L2AP: Fast Cosine Similarity Search With Prefix L-2 Norm Bounds

David C. Anastasiu and George Karypis
Department of Computer Science and Engineering University of Minnesota

IEEE/2014

Paulo Henrique da Silva

Prof. Dr. Wellington Santos Martins

{paulohsilva, wellington}@inf.ufg.br

Similarity Search

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

26 September, 2018

Seminários

Sumary

- Introduction
- Related Works
- Data and Methods
- Conclusions
- Future Works

Introduction

- Context
  - The importance of the text mining has increased significantly in recent years due to the enormous growth of information generated in digital media. Today there is an immense volume of data generated in the most varied formats, that is, in a destructured form and that needs to be structured, analyzed and indexed to extract relevant information

Introduction

- Problem
  - Given a dataset *R*, a query *S*, a function *sim(x,y)* and a threshold *t*, similarity search finds all objects in the dataset with a similarity value of at least *t* when compared to query *S*
  - High dimensional sparse datasets
  - Calculated by a distance function - Cosine similarity
  - Inverted index
  - All Pairs Similarity Search - $O(n^2)$

Introduction

- Motivation
  - Relevance in many real-world applications
    - Near-duplicate document detection
    - Clustering
    - Collaborative fitering
    - Recomender systems (e.g., books or movies)

Introduction

- Objectives
  - Propose new filtering techniques
  - Prune candidates using value-based bounds (ℓ^2-norm)
  - Reduce inverted index
  - Framework candidate generation and verification

Related Works

- Chaudhuri et al – Formalizy prefix-filtering
- Bayardo et al – Developed additional pruning strategies
- Xiao et al – Introduce postional filtering
- Ribeiro and Härder – Minimize the size of the inverted index
- Lee et al – Introduce length filtering and suffix filtering

Data and Methods

- Datasets
  - Textual datasets (corpus)
  - Represented as *tf-idf* weighted vectors
  - Evaluated L2AP and L2AP-approx methods against state-of-the-art

Dataset	n	m	nnz
RCV1	804414	43001	61e6
WikiWords500k	494244	343622	197e6
WikiWords100k	100528	339944	79e6
TwitterLinks	146170	143469	200e6
WikiLinks	1815914	1648879	44e6
OrkutLinks	3072626	3072441	223e6

L2AP: Fast cosine similarity search with prefix L-2 norm bounds

Data and Methods

- L-2 Norm Bounds (Variable-Length Prefix Filter)
  - Require at least 2 common tokens in the prefixes
  - Prefix size for a set r: $prefix_2(r) = \left\lfloor (1 - \tau_c^2)|r| \right\rfloor + 1$
  - Considering $\tau = 0.65$, $prefix_2(r) = 5$ and $prefix_2(s) = 3$
  - The pairs is pruned

ℓ-Prefix Filter ($\ell = 2$)

Sandes et al. 2017

Data and Methods

- TF-IDF
  - TF – measure of how important the term is to the document
  - IDF – measure the importance of the term in the corpus

$$\text{tfidf} = \text{tf} \cdot \text{idf}$$

$$\text{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$\text{idf}_i = \log \frac{|D|}{|\{d : d \ni t_i\}|}$$

www.google.com

- Binary → count → weight matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

|V|

Each document is now represented by a real-valued vector of tf-idf weights ∈ R|V|

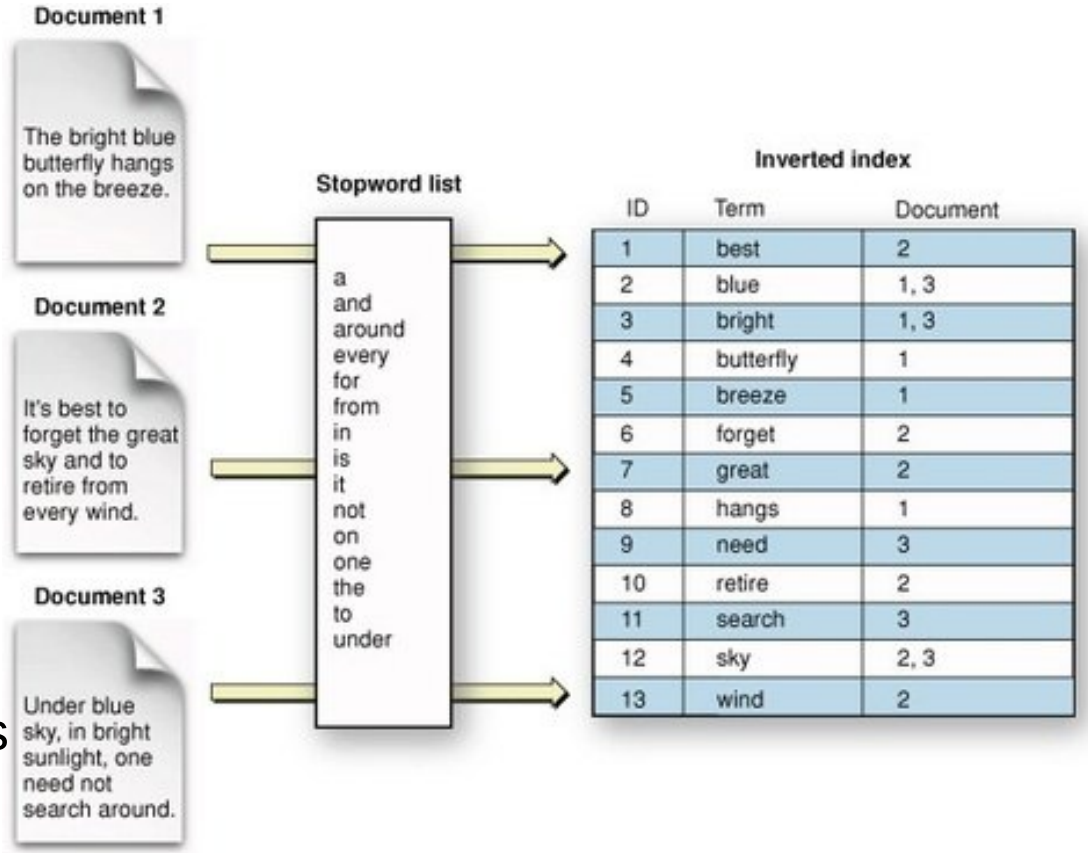
Data and Methods

- Cosine Similarity
  - Measure of similarity between two vectors
  - Terms are axes in the vector space
  - Documents are the vectors
  - High dimensionality and sparsity

$$COS(r,s) = SIM(r,s) = \frac{\sum_{i=1}^{n} r_i.s_i}{\sqrt{\sum_{i=1}^{n} r_i^2} \cdot \sqrt{\sum_{i=1}^{n} s_i^2}}$$

Data and Methods

- Inverted index
  - Links a word to a document
  - Fetch the document
  - Tokenize document
  - Remove stop words
  - Stem to root word
  - Record document Ids
  - Merge/store the words



https://www.quora.com

Conclusions

- Introduction of new filtering strategies
- Drastic reductions in the inverted index size
- Significant speedups over all exact baselines methods – AllPairs, MMJoin
- BayesLSH-Lite approximate candidate pruning cannot improve significantly over the exact pruning strategies

Future Works

- Evaluate the efficiency of ℓ^2-norm with others similarity function (Dice and Tanimoto)
- Apply this method with related problems such as Nearest Neighbor or k-Nearest Neighbor search
- Scaling up the number of threads and processors

Questions?

L2AP: Fast Cosine Similarity Search With Prefix L-2 Norm Bounds

David C. Anastasiu and George Karypis
Department of Computer Science and Engineering University of Minnesota

IEEE/2014

Paulo Henrique da Silva

Prof. Dr. Wellington Santos Martins

{paulohsilva, wellington}@inf.ufg.br

Similarity Search

UFG
UNIVERSIDADE FEDERAL DE GOIÁS

26 September, 2018

Seminários