Microsoft

Large-scale Retrieval – Theory to Practice

Kiran Shiragur
Senior researcher
Microsoft Research India

MSR India, MSR Redmond, Microsoft Product teams and Academic collaborators (MIT, Cornell, UW-Madison)

Traditional retrieval problem

Popular approach: Dense retrieval

Objects
(documents, ads, etc)

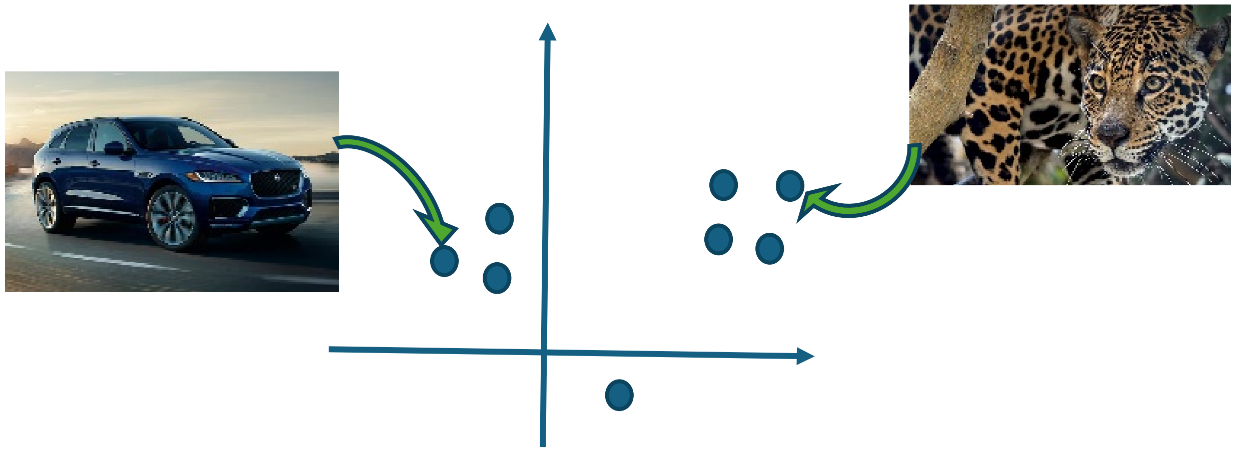
Feature Vectors

Vectors in a high dimensional space

Embedded Vectors

Data structure to support efficient vector search

Focus



DiskANN Impact

Bing

Ads, MSAN
500m$ Incremental Annual Revenue

Web Index
2% Fidelity Gain
25% Machine Savings

M365 Indices
40% COGS savings

First-Party Users

Pinecone "PGA is based off Microsoft's FreshDiskANN"

cassandra "It is inspired by DiskANN, a disk-backed ANN"

"Timescale Vector speeds up ANN search ... inspired by the DiskANN algorithm"

External Adaptations

Microsoft SQL Server

Azure

Windows 11

Customer Facing Offerings

NeurIPS'23 Competition Track:
Big-ANN

Supported by Microsoft Pinecone aws zilliz

RESEARCH-ARTICLE

Worst-case performance of popular approximate nearest neighbor search implementations: guarantees and limitations

AUTHORs: Piotr Indyk, Haike Xu Authors Info & Claims

NIPS '23: Proceedings of the 37th International Conference on Neural Information Processing Systems
Article No.: 2891, Pages 66239 - 66256

Published: 10 December 2023 Publication History

Academic Impact

What next?

1) Better theoretical understanding of vector search algorithms

2) Entering a new phase in vector search with new variants

New directions

Focus

Diverse search

Embeddings are evolving (Multi-vector)

Filter search, Online search, Distributed search, Page search

Vector search data structure to support new requirements?

Motivating questions

1) Better theoretical understanding of DiskANN

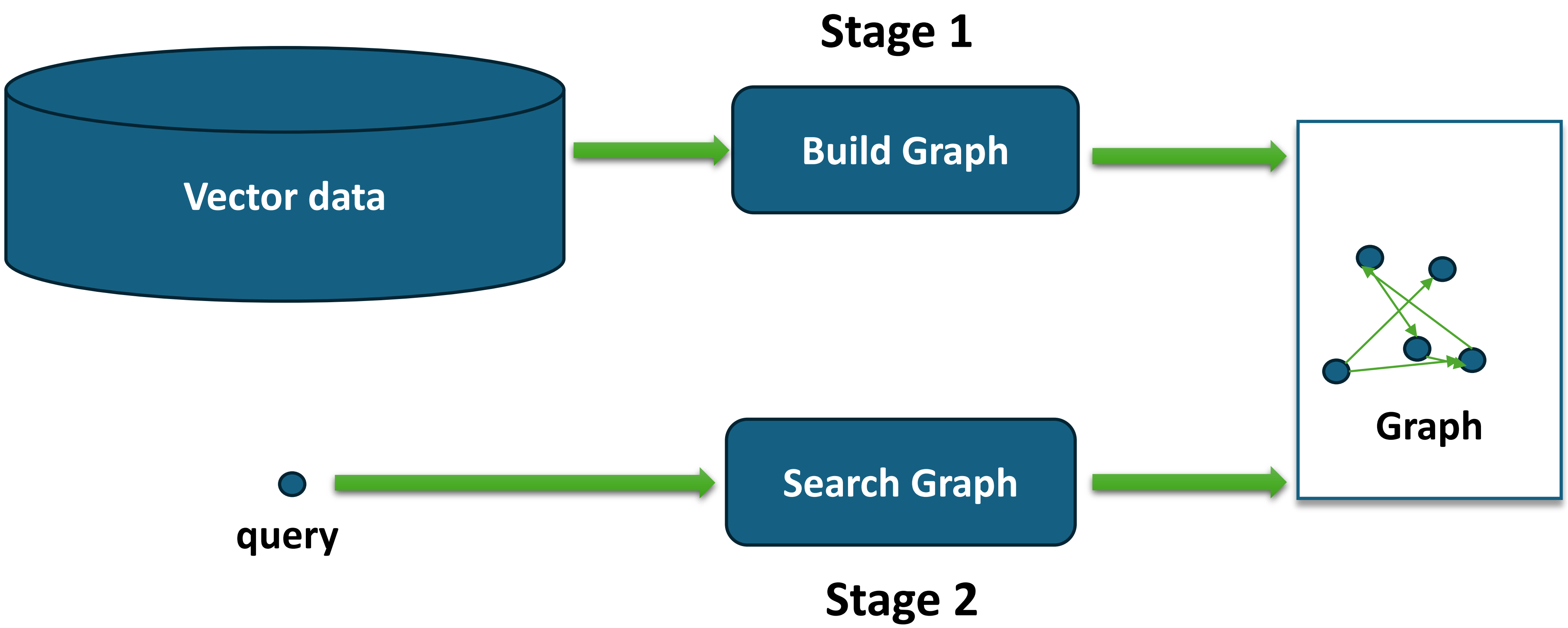
2) Modify DiskANN to accommodate diversity?

Improved analysis of DiskANN

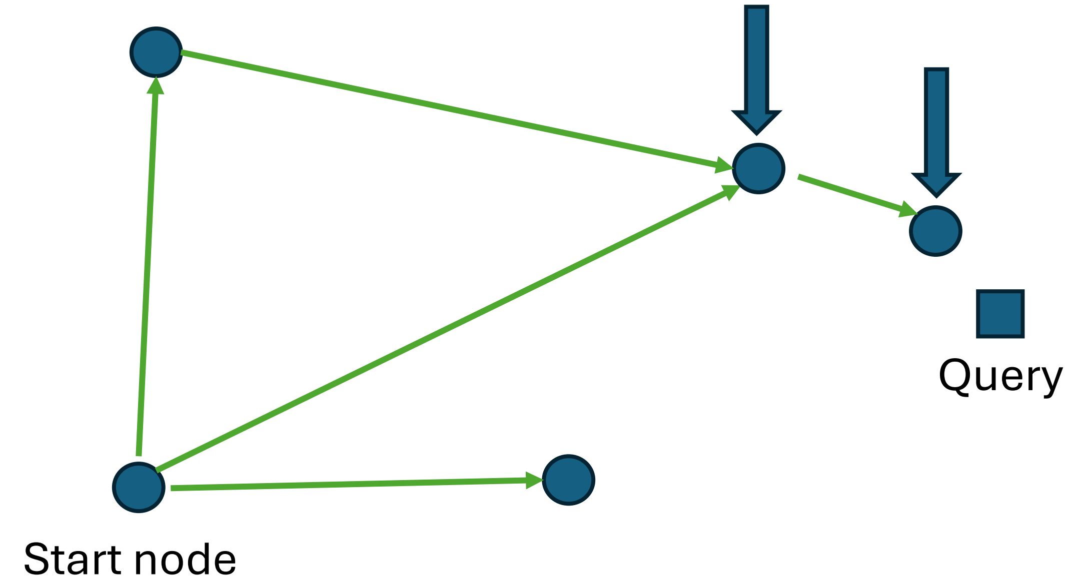
Sort Before You Prune: Improved Worst-Case Guarantees of the DiskANN Family of Graphs (**ICML**, Under Review)

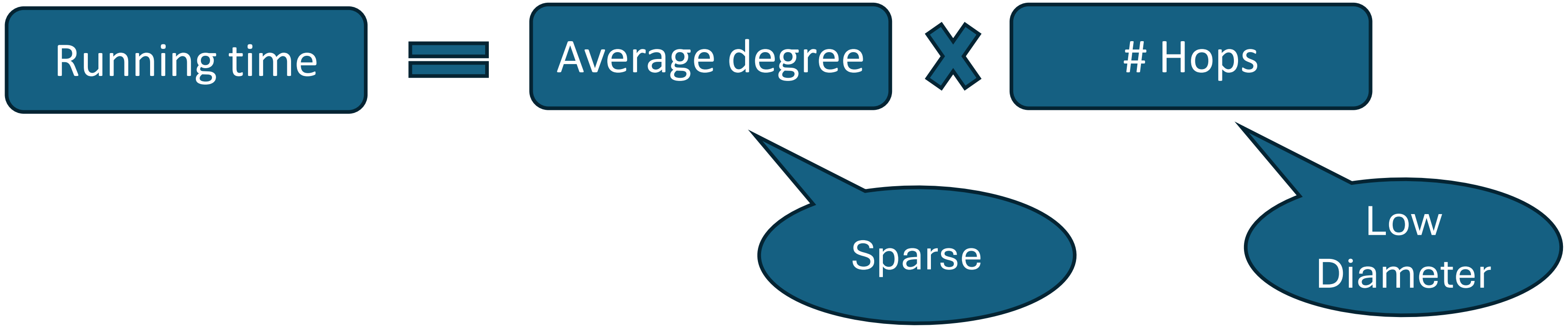
Siddharth Gollapudi, Ravishankar Krishnaswamy, KS, Harsh Wardhan

Graph based approach (DiskANN)



Greedy search



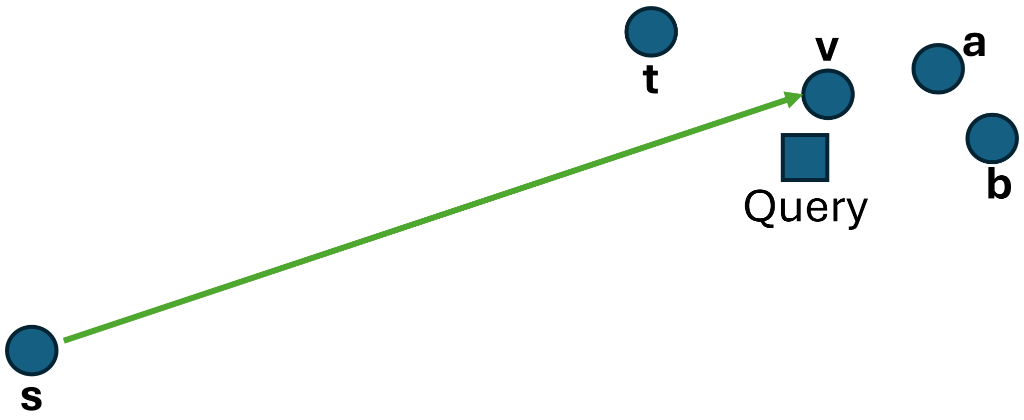


Graph – Desired Properties

Low average degree

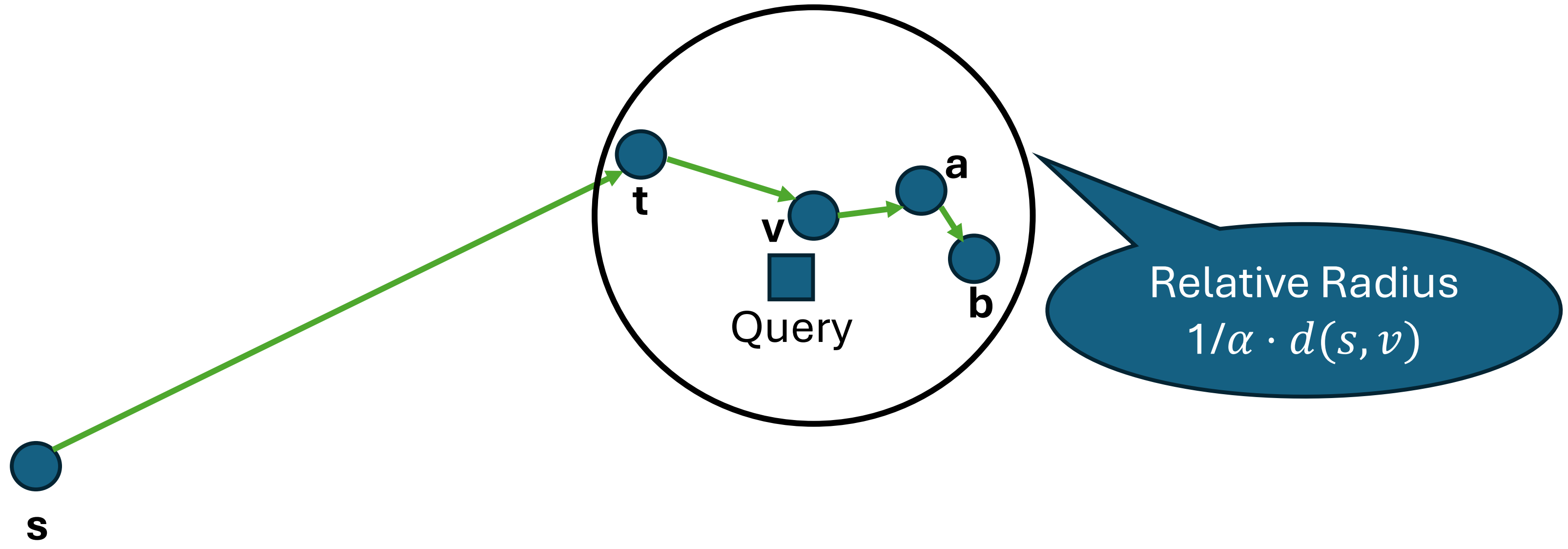
Low diameter

Greedy search should reach good approximate solution



Linear search!

α – Reachable Graph



$for\ all\ \boldsymbol{s}, \boldsymbol{v}\ there\ exists\ \boldsymbol{t}\ such\ that$

$d(v, t) \leq \frac{1}{\alpha} \cdot d(s, v)$

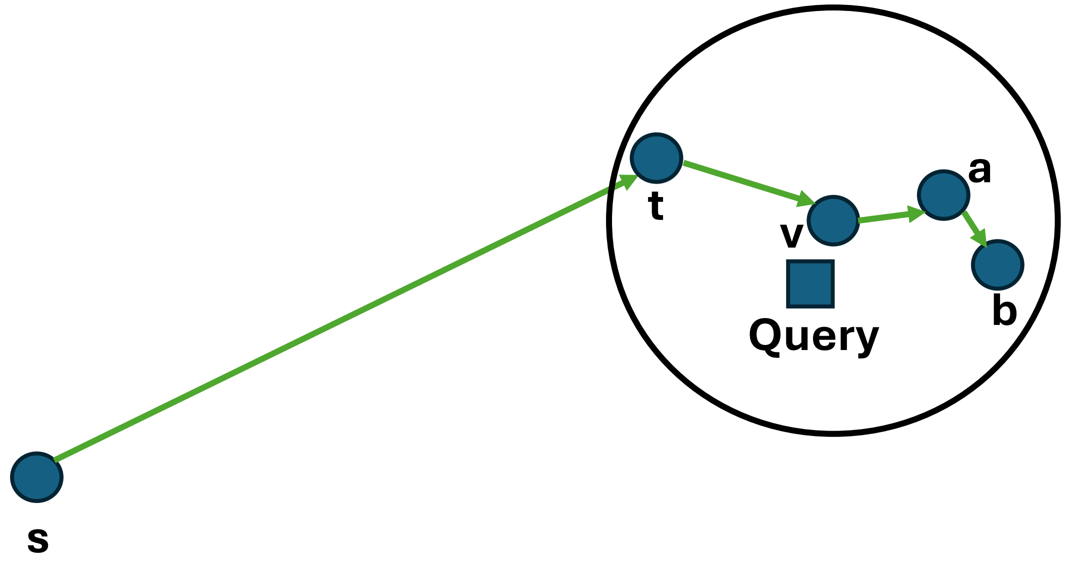
$(s, t)\ edge\ exists$

[$\boldsymbol{Indyk\ Xu, 23}$]

Maximum degree$\leq \alpha^d$, where d is doubling dimension

Maximum diameter $\leq \log_\alpha(\cdot)$

Local optimum solution



$$d(s,q) \leq d(t,q)$$

$s\ is\ local\ optimum$

$$d(t,v) \leq \frac{1}{\alpha} \cdot d(s,v)$$

$\alpha - reachable\ property$

$$\max \frac{d(s,q)}{d(v,q)}$$

Worst case
approximation ratio

Local optimum solution ($\alpha -$ Reachable graph)

$\max \frac{d(s,q)}{d(v,q)}$

$v \text{ is global optimum}$

$d(s,q) \leq d(t,q)$

$s \text{ is local optimum}$

$d(t,v) \leq \frac{1}{\alpha} \cdot d(s,v)$

$\alpha - \text{reachable property}$

$d(\cdot,\cdot) \text{ is a metric}$

metric property

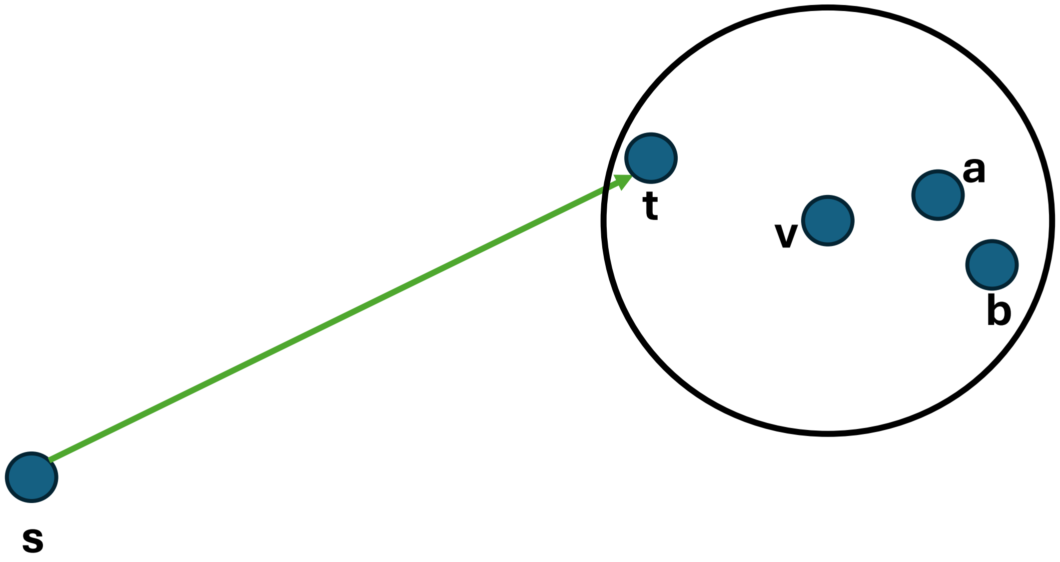
$[\boldsymbol{Indyk\ Xu, 23}]$

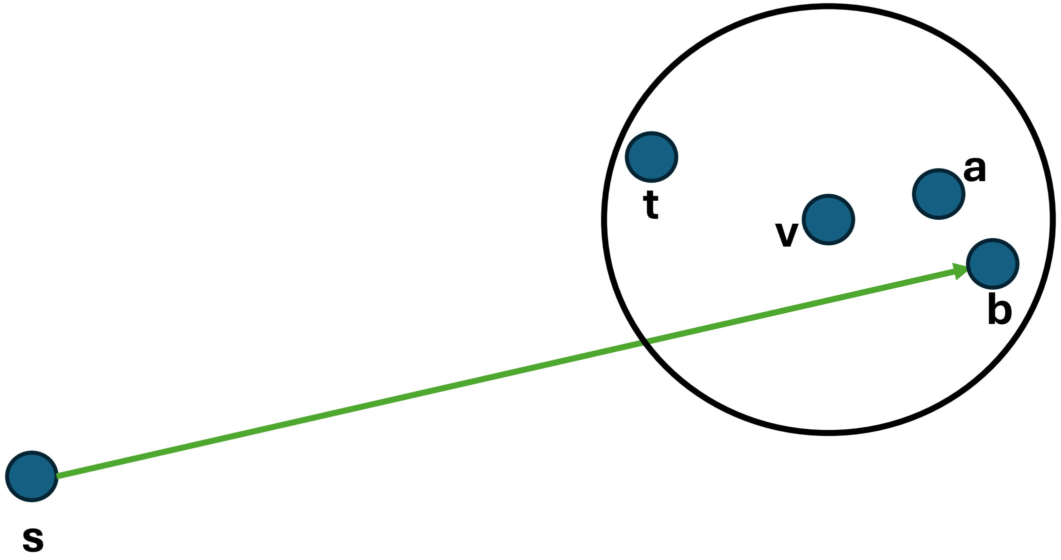
$\text{Worst case ratio} \leq \frac{\alpha+1}{\alpha-1}$

Tight!

Running time $O(\alpha^d \cdot \log_\alpha(\cdot))$

$\alpha -$ sorted Reachable Graph





$d(s,t) \leq d(s,v)$

$d(s,b) \geq d(s,v)$

Local optimum solution ($\alpha -$ sorted Reachable)

$$\max \frac{d(s,q)}{d(v,q)}$$

v is global optimum

$$d(s,q) \leq d(t,q)$$

s is local optimum

$$d(v,t) \leq \frac{1}{\alpha} \cdot d(s,v)$$

$\alpha -$ reachable property

$d(\cdot,\cdot)$ *is a metric*

metric property

$$d(s,t) \leq d(s,v)$$

sorting property

Our results

Worst case approximation ratio

$$\frac{\alpha}{\alpha - 1}$$ **Euclidean metric**

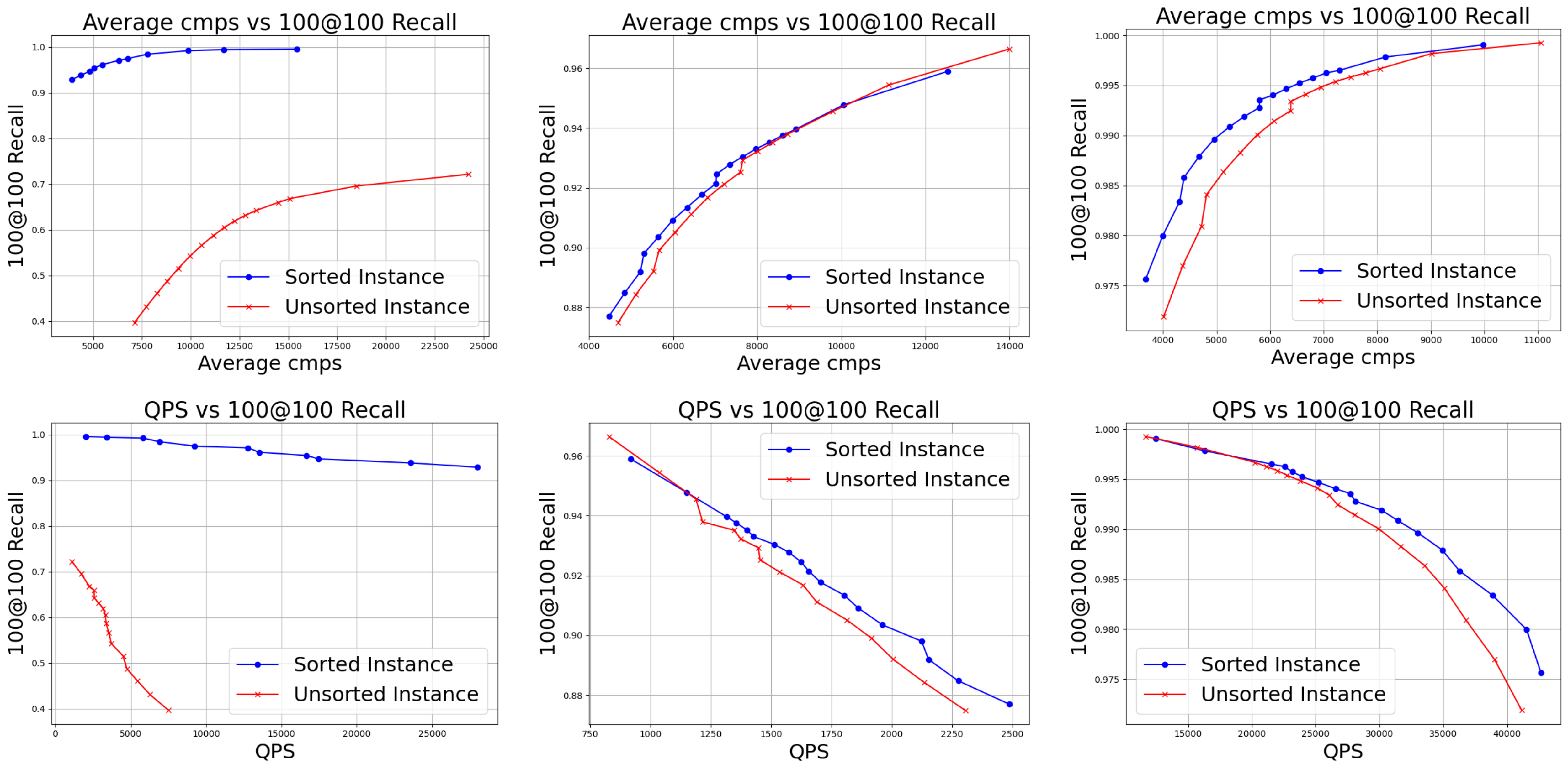
Beam search

$$\frac{\alpha}{\alpha - 1}$$ **Euclidean metric**

$$\frac{\alpha + 1}{\alpha - 1}$$ **Worst case metric**

All results are tight!

Experiments



Proof for Euclidean metric

$\max \frac{d(s,q)}{d(v,q)}$

$\max \frac{||s-q||_2}{||v-q||_2}$

v is global optimum

$||s-q||_2 \leq ||t-q||_2$

$d(s,q) \leq d(t,q)$

s is local optimum

$||v-t||_2 \leq \frac{1}{\alpha} \cdot ||s-v||_2$

$d(v,t) \leq \frac{1}{\alpha} \cdot d(s,v)$

α − reachable property

ℓ_2 *metric*

$d(\cdot,\cdot)$ *is a metric*

metric property

$||s-t||_2 \leq ||s-v||_2$

$d(s,t) \leq d(s,v)$

sorting property

Proof for Euclidean metric

SDP!

$$\max \frac{d(s,q)}{d(v,q)}$$

$$\max \frac{||s-q||_2^2}{||v-q||_2^2}$$

v *is global optimum*

$||s-q||_2^2 \le ||t-q||_2^2$

$$d(s,q) \le d(t,q)$$

s *is local optimum*

$||v-t||_2^2 \le \frac{1}{\alpha} \cdot ||s-v||_2^2$

$$d(v,t) \le \frac{1}{\alpha} \cdot d(s,v)$$

$\alpha -$ *reachable property*

ℓ_2 *metric*

$$d(\cdot,\cdot) \text{ is a metric}$$

metric property

$||s-t||_2^2 \le ||s-v||_2^2$

$$d(s,t) \le d(s,v)$$

sorting property

Weak Duality!

Motivating questions

1) Better theoretical understanding of DiskANN

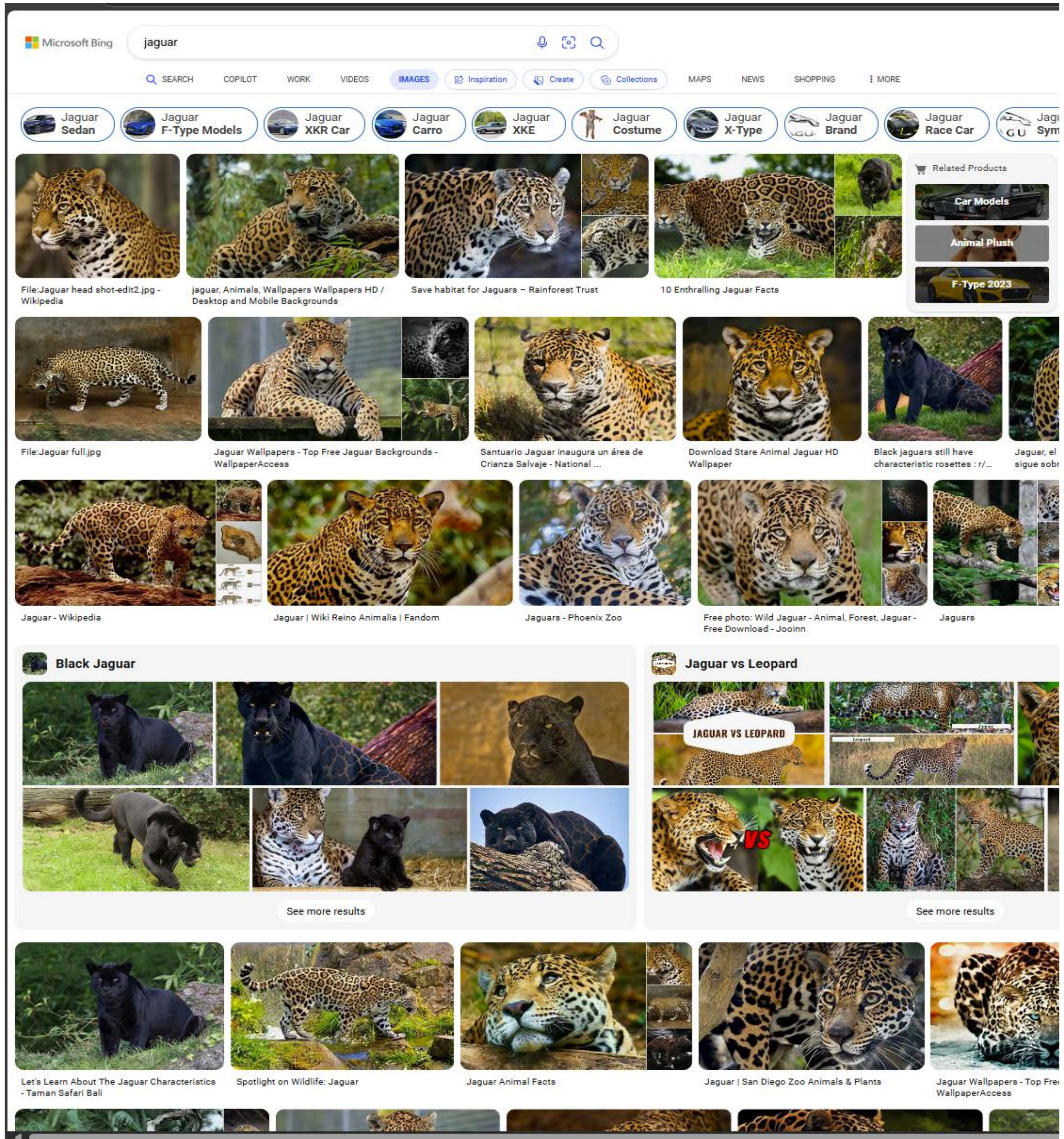
2) Modify DiskANN to accommodate diversity?

Diversity in Vector Search

Graph-based Algorithms for Diverse Similarity Search (**ICML**, Under Review)

Piotr Indyk (MIT), **Ravishankar Krishnaswamy (MSRI), Sepideh Mahabadi (MSR Redmond), KS, Haike Xu (MIT)**

Diversity in Search

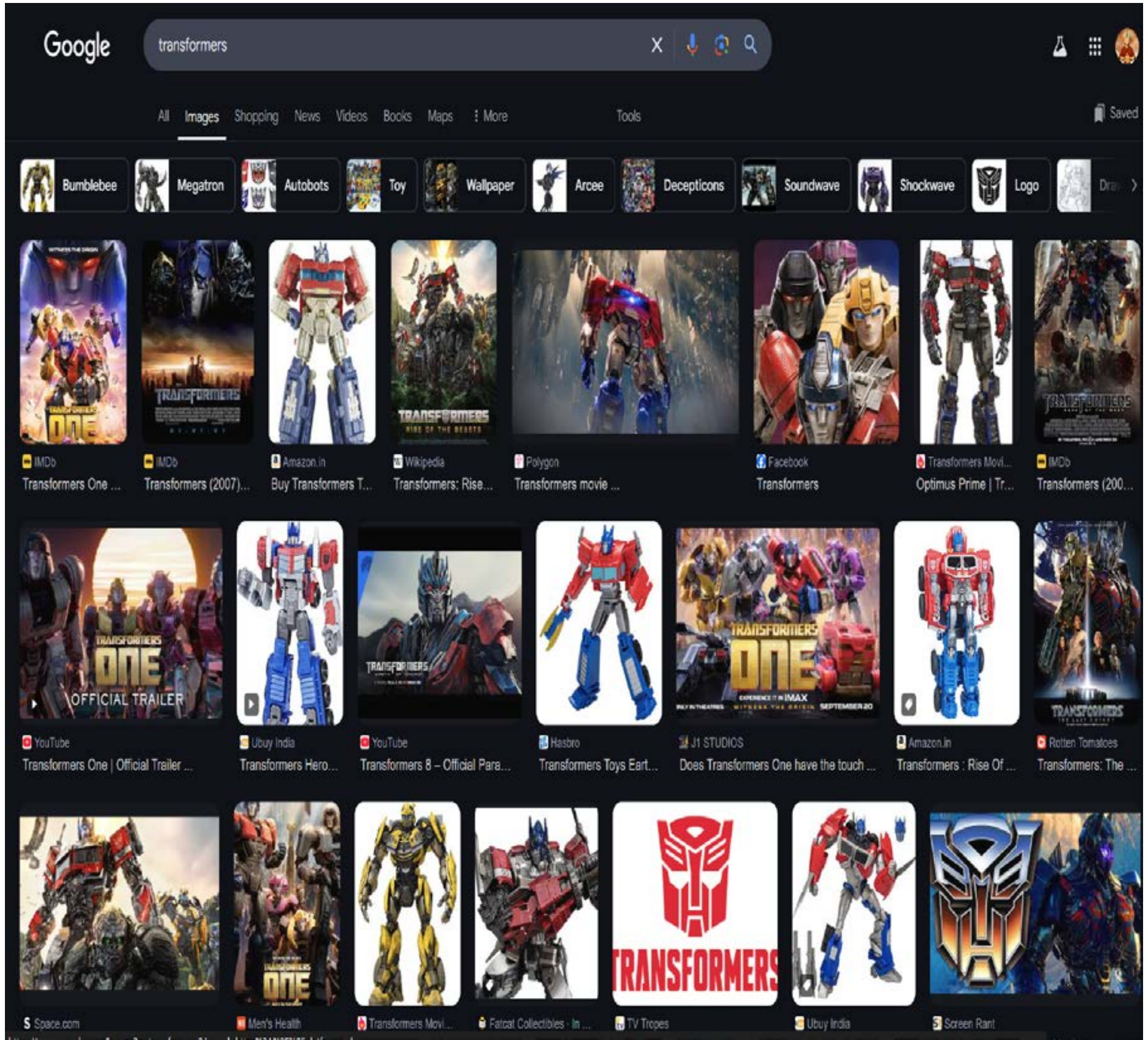


SEJ · News

Google Announces Site Diversity Change to Search Results

SEJ STAFF
Roger Montti
June 6, 2019 · 4 min read

997 SHARES **12K** READS



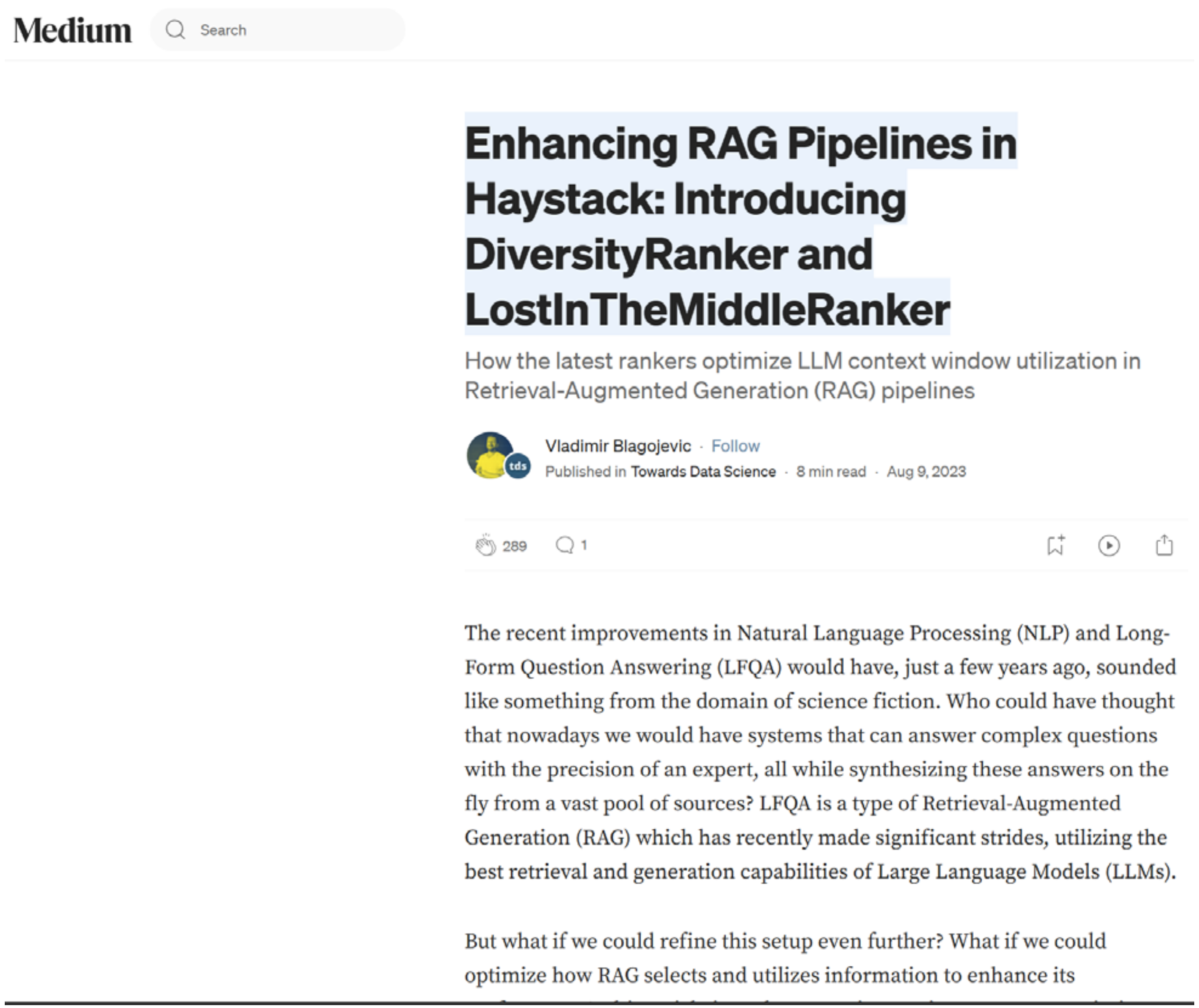
Diversity in RAG

Write a 2-pager on the history of ICTS

Author

Timeline

Project topics



Medium

Search

Enhancing RAG Pipelines in Haystack: Introducing DiversityRanker and LostInTheMiddleRanker

How the latest rankers optimize LLM context window utilization in Retrieval-Augmented Generation (RAG) pipelines

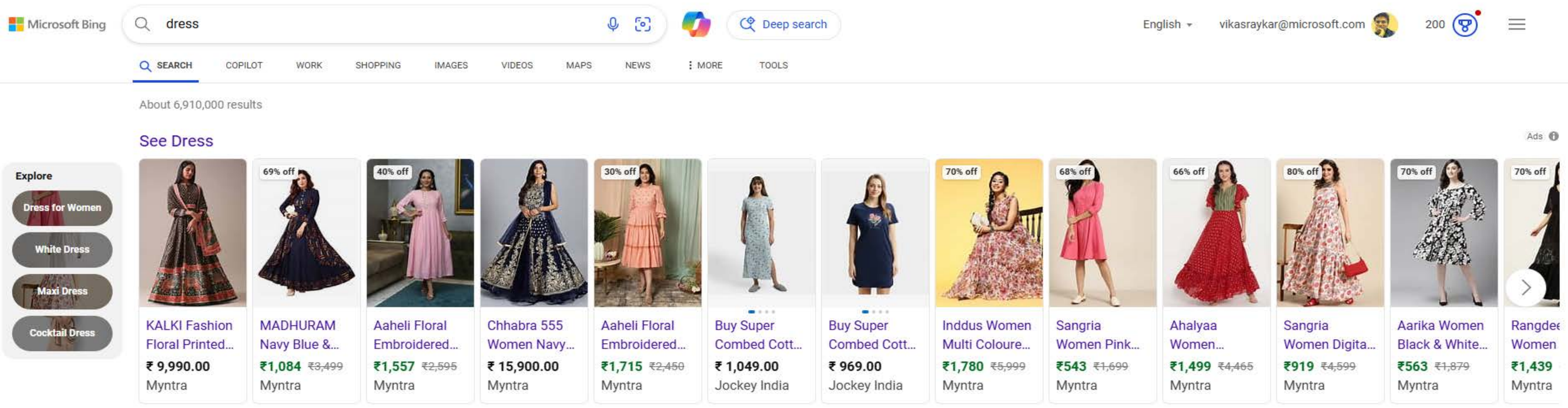
Vladimir Blagojevic · Follow
Published in Towards Data Science · 8 min read · Aug 9, 2023

289 1

The recent improvements in Natural Language Processing (NLP) and Long-Form Question Answering (LFQA) would have, just a few years ago, sounded like something from the domain of science fiction. Who could have thought that nowadays we would have systems that can answer complex questions with the precision of an expert, all while synthesizing these answers on the fly from a vast pool of sources? LFQA is a type of Retrieval-Augmented Generation (RAG) which has recently made significant strides, utilizing the best retrieval and generation capabilities of Large Language Models (LLMs).

But what if we could refine this setup even further? What if we could optimize how RAG selects and utilizes information to enhance its

Diversity in Ads



Retain small sellers

Advertiser fairness

Enhance user experience

Question

Can we build a data structure to support diverse vector search?

Definition of Diversity?

Single attribute

Definition of diversity depends on context

Objects

Relevance vector v_i

Color c_i

Seller (Ads)

Timeline, Topics (RAG)

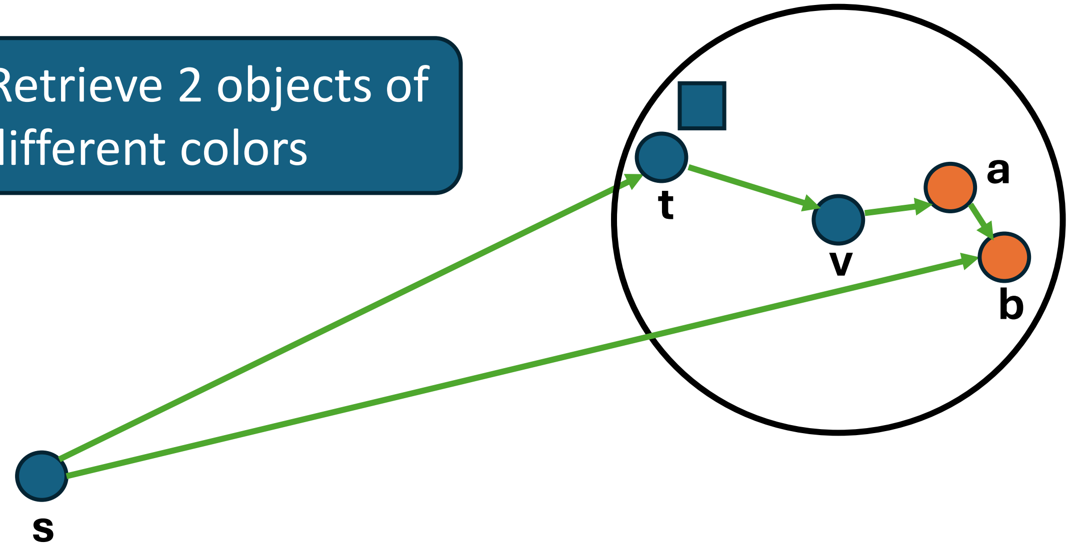
Query

Retrieve k=100 relevant objects

At most 10 objects from same color

Construction of Graph (Diverse DiskANN)

Goal: Retrieve 2 objects of different colors



Optimal: **{t,a}**

Naïve Greedy + Naïve Graph: **{t,v}**

Diverse Greedy + Naïve Graph: **{t}**

Diverse Greedy + Diverse Graph: **{t,b}**

Good approximate solution

Diverse Greedy Search

While searching we traverse this graph while satisfying diversity constraint.

Our results

Our algorithm returns a set of points $v_1, v_2, \dots, v_k$

Diversity constraints are satisfied

Approximation ratio

$$\frac{d(v_i, q)}{d(opt_i, q)} \leq \frac{\alpha + 1}{\alpha - 1}$$

Running time

$$O(m * f(\alpha))$$

$m \text{ is diversity parameter}$

General case

Objects

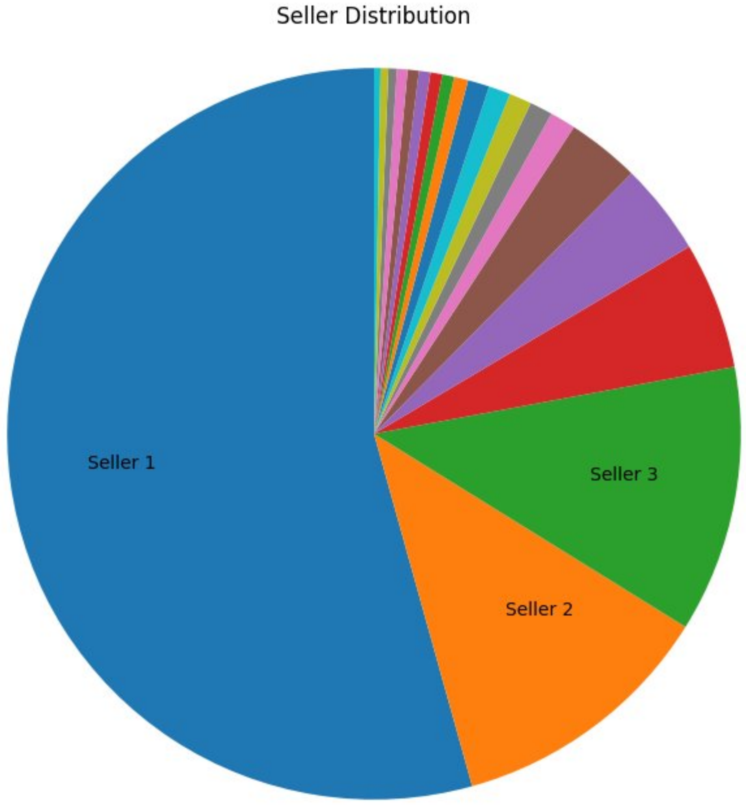
Relevance vector v_i

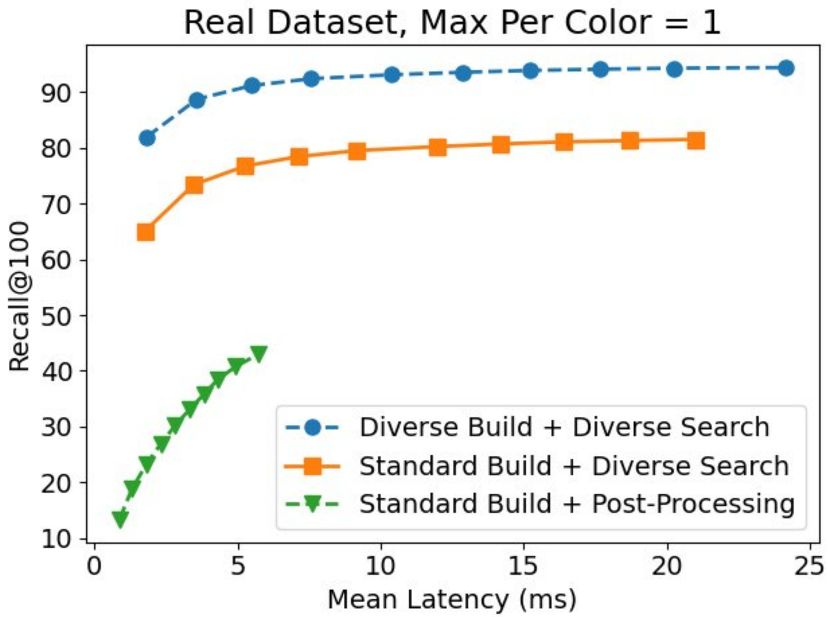
Diversity vector u_i

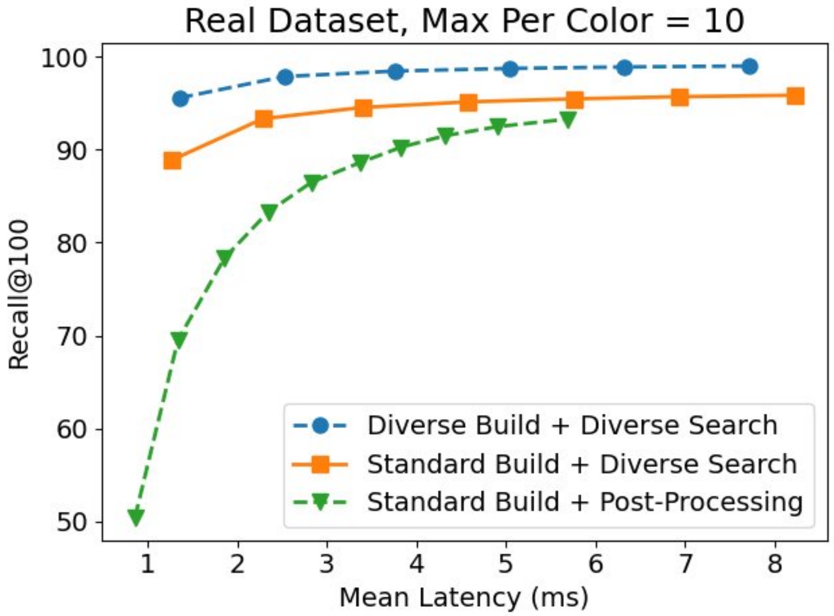
Experimental results

Product ads dataset

20 Million vectors, 5000 queries, 64-dimension embeddings







Standard DiskANN Build + Standard DiskANN Search + Post-processing

Standard DiskANN Build + Diverse DiskANN Search

Diverse DiskANN Build + Diverse DiskANN Search

New directions

Diverse search

Embeddings are evolving (Multi-vector)

Filter search, Online search, Distributed search, Page search

Wider applicability

Provable algorithms

Promising experiments!

First provable graph algorithms

References:

1) Sort Before You Prune: Improved Worst-Case Guarantees of the DiskANN Family of Graphs (**ICML**,2025)
2) Graph-based Algorithms for Diverse Similarity Search (**ICML**, 2025)
3) α-Reachable Graphs for Multivector Nearest Neighbor Search (**ICML**, **VecDB** workshop 2025)

Thank you!

RF position at MSR