Article retrieval system

https://github.com/Rebryk/SPbAU-IR

Vsevolod Stepanov
Yurii Rebryk
tehnar5@gmail.com
y.a.rebryk@gmail.com

ABSTRACT

This is our report for Information Retrieval project.

1 PROJECT DESCRIPTION

We are developing a system for searching scientific articles.

The data for our search engine is collected from online libraries like arxiv.org and from publishers that have information about their articles available online (like springer.com).

Firstly, we crawled `20k` pages from these sources and saved them for future processing.

Then, we filtered irrelevant pages (those that doesn't contain an article description) and processed the rest of the documents. For each relevant page, we extracted article's title, abstract, list of authors, date of publishing, etc. Also, we store a link for a PDF version of an article, if it's available.

After that, we are going to implement a search engine for searching among these articles. **We will discuss it later**.

The key feature of our service will be creating some kind of graph over authors/articles and their references to each other. **We will discuss it later**.

2 SYSTEM DESIGN

We use `Python 3` for development and `PostgreSQL` database.

Articles, processed articles' abstracts and inverted index are stored in a filesystem. Articles' metadata and PAF are stored in the database.

3 DATA ACQUISITION

For data acquisition we implemented a web crawler.
The key details of out implementation are:

(1) **Politeness policy**
We follow the constraints defined in `robots.txt`. We do not visit excluded pages, do not store page if there is a `noindex` meta tag, and of course, do not spam a website with a lot of queries. Even if a delay is not specified in `robots.txt` we use value defined in the configuration file (default `delay` is `50ms`).

(2) **Distributed crawling**
The problem with Python's multithreading is that only one thread can be run at a time because of `GIL`, so we use multiprocessing instead. Each crawler runs in its own process and several crawlers can be run simultaneously. Crawlers do not exchange URLs. This solution is suitable in our case because:
   (a) There are not so many data sources with scientific articles so we can just list them.
   (b) We are not interested in links leading to another domain, because, most likely, this domain will not contain any relevant documents.

   Also it gives us a few advantages:
   (a) Simplicity of implementation. It is more difficult to make a mistake.
   (b) Easy to run on different computers, because communication is not needed.
   (c) Each crawler has its own set of allowed hosts (we define these hosts in the configuration file). It prevents the same page to be downloaded several times.

We store each HTML page we crawled in a filesystem and put some page's metadata (like its URL, hash of page content, date of last page modify) in database. Duplicated pages (with equal hashes) are ignored.

4 DATA PROCESSING AND STORAGE

As we have a few data sources and we need a specific information extracted from crawled documents (like article's title, abstract and list of authors), we implemented a separate data processor for each data source.

Data processing is done as follows:

(1) Firstly, we filter out pages that doesn't contain an article.
(2) Then, we extract relevant parts of remained pages and transform it to a plaintext, removing all the HTML tags.
(3) After that, article's abstract is processed for future indexing, stemming its words and removing stopwords and punctiation

Raw and processed article's abstract are stored in the filesystem, other information (like title, list of authors, link to PDF if available) is put into database

We use `BeautifulSoup` for webpage parsing, and `NLTK` for abstract processing.

4.1 Indexing

4.1.1 Page attribute file. For each article, we store its attributes: the title, the path to file with abstract, the path to file with processed abstract, the number of words, the link to pdf and so on.

This information we store in the database.

4.1.2 Inverted index. We built an inverted index. For each word, we store a list with documents that contain this word. And for each document, we store the number of occurrences of this word in the document with positions of these occurrences.

The document lists are sorted by the number of occurrences. Also, we use gap values to store positions of these occurrences for better compression.

This index can be build by several processes simultaneously. We split all articles by their `id` into groups, and each group is processed by a separate process. After that, all indices are merged into one index.

We store this index in a filesystem. We use `gzip` for compression.

5 RANKING

We implemented TF-IDF with cosine similarity for ranking.

We used inverted index constructed previously to calculate TF-IDF.

To improve search perfomance, for each document its vector is constructed once and then stored in memory. Also, to decrease memory usage and improve search perfomance even further, vectors are stored in a sparse form.

6 FEATURES

6.1 Article similarity graph

We want to plot aricles as points on 2D plane in a way that articles that are "similar" would be close to each other as points on plane.

To measure "similarity" of articles we're going to use doc2vec. Unfortunatelly, training doc2vec takes some time, so for now we implemented more simple approach. We use pretrained word2vec to get a vector for each term in an article and then use mean value of these vectors as a document's vector.

Later, we're going to compare this two approaches (but most likely, trained doc2vec would be better)

6.2 Authors graph

The same feature will be implemented for authors.

For each author we'll calculate mean vector of his articles and then visualize it the in same way as for articles.