MÉTODOS ICONOGRÁFICOS DE OBSERVACIÓN, EXPLORACIÓN Y COMUNICACIÓN APLICADOS A LA MINERÍA DE TEXTOS

Cesari, M.[1], Rodríguez, D. [1], Rancán, C. [1], Merlino, H. [1, 2], Britos, P.[1, 2], García-Martínez, R.[1, 2]

[1]Centro de Ingeniería del Software e Ingeniería del Conocimiento. Instituto Tecnológico de Buenos Aires
[2]Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires

{drodrigu, crancan, hmerlino, pbritos, rgm}@itba.edu.ar

1. INTRODUCCION

La lingüística computacional es la ciencia que trata de la aplicación de los métodos computacionales en el estudio del lenguaje natural (Gelbukh and Bolshakov, 1999). El objetivo más importante es la comprensión del lenguaje, es decir, la transformación del lenguaje hablado o escrito a una representación formal del conocimiento, como por ejemplo una red semántica. Algunas de estas otras áreas de investigación son procesamiento de voz, generación de texto y procesamiento de texto.

El procesamiento automático de textos es una de las áreas mas importantes dentro de esta area. El mismo considera una gran diversidad de tareas, como la separación de palabras, y tareas de minería de texto (categorización, clasificación de textos, clustering, descubrimiento de patrones, tendencias, desviaciones, etc.).

La minería de texto es la más reciente área de investigación del procesamiento de textos. Ella se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una compilación de textos, es decir, la minería de texto es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún documento textual, pero que surgen de relacionar el contenido de varios de ellos (Hearst, 1999; Kodratoff, 1999). Tiene como objetivo principal la búsqueda de conocimiento útil en enormes colecciones de documentos estructurados y no-estructurados (e-mails, actas, libros, artículos, discursos, encuestas, etc.). Los problemas a abordar pueden surgir del estudio de textos (comparación de estilos, atribución de autor, búsqueda documental, etc.) o ser de naturaleza no textual, pero cuyo tratamiento lleve a considerar ciertos textos como datos portadores de información (será el caso en psicología y sociología con las entrevistas en profundidad y tests, en politología con los discursos, programas políticos y artículos periodísticos, etc.). Entre los textos se encuentran las opiniones de respuestas abiertas de encuestas. El tratamiento de estos tipos de texto, se enriquece con la información complementaria obtenida con las respuestas al cuestionario estructurado. Una de las herramientas de la minería de texto es el "Cartografiado de Texto", que nos permite extraer unidades en los textos, enriquecer la lexicometría con los métodos de análisis multivariado y aplicar las herramientas de visualización a las tablas léxicas o volúmenes de datos lingüísticos. Estas herramientas de visualización involucran técnicas estadísticas de análisis léxico, técnicas estadísticas de exploración multivariada y técnicas de Inteligencia Artificial como mapas autoorganizados de Kohonen.

2. DESCRIPCIÓN DEL PROBLEMA

El tesoro más valioso de la raza humana es el conocimiento. Gran parte de este conocimiento existe en forma de lenguaje natural: libros, periódicos, informes técnicos, encuestas de opinión, etcétera. La posesión real de todo este conocimiento depende de nuestra habilidad para hacer ciertas operaciones con la información. Muchos datos que el investigador se ve obligado a procesar provienen de textos, para obtener datos relevantes de un texto es necesario sistematizar el conjunto de la información contenida en el mismo y para esto hace falta ciertos principios y técnicas de

análisis. La minería de texto provee de estos principios y técnicas, se enfoca en el descubrimiento de patrones interesantes y nuevos conocimientos en un conjunto de textos, es decir, su objetivo es descubrir cosas tales como tendencias, desviaciones y asociaciones entre "grandes" cantidades de información textual.
Existen grandes volúmenes de Información textual organizados en documentos (Corpus), internamente poco estructurados, esto lleva a que el análisis clásico de datos textuales no sea económico y consuma muchos recursos en especialistas y tiempo. Este tipo de procesamiento masivo de la información plantea mayor volumen de parámetros y variables. Esta situación ha motivado el desarrollo de nuevas metodologías con técnicas y paradigmas existentes, y la integración de los métodos de análisis que faciliten el proceso de exploración de datos textuales.
En este argumento se plantea la necesidad de contar una metodología que permita completamente la preparación, el tratamiento, el análisis y visualización de información apreciable de grandes volúmenes de datos textuales.
El Cartografiado de Texto, constituye una nueva estrategia de comunicación de la información aportada por la observación de un sistema estudiado y la sistematización del gran conjunto de datos textuales, de modo que la "información contenida y su estructura de dependencia", pueda representarse gráficamente y comunicarse eficazmente. El Cartografiado, permite brindar una representación de toda la estructura de la información en un sólo gráfico, aunque los datos sean numéricos, alfanuméricos o textuales y además las relaciones entre ellos, lo que permite brindar un diagnóstico a través de la imagen de los mismos, una rápida y completa comunicación y la interpretación clara de toda la información contenida en su estructura.

3. ABORDAJE DE LA SOLUCION

Para poder llevar adelante la solución al problema planteado, se seguirán los siguientes pasos:

1. Definición de un marco teórico que presente en forma sistemática la integración de las distintas técnicas estadísticas de análisis léxico existentes, técnicas estadísticas de exploración multivariada de reciente utilización y técnicas de Inteligencia Artificial como mapas autoorganizados de Kohonen aportadas por la minería de datos; y utilizarlas en el trazado de una metodología para la exploración y diagnóstico por imagen de datos textuales.
2. Comparación de herramientas lingüísticas, estadísticas, e inteligentes permiten la extracción, la comparación y el mapeo (Cartografiado) de los contenidos en textos.
3. Aplicación la metodología de Cartografiado de Texto propuesta, a Casos de Ejemplo (estudios de textos literarios, análisis de respuestas abiertas de encuestas, estudios de test psicológicos,…).

3.1. Estado de avance

3.1.1. Propuesta metodológica

3.1.1.1. Elaboración de documentos léxico métricos

- **Preparación del documento para el registro de los datos textuales.** Edición del corpus:

  Componentes posibles del corpus: *narraciones, artículos periodísticos, informes, desgrabaciones de entrevistas y grupos, respuestas libres a preguntas abiertas, y variables sociodemográficas, socioeconómicas, actitudinales, que tipifican o segmentan las entrevistas o grupos, variables que actúan como predictores - variable independiente- , del criterio -variable dependiente.*

- **Estudio de las unidades estadísticas (formas, lemas, segmentos)** Segmentación del texto en unidades.

  La segmentación del corpus textual implica diferenciar las unidades elementales: *la forma gráfica (una secuencia de letras comprendidas entre dos espacios), el lema (todos los vocablos que cuentan con una misma raíz y con significado equivalente, es decir, una familia de palabras), los segmentos repetidos (una secuencia de dos o más palabras que aparecen más de*

una vez en un corpus de datos textuales), los cuasi segmentos (palabras que aparecen en una determinada secuencia pero que presentan alguna diferencia en el género o número).

- **Estudio de la riqueza de vocabulario (frecuencia de segmentos repetidos).** Construcción del vocabulario del texto.

  Este se presenta en una tabla (Glosario) de orden léxico métrico donde se muestra el número identificatorio de cada palabra, la palabra del glosario del corpus, la frecuencia de aparición y la longitud de la unidad medida en número de caracteres.

3.1.1.2. Análisis y cartografiado

Nos permiten dos tipos de aplicaciones:

- text mining, para buscar y extraer información significativa y clasificada (sobre las diversas entidades lingüísticas);
- text mapping, para explorar gráficamente las relaciones entre temas y palabras clave;

- **Armado de la Tablas léxicas.** formar una tabla de contingencia (Respuestas*formas) o sea una "tabla léxica básica" y una tabla de contingencia (Formas*textos) o sea una "tabla léxica agregada".
- **Análisis multivariado de datos textuales.** Aplicación del [1]Análisis Factorial de Correspondencias, sobre las tablas lexicográficas o la Clasificación Automática (Clasificación jerárquica ascendente) de las formas lexicales y textos.
  - **Identificación de Los "Textos característicos",** Selección de frases enteras características de cada texto, escogidas según un cierto criterio como representantes del texto.
  - **Identificación de frases modales.** Obtención de Tipologías o grupos a partir de respuestas y de textos. Asociación de variables estructuradas, al análisis de las tablas léxicas permitiendo la clasificación según los léxicos empleados y las modalidades escogidas en las variables.
  - **Visualización de los resultados del Análisis multivariado.** Representación de la distribución del corpus lexicográfico mediante Mapas preceptúales. Utilización del Análisis de Correspondencias para la representación gráfica de la información contenida en las Tablas léxicas.
  - **Análisis discriminante textual.** Predicción de variables léxicas objetos de estudio (opiniones, actitudes, predisposiciones, perfil de imagen, etc.) a partir del texto. Aplicación del Análisis Factorial Discriminante de los métodos multivariados.
- **Aplicación de los [2]Mapas Autoorganizados de Kohonen (SOM):** Clasificación de documentos y Creación de mapas de un corpus

3.1.2. Algoritmos a utilizar

Para poder efectuar los procedimientos enunciados en el esbozo de la metodología, de forma eficiente, se ha escogido los principales algoritmos que serán expuestos:

- Codificación del [3]Corpus.
- Ordenamiento lexicográfico.
- Recorrido de un Árbol binario léxico.
- Árbol binario de Búsqueda del vocabulario de un corpus
- Árbol Binario de Búsqueda de prefijos

---

[1] La aplicación del Análisis Factorial en el campo de análisis de datos textuales, se centra, principalmente, en el Análisis Factorial de Correspondencias, algoritmo estadístico desarrollado por Jean Pau Benzécri (1973, 1976). Se trata de un método descriptivo (no explicativo) que se clasifica entre los métodos multivariables de interdependencia y permite visualizar los datos (que pueden ser cualitativos o cuantitativos) mediante la representación de una nube de puntos en un espacio de dimensiones reducidas, en función de las distancias euclidianas entre los puntos.

[2] T. Kohonen presentó en 1982 un sistema con un comportamiento semejante al del cerebro. Se trataba de un modelo de red neuronal con capacidad para formar mapas de características de manera similar a como ocurre en el cerebro.

[3] Colección completa de textos

- Árbol Binario de Búsqueda de segmentos
- Construcción implícitas de particiones
- Detección de cadenas repetidas.
- Construcción de sub espacios invariantes de la matriz de datos textuales. Análisis Factorial de Correspondencias. Análisis Factorial Discriminante.
- Clasificación jerárquica ascendente
- Concordancia de formas gráficas.
- Criterio del Valor de Test para la significación estadística en la exploración de datos.
- Clasificación y creación de Mapas autoorganizados del corpus (mapa de Kohonen)

4. FORMACIÓN DE RECURSOS HUMANOS

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando: dos tesistas de maestría de Ingeniería del Software y un tesista de grado.

5. CONCLUSIONES

Los métodos del *Cartografiado de Texto*, proporcionan herramientas extraordinarias para poder extraer la información contenida en textos. Cuando se trata de comprimir miles de palabras en unos resultados concisos, siempre hay *una simplificación que puede producir alguna deformación*. Por otra parte, como manifiesta L. Lebart, cada análisis textual es una verdadera investigación.
El objetivo principal del "Cartografiado de la información", es la construcción de un nuevo "lenguaje de la información". Se trata de realizar gráficos de amplios conjuntos de datos donde las personas, los entes, los objetos o el medio a describir se trasforman en representaciones sobre un plano. La metodología propuesta permite:

- Utilizase como una *aplicación general* que permita una *lectura fácil* de la información que contiene, ya que la regla de interpretación es la de la "proximidad de los puntos representados".
- El método algorítmico que aplica su transformación, tiene el papel de *instrumento de observación*, sistematizando los volúmenes de datos y proporcionando imágenes a partir de una realidad.
- *Utilizar las facultades de percepción humana cotidianamente utilizadas*. Sobre los gráficos se "ve" con los ojos y el misterioso análisis iconográfico que nuestro celebro hace de una imagen: las agrupaciones, oposiciones y tendencias, imposibles de discernir directamente sobre una tabla de datos, inclusos después de un examen prolongado.
- *Diagnosticar situaciones* debido a que las tablas de datos son precisamente un obstáculo para su lectura fácil y su asimilación directa; el "cartografiado de la información contenida" se ofrece mediante una panorámica excepcional, permitiendo una crítica particular de la realidad para el usuario. Las figuras dadas por los gráficos presentan constataciones, inferencias, estimaciones, entrañan conjeturas, y por esto constituyen preciosos instrumentos de análisis y comunicación simultáneamente.
- *Conocer la "realidad":* uno de los principales problemas con los que se enfrenta todo periodista, gobernante, político o investigador, es la "conceptualización" del medio en donde se desarrolla; es decir, "lograr sintetizar afirmaciones generalizables a una situación determinada". Es aquí donde precisamente el servicio propuesto tiene su máxima aportación.
- Medir ciertos aspectos intrínsecos del medio real y transformarlos a un "*espacio de información básico*" que produce un modelo simulado, que es imagen actualizada de esa realidad. En ese sentido, esto constituye principalmente el Servicio de Cartografiado.
- Permitir *exhibir aspectos que se escapan a la observación directa:* propone ir más lejos de las apariencias de los datos: "el Servicio de cartografiado de la información" establece un compromiso entre el poder explicativo y la simplicidad; cumple una función de transferencia iconográfica y su contribución más importante es hacer viva la estructura de la información y trasmitirla a todos los usuarios por igual.

- Permitir crear un *vínculo*, entre la prestación de consultoría a través de "mapas de indicadores estadísticos" con el debate social, la argumentación y justificación de las decisiones ejecutivas y la comunicación eficiente de la información al medio.

La metodología propuesta, constituye una nueva *estrategia de representación gráfica* de la información, aportada por una observación de los multiatributos de un medio o sistema estudiado y la sistematización del gran conjunto de datos aportados, de modo que la "información contenida y su estructura de dependencia", pueda representarse gráficamente y comunicarse eficazmente. Aunque se ha expuesto una *guía metodológica de análisis*, ésta no es totalmente automática, el investigador dispone de muchas opciones y tiene que tomar decisiones no excluyentes o realizar el análisis de varias formas diferentes para comparar los resultados.

6. REFERENCIAS

GELBUKH AND BOLSHAKOV (1999), Avances en Análisis Automático de Textos. Proc. Foro: Computación, de la Teoría a la Práctica. IPN, Mexico City, May 26 – 28, 1999.

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.

KODRATOFF (1999), Knowledge Discovery In Texts: A Definition And Applications, Proc. Of The 11th International Symposium On Foundations Of Intelligent Systems (ISMIS-99), 1999.