Fundamentos de Data Mining
Tarea 1

Alumno: Diego Andrés Palma Sánchez
Profesor: John Atkinson

Noviembre de 2015

1. Introducción

En el presente informe se describe la metodología utilizada para detectar relaciones relevantes en una base de datos textual. En particular, los documentos escogidos fueron extraídos de la base de datos científica *scielo*[1]. Se resolvió el problema utilizando una técnica de minería de datos para la generación de *Reglas de asociación*[1]. Para encontrar las reglas de asociación se utilizó el algoritmo APRIORI descrito en [1]. Posteriormente, para evaluar la calidad de las reglas obtenidas se utilizó una medida de *Disimilaridad Semántica* entre el antecedente y el consecuente de cada regla. El supuesto utilizado es *"entre más distantes sean las partes de la regla, más interesante podría ser la relación"*. Para realizar este cálculo, se utilizó una técnica de análisis semántico basada en bases de datos textuales (corpus) conocida como *Latent Semantic Analysis* (LSA)[2]. Finalmente se hicieron experimentos variando la cantidad de documentos para minar las reglas, los parámetros requeridos para aplicar el algoritmo apriori y el umbral de disimilaridad semántica.

2. Minando Reglas de Asociación desde Scielo

Para experimentar se obtuvieron documentos desde *Scielo* que fuesen de tópicos diferentes. En particular se consideraron 3 temas, unos más específicos y otros más generales con el fin de observar si se generan distintos tipos de reglas. Los tópicos considerados, y la cantidad de documentos extraídos fueron:

- Cáncer: 240 documentos
- Agua: 859 documentos
- Innovación: 900 documentos

Cada corpus fue procesado de manera de eliminar signos de puntuación, palabras irrelevantes (stopwords), y se utilizó un modelo de bolsa de palabras para entrenar LSA [2]. Finalmente se obtuvo una representación de *keywords* de cada documento para posteriormente aplicar el algoritmo APRIORI y encontrar diferentes reglas de asociación. Con fines de simplificar el problema, se consideraron reglas del tipo.

[1] http://www.scielo.cl/scielo.php

IF A THEN B

Para poder interpretar las reglas obtenidas, se obtuvo un mapeo de las palabras a los documentos donde aparecen. Finalmente, se realizaron los siguientes experimentos:

1. Mantener constantes el número documentos de entrada a todo el proceso, los umbrales, y registrar las reglas generadas.
2. Incrementar el número de documentos de entrada a todo el proceso (en múltiplos de 10), mantener constantes los umbrales y registrar las reglas generadas.
3. Mantener fijo el número de documentos, incrementar el umbral de support, y registrar las reglas que se generan.
4. Incrementar en pasos de 2 % el umbral de similaridad, manteniendo constantes el support y la cantidad de documentos de entrada, y luego registrar las reglas generadas.

Entre los archivos entregados se encuentra el detalle de las implementaciones, tanto de los algoritmos como la ejecución de los experimentos; de igual forma el registro de las reglas obtenidas junto a los parámetros utilizados. El resto de esta sección se dividirá en subsecciones comentando los resultados y discutiendo las reglas obtenidas. El detalle de todas las reglas encontradas y los parámetros para cada experimento puede encontrarse en la carpeta "*runs*" de cada tópico.

2.1. Cáncer

2.1.1. Experimento 1

No se encontraron reglas que cumplieran con la condición del umbral de disimilaridad de 90 % entre el antecedente y el consecuente. Se consideró trabajar con 240 documentos y un support mínimo del 70 %. Las reglas más relevantes encontradas fueron:

Cáncer → Pacientes
Cáncer → Estudio
Pacientes → Años

La regla de asociación entre cáncer y pacientes no aporta conocimiento relevante. Los documentos en general tratan de casos clínicos, o estudios de la evolución de un cáncer (biliar, pulmonar, por ejemplo) en pacientes de distintas edades. De ahí se explica la relación encontrada entre cáncer y estudio, y pacientes. La relación entre pacientes y años puede ocurrir debido a las siguientes razones: Hay documentos que hablan sobre el seguimiento de un paciente con cáncer y la evolución de la enfermedad con el pasar de los años. Por otro lado, también en algunos casos clínicos se menciona que pacientes de una cierta edad comienzan a mostrar indicios de algún cáncer. En este contexto, se puede inferir que en Chile las personas que van a hacerse controles médicos generalmente se encuentran en una edad adulta; varios estudios señalaban que eran personas mayores de 35 años. También, que a partir de esa edad, hay una gran cantidad de casos de personas que padecen cáncer.

2.1.2. Experimento 2

Variando la cantidad de documentos en el corpus, se consideró extraer reglas de asociación utilizando 40, 80, 120, 160, 200, 240 documentos (Support 70 %, y umbral de disimilaridad del 90 %). En los primeros casos se encontraron algunas reglas nuevas, pero poco relevantes, tales como:

Casos → Pacientes
Cáncer → Forma

La primera regla se debe a que una gran cantidad de los documentos dentro del corpus son casos clínicos que hablan sobre pacientes con cáncer. En el segundo caso, la relación entre forma y cáncer puede deberse que se hacen seguimientos a personas con cáncer y se monitorean los tumores, su evolución, y otros indicios que aparecen en pacientes con dicha enfermedad. La forma es importante de acuerdo al tipo de cáncer que la persona padezca, en los documentos se habla de cáncer a la vesícula, cáncer pulmonar, cáncer pancréatico, entre otros.

2.1.3. Experimento 3

En este experimento aparecen las reglas con mayor disimilaridad semántica entre el antecedente y el consecuente (entre 60 % y 75 %). Lo que se hizo fue variar el support utilizando los siguientes valores: 30 %, 50 %, 70 %, 90 %.

La mayor cantidad de reglas fue encontrada con el support del 30 %, esto se debe a que hay menores restricciones para los ítemsets. Se encontraron reglas como:

Pacientes → Población
Células → Pacientes
Estudio → Mortalidad
Cáncer → Meses
Tumores → Pacientes
Tratamiento → Diagnóstico

La primera regla no es relevante considerando que en el texto se hace referencia al aumento de diagnósticos de cáncer en la población mundial durante los últimos años. En el caso de la regla que relaciona células con pacientes, se explica porque en los casos clínicos se hacen seguimientos a pacientes que están siguiendo un tratamiento de quimioterapia para controlar la evolución de células cancerosas.

La relación entre estudio y mortalidad da un indicio en que los estudios recientes en salud, han mostrado que el cáncer ha tenido una tasa de mortalidad mayor en los últimos años.

La relación entre cáncer y meses no es tan clara. Algunos documentos explican la evolución del cáncer en meses de tratamiento. Otros discuten acerca de la expectativa de vida de personas que sufren esta enfermedad. Otros simplemente hablan de estudios que duraron meses.

Finalmente la relación entre tumores y pacientes es trivial. Los pacientes con cáncer tienden a tener tumores que dan indicios de su enfermedad. A partir de estos antecedentes, se diagnostica la enfermedad y se comienza un tratamiento en caso de no ser demasiado tarde, según los estudios.

2.1.4. Experimento 4

Como se vió en el experimento 1, para un support del 70 % utilizando todos los documentos del corpus, se obtienen reglas con disimilaridad entre antecedente y consecuente que varían desde 20 % hasta 50 %. Por tanto, se varío el umbral de disimilaridad desde 38 % hasta 46 % en pasos de a dos (para más detalle ver los archivos de salida de los runs). Se encontró que a medida que aumenta el umbral requerido, aparecen menos reglas. Las reglas que aparecen son similares a las encontradas previamente.

2.2. Agua

2.2.1. Experimento 1

Para el primer experimento no se encontraron reglas que superaran el umbral de disimilaridad del 90 %. Sin embargo, se registraron las reglas de los itemsets que aparecieron con la configuración establecida (Support del 90 %). Varias de las reglas encontradas no son realmente conocimiento nuevo, por ejemplo reglas como:

$$\text{Agua} \rightarrow \text{Análisis}$$
$$\text{Agua} \rightarrow \text{Estudio}$$

No aportan conocimiento nuevo, debido a que los documentos de las cuales fueron extraídas hablan de diferentes estudios relacionados con el agua (como por ejemplo en agricultura, en ciencias de la tierra y ecosistemas). Sumado a esto, un análisis de datos es lo que se realiza en varios estudios. La disimilaridad semántica de los antecedentes con los consecuentes está cercana a 0.46 en ambas reglas.

Otras reglas encontradas por ejemplo son:

$$\text{Agua} \rightarrow \text{Mayor}$$
$$\text{Agua} \rightarrow \text{Universidad}$$

Tampoco aportan conocimiento, pues son palabras que aparecen prácticamente en todos los documentos extraídos; universidades son las que realizan estudios y la palabra mayor hace referencia a resultados obtenidos, o bien a la relevancia que tendría el estudio especificado.

En este experimento no se encontraron reglas interesantes.

2.2.2. Experimento 2

Variando la cantidad de documentos se observa un comportamiento similar a lo ocurrido con el tópico "Cáncer", es decir, aumenta la cantidad de reglas encontradas utilizando menos documentos. La razón de ello es que, en este caso, al ser menos el total, los valores de support encontrados para cada ítemset son mayores. Este comportamiento se debe al corpus en sí y no a una "propiedad" del algoritmo, pues, como se mostrará más adelante, en otros tópicos ocurre un comportamiento diferente. A pesar de lo mencionado, no se observan reglas interesantes.

2.2.3. Experimento 3

Se consideraron los siguientes niveles de support mínimo: 30 %, 50 %, 70 %, 90 %. Se observa que cuando se reduce el umbral del support, se encuentran más reglas. La justificación de esto es simple estadística, si se reducen las restricciones para frecuencia de aparición de un ítem, entonces éste se considerará como ítem frecuente. Las reglas con mayor disimilaridad semántica (sobre el 65 %) se encontraron considerando un support de un 30 %. Entre estas reglas destacan:

$$\text{Temperatura} \rightarrow \text{Agua}$$
$$\text{Norte} \rightarrow \text{Agua}$$
$$\text{Campo} \rightarrow \text{Agua}$$
$$\text{Ingeniería} \rightarrow \text{Agua}$$

La primera regla se da, debido a que cierta fracción de los documentos hablan acerca de estudios térmicos del agua, tanto en el área agrícola y de vegetación, como en áreas de ingeniería

tales como extracción de Fe, o estudios relacionados al monitoreo de aguas superficiales y subterráneas (con fines de realizar estudios ambientales). La relación explica que existe una gama de técnicas basadas en temperatura tales como: termografía, o técnicas de extracción basadas en estudio del calor latente de suelos y termodinámica.

La asociación de norte y agua se le da una interpretación que tiene que ver con el contexto minero en Chile. Una parte de los documentos habla sobre la actividad minera en el norte de Chile y cómo han sido afectadas las napas subterraneas a través de la extracción de agua para procesos mineros. Estos documentos están relacionados a estudios ambientales y problemas sociales y ambientales que están causando hoy en día las mineras que operan al norte del país.

La asociación campo con agua es ambigüa. Algunos documentos hacen referencia al concepto de campo de investigación. Otros documentos hablan de un proceso de solidificación y estabilización de componentes químicos en matrices sólidas de hormigón, en este contexto campo hace referencia a campos vectoriales de velocidad, relacionados con el proceso de solidificación.

Ingeniería y agua es una relación que no aporta conocimiento nuevo si se considera el corpus analizado. La razón de ello es que una cantidad no menor de documentos hacen referencia al uso del agua en procesos de ingeniería: Ingeniería agrícola, técnicas de estabilización de componentes químicos en hormigones, técnicas de monitoreo de aguas subterráneas, etc.

2.2.4. Experimento 4

Como se mencionó anteriormente, no se obtuvo ninguna regla cuyo umbral de disimilaridad entre el antecedente y el consecuente fuese mayor a 90 %. Se probó con umbrales entre 40 % y 50 %, manteniendo un support de 70 % y los 859 documentos. Se observó que a medida que el umbral de disimilaridad aumenta, la cantidad de reglas encontradas es menor. La calidad de las reglas es similar a las encontradas en experimentos anteriores.

2.3. Innovación

2.3.1. Experimento 1

Al igual que en los casos anteriores, no se encontraron reglas que superaran el umbral de disimilaridad del 90 %. Sin embargo las "mejores" reglas en términos de esta métrica, se encuentran cerca del 60 %, valor superior al obtenido en los documentos anteriores. Algunas de las reglas obtenidas, que podrían ser interesantes:

$$\text{Desarrollo} \rightarrow \text{Información}$$
$$\text{Innovación} \rightarrow \text{Investigación}$$
$$\text{Investigación} \rightarrow \text{Proceso}$$

La regla desarrollo e información aparece debido a que varios documentos hablan sobre el desarrollo de procesos industriales, manufactura, y similares mediante el uso de tecnologías de información. En particular, esta regla entrega un indicio que se está aplicando un uso innovador de tecnologías de información para el desarrollo industrial. De igual forma, se relaciona innovación con investigación por las nuevas tecnologías que están apareciendo en el ámbito industrial. La relación entre investigación y proceso es difusa. Algunos documentos hablan de las mejoras en procesos industriales debido a investigaciones recientes. Otros hablan de investigación como un proceso científico ligado a la innovación. Cabe destacar que el corpus cuenta con un amplio temario, por lo que las relaciones son un tanto genéricas.

2.3.2. Experimento 2

El tema escogido "innovación" es amplio, puesto que las recientes investigaciones científicas intentan innovar en diferentes áreas. No se encontraron reglas relevantes que fuesen diferentes a las encontradas en el experimento anterior. Sin embargo, al aumentar la cantidad de documentos se encontró que aumentaba la cantidad de reglas encontradas.

2.3.3. Experimento 3

Para este corpus, el experimento se realizó utilizando support mínimos de 65 %, 70 %, 75 % y 80 %. Disminuyendo el support mínimo aparecieron reglas nuevas con una disimilaridad semántica entre el antecedente y el consecuente de aproximadamente 70 %. Estas reglas son del tipo:

$$\text{Conocimiento} \rightarrow \text{Investigación}$$
$$\text{Innovación} \rightarrow \text{Conocimiento}$$

El resto de las reglas encontradas son similares a las vistas en experimentos previos y no aportan conocimiento nuevo relevante. La regla de asociación que relaciona conocimiento con investigación no es clara. Hay documentos que hablan de salud pública en el contexto social, que relacionan investigaciones innovadoras para reducir la desigualdad en la salud. Se menciona que los factores determinantes que crean la brecha entre la salud pública y privada sean de conocimiento público. Por otro lado, otros documentos que hablan de salud, hablan sobre investigaciones para el tratamiento de enfermedades como Alzheimer, u otras enfermedades relacionadas con la mente, las que entre sus efectos, pueden causar pérdida de conocimiento. Considerando esto, la regla que relaciona innovación con conocimiento viene desde los mismos documentos.

2.3.4. Experimento 4

Al reducir el umbral de disimilaridad y variarlo desde 38 % hasta 50 % , se observa que mientras aparecen reglas nuevas que cumplen con el requerimiento mínimo. Sin embargo, no aparecen reglas que no se hayan visto antes.

3. Otros Experimentos

En la sección previa se utilizó un modelo de bolsa de palabras, eliminando palabras irrelevantes (stopwords). Para la representación de cada documento como ítemset se escogieron las palabras que aparecieran más de una vez en cada documento. Sin embargo, esta no es una representación de las keywords o palabras clave que considere todo el corpus, por lo que se realizó otro experimento extrayendo keywords utilizando la medida estadística *term frequency - inverse document frequency*, la cual intenta reflejar la importancia que tiene una palabra en un documento dentro de un corpus. Luego, se aplicó el algoritmo APRIORI con un support mínimo del 70 %, sin encontrar ninguna regla. Inspeccionando el conjunto de datos, se observa que no hay "palabras claves" que aparezcan en un 70 % de los documentos. Se probó reducir el umbral de support, y aumentar la cantidad de "palabras claves", aún así, sólo comenzaron a aparecer reglas una vez el support fue tan bajo como 5 %. Esto ocurrió en todos los corpus utilizados. De igual forma, se encontraron reglas triviales que no aportan conocimiento nuevo, tales como:

$$\text{Buenos} \rightarrow \text{Aires}$$
$$\text{Habana} \rightarrow \text{Cuba}$$
$$\text{Caudal} \rightarrow \text{Flujo}$$

Otras reglas que aparecieron tienen una mayor relevancia:

$$\text{Cáncer} \rightarrow \text{Virus}$$

Mapeando el antecedente y el consecuente a los documentos donde aparecen, esta relación se debe a que son casos clínicos de personas con un historial médico de enfermedades atribuidas a víruses, que posteriormente fueron afectadas por cáncer. Sin embargo, la cantidad de documentos donde se observa esta asociación es pequeña en comparación al total de documentos.

Para reducir la variabilidad de las palabras también se probó realizar stemming. Sin embargo, no cambia el hecho que al extraer keywrods con la técnica tf-idf, no es posible encontrar reglas con un support alto.

4. Conclusiones

No es trivial extraer reglas de asociación desde una base de datos textual. La razón es que si se considera como ítem las palabras, dependerá del contexto en que aparezca y el significado cambia si la variabilidad de los tópicos dentro del corpus es amplia.

Otro punto importante es qué medidas utilizar. La medida de disimilaridad con la que se trabajó está fuertemente ligada al corpus de donde se obtuvieron las reglas, de igual forma el support.

Las reglas más novedosas se encontraron al reducir el support. Sin embargo, existe un tradeoff con el tiempo y espacio requerido para aplicar el algoritmo APRIORI. Si se reduce mucho el support mínimo, podría volverse poco práctico el uso de este algoritmo, debido a la gran cantidad de ítemes posibles (palabras).

Para afrontar el problema de la variabilidad de las palabras, se puede utilizar técnicas como lemmatizar o stemming. Ello permitiría reducir la cantidad de ítemes posibles, y aplicar el algoritmo con umbrales de support más bajos; de esta forma se podrían encontrar reglas más interesantes.

Referencias

[1] Jiawei Han, Micheline Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[2] T. Landauer, S. Dumais, "Introduction to Latent Semantic Analysis", in: Discourse Procesess 25, pp. 259-284, 1997.