CS47300 Fall 2020 Assignment 1 Solutions
Posted 11 October 2020.

1 Ad-Hoc Retrieval

Describe briefly an example of an ad-hoc information retrieval system you have used that is not web search. Include in your description:

- The corpus from which documents are retrieved. Give an example of an object in the corpus.
  An example corpus is the database of research papers, books, articles and more electronically accessible through Purdue libraries. An example of an object in the corpus is "The Thermalization, Condensation and Flickering of Photons." Journal of Physics B: Atomic, Molecular and Optical Physics by Klaers, Jan.

- The structure of a query. Give an example query.
  The structure of the query is the title/name of the book, paper, article or author that we are searching for. For example, I would expect the query query, "Thermalization, Condensation and Flickering of Photons." to return the article above

- How the retrieved objects are presented.
  The retrieved information is returned as a list of all the articles and books related to the query with the title, name of the author and a brief description about whether it is available for free, peer reviewed and more.

2 Filtering as Information Retrieval

Give an example of an information retrieval system you have experienced utilizing filtering, rather than ad-hoc retrieval. Try to answer:
There are many examples, I'll take products filtering on Amazon as example. On Amazon, you are shown product recommendations even if you don't give a query.

- What corpus is said algorithm filtering on?
  All products on Amazon

- What is most likely to be the type of filtering technique applied?
  There are a number of techniques. They are known for collaborative filtering (which we'll discuss later), but they also filter based on your past purchases or searches, as well as explicit information you provide in your provide (make and model of vehicle, type and number of people or pets in your household.

- Why does filtering makes sense in this application?
  Because sometimes when users browse products, they only know what type of products they want to buy and compare a lot of products before make decisions. In this case, users may not exactly know the keywords for query in ad-hoc retrieval.

- How are queries in the filter specified? Do they fit the traditional notion of a "query"?
  Multiple methods: past history, profile settings, perhaps things Amazon wants to sell, or simply through indication of a category of interest. Category is different from ad-hoc retrieval because in the category is known in advance, and thus relevance can be established through other means (e.g., manual assignment.

3 Full Text Indexing

Give an example of an Information Retrieval system (it need not be ad-hoc information retrieval) where full text indexing would be expected to work better than controlled vocabulary indexing. Briefly describe why:

There are many such examples. I'll use the idea of public library search, but the arguments likely apply to web search as well.

- Full text indexing is expected to perform better than controlled vocabulary indexing, and

  In these cases, we may have documents that do not match any pre-conceived notion of the vocabulary, so a controlled vocabulary may not work for the entire corpus (even assuming we have an easy means of assigning it.) Furthermore, those issuing queries may have information needs that we do not anticipate, and thus cannot be expressed well in the controlled vocabulary.

- Give a drawback to full text indexing in this given application, or reason about why there is no drawback.

  Full text indexing is likely to result in low precision. With a controlled vocabulary (e.g., the topics associated with the Dewey Decimal System in a library catalog), the words are limited to ones that provide a reasonable way of semantically distinguishing documents. Words that may carry semantic meaning (and thus be selected to be in a query), but apply to a wide variety of topics, would not be part of the controlled vocabulary. The key issue people had here was not thinking enough about how a controlled vocabulary could be helpful in their application (even though full-text would be generally better.)

4 Stemming and Stopwords

Perform stemming and stopword removal on the following body of text:

"The quick brown fox jumped over the lazy dog."

For each modification you make to the phrase, include either the rule you used for stemming or whether the word is on the stopword list.

Final output sentence: "quick brown fox jump over lazy dog". Stopword: "the", Stemming: remove "ed" at the end of a word. There may be more aggressive stemmers that might modify lazy or over, but most of these word would be unchanged by most stemmers.

5 Inverted Lists

Create an inverted list from the following 3 documents:

D1: "The quick brown fox jumped over the lazy dog."

D2: "But why did the fox jump over the dog?"

D3: "Calling the dog lazy is not very nice."

When creating this inverted list, assume the ad-hoc system generating the inverted list simply treats words independently, and applies a greater importance to less common words.

To make the answer shorter, I was a bit more aggressive with the stopword list here. Stopwords: "the", "over", "but", "why", "did", "not", and "very"

Stemming: "*ed" → "*" and "*ing" → "ing"

See the table below. Note that, depending on stopwords and stemming, the number of terms might vary.

Term ID	Term	Documents
1	quick	D1
2	brown	D1
3	fox	D1, D2
4	jump	D1, D2
5	lazy	D1, D3
6	dog	D1, D2, D3
7	call	D3
8	nice	D3

6 Relevant Real-World Applications

The infamous “YouTube Recommended Algorithm” has been the subject of many a YouTube content creator’s woes, and the subject of many a viewer’s comment (“Like this comment if this video was in your recommended 9 years later”). Where does the YouTube Recommended Algorithm fit into the IR field? Is it most likely Ad-Hoc or Filtering? And what kind of ad-hoc/filtering algorithm is it? Provide reasoning for all assertions.

7 “Right” and “Wrong” in Information Retrieval

If one were to look up “Vaccinations Dangerous”, what would one expect to be the subject matter of the majority of the resulting documents? Now, if one were to look up “Vaccinations Not Dangerous”, how would that change your results? What does this indicate about the ‘correctness” of information retrieval systems?

My information need for the first query would probably be articles talking about hazards or side-effects of vaccines, whereas the second would be articles claiming that reports of problems with vaccines were erroneous. However, typical ad-hoc retrieval systems would treat “not” as a stopword, and give me the same results for both queries.