas impuestas por el escenario de extracción. Se suelen conseguir dos enfoques diferentes para resolver tales dependencias:
    - Reconocimiento de patrones: La mayoría de los sistemas de extracción de la información utilizan este enfoque. La simplificación sintáctica permite reducir el procesado semántico a coincidencia de patrones, donde patrones específicos del escenario, también llamados patrones de extracción o reglas de EI, se utilizan para identificar dependencias entre los constituyentes. De hecho, tales reglas de EI son conjuntos de decisiones de resolución de ambigüedades para ser aplicadas durante el proceso de análisis completo. Pueden ser vistos como un conjunto de expectaciones semántico-sintácticas de las diferentes tareas de extracción. Por un lado, algunas reglas de EI permiten identificar propiedades de entidades y relaciones entre tales entidades. En general esto se consigue mediante el uso de información sintáctico-semántica sobre nombres y modificadores. Por otro lado las reglas de EI que usan relaciones predicado-argumento (Objeto, sujeto, modificadores) permiten identificar eventos entre entidades. La representación de estas reglas IE difieren grandemente entre los diferentes sistemas de EI.
    - Relaciones gramaticales: Generalmente, la estrategia de coincidencia de patrones requiere una proliferación de reglas de EI específicas para la tarea, con variantes explícitas para cada forma verbal, variantes explícitas para diferentes cabeceras léxicas. En vez de usar reglas de EI, un modelo sintáctico más flexible consiste en definir un conjunto de relaciones gramaticales entre entidades como relaciones generales(Sujeto, objeto y modificador), algunas relaciones de modificador especializadas(Temporales y de localización) y relaciones para argumentos mediados por sintagmas proposicionales entre otros. De forma similar a las gramáticas de dependencia, se construye un grafo siguiendo reglas generales de interpretación para las relaciones gramaticales. Los fragmentos

previamente detectados son nodos dentro de tal grafo, siendo las relaciones entre ellos las aristas etiquetadas

- **Análisis del discurso.** Los sistemas de EI generalmente proceden representando la información extraída de una frase como plantillas parcialmente rellenas o como formas lógicas. Tal información puede ser incompleta debido a la ocurrencia de elipsis, y algunas veces puede referir a las mismas entidades en presencia de coreferencia, anáfora. La meta principal del proceso de Análisis del discurso es la resolución de estos aspectos semánticos. Los sistemas que funcionan con plantillas parciales hacen uso de algún procedimiento de fusión para esa tarea. Sin embargo, trabajar con formas lógicas permite que los sistemas de EI usen procesos de interpretación semántica tradicional
- **Generación de plantillas de salida,** para traducir las interpretaciones finales en el formato deseado. La generación de plantillas intenta mapear las informaciones hacia el formato de salida deseado. Sin embargo, algunas inferencias pueden ocurrir en esta fase debido a restricciones específicas de dominio en la estructura de salida, como los siguientes casos:
  - Huecos de salida que toman valores de un conjunto predefinido.
  - Huecos de salida forzadas a ser instanciadas.
  - Clases de información extraída que generan un conjunto de plantillas de salida diferentes.
  - Huecos de salida que tienen que ser normalizados. Por ejemplo, fechas, productos que deban ser normalizadas con un código de una lista estándar

3. Extracción de terminología (Automatic Terminology Extraction)

3.1. Definición y objetivos

La extracción de terminología es el proceso mediante el cual se seleccionan de un texto o conjunto de textos unidades candidatas a constituir términos. No hay que confundir la extracción de terminología con la creación de un glosario terminológico a partir de un texto y de una base de datos terminológica. En el caso de la extracción automática de terminología, intentamos descubrir los términos más relevantes sin conocer previamente estos términos. En cambio, en el segundo caso, buscamos qué términos de una base de datos terminológica están presentes en un determinado texto y por lo tanto los posibles términos son conocidos a priori.

La extracción automática de terminología es una aplicación de la lingüística computacional muy interesante para la actividad del traductor tanto en la fase de preparación de un proyecto, como posteriormente, una vez finalizado. En la fase de preparación de un proyecto podemos descubrir los términos más relevantes y unificar los equivalentes de traducción. Esta posibilidad es especialmente interesante en proyectos grandes donde participan diferentes colaboradores. También es interesante extraer terminología a partir de proyectos ya finalizados, para poder recopilar entradas terminológicas que se puedan utilizar en futuros proyectos.

3.2. Metodología

Como bien se indica en el artículo de Anselmo Peñas, Felisa Verdejo y Julio Gonzalo, "*Terminology Retrieval: towards a synergy between thesaurus and free text searching*", la construcción de un diccionario de sinónimos requiere coleccionar un conjunto de términos salientes. Esta es una tarea que combina enfoque deductivo e inductivo. Los procedimientos deductivos funcionan analizando vocabularios, sinónimos e índices existentes para diseñar un nuevo diccionario de sinónimos para el alcance, estructura y nivel de especificación deseados; el enfoque inductivo analiza los vocabularios del mundo real en los repositorios de documentos para identificar términos y actualizar las terminologías. Esto se divide en tres pasos principalmente:

1. Extracción de términos via análisis morfológicoparte de etiquetado del discurso y análisis. Se distingue entre términos de una palabra(términos monoléxicos), y términos de varias palabras, extraídos con distintas técnicas

2. Valoración de peso de los términos con información estadística, midiendo la relevancia del término en el dominio.

3. Selección del término. Ranking del término y truncando las listas por umbrales de peso.

Estos pasos requieren uno previo en el que el corpus relevante sea identificado, automáticamente recolectado y preparado para la tarea de recuperación de la terminología.

3.3. Extracción de terminología a partir de la web

El sistema **Website Term Browser,** aplica técnicas de NLP para realizar automáticamente las siguientes tareas:

1. **Extracción de Terminología e indexación de una colección de textos multilingüe.** La colección de documentos es procesada automáticamente para obtener una lista grande de frases terminológicas. La detección de frases en la colección se basa en patrones sintácticos. La selección de frases se basa en la frecuencia del documento y en la inclusión de frases

2. **Procesamiento interactivo de consultas en lenguaje natural y recuperación**

   a) Las palabras de la búsqueda lematizadas se expanden con palabras semánticamente relacionadas en el lenguaje de la consulta, y todos los lnguajes objetivo usando la base de datos léxica EuroWordNet y algunos diccionarios bilingües

   b) Se recuperan algunas de las frases que contienen algunas de las palabras expandidas. El número de palabras expandidas es normalmente alto y el uso de palabras semánticas relacionadas (tales como sinónimos) producen mucho ruido. Sin embargo a recuperación y ordenamiento de términos vía información en frases descarta la mayor parte de las combinaciones inapropiadas de palabras, tanto en el lenguaje origen como en el lenguaje destino.

   c) A diferencia de la recuperación inter-lingüística por colección, donde la información de las frases se usa sólo para seleccionar la mejor traducción de acuerdo a su contexto, en este proceso todas las frases resultantes se guardan para el proceso interactivo de selección

   d) Los documentos también son ordenados de acuerdo a la frecuencia y cobertura de las frases relevantes que contienen.

3. **Navegación por proposiciones considerando variaciones morfo-sintácticas, semánticas e inter-lingüísticas de la consulta.** Se presentan dos tipos de información: Un rango de frases que salen en la colección y que son relevantes para la consulta del usuario y un rango de documentos. Las frases en los diferentes lenguajes se muestran organizadas por una jerarquía de acuerdo a:

   a) Número de términos expandidos contenidos en la frase.

   b) Aparición de la frase de acuerdo a su peo como expresión terminológica . Este peso se reduce a la frecuencia dentro de la colección de documentos Si no hay un corpus multidisciplinar con el que comparar.

   c) Inclusión de frases. Para propósitos de presentación un grupo de frases conteniendo subfrases se presentan como incluidas por la subfrase más frecuente en la colección. Eso ayuda a navegar por el espacio de frases de manera similar a una jerarquía de temas.

3.4. Problemática asociada al lenguaje natural

Los problemas principales que surgen en el lenguaje natural son los siguientes:

- Pérdida de cobertura debida a patrones sintácticos no exhaustivos y etiquetado incorrecto de parte del discurso
- Pérdida de cobertura debido a una lematización incorrecta de componentes de frases en el texto.
- Pérdida de cobertura debida a una incorrecta expansión, lematización y traducción de las palabras de la consulta e incorrecto descarte en la selección de frases y en la clasificación de los términos
- Falta de coincidencias causadas por acentos y mayúsculas.

4. Similitud, clasificación, clustering

4.1. Definición de cada uno. Semejanzas y diferencias.

- **Clasificación:** Es un tipo de análisis de datos, pueden ser usados para clasificar datos y los que se usan para predecir tendencias. La clasificación de datos predice clases de etiquetas. Otra técnica es la predicción de datos que predice funciones de valores continuios. Aplicaciones típicas incluyen análisis de riesgo para préstamos y predicciones de crecimiento. Algunas técnicas para clasificación de datos incluyen: clasificación bayesianas. K-Nearest Neighbor, algoritmos genéticos, entre otros.
- **Clustering:** En el ámbito de la web podemos decir que se han hecho diversos estudios orientados principalmente a realizar agrupamientos por contenido. Por ejemplo, cuando hacemos búsquedas por temas denominados motores de búsqueda los cuales indexan archivos almacenados en los servidores Web de los cuales podemos citar el sistema Grokker. Grokker es un sistema de búsqueda que permite realizar búsquedas en la base de datos de Yahoo!, en la tienda de libros Amazon y en Librería Digital ACM. Los resultados se agrupan por similitud de contenidos y también se pueden presentar de forma gráfica, en forma de esferas (clusters) agrupando temáticas.

Las técnicas de clustering son técnicas de clasificación no supervisadas de patrones (observaciones, datos o vectores de característicos) en grupos o clusters. Estas técnicas han sido utilizadas en diversas disciplinas y aplicadas en diferentes contextos, lo cual refleja una gran utilidad en el análisis experimental de datos.

El agrupamiento se ha incluido dentro del ámbito de la inteligencia artificial encuadrándose dentro del aprendizaje no supervisado, por última la Minería de Datos recoge el agrupamiento como una de las clases de problemas a tratar dentro de su ámbito y recupera las técnicas y metodologías previamente desarrolladas extendiéndolas al volumen de datos que se procesan en este campo. De forma más general, podemos definir el clustering como el proceso de clasificación no supervisada de objetos.

- **Clustering vs clasificación**: En primer lugar es importante distinguir entre *agrupamientos* o *clasificaciones no supervisadas* y *análisis discriminante* o *clasificación supervisada*. En el primer caso no se tiene ninguna información relacionada con la organización de los ítems en los grupos o clases y el objetivo es encontrar dicha organización en base a la proximidad entre ítems. Casi no existe información previa acerca de la estructura y la interpretación de las clases o grupos obtenidos es realizada posteriormente por el analista. En el segundo se posee información de qué clase pertenece cada ítem y lo que se desea determinar cuáles son los factores que intervienen en la definición de las clases y qué valores de los mismos determinan estas. Se puede clasificar el agrupamiento y la clasificación en general según distintos criterios.

Un ejemplo claro de agrupamiento sería la búsqueda de grupos de clientes de una entidad bancaria utilizando para ellos los datos de la cuenta corriente: edad, dirección, nivel de renta... etc. Y un ejemplo de clasificación sería encontrar los elementos que determinan la aparición de cáncer de pulmón analizando datos de, edad, calidad de vida, nivel económico... etc. tanto de personas enfermas como sanas.

Segundo, podemos decir que la tarea de clasificar o clasificar objetos en categorías es una de las actividades más comunes y primitivas del Hombre y viene siendo identificada en función de grandes volúmenes de información en diversas areas.

Intuitivamente, dos ítems o variables pertenecientes a un grupo válido deben ser más parecidos entre si que aquellos que esten en grupos distintos y partiendo de esta idea se desarrollan las técnicas de agrupamientos. Estas técnicas dependes claramente del tipo de datos que se está analizando, de qué medidas de semejanzas se estén utilizando y de qué clase de problema se esté resolviendo.

En un sentido más correcto, el objetivo es reunir un conjunto de objetos en clases tales que el grado de asociación natural para cada individuo es alto con los miembros de su misma clase y bajo con los miembros de las otras clases. Lo esencial del análisis de agrupar se enfoca entonces a cómo asignar un significado a los términos, grupos naturales y asociación natural, donde natural usualmente se refiere a estructuras homogéneas y bien sepradas.

4.2. Finalidad de cada uno.

- **Clustering**: El objetivo del clustering consiste en identificar grupos distintos en un conjunto de datos. La idea básica del clustering basado en modelos es la aproximación de la densidad de datos por un modelo de mezcla, por lo general una mezcla de gaussianas, y para estimar los parámetros de las densidades de los componentes, las fracciones de mezcla, y el número de componentes de los datos. El número de grupos distintos en los datos entonces se toma como el número de componentes de la mezcla, y las observaciones se dividen en grupos (las estimaciones de los grupos) utilizando la regla de Bayes. Si los grupos están bien separados y mirar de Gauss, a continuación, las agrupaciones resultantes de hecho tiende a ser distinta", " en el sentido más común de la palabra - contiguos, zonas densamente pobladas del espacio de características, separadas por contiguos, regiones relativamente vacías. Si los grupos no son de Gauss, sin embargo, esta correspondencia puede romper; un grupo aislado, con una distribución no elíptica, por ejemplo, puede ser modelado por no uno, sino varios componentes de la mezcla, y los grupos correspondientes ya no estar bien separados . Se presentan los métodos para evaluar el grado de separación entre los componentes de un modelo de mezcla y entre los grupos correspondientes. También proponemos un algoritmo para la poda del árbol de racimo generado por la agrupación jerárquica basada en modelos. El algoritmo se inicia con el árbol correspondiente al modelo de mezcla elegida por el Criterio de Información Bayesiano. A continuación, se funde progresivamente las agrupaciones que no parecen corresponder a los distintos modos de la densidad de datos.

- **Clasificación:** intenta asignar un elemento de datos a una categoría predefinida basada en un modelo creado a partir de datos de entrenamiento pre-clasificados (aprendizaje supervisado). Términos más generales, tanto la agrupación y clasificación están bajo el área de descubrimiento de conocimiento en bases de datos o data mining. La aplicación de técnicas de minería de datos de contenido de páginas Web que se conoce como minería de contenido web que es un nuevo subárea de la minería web, parcialmente construida sobre el terreno establecidos de recuperación de información

4.3. Usos y aplicaciones

- **Clustering**: Dentro del área de la Minería Web de Uso podemos encontrar diversos estudios relacionados principalmente en agrupamientos por contenido, siendo este uno de las principales área donde se utiliza el clustering en la Web. Por ejemplo podemos nombrar algunos buscadores que utilizana esta técnica para realizar agrupamiento o clustering por contenido como Vivisimo, Grokker, Clusty, iBoogie.

  Con esto podemos decir que existen diferentes sistemas que se preocupan de saber cuáles son las características del usuario relacionado principlamente en el contenido que el usuario visita o los temas que se relacionan con su navegación.

  Por esta razón surge una necesidad, la necesidad de agrupar las páginas de los usuarios para saber cuáles son las páginas más representativas, también un segundo enfoque relacionado con la agrupación de las sesiones de usuarios, ya que a partir de esta agrupación podemos identificar grupos de usuarios con ciertas carcterísticas, preferencias y/o intereses en su navegación. Lo cual nos permitirá realizar un estudio demográfico y también obtener diferentes perfiles que representen a los conjuntos de las características de los usuarios. Realizando estas agrupaciones podemos de alguna manera entregar una mejor información al usuario durante su navegación.

La figura nos muestra un enfoque general de lo que hemos planteado hasta estos momentos. Esta representación que muestra la figura esta hecho en un pseudo-lenguaje que nos permitirá ver todo lo relacionado con la partición inicial de los datos, pasando por la técnica de agrupamiento tanto para las páginas como para las sesiones y finalmente la validación de los resultados que es un punto de suma importancia al momento de obtener los resultados.

- **Clasificación**: El método de clasificación se basa en ingenuas Bayes. Algunos trabajos notables que se ocupan de abrillantar de búsqueda en la web incluyen, que describe dos métodos particional, que es un enfoque de agrupación jerárquica. Nahm y Mooney [NMoo] se describe una metodología que puede ser la extracción de información y minería de datos se combinaron para mejorar unos a otros, la extracción de información proporciona el proceso de minería de datos con acceso a los documentos de texto (text mining) y en vez de minería de datos proporciona reglas para el ganado porción de extracción de información para mejorar su rendimiento.

5. Áreas de investigación relacionadas

La minería de contenido y minería de texto, están directamente relacionadas con las áreas de investigación referentes a los siguientes enfoques:

- Recuperación de Información y Extracción de Información
  - La IR y la web mining tienen diferentes objetivos, es decir la web mining no busca remplazar este proceso. La web mining pretende ser utilizada para. La recuperación de información es altamente popular en grandes empresas del mundo web, las cuales hacen uso de este tipo de sistemas, la máquinas de búsqueda (google y altavista), directorios jerárquicos (yahoo) y otros tipos de agentes y de sistemas de filtrado colaborativos.
  - La diferencia principal, independientemente de las técnicas que usan, que existe entre la Recuperación de la información y la Extracción de la Información recae principalmente en que uno recupera documentos relevantes de una colección y la otra recupera información relevante de dichos documentos. La IE se centra principalmente en la estructura o la representación de un documento mientras que la IR mira al texto en un documento como una bolsa de palabras en desorden.
  - Las principales categorías de la Web Text mining son Text Categorization, Text Clustering, association analysis, trend prediction.
    - Text Categorization: dada una predeterminada taxonomía, cada documento de una categoría es clasificada dentro de una clase adecuada o más de una. Es más conveniente ó fácil realizar búsquedas especificando clases que buscando en documentos. Actualmente existen varios algoritmos de text categorization, dentro de los cuales encontramos, K-nearest, neighbor-algorithm y naive bayes algorithm.
    - Text Clustering: el objetivo de esta categoría es el de dividir una colección de documentos en un conjunto de clusteres tal que la similitud intra-cluster es minimizada y la similitud extra-cluster es maximizada. Podemos hacer uso de text clustering a los documentos que fueron extraidos por medio de una máquina de búsqueda. Las busquedas de los usuarios referencian directamente a los clusters que son relevantes para su busqueda. Existen dos tipos de text clustering, clustering jerarquico y clustering particional (G-HAC y k-means)

- Desde el punto de vista de Bases de Datos
    - El objetivo principal que tiene la web content mining desde el punto de vista de BD es que busca representar los datos a través de grafos etiquetados.

Pero también, está relacionado con las siguientes áreas:

- Minería de Estructura Web (Web Structure Mining)
- Minería de Uso Web (Web Usage Mining)
    - Reglas de asociación.
    - Patrones de secuencia.
    - Clustering.

6. Conferencias internacionales

Algunas de las conferencias internacionales que abordan el tema de la minería de uso de la web, son las siguientes:

- International Conference on Databases Theory (ICDT)
- Internacional Conference on Very Large Data Base IBM Almaden Research Center
- International World Wide Web Conference
- Conference on Artificial Intelligence (AAAI198)
- International Conference on Machine Learning (ICML)
- International Conference on Distributed Computing Systems
- European Conference on Machine Learning (ECML-98)
- International Conference Machine Learning
- International Conference on Knowledge Discovery and Data Mining
- International Computer Software and Applications Conference on Prolonging Software Life

Referencias

[1] Marti A. Hearst. Untangling Text Data Mining. Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, June 20-26, 1999 (invited paper).

[2] Turmo, Jordi. Information Extraction, Multilinguality and Portability. Revista Iberoamericana de Inteligencia Artificial, N.22, vol. 5, Invierno 2003.

[3] Peñas, A., Verdejo, F. and Gonzalo, J. Terminology Retrieval: towards a synergy between thesaurus and free-text searching. In F.J. Garijo, J.C. Riquelme and M. Toro editors, Advances in Artificial Intelligence - IBERAMIA 2002, LNAI 2527, Lecture Notes in Computer Science. Springer-Verlag, 2002.

[4] Escobar-Jeria V., Martin-Bautista M., Sanchez D. y Vila M.A (2006) Minería web: Aplicaciones con lógica difusa.

[5] Etzioni 0. (1996) The world wide web: Quagmire or gold mine.

[6] Tan A. (1999) Text mining: Promises and challenges.

[7] Zadeh L. (1975) The concept of linguistic variable and its application to approximate reasoning..

[8] Víctor Heughes Escobar Jeria (2007) Minería Web de Uso y Perfiles de Usuario: Aplicaciones con Lógica Difusa.

[9] García F. y Gil A. (2002) Personalización y recomendación en aplicaciones de comercio electrónico.

[10] Kim K. y Cho S. (2001) Personalized mining of web documents using link

[11] Kandel A. Fuzzy techniques in pattern recognition.

[12] Molina L. (2002) Data mining: torturando a los datos hasta que confiesen.