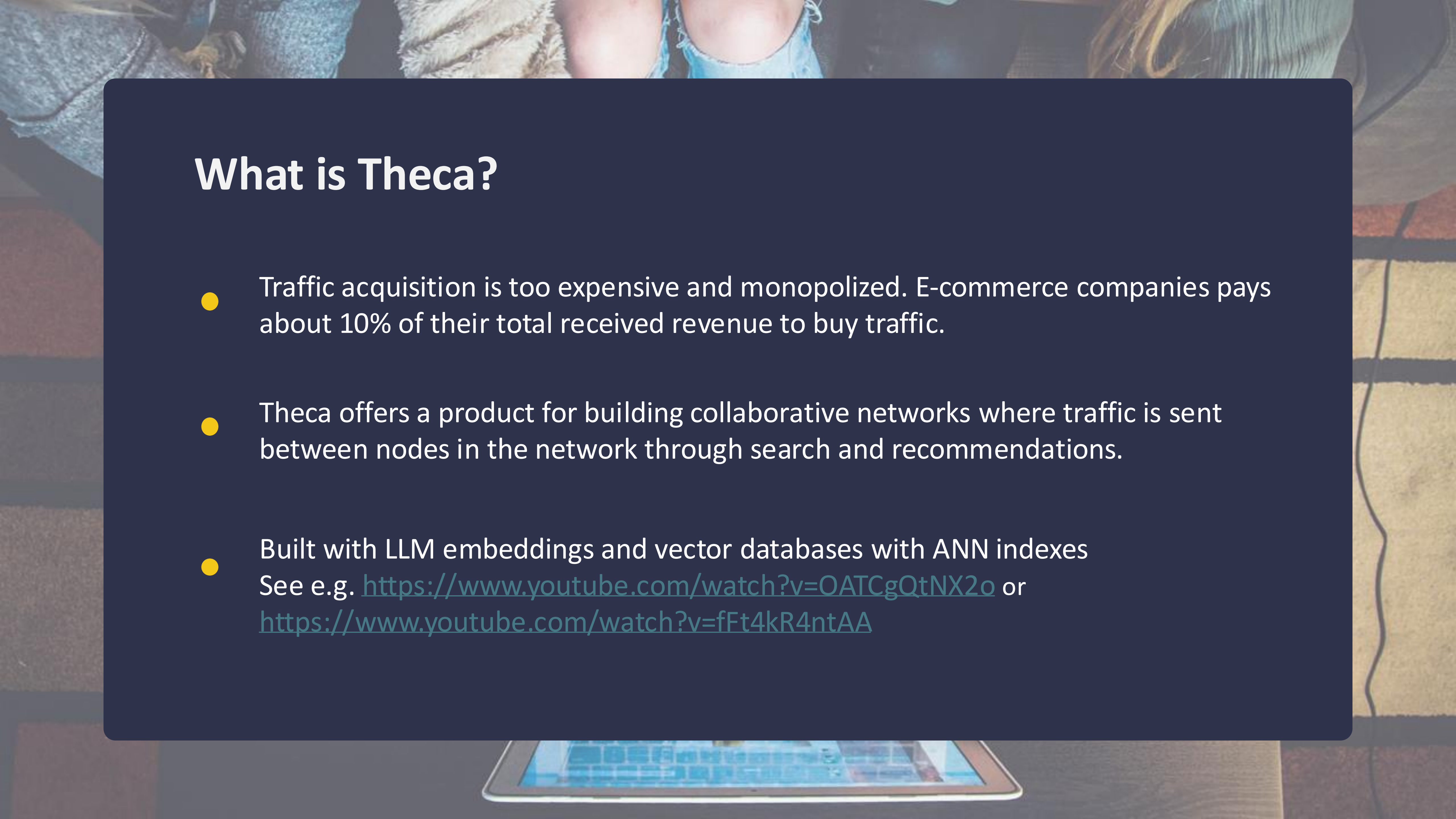
theca

Rasmus.Ros@theca.com
Tech Lead, PhD

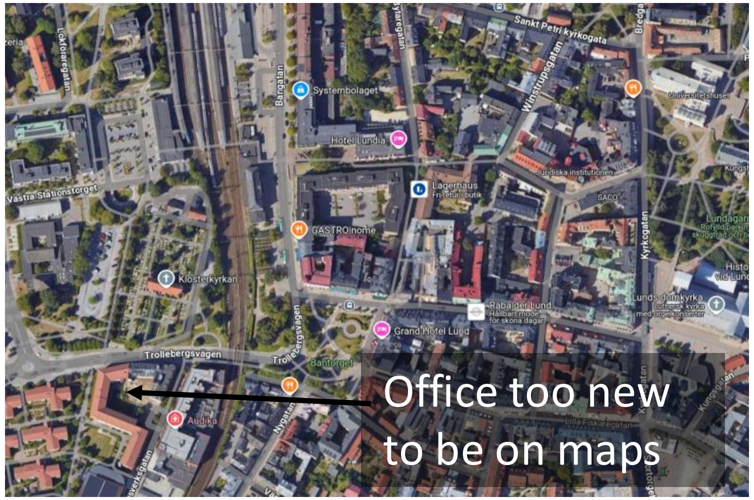


What is Theca?

- Traffic acquisition is too expensive and monopolized. E-commerce companies pays about 10% of their total received revenue to buy traffic.
- Theca offers a product for building collaborative networks where traffic is sent between nodes in the network through search and recommendations.
- Built with LLM embeddings and vector databases with ANN indexes
  See e.g. https://www.youtube.com/watch?v=OATCgQtNX2o or https://www.youtube.com/watch?v=fFt4kR4ntAA

About us

- Started 2022-08 now with 7 employees
- Strong finances from founder



Office too new to be on maps

- We offer three master's thesis
- All of them are problems close to our hearts, no toy problems!

- If you are interested, email us at Careers@theca.com

theca

Thesis Opportunity

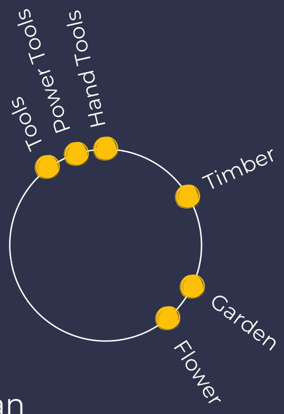
Scalable Optimization of Product Categories Using Spherical Codes and LLM Embeddings

Problem Description

Semantic search using large language models (LLMs) combined with an approximate nearest neighbor (ANN) index is the current state-of-the-art in search technology. This approach creates vector embeddings for documents. When users search, their query is also embedded, and the system retrieves the most relevant results based on semantic similarity through ANN. This method is superior to traditional keyword-based search systems in capturing user intent, as it considers word order and understands synonyms. However, keyword-based search retains an advantage in performance for advanced filtering options where users can filter results by product categories or other attributes.

This thesis explores encoding product category taxonomies into an n-dimensional spherical code for improved ANN indexing. The distances between category embeddings are optimized based on semantic similarity. For instance, "Power Tools" and "Hand Tools" should be placed close. This approach allows the filter to act as a soft ranking function, retrieving relevant results even if the category and query mismatch. For example, searching for "iPhone" under "Tablets" could still return iPads or iPhones.

What You'll Work On



The research will focus on solving a constrained optimization problem, using vector embeddings of product categories from an online store. The output will be optimized codes for each category node. A key challenge is scaling the solution for large datasets, and various optimization methods will be tested for efficiency. Possible extensions include:

- Compare the approach against baselines (such as computing the mean embedding vector of each category and averaging that with the query vector),
- Analyzing the theoretical properties of the optimization problem and its solution,
- Compare with model-based solutions where the codes are learned from data,
- Develop methods to automatically determine the optimal number of dimensions for the hypersphere.

Sounds like something for you?

Let's get in touch and talk more, email us at **Careers@theca.com**

theca

Thesis Opportunity

Closed-loop RAG for Content-based Recommendations

Problem Description

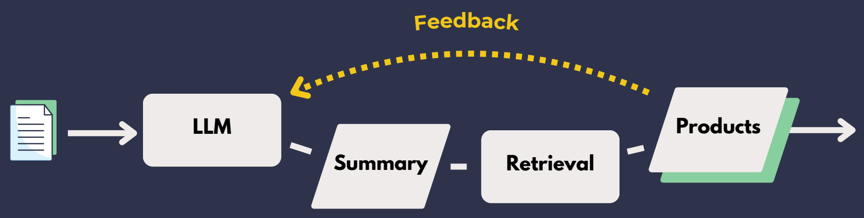
Generative Large Language Models (LLMs) excel in various natural language processing tasks but face issues like hallucinations, outdated knowledge, and untraceable reasoning. Retrieval Augmented Generation (RAG) addresses these flaws by integrating LLMs with information retrieval, enhancing accuracy and reliability through up-to-date, source-traceable data supplied during response generation.

This thesis introduces a closed-loop RAG system for content-based recommendations in e-commerce. It takes text from inspirational content (e.g., blogs) as input and outputs a list of related products. This output is fed back into the model to refine retrieval, repeating the process until the final selection. That is, each iteration combines the original text and previous product selections into a unified input for the LLM, which then generates a more targeted search query for improved recommendations. The processed query can be stored to enable instant, dynamic retrieval when users browse the content.

What You'll Work On

The goal of this thesis is to build a functional closed-loop RAG system for product retrieval in e-commerce, optimize each component for performance and accuracy, and evaluate its performance. Additional points that may be addressed include:

- Determining iteration stopping criteria,
- Prompt engineering for initial and iterative steps,
- Evaluating different commercial and local LLMs for generation,
- Identifying types of data to feed into the system (e.g., product categories).



Sounds like something for you?

Let's get in touch and talk more, email us at **Careers@theca.com**

theca

Thesis Opportunity

Evaluating re-ranking methods in LLM search

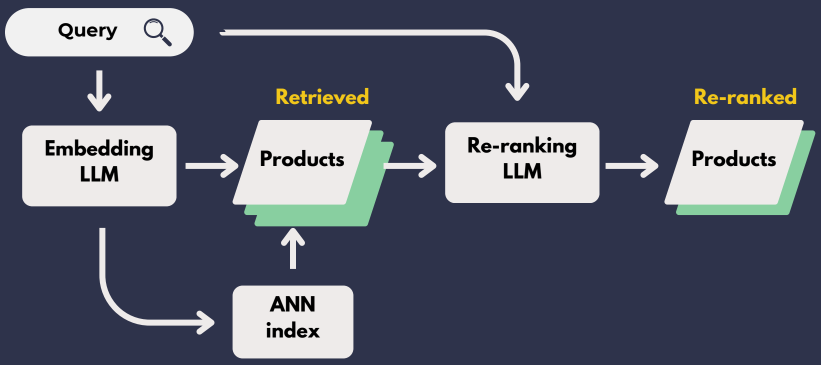
Problem Description

Semantic search has become an effective way to retrieve text based on meaning, rather than exact keyword matches. A typical semantic search system comprises several components: a Large Language Model (LLM) trained on extensive data, an approximate nearest neighbor (ANN) index of text embedding vectors, and a retrieval process that finds items similar to a query. These items are similar not just by keywords, but by their underlying meaning. However, the retrieval process may be improved with additional post-processing in a re-ranking step.

With re-ranking, we initially retrieve more documents than are ultimately displayed to users and reorder them through an additional process, finally selecting and presenting only the top results. A common example of re-ranking is a probabilistic classifier that performs expensive pair-wise comparisons of the query and each product. Another example is the use of Matryoshka embeddings, where embeddings are truncated into a smaller version used for indexing and a full-sized version that is used for final comparisons.

What You'll Work On

The main objective of the thesis is to research and implement re-ranking methods. This entails designing a system for evaluation of re-ranking methods and ensuring the methods scale to large datasets and production environments. The evaluation needs to consider a variety of contexts such as the processing time budget for a ranking, whether to use GPUs or not, the size of embeddings, the number of documents in the data sets, combinations of re-ranking methods, etc.



Sounds like something for you?

Let's get in touch and talk more, email us at **Careers@theca.com**