

Minería de texto

Recuperación y organización de la información

Minería de texto | Descargas

Minería de texto

La mayor parte del conocimiento humano está representado en lenguaje natural. Para poder acceder a dicho conocimiento es necesario poder contestar a estas preguntas:

- ¿Cómo buscamos la información?
- ¿Cómo comparar fuentes de información diferentes y sacar conclusiones?
- ¿Cómo manejamos los textos para, por ejemplo, traducirlos o editarlos?

El objetivo de esta página es introducir los conceptos y tareas básicas del procesamiento de textos utilizando lo que se ha venido a denominar la **Minería de texto**. La **Minería de texto** consiste en la búsqueda a partir de técnicas de aprendizaje automático de regularidades o patrones que se encuentran dentro de un texto.

Lingüística computacional

La Lingüística computacional es la ciencia que estudia la aplicación de técnicas computacionales a la comprensión del lenguaje. Esta ciencia se apoya en otras

dos, que son la inteligencia artificial y la lingüística. Su principal meta es la comprensión automática de textos mediante una representación formal del mismo.

La minería de texto es una de las muchas ramas de la lingüística computacional.

Técnicas de minería de texto

La minería de texto es el proceso encargado del descubrimiento de conocimiento que no existe en el texto, pero que surge al relacionar el contenido de varios textos.

La minería de texto se divide en dos etapas que son el pre-procesamiento y una etapa de descubrimiento. Dependiendo del tipo de métodos utilizados en la etapa de pre-procesamiento se genera una representación distinta del contenido del texto.

A continuación se describen algunos de estos métodos.

Técnicas clásicas

Las técnicas clásicas en minería de texto se estructuran básicamente en tres etapas:

- Etapa de preprocesamiento: Es el proceso mediante el cual los textos se transforman en algún tipo de representación estructurada que facilite su análisis

- Etapa de representación: La representación depende de la técnica de prerocesamiento utilizada y determinarán cuál será el algoritmo de descubrimiento a utilizar.
- Etapa de descubrimiento: Son algoritmos que a partir de una representación estructurada de la información, son capaces de descubrir regularidades en los textos.

Como se puede observar, todas las etapas están muy interrelacionadas, así pues, la primera etapa condiciona el descubrimiento de los patrones que la minería de texto puede realizar.

Las técnicas más usadas en minería de texto son los vectores de temas que muestran el nivel temático del texto, la secuencia de palabras que permite descubrir patrones en el texto y las tablas de datos que permite descubrir interrelaciones entre entidades.

Grafos conceptuales

Los grafos computacionales son una técnica de representación en minería de texto muy potente. Con las anteriores técnicas no se podían responder a preguntas como: ¿Cuál es la opinión mayoritaria de los españoles sobre la guerra? (Consensos) ¿Ha habido un cambio de actitud del gobierno de España ante la guerra? (Tendencias) ¿Alguien opina de forma distinta a la mayoría? (Desviaciones).

La técnica de grafos conceptuales aportan mayor semántica. Un grafo

conceptual es un grafo bipartito que tiene dos tipos de nodos, conceptos y relaciones conceptuales. En la figura se puede observar la pinta que tiene el grafo conceptual de la frase: Bush critica a Zapatero.



Los grafos se comparan utilizando conocimiento del dominio como diccionarios de sinónimos y jerarquías de conceptos. Se realiza una operación de intersección entre dos grafos para dar un resumen de ambos y a dicho resumen se le valora con una puntuación que indica el grado de similitud entre ambos textos.

La agrupación de dos o más grafos permite descubrir la estructura oculta de la colección de textos. Para agrupar los grafos se pueden utilizar técnicas de agrupamiento como las estrategias colaborativas, el agrupamiento en k medias o Comweb que no se explicarán aquí. Si se desea más información acerca de estas técnicas consultar:

- Machine Learning: An Artificial Intelligence Approach Vol I-IV [Michalski and Teccuci, 1994]
- Machine Learning de Tom Mitchell [Mitchell, 1997]

Todas estas técnicas se basan en suministrar a los algoritmos un conjunto de ejemplos a partir de los cuales generan las agrupaciones.

La programación lógica inductiva ha sido aplicada en la minería de texto en múltiples ocasiones. La ventaja que posee es su capacidad de representación basada en lógica de segundo orden, que permite generalizar conceptos y descubrir definiciones de conceptos de forma automática. Normalmente se utiliza Prolog para programar las herramientas. Esta técnica permite introducir conocimiento a priori del dominio en forma de definiciones mediante predicados relacionados.

Esta técnica requiere no sólo de un conjunto de entrenamiento con ejemplos sino también de las relaciones ya descubiertas por el diseñador y basadas en cláusulas de Horn. Por ejemplo si queremos descubrir la definición del término $\text{abuelo}(x,y)$ a partir de los conceptos de $\text{padre}(x,y)$ y $\text{madre}(x,y)$ descritos de forma extensional, es decir mediante aquellas parejas de valores que cumplan la relación de ser padre y madre y ejemplos de relaciones de abuelos. $\text{padre} = \{\text{pedro,juan}\}, \{\text{juan, luis}\}, \{\text{enrique,maria}\}$ $\text{madre} = \{\text{maria,elisa}\}$ $\text{abuelo} = \{\text{pedro,luis}\}, \{\text{enrique,elisa}\}$ Generaría la siguiente definición $\text{abuelo}(x,z) = \text{padre}(x,y) \text{ and } \text{madre}(y,z)$ or $\text{abuelo}(x,z) = \text{padre}(x,y) \text{ and } \text{padre}(y,z)$

Se utiliza como algoritmo FOIL y algunos de sus derivados. Se puede consultar más información sobre FOIL y la programación lógica inductiva en: Aprendizaje Automático [Borrajó and Boticario and Isasi, 2006].

Programación genética

Basándose en la representación utilizada en la programación lógica inductiva se

peuden utilizar herramientas de programación genética para el tratamiento de la minería de texto.

La programación genética es un método de generación de programas para ordenador de forma automática con insperación evolutiva, partiendo de programas muy simples que mediante el cruce de unos con otros y procesos de mutación aleatoria, permiten generar programas más y más aptos en la realización de la tarea que se le describe. La aptitud de los programas se mide de forma numérica mediante una función denominada de fitness. Algunas extensiones de la programación genética permiten describir nuevas primitivas a partir de las primitivas inicialmente descritas.

Básicamente, la idea consiste en introducir como primitivas las relaciones expresadas como clausulas de Horn y utilizar un sistema que permita utilizar las denoinadas ADFs, que no son más que evoluciones paralelas de otras primitivas que se pueden utilizar en la definición principal, para generar definiciones muy resumidas de los conceptos. La función de fitness deberá medir el número de ejemplos que se cubren con la definición de cada uno de los individuos generados en cada generación, ponderandose con la longitud de la definición para guiar al algoritmo a soluciones sencillas y descubrimiento de conceptos intermedios. Esta técnica ha tenido problemas con definiciones recursivas debido a que la programación genética tiene problemas de eficiencia con primitivas recursivas. Si se conoce que la naturaleza de la definición a encontrar es recursiva, probablemente sea mejor solución las técnicas anteriormente descritas.