UNIVERSIDAD DE CONCEPCIÓN
FACULTAD DE INGENIERÍA

Tópicos en manejos de grandes volúmenes de datos
Docente: Cecilia Hernández
Autores: Bastián Gómez
Jesús Gómez
Rodolfo Vergara

Proyecto 1: Estimadores de Cardinalidad

Para la realización del proyecto 1 se implementaron 2 códigos que se utilizan para estimar la cardinalidad de un conjunto de datos, en este caso, 31-mers obtenidos de 2 documentos con genomas, los códigos implementados fueron PCSA y HLL que por sus siglas en inglés significan Probabilistic Counting with Stochastic Averaging e Hyper Log Log respectivamente, estos se realizaron basados en la documentación y pseudocódigos vistos en la asignatura.

PCSA trabaja usando un bitmap, primero se procesan los k-mer con 1 hash para determinar en que "bucket" del bitmap son asignados y otro para independizar los resultados pues no por 2 k-mers tener elementos parecidos deben ser asignados al mismo bucket pues en caso de tener un universo con muchos elementos similares esto podría causar clustering por lo que usamos hashings para evitar esto. Posterior al procesado, se introduce en el bucket correspondiente un número que corresponde a la posición del primer '0' de derecha a izquierda y finalmente cuando todos los elementos sean procesados, de cada bucket se obtiene la posición del primer 0 de derecha a izquierda y se promedia, por ultimo se realiza un ajuste de error y se retorna la cardinalidad estimada.

Este algoritmo en cuanto a espacio usa en nuestra implementación únicamente 4M bytes donde M es el tamaño del vector que se usa como bitmap, por su parte el error está dado por la formula: e= +-0.78/sqrt(M)

HLL por su parte usa un procedimiento similar, procesando los datos con 1 hash para determinar mediante los primeros bits su ubicación en una tabla que usará como sketch y con los bits restantes se obtiene la posición del primer 0 de izquierda a derecha y se almacena en el bucket asignado, sin embargo si un elemento posterior obtiene un número menor al calcular la posición, se actualiza manteniendo siempre el mayor, finalmente se obtiene la media armónica de los elementos almacenados en la tabla y se realiza un ajuste de error en base al estimado obtenido para finalmente retornar la cardinalidad estimada.

Análogo a PCSA este algoritmo usará 4M bytes donde M es el tamaño del vector usado como tabla para el sketch, M estará dado por 2^b donde b serán el número de bits utilizados en cada bucket y

el error estará dado por e= +-1.04/sqrt(M). Cabe mencionar que se implementaron dentro de HLL las operaciones de unión e intersección de conjuntos.

En cuanto al análisis experimental, se realizaron pruebas cambiando el M en PCSA y el b en HLL bajo 2 documentos distintos obteniendo los siguientes resultados:

Para el documento 1 el cual tiene un peso estimado de 3GB se obtuvo:

M	PCSA
64	1497473318
128	1489386099
256	1677825639
512	1637433762
1024	1630796960

b	HLL
8	1822585473
9	1748330723
10	1794107355
11	1739569656
12	1758431877

Para el documento 2 con un peso estimado de 2GB:

M	PCSA
64	852652798
128	981564717
256	1016729954
512	1054581740
1024	1088672001

b	HLL
8	1108930436
9	1086206871
10	1126061669
11	1089339150
12	1078989014