Recuperación de Textos

Índices de Firmas

Una alternativa a los índices de inversión de textos son los archivos de firmas (s*ignature files*).

Una firma digital de un archivo es cualquier secuencia de bytes de tamaño fijo, es decir sin relación con el tamaño del archivo, que se obtenga mediante una función que procese su contenido (en archivos de texto, sus términos). La razón de que se llame firma digital o sólo firma, es que su propósito principal es el de validación de integridad; de este modo, un CRC (control de redundancia cíclica) de un archivo podría considerarse una firma: permite validar la integridad del archivo luego de su transmisión o recuperación de un dispositivo de almacenamiento, que pudo verse afectada ya sea por problemas técnicos o por un virus.

Para recuperación de textos, la idea es que la firma de un documento sirva para verificar si un término cualquiera puede estar contenido en el mismo. Para esto se determina el tamaño en bits B que tendrá la firma, y se escoge una serie de funciones de dispersión (*hash*) que, aplicadas a cualquier término, devuelvan un entero entre 0 y B-1, determinando los bits que se pondrán en uno en la firma. Es inevitable que más de un término de un documento establezca el mismo bit de su firma en 1, por lo que así como las firmas no son infalibles para validar integridad, tampoco lo son para asegurar si un término está contenido en un documento: todo documento que tenga bits en 1 en las mismas posiciones que en la firma de un término de consulta se debe recuperar e inspeccionar en busca del término antes de agregarlo como resultado de la consulta...

Archivos de Firmas (Bitstring Signature Files)

Un índice de firmas es un archivo secuencial con un registro con la firma correspondiente a cada documento.

Por ejemplo, para los documentos de avisos clasificados de autos y camionetas, se obtiene firmas de 24 bits usando tres funciones de dispersión para cada término:

Término	h1	h2	h3
vend	21	10	16
auto	3	9	18
camioneta	9	5	3
usad	9	10	20
excelente	9	20	0
oferta	17	7	23
segund	22	5	18
mano	1	3	23
ocación	15	14	5
permut	3	12	22
caminoeta	3	11	7

Número de Documento	Contenido	Firma
1	Vendo autos y camionetas	`00010100 01100000 10100100`

Número de Documento	Contenido	Firma
2	Autos usados	00010000 01100000 00101000
3	Excelente oferta de camionetas	10010101 01000000 01001001
4	Autos de segunda mano	01010100 01000000 00100011
5	Autos y camionetas de ocación	00010100 01000011 00100000
6	Permuto auto por caminoeta	00010001 01010000 00100010
7	Autos y más autos	00001000 01000000 00100000

Para consultar, por ejemplo, documentos que contengan los términos de “camionetas usadas”, se construye la firma de la consulta fc=00010100 01100000 0001000 y luego se recupera la firma de cada documento fd, y si (fd and fc)=fc se considera el documento como candidato para la respuesta, aunque hay que validar si contiene o no todos los términos de la consulta.

Número de Documento	Contenido	Firma
1	Vendo autos y camionetas	00010100 01100000 10100100
2	Autos usados	00010000 01100000 00101000
3	Excelente oferta de camionetas	10010101 01000000 01001001
4	Autos de segunda mano	01010100 01000000 00100011
5	Autos y camionetas de ocación	00010100 01000011 00100000
6	Permuto auto por caminoeta	00010001 01010000 00100010
7	Autos y más autos	00001000 01000000 00100000

Como no hay ningún documento candidato se debe intentar otra consulta. Por ejemplo si sólo se buscara documentos para el término camioneta, los documentos candidatos son 1, 3, 4, 5, debiendo descartarse el 4.

Archivos de Porciones de Firmas (Bitslice Signature Files)

Para resolver una consulta con un archivo de firmas, es necesario recorrer todo el archivo. Para reducir el costo en accesos a disco se puede almacenar las firmas transpuestas: cada registro i del archivo de transposición representa los bits en la i-ésima posición de la firma de todos los documentos; de manera que para determinar los documentos que puedan contener un término t, sólo sería necesario recuperar tantos registros cuantos unos tenga la firma de t (a lo sumo la cantidad de funciones de dispersión que se empleen para construir las firmas).

Los registros de porciones de firma del cado de ejemplo serían:

Número de Porción	Porción
0	0010000
1	0001000
2	0000000

Número de Porción	Porción
3	1111110
4	0000001
5	1011100
6	0000000
7	0010010
8	0000000
9	1111111
10	1100000
11	0000010
12	0000000
13	0000000
14	0000100
15	0000100
16	1000000
17	0010000
18	1101111
19	0000000
20	0110000
21	1000000
22	0001010
23	0011000

Entonces, para determinar qué documentos son candidatos a contener el término “camioneta”, como la firma del término tiene unos en las posiciones 3, 5 y 9, se recupera las porciones de firma correspondientes a estas posiciones `1111110, 1011100 y 1111111`, y la conjunción de las porciones resulta con unos en los documentos candidatos: `1011100` (los documentos cuyos números corresponden a las posiciones de los bits en 1: 1, 3, 4 y 5).

Archivos de Porciones de Firmas de Grupos de Documentos (Blocked Signature Files)

Un problema particular de los archivos de porciones de firma es el tamaño de las porciones, que tienen un bit por documento; entonces si la cantidad de documentos indexados es muy grande, los registros del archivo de porciones resultan gigantes. Este problema se puede atenuar agrupando

documentos en bloques, de manera que cada bit de cada porción de firma corresponda a B documentos, siendo B el factor de agrupación. La longitud de las porciones se divide entonces por B, pero para mantener baja la densidad de las porciones (pocos unos) se debe aumentar el tamaño de las firmas. Para reducir la probabilidad de que un bloque contenga todos los términos de una consulta pero ningún documento del bloque quede en el resultado, los números de documento deben mapearse distinto en cada porción; esto es, los documentos deben asignarse a distintos bloques en cada porción. El número n de mapeos distintos de documentos debe ser submúltiplo del tamaño en bits tf de la firma de los documentos; cada mapeo se aplica a tf/n porciones.

Si bien el caso de ejemplo no es adecuado para graficar esta técnica dada la poca cantidad de documentos, supóngase que se decide agruparlos de a dos: las porciones se reducirían a 4 bits, y habría que mapear los documentos en grupos distintos en cada porción: por ejemplo, si el documento d estuviera en la posición i de una porción completa, podría asignarse la posición (i+1) % d en la porción siguiente. La tabla del ejemplo anterior quedaría con cada porción rotada un bit a derecha y luego los pares de bits 1-2, 3-4 y 5-6 reemplazados por la disyunción entre ambos. Pero para obtener resultados similares que con los documentos sin agrupar, habría que duplicar el tamaño de las firmas.