Tarea No. 1: Fundamentos de Data Mining

Profesor: John Atkinson

Fecha Publicación: 15 de Octubre del 2015
Fecha Entrega: 12 de Noviembre del 2015

1 Objetivo

El objetivo de esta tarea es introducir métodos basicos de minería de datos con el fin de descubrir patrones relevantes extrayendo **reglas de asociación**.

2 El Problema

Se desea detectar relaciones relevantes a partir del análisis de documentación en Español desde artículos científicos contenidos en la base de datos científica **scielo** (`http://www.scielo.cl/scielo.php`). Para esto, Ud debe utilizar una técnica de minería de datos para la generación de **Reglas de Asociación**. Es decir, debe implementar el algoritmo correspondiente para poder extraer reglas de asociación desde los datos mencionados. Para ello, considere los siguientes aspectos:

1. Las reglas de asociación generalmente operan sobre bases de datos relacionales. Para ser aplicadas sobre bases de datos textuales, Ud. debe establecer una representación básica adecuada para transformar el texto en un conjunto de atributos. En este esquema, se sugiere primero la extracción las keywords de los textos, y luego aplicar un método de generación de reglas de asociación. Note entonces que deberá llevar a cabo un preprocesamiento de los documentos, por ejemplo, para eliminar palabras inútiles (preposiciones, artículos), filtrado y posteriormente extraer las keywords. Una lista de palabras irrelevantes y que por tanto deberían ser descartadas del análisis (*stopwords*) se encuentra disponible en: `http://www.inf.udec.cl/ atkinson/stoplist.txt`

2. Un problema con las medidas de *Support* y *Confidence* utilizadas en algoritmos de generación de reglas de asociación es que debido a su naturaleza estadística, en muchas ocasiones sesgan el análisis produciendo patrones irrelevantes. Para enfrentar parcialmente dicho problema, Ud. debe utilizar una medida adicional que permita evaluar más precisamente el grado de novedad o sorpresividad de una regla generada. Específicamente, se medirá la **Disimilaridad Semántica** entre el antecedente y consecuente de cada regla. La medida se podría interpretar como *"entre más distantes sean las partes de la regla, más interesante podría ser la relación"*.

   Para realizar el cálculo anterior, Ud. debe utilizar una técnica de análisis semántico basada en bases de datos textuales (corpus) conocida como *Latent Semantic Analysis* (LSA), la cual lleva a cabo por medio de métodos de descomposición matricial (ej. SVD), un proecso de reducción dimensional. Luego, LSA es capaz de generar vectores de que representan "sentidos"' de las palabras y que posteriormente permitiran comparar la similaridad entre conceptos. Una vez que se dispone de estos vectores para cada palabra de interés, la similaridad entre dos palabras puede calcularse fácilmente, por ejemplo, utilizando una función del tipo *coseno* entre vectores.

   En resumen, LSA se debe entrenar inicialmente con un conjunto de documentos que se le proporcionan como entrada. Posteriormente, el método entregará los vectores relevantes a cada concepto.

   Información básica acerca de LSA puede encontrarse en la dirección: `http://lsa.colorado.edu/`. Una buena versión pública de LSA se denomina *InfoMap*, la cual puede obtenerse del sitio:

   `http://infomap-nlp.sourceforge.net/`.

En base a lo anterior, el trabajo puede enfocarlo dividiéndolo en las siguientes etapas:

1. Entrenar LSA (vía InfoMap) utilizando un set original de al menos 1000 documentos de **scielo** para obtener los vectores de similaridad de todas las palabras relevantes.

2. Preprocesar y filtrar los documentos originales para obtener la representación de keywords de cada uno de ellos (cada set de keywords de un documento es análogo a los atributos de una tupla de una base de datos relacional). Puede utilizar herramientas como perl, php, phyton, etc para realizar este filtrado.

3. Aplicar el algoritmo APRIORI para obtener las mejores reglas de asociación a partir del paso (2) (asuma un umbral de **Support** de 70%).

4. Evaluar la disimilaridad semántica de las reglas obtenidas utilizando los vectores del paso (1) (asuma un umbral de disimilaridad del 90%). La salida final será el set de reglas de asociación que tienen el mayor grado de disimilaridad. Para calcular esta medida, asuma que una regla tiene la forma: **IF A THEN B**

   Utilizando los vectores del paso (1), la imilaridad de cada regla obtenida podría estimarse como: $sim = coseno(vector(A), vector(B))$, entonces la disimilaridad se computa como $1 - sim$ (el coseno varía entre -1 y 1, siendo 1 el valor máximo cuando el ángulo entre dos vectores es cero!!).

Una vez implementados los algoritmos correspondientes, Usted deberá ejecutar los programas para generar las reglas filtradas según los criterios previamente descritos. Para ello, se deberán realizar los siguientes experimentos:

- Mantener constantes el número de entrada a todo el proceso, los umbrales, correr el programa, y registrar las reglas que se generan.
- Incrementar el número de documentos de entrada a todo el proceso (en múltiplos de 10), correr el programa, y registrar las reglas que se generan (mantenga constante los umbrales).
- Mantener fijo el número de documentos, incrementar el umbral de support, correr el programa, y registrar las reglas que se generan (mantenga constante el umbral de disimilaridad).
- Incrementar en pasos de 2% el umbral de disimilaridad, y registrar las reglas que se generan (mantenga constante el número de documentos y el umbral de support).

Finalmente, Ud. deberá entregar lo siguiente:

1. El código fuente de los programas implementados (no incluye InfoMap).
2. Un archivo (texto) que contenga las reglas de asociación finalmente filtradas utilizando la medida de disimilaridad y su correspondiente valor, para cada uno de los experimentos (puede entregar un archivo diferente por cada experimento que realizó). Debe especificar claramente los parámetros que utilizó en los diferentes "runs".

La tarea será evaluada principalmente considerando la calidad y relevancia de las reglas de asociación obtenidas.