Tema 6
Actividad 1

Autor
Juan José Méndez Torrero

Usar la librería SPMF para aplicar un algoritmo de patrones secuenciales

Introducción

Para la realización de esta actividad, se ha seleccionado el conjunto de datos BIBLE, el cual se puede encontrar junto con la documentación del programa SPMF. Junto con esto, se ha elegido utilizar el algoritmo PrefixSpan para realizar el análisis de los patrones secuenciales frecuentes.

El conjunto de datos seleccionado para esta actividad es una conversión de la biblia a una base de datos secuencial, donde cada palabra es un ítem, con lo que, para realizar un análisis más coherente, se ha decidido descargar la versión con los valores reales de cada ítem dentro del conjunto de datos.

Algoritmo PrefixSpan

Este algoritmo de patrones secuenciales nos permite encontrar patrones secuenciales frecuentes dentro de un conjunto de datos, basándose en el crecimiento de patrones (*Pattern-Growth)*. Para su configuración, se ha decidido utilizar un soporte mínimo del 10%, ya que un soporte mayor nos devolvería una cantidad muy pequeña de patrones frecuentes.

Una vez ejecutado el algoritmo sobre el conjunto de datos, como se puede observar en la Figura 1, podemos ver que se ha demorado alrededor de 553 milisegundos, y que ha encontrado un total de 341 patrones secuenciales frecuentes.

```
============= PREFIXSPAN 0.99-2016 - STATISTICS =============
Total time ~ 553 ms
Frequent sequences count : 341
Max memory (mb) : 352.85247802734375
minsup = 2390 sequences.
Pattern count : 341
===================================================

Post-processing to show result in terms of string values.
Post-processing completed.
```

Figura 1

La Figura 2 muestra algunos de los patrones secuenciales frecuentes que se han encontrado tras ejecutar el algoritmo PrefixSpan sobre el conjunto de datos seleccionado.

```
the -1 the -1 of -1 the -1 of -1 the -1 #SUP: 2657
the -1 the -1 of -1 and -1 #SUP: 3932
the -1 the -1 of -1 and -1 the -1 #SUP: 2816
the -1 the -1 of -1 of -1 #SUP: 3886
the -1 the -1 of -1 of -1 the -1 #SUP: 2838
the -1 he -1 #SUP: 2834
the -1 to -1 #SUP: 4611
the -1 to -1 the -1 #SUP: 2963
the -1 lord -1 #SUP: 5365
the -1 lord -1 the -1 #SUP: 3084
the -1 lord -1 and -1 #SUP: 2796
the -1 lord -1 of -1 #SUP: 2555
the -1 be -1 #SUP: 2662
the -1 unto -1 #SUP: 3123
the -1 that -1 #SUP: 4686
the -1 that -1 the -1 #SUP: 2989
the -1 and -1 #SUP: 10879
the -1 and -1 the -1 #SUP: 7379
the -1 and -1 the -1 the -1 #SUP: 4632
the -1 and -1 the -1 the -1 the -1 #SUP: 2912
the -1 and -1 the -1 the -1 of -1 #SUP: 2721
the -1 and -1 the -1 and -1 #SUP: 3769
```

Figura 2

Como se puede observar, se han encontrado patrones secuenciales bastante interesantes, como puede ser el patrón secuencial *{"the", "Lord"}* con un total de 5365 apariciones.

Conclusiones

Si utilizáramos un soporte mínimo mucho menor como configuración para este algoritmo, sería bastante costoso, computacionalmente, encontrar patrones secuenciales frecuentes. Aún así, de los resultados se puede observar que una de las palabras más utilizadas dentro de la biblia en su versión inglesa es *the*, y que, la palabra L*ord* aparece un total de 5774 veces dentro de la biblia. Además, la palabra Lord seguida de las preposiciones *the, and* y *of* aparecen como mínimo 2694 veces, como se muestra en la Figura 3.

```
lord -1 #SUP: 5774
lord -1 the -1 #SUP: 3256
lord -1 and -1 #SUP: 2965
lord -1 of -1 #SUP: 2694
```

Figura 3