Leitura do artigo

ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction

(Keshav Santhanam et al)

Conceitos

Late interaction: queries e documentos são representados em multi-vetores. A relevância entre eles é estimada com algum cálculo da interação entre esses dois conjuntos de vetores.

MaxSim: Operador usado no ColBERT para esse cálculo:

$$S_{q,d} = \sum_{i}^{N} max_{j=1}^{M}(Q_i D_j^T)$$

Query (Q): N vetores
Documento (D): M vetores

0. Inicia o score = 0
1. Para cada vetor da query, calcula a similaridade de cosseno entre esse vetor e todos os vetores dos documentos.
2. Pega o máximo de (1) e acumula o score
3. Repete os passos 1 e 2 para todos os vetores da query

Contribuições do artigo

Não li o artigo do primeiro ColBERT, mas fiquei com a impressão que trata-se do mesmo modelo. O que difere é a forma de armazenagem dos pesos dos documentos.

No ColBERTv2 em vez de guardar os vetores completos, o espaço é dividido em clusters e guarda-se apenas o índice do cluster mais próximo e a diferença entre o vetor e o centro do cluster (resíduo).

Contribuições do artigo

Com vetores de 128 dimensões, com 2 bytes por dimensão o ColBERTv1 representa cada vetor com 256 bytes.

O ColBERTv2 faz uma redução de dimensionalidade dos vetores e depois separa o espaço em clusters. Cada cluster é indexado com 4 bytes (sendo possível indexar até 2^{32} = 4.294.967.296 clusters). E cada dimensão do vetor de resíduo é representada por 1 ou 2 bits, o que significa no total 16 ou 32 bytes (total de bits por dimensão * 128 dimensões / 8 bits/byte).

No total, cada vetor passa a ser representado por 20 ou 36 bytes.

Dúvidas

O que muda do ColBERTv1 para o ColBERTv2 é só a representação dos dados mesmos?

O que ele chama de redução de dimensionalidade é a mesma redução de dimensionalidade no contexto de PCA/SVD ou é uma espécie de quantização dos resíduos em poucos níveis (similar a modulação de pulso)?

Na proposta de índice invertido deles o que é a chave e o que são os valores?

Obrigado

Leandro Carísio

carisio@gmail.com