Estimativa de Cardinalidade da Interseção de Conjuntos Utilizando as Estruturas *MinHash* e *HyperLogLog*

Juan Pedro Alves Lopes, Paulo Eustáquio Duarte Pinto, Fabiano de Souza Oliveira
IME/DICC, Universidade do Estado do Rio de Janeiro (UERJ)

Resumo

Apresentamos uma técnica para estimativa da cardinalidade da interseção de conjuntos. Isto é, Dados conjuntos $A_1, A_2, \ldots, A_n$, o objetivo é estimar $|A_1 \cap A_2 \cap \cdots \cap A_n|$. Usando as estruturas **MinHash** e **HyperLogLog** é possível obter, com complexidade de memória sublinear, uma aproximação (ϵ, δ), isto é, com erro relativo menor que ϵ com uma probabilidade fixa $1 - \delta$. Esta técnica mostra-se superior a outras anteriormente descritas por ter erro relativo apenas à cardinalidade interseção dos conjuntos, ou seja, independente da cardinalidade dos conjuntos originais.

Introdução

Estimar a cardinalidade da interseção entre múltiplos conjuntos é um problema importante para diversas aplicações. Embora haja algoritmos determinísticos triviais para calcular este valor, normalmente eles exigem ter os conjuntos acessíveis em memória ou a execução de múltiplas operações de entrada e saída para manipulá-los em disco.

Muitas vezes, especialmente em aplicações que geram uma grande quantidade de dados (ex.: conjunto de logs de visitas a grandes portais), os conjuntos de interesse não cabem na memória de um único computador ou estão distribuídos geograficamente em múltiplos servidores, tornando os algoritmos clássicos custosos demais para serem utilizados na prática.

Neste trabalho, apresentamos uma técnica paralelizável que combina as estruturas de dados *MinHash* [1] e *HyperLogLog* [2] para permitir uma estimativa da cardinalidade da interseção entre múltiplos conjuntos.

Referências

[1] A. Z. Broder, "On the resemblance and containment of documents," in *Compression and Complexity of Sequences 1997. Proceedings*, pp. 21–29, IEEE, 1997.

[2] P. Flajolet, É. Fusy, O. Gandouet, and F. Meunier, "Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm," *DMTCS Proceedings*, vol. 1, no. 1, 2008.

[3] P. Flajolet and G. N. Martin, "Probabilistic counting algorithms for data base applications," *Journal of computer and system sciences*, vol. 31, no. 2, pp. 182–209, 1985.

CNMAC2016
XXXVI CONGRESSO NACIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL

Descrição da técnica

A técnica consiste em computar as estruturas *MinHash* e *HyperLogLog* para todos os conjuntos. É trivial computar o *HyperLogLog* da união dos conjuntos a partir das estruturas computadas, portanto apenas manipulando a definição do índice de *Jaccard*, estima-se a cardinalidade da seguinte forma:

$$|A_1 \cap A_2 \cap \cdots \cap A_n| = \underbrace{J(A_1, A_2, \ldots, A_n)}_{\text{estimado por } \textit{MinHash}} \times \underbrace{|A_1 \cup A_2 \cup \cdots \cup A_n|}_{\text{estimado por } \textit{HyperLogLog}}.$$

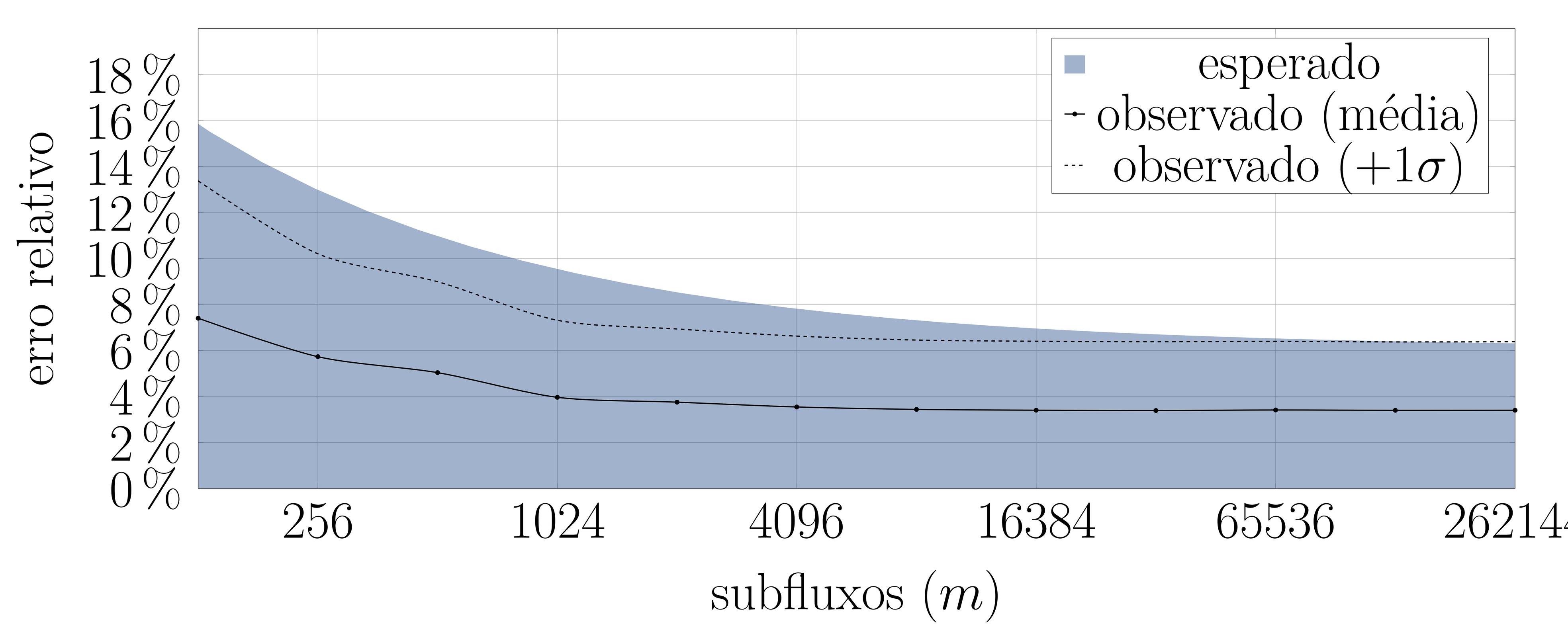
O erro da técnica pode ser derivado a partir dos erros relativos de ambas as estruturas. Sejam ϵ_M e ϵ_H os erros relativos na estimativa de *MinHash* e *HyperLogLog*, e ϵ o erro relativo da estimativa da interseção, isto é,

$$|A_1 \cap A_2 \cap \cdots \cap A_n| \times (1+\epsilon) = J(A_1, A_2, \ldots, A_n) \times (1+\epsilon_M) \times |A_1 \cup A_2 \cup \cdots \cup A_n| \times (1+\epsilon_H).$$

Logo,

$$\epsilon = \epsilon_M + \epsilon_H + \epsilon_M \epsilon_H.$$

Na figura abaixo, o resultado de um experimento variando o número de subfluxos (m) da estrutura *HyperLogLog*, com número de elementos na assinatura *MinHash* fixo ($k = 2048$).



MinHash

Permite estimar a semelhança entre conjuntos através da aproximação do coeficiente $J(A, B)$ de similaridade de *Jaccard* [1], definido para dois conjuntos A e B, como:

$$J(A_1, A_2, \ldots, A_n) = \frac{|A_1 \cap A_2 \cap \cdots \cap A_n|}{|A_1 \cup A_2 \cup \cdots \cup A_n|}.$$

A estrutura baseia-se na observação de que, dada uma função de hash h, sendo $h_{\min}(A) = min_{x \in A} h(x)$, então $\Pr[h_{\min}(A) = h_{\min}(B)] = J(A, B)$, denotando assim um estimador não-enviesado do índice de Jaccard. É possível mostrar que é preciso utilizar k funções de hash, de forma que o erro da estimativa seja menor que ϵ, com confiança $1 - \delta$, satisfazendo:

$$k \geq \frac{2 + \epsilon}{\epsilon^2} \times \ln(2/\delta).$$

HyperLogLog

Permite estimar o número de elementos distintos em um fluxo de dados, utilizando memória sublinear [3]. A estrutura baseia-se na observação do padrão de bits do *hash* dos elementos do conjunto. Note que a probabilidade do hash iniciar com um certo número de bits zero é dada por

$$\Pr[h(x) = 0^{p-1}1\ldots] = 2^{-p}$$

O algoritmo consiste em particionar o fluxo em m subfluxos disjuntos e, para cada um, observar o maior prefixo $0^{p-1}1$, indicativo de que a cardinalidade naquele fluxo é, com alta probabilidade, da ordem de 2^p. Quanto maior o valor de m, mais precisa se torna a estimativa.

Mostra-se que o erro relativo padrão do estimador é igual a $1.04/\sqrt{m}$, com uma distribuição aproximadamente gaussiana.