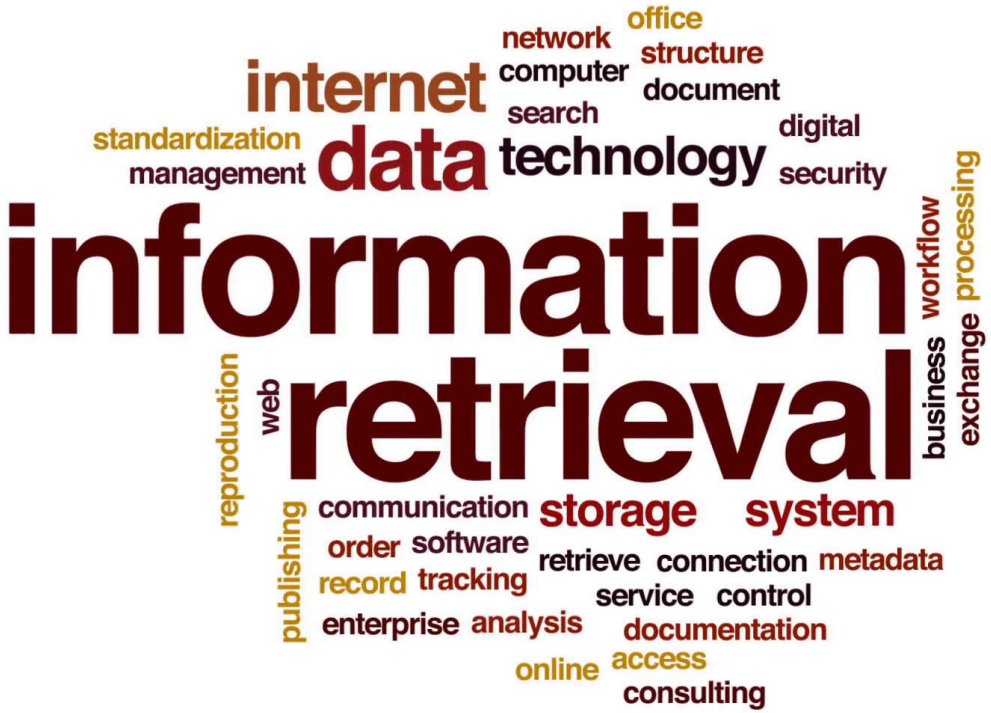


Extending Search Engines with Synonyms

LAMPROS LOUNTZIS

Table of Contents

INTRODUCTION

The problem

Problem statement

To answer users' questions as precisely as possible, we should examine their queries from many perspectives!

When we search for something on the web, we usually give keywords or a description of what we want to find.

But that's not always the case! There are times when we're careless of the words we choose to use in our searches, ambiguous, or even indifferent of the vocabulary we use.

Context

Query expansion is a method which consists of adding terms to the user's query.

Thesaurus is a collection of synonyms and sometimes antonyms of words.

Embeddings are vectors that encode the meaning of words such that words that are close in the vector space have similar meaning.

Synonymity is the **semantic relation that holds between two words** that can express the same meaning.

Challenges deep-dive

Challenge 1

User language & data

- What kind of **language** should we be able to handle (formal or everyday vocabulary) ?
- How can we handle terms that have a **different meaning** based on context ?
- Is the search engine **topic specific** or is it a **general-purpose** IR system ?

Challenge 2

Query extension method

- The method is **heavily dependent** on the users' queries and the data.
- What kind of **information** should be included in a thesaurus or the embeddings (e.g., formal terms, abbreviations) ?
- How **big** the thesaurus or the embeddings should be ?

Challenge 3

Performance

- The IR system's performance is **dependent** on the method used to extend the queries.
- Can the **volume** of the thesaurus or the embeddings impact the system's performance ?
- How quickly can we deliver results ?
- The method used should not cause system degradation!

SEARCH ENGINE

Data and IR system

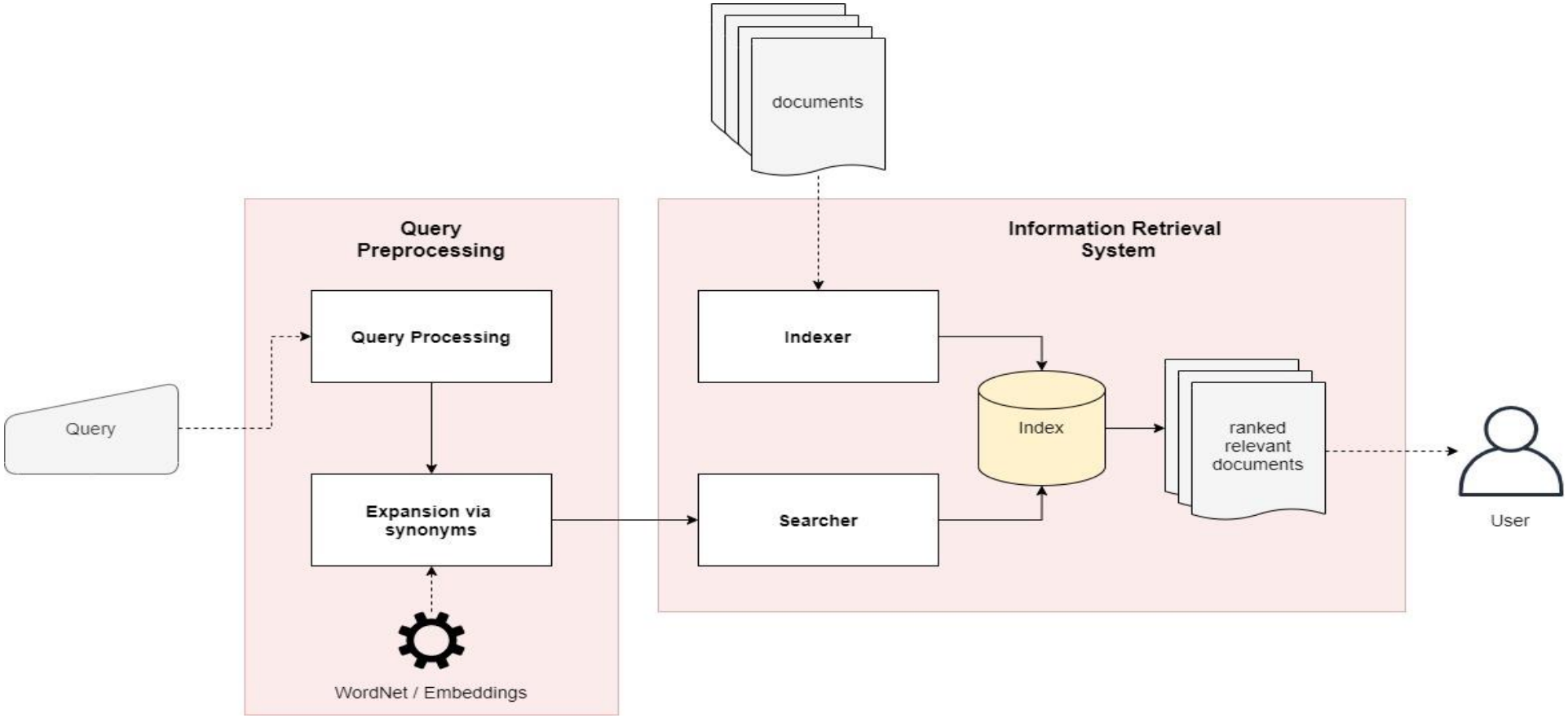
- The Information Retrieval system (search engine) was created using the **Apache Lucene** library, in **Java**.
- We're using the **CISI** corpus (by the Information Retrieval Group of University of Glasgow) which consists of text data about 1,460 documents and 112 associated queries.
- From the documents and queries we only keep the **ID**, the **title** and the **abstract** fields.
- The documents have two searchable fields: the title and the abstract. Thus, we perform **multifield document retrieval** based on the fields mentioned above.
- The data initially are **cleaned** by:
  - **removing symbols** and
  - **lowercasing** the text (optional).
- Further analysis on data is carried by the **EnglishAnalayzer**, which:
  - **tokenizes** the text based on grammar (implements Word Break rules from Unicode Text Segmentation algorithm),
  - **removes possessives** from words and **stopwords**,

Data and IR system

- **lowercases** the text,
- **stems** the text using the **Porter stemming algorithm**.

- Especially at **query time**, the **EnglishAnalayzer** is extended to handle **query expansion with synonyms**, either by using a thesaurus or word embeddings.
- For query-document similarity and scoring we are using **Okapi BM25 Similarity**, which is state-of-the-art in probabilistic information retrieval.
- To evaluate our IR system, we have used the **trec_eval** tool and its metrics.
- Specifically, the search engine is evaluated on **the top k retrieved documents**, and we consider the following **metrics**:
  - **precision**,
  - **recall**,
  - **mean average precision (MAP)**.

IR system architecture



Query extension methods

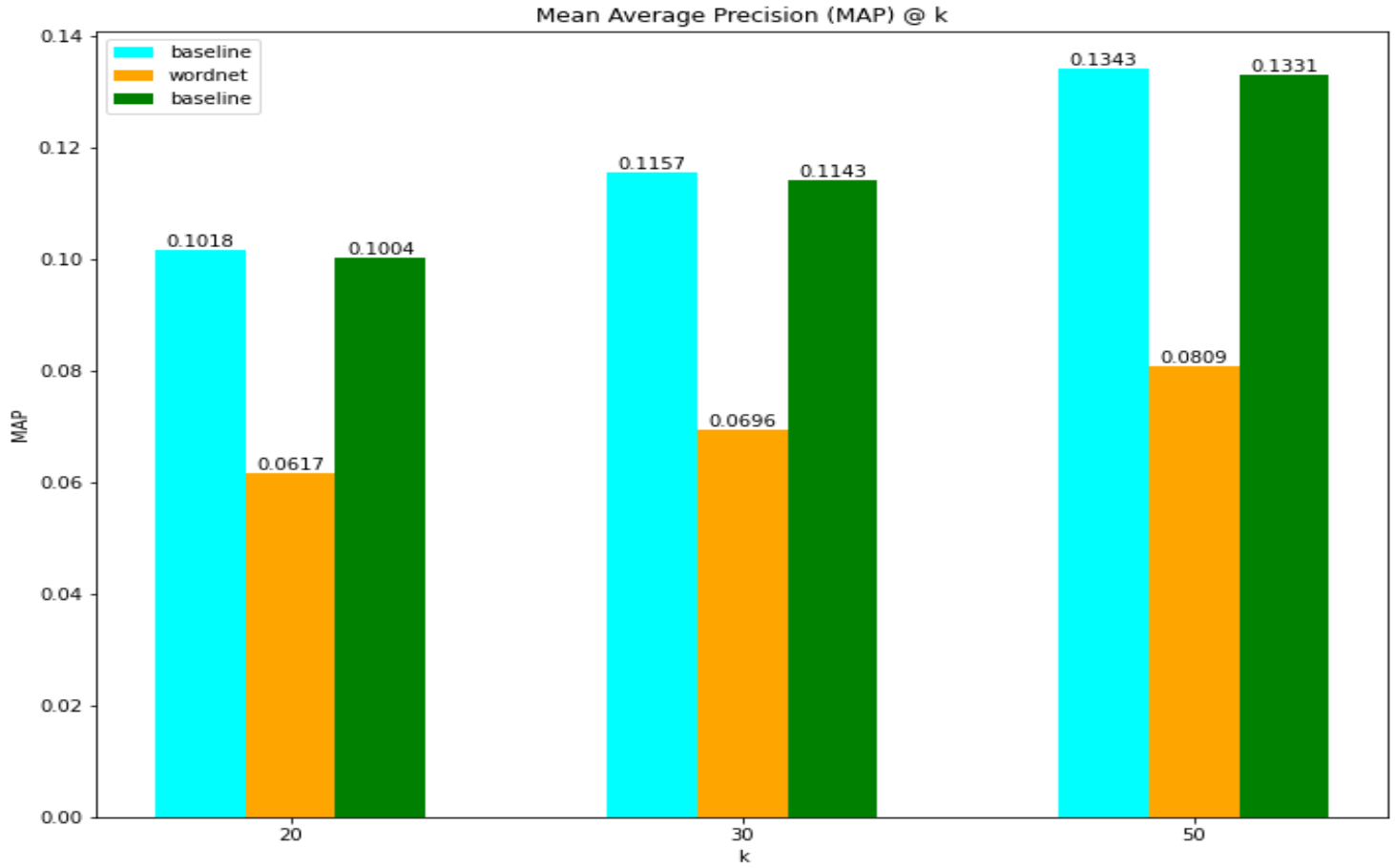
- To capture the meaning of the user's query, we'll **expand** it using the **synonyms** of the terms included. The new query will be a conjunction of the initial terms and their synonyms.
- Two methods are going to be used:
  - expansion through a **thesaurus**,
  - expansion through **embeddings**.
- Thesaurus:
  - We'll use **WordNet** which is a lexical database in which nouns, verbs, adjectives and adverbs are grouped into sets of **cognitive synonyms (synsets)**, each expressing a distinct concept.
  - Synsets are interlinked by means of conceptual-semantic and lexical relations.
- Embeddings:
  - We'll use **fastText** which is an open-source library that allows to learn text representations and text classifiers.
  - The **pre-trained word vectors** used are **wiki-news-300d-1M.vec**.
  - Synonym terms are close in the vector space, so we'll choose terms that minimize the cosine distance.
  - Thus, a **similarity threshold of 0.98** has been chosen.

RESULTS

Results

	Baseline			WordNet			Embeddings		
	k = 20	k = 30	k = 50	k = 20	k = 30	k = 50	k = 20	k = 30	k = 50
MAP	0.1018	0.1157	0.1343	0.0617	0.0696	0.0809	0.1004	0.1143	0.1331
Precision @ k	0.2638	0.2320	0.1924	0.1796	0.1632	0.1426	0.2579	0.2276	0.1913
Recall @ k	0.1816	0.2392	0.3090	0.1332	0.1669	0.2283	0.1663	0.2336	0.3081

Results


Mean Average Precision (MAP) @ k
baseline
wordnet
baseline
MAP
k
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
20
30
50
0.1018
0.0617
0.1004
0.1157
0.0696
0.1143
0.1343
0.0809
0.1331

Discussion

- To begin with, as the **number of top retrieved documents increases**, **precision decreases**. This is because more irrelevant documents (lesser relevancy value with respect to the query) are retrieved.
- Since **precision decreases**, **recall increases**, because these two measures are inversely proportional.
- Also, the **mean average precision increases while k (number of top retrieved documents) increases**. That is because there are more true positives (relevant documents with respect to the query) leading the set of retrieved documents. One could say that the sorting by the similarity function is better.

We expected that our IR-system extended with synonyms would have a greater performance. However, it performs poorly 😢 **Why?**

1. Query expansion is often effective in **decreasing precision**. More documents are retrieved due to synonym terms, However not all of them have a high relevancy value.
2. To add, **mean average precision is decreased** since the true positives are lesser.
3. One thing we can clearly observe is that the **use of WordNet has made our search engine perform worse** than our baseline model. What we can speculate is that the use of every synonym for each term acts as **noise** during retrieval. Also, thesauri and dictionaries give far too **little coverage of the rich domain-particular vocabularies** of most scientific fields We could do, is make use of a **domain specific thesaurus**.

Discussion

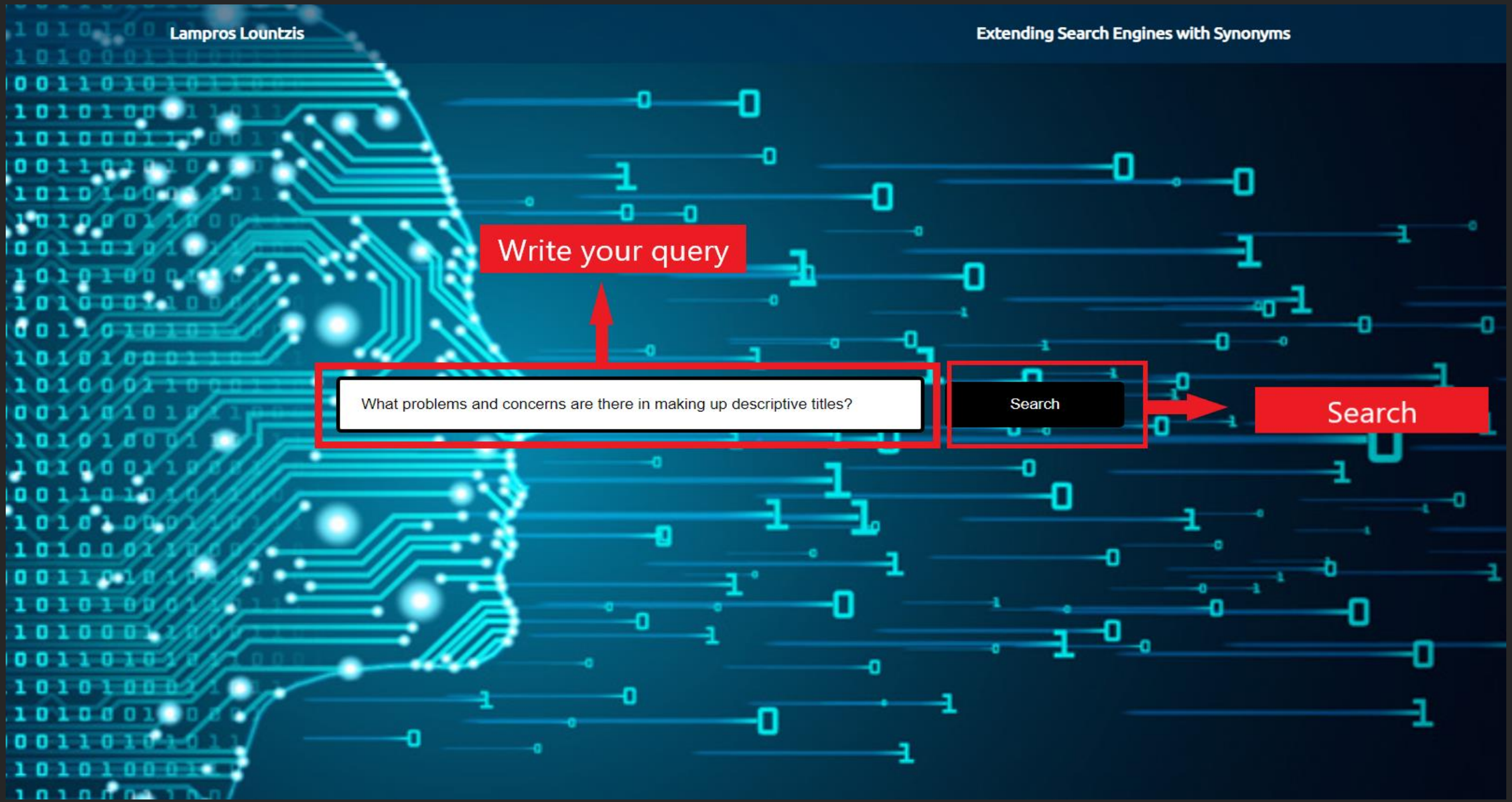
4. Contrary to WordNet, the IR-system that leverages **word embeddings seems to be perform as great as the baseline model**. One reason, is that we have set a **threshold to the selection of the synonyms** (0.98 similarity and above) and thus we get rid of terms that have lost the initial meaning (noise). Another reason, is that fastText word embeddings contain **more terms** (1M terms) than a thesaurus. They contain everyday vocabulary, more scientific terms, etc. so it guaranteed that most query terms will have close synonyms. In general, **building word embeddings (e.g. with Word2Vec) is easier than building a thesaurus or dictionary** since it requires constant update and manual creation of relationships (synonymity, antonymy).

- Overall, **query expansion through synonyms was less successful**.

<u>Ideas and further discussion:</u>

- One way we could understand the users' queries better could be if through **users' feedback**.
- **Relevance feedback**, is a feature of some IR-systems and the idea behind it is to take the results that are initially returned from a given query, to gather the user feedback, and to use information about whether those results are relevant to perform a new query.
- There are three types of feedback: **explicit feedback**, **implicit feedback** and **pseudo-feedback**.

DEMO


Lampros Lountzis
Extending Search Engines with Synonyms
Write your query
What problems and concerns are there in making up descriptive titles?
Search
Search

term integrated activities extending over periods of days or years receive no attention.

Score: 8.395382

ID: 236

Title: Book Publishing: What it Is, What it Does

We speak of book publishing as an industry and as a profession. Both designations are certainly appropriate. Book publishing is a business conducted, for the most part, for profit. But its practitioners - at least those who do it honor - have motivations that transcend their profit interest. They know that books are no mere commodity, no mere items for consumption that leave their readers much as they find them. Books, like other vehicles of information and sources of entertainment can change, influence, elevate, demean, exalt, or depress those who expose themselves to them. What books are and can be depends heavily on the judgement, integrity, taste, and acumen of those who select and produce them - their publishers.

Score: 7.951471

Result (retrieved document)

ID: 60

Title: Information Science: What Is It?

In seeking a new sense of identity, we ask, in this article, the question: What is information science? What does the information science do? Tentative answers to these questions are giv

discussion that will help clarify the nature of our field and our work..

Score: 7.8107715

ID: 611

Thank you 😃