Trabalho apresentado no CNMAC, Gramado - RS, 2016.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

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

Juan Pedro Alves Lopes[1]
Programa de Pós-Graduação em Ciências Computacionais, IME/UERJ, Rio de Janeiro, RJ
Paulo Eustáquio Duarte Pinto[2]
Departamento de Informática e Ciência da Computação, IME/UERJ, Rio de Janeiro, RJ
Fabiano de Souza Oliveira[3]
Departamento de Informática e Ciência da Computação, IME/UERJ, Rio de Janeiro, RJ

1 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 devido a aparição de aplicações na Internet 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.

2 Descrição da técnica

Dados conjuntos $A_1, A_2, \ldots, A_n$, o objetivo é estimar $|A_1 \cap A_2 \cap \cdots \cap A_n|$. Para tanto, aproveitamos os resultados obtidos pelo uso de duas estruturas: *MinHash* e *HyperLogLog*.

A estrutura *MinHash* [1] permite estimar o índice de *Jaccard* $J(A_1, A_2, \ldots, A_n)$ entre múltiplos conjuntos, utilizado para medir o grau de similaridade entre tais conjuntos. Isto é,

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

[1]me@juanlopes.net; parcialmente financiado pela FAPERJ.
[2]pauloedp@ime.uerj.br; parcialmente financiado pela FAPERJ.
[3]fabiano.oliveira@ime.uerj.br; parcialmente financiado pela FAPERJ.

MinHash baseia-se na observação dos menores valores resultantes da aplicação de uma função de hash a cada elemento do conjunto. A comparação entre os menores valores de cada conjunto define variáveis de Bernoulli com probabilidade $p = J(A_1, A_2, \ldots, A_n)$.

A estrutura *HyperLogLog* [2] permite estimar a cardinalidade de conjuntos utilizando memória sublinear. Ela baseia-se na observação do padrão de bits do hash de cada elemento para definir um conjunto independente de estimadores de cardinalidade.

Uma característica importante de ambas as estruturas é a possibilidade de derivar a estrutura relativa à união entre conjuntos computados anteriormente, sem precisar recorrer aos elementos originais de cada conjunto.

Assim, 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,

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

Este resultado mostra que, ao contrário de técnicas baseadas no princípio de inclusão-exclusão, o erro é relativo apenas à cardinalidade da interseção, ou seja, não é sensível às cardinalidades dos conjuntos originais.

3 Conclusão

Neste trabalho, introduzimos uma técnica para estimar a cardinalidade da interseção entre conjuntos que requer apenas uma representação compacta (sublinear) dos conjuntos em memória, cuja computação pode ser facilmente paralelizada.

Como resultado importante, mostramos que o erro relativo desta técnica não depende das cardinalidades dos conjuntos originais, o que constitui uma vantagem, visto que estas podem ser ordens de grandeza maiores que aquela da interseção.

Referências

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

[2] P. Flajolet, E. Fusy, O. Gandouet, e F. Meunier, Hyperloglog: the analysis of a near-optimal cardinality estimation algorithm, *DMTCS Proceedings*, (1), 2008.