

# Development of a Search Engine and Applications in IR

Robert Pinsler  
WKWSCI, NTU  
N1509281G@e.ntu.edu.sg

Yiding Liu  
SCE, NTU  
LIUY0130@e.ntu.edu.sg

Yitong Guan  
SCE, NTU  
GUAN0049@e.ntu.edu.sg

Jenn Bing Ong  
SCE, NTU  
ONGJ0063@e.ntu.edu.sg

## ABSTRACT

### 1. INTRODUCTION

### 2. SYSTEM OVERVIEW

The search engine is able to index and search publication records listed in the dblp computer science bibliography (DBLP). It is built on top of PyLucene, a Python extension of the open-source software library Lucene that provides text indexing and search capabilities.

Publication records from DBLP are provided in a single XML file. The dataset comprises of various kinds of record types, such as articles published in a journal or magazine (*article*), papers published in a conference or workshop proceedings (*inproceedings*), books, and PhD theses. From those, only documents that are classified as *article* or *inproceedings* are considered.

### 3. INDEXING

To index a record, it first has to be extracted from the XML file before it can be stored. This can be done conveniently using an XML parser. Due to the large file size of the dataset, reading in the whole XML tree at once is impractical. Instead, we use a SAX-like parser from the *lxml* Python library that sequentially reads the document and emits events when it encounters certain XML tags. This allows to only react to `<article>` and `<inproceedings>` tags, thereby ignoring record types we are not interested in. This makes it very fast and memory-efficient. During event processing, it is possible to access the current element, i.e. an *article* or *inproceedings* record, and its children, i.e. attributes of the record such as title or year of publication. As some of the publication titles use basic HTML text formatting, HTML tags are removed prior to indexing.

Each record is indexed as a Lucene *document*, comprising of the following *fields*: id, title, authors, year and venue (journal or booktitle, depending on the record type). All

fields are treated as strings<sup>1</sup>. Field contents are stored in the index, which allows to retrieve them later during search. This is especially useful to enhance search results with additional information when presented to the user. The values of the fields *id* and *year* are stored as-is. All other fields are tokenized ( $f_{\text{token}}$ ) and further processed, including lowercase conversion ( $f_{\text{low}}$ ), stemming ( $f_{\text{stem}}$ ) and removal of stopwords ( $f_{\text{stop}}$ ). In particular, we use the Porter stemmer and the default stopwords list of Lucene, which consists of 33 common English words. Table 1 summarizes the analysis applied to each field.

In order to evaluate the impact of additional token processing onto the index, different experiments are conducted. We quantify the speed of indexing in seconds,  $t_{\text{ind}}$ , as well as the number of terms in the *title* vocabulary,  $|V_{\text{title}}|$ , when applying lowercase conversion, stemming and/or removal of stopwords on the *title* field. The results are shown in Table 2. [...]

Table 1: Analysis of document fields

Field	$f_{\text{token}}$	$f_{\text{low}}$	$f_{\text{stop}}$	$f_{\text{stem}}$
id				
year				
authors	X	X	X	
venue	X	X	X	
title	X	X	X	X

Table 2: Evaluation of token processing on *title* field

$f_{\text{low}}$	$f_{\text{stop}}$	$f_{\text{stem}}$	$t_{\text{ind}}$	$ V_{\text{title}} $
			0.123	1337
X			0.123	1337
	X		0.123	1337
		X	0.123	1337
X	X	X	0.123	1337

### 4. SEARCH

The search module makes use of the document index to retrieve the most relevant documents to a given query. It supports free text keyword queries on any combination of attributes. By default, all attributes are considered. Phrase queries are supported using double quotation marks (e.g.

<sup>1</sup>The *year* attribute may also be stored as integer, allowing for features like range searches. However, this would require additional effort during search that we try to avoid.

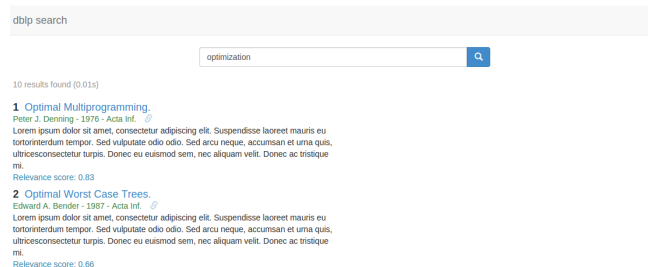
“information retrieval”) [...]. The search returns the  $N$  most relevant documents according to Lucene’s internal scoring function, where  $N$  is a configurable parameter. It also measures the time needed to retrieve the results.

## 5. EVALUATION

## 6. USER INTERFACE

There are two ways to interact with the search engine: via a text-based command line interface or via a web UI. The web UI is built with Flask, a lightweight Python web framework that leverages Jinja2 as a templating engine. This allows to easily create HTML documents from within Python. Additionally, we incorporate the web framework Bootstrap to achieve responsive web design. Figure 1 depicts the search results for an example query.

**Figure 1: Web-based search user interface**



## 7. APPLICATIONS IN IR

### 7.1 Popular research topics

### 7.2 Similar publication venues

## 8. CONCLUSIONS