CS 6140: Data Mining Assignment 5

Frequent Items

Yash Gangrade (u1143811), MS in Computing 27th February 2019

Contents

1 Streaming Algorithms

1.1 Run the Misra-Gries Algorithm on streams S1 and S2

Ans The output from the Misra-Gries Algorithm is summarized in the table below:

S1	**S1**	**S2**	**S2**
Label	**Counter**	**Label**	**Counter**
a	355715	a	899790
c	475715	p	1
u	1	e	0
b	625715	x	1
m	1	u	0
i	1	c	607161
z	1	b	406116
g	1	o	0
r	0	g	1

Next, to find the objects which occur more than 20% of the times, we find the minimum counter of such objects. For S1, we have $m = 3,000,000$ and for S2, we have $m = 4,000,000$ and k = 10. Also, the minimum counter is defined by

$$f_q = \frac{20}{100} \cdot m - \frac{m}{k} = \frac{m}{10}$$

Therefore, minimum frequency counters for S1 and S2 are:

$$f_1 = 300000, f2 = 400000$$

So, the objects/labels with counter greater than equal to f_1, f_2 are the objects which *might* occur more than 20% of the time.

- Stream S1 -> a, b, c
- Stream S2 -> a, b, c

Secondly, to find the objects which must occur more than 20% of the time, we find the objects with the counter greater than $\frac{20}{100} * m$ so we have, $f_1^{'} = 600000, f_2^{'} = 800000$. Therefore, the objects with counter greater than equal to $f_1^{'}, f_2^{'}$ are the objects which *must* occur more than 20% of the time.

- Stream S1 -> b
- Stream S2 -> a

1.2 Build a Count-Min Sketch for streams S1 and S2

Ans The estimated counts for the objects are as follows:

Object	**Count S1**	**Count S2**
a	510000	1108483
b	826759	677971
c	724034	942737

Now, we will analyze it for each of the streams. The objects which might occur for more than 20% of the times will have counter at least 600000 for S1 and 800000 for S2. Therefore, such objects are:

- Stream S1 -> b, c
- Stream S2 -> a, c

1.3 Change in implementation if Twitter dataset is used

Ans If the object of the stream is a word the basic principle would remain same, we will now have each object as a word and we will keep a count of frequency of a word rather than a alphabet. The first difference would be in the number of possible objects. One other change we need is that we would need to convert the stream of words into a list of words which would be iterable over each word instead of character as before, or we can also convert the words into n-grams and then apply heavy filters. It might assist in reducing the memory required by putting a bound on n. In case of words over English alphabets its far more than the number of alphabets. Hence n will be a very large number. The size of each object will be $k_i = \log n$. Thus, to count the number of objects we have seen, we require $\log m$ space.
In the Misra-Gries algorithm, in order to achieve a maximum error of ϵm we need $1/\epsilon$ counter. In case of words the ϵ would be very small and hence number of counters k will be large so the space required will be more.
In case of Count- Min sketch the total space requirement is $t(k \log m + \log n)$. This will also be higher in case of objects as words.

1.4 One advantage of Count-Min Sketch over the Misra-Gries algorithm

Ans There are several advantages of Count-Min Sketch over the Misra-Gries algorithm. Two of them are described below:

- One advantage is that Count-Min Sketch uses several multiple different hash functions for one character/word object and then takes the minimum of those hash functions leading it to be more accurate, while the Misra-Gries only has one vector, the Count-Min Sketch uses a Matrix.
- One of the advantages is in case of Turnstile Model. In the turnstile model each update is of the form $< i, c >$, so that a_i is incremented by some (maybe negative) integer c. In the "strict turnstile" model, a_i at any time cannot be negative. The Count-Min has the same guarantees in the turnstile model, but Misra-Gries does not have the same gurantees.