Week 7 – Information Retrieval WS 15/16

Johannes Jasper(755292)
Norman Rzepka(754644)

December 2, 2015

Assignment 1

a)

Have you ever used any other search engine except for Google? Try using Yippy and Carrot2 with patent-related terms. These search engines provide a list of relevant topics to the query terms along with the corresponding pages of each topic. What do you think about the clustering results? Are the topics of the clusters sufficiently relevant to your query?

For generic queries (such as 'mobile device') the sites provide a wide range of related topics and categories. These help to specify the query and narrow it down. One has to filter the suggestion, though, since some suggested topics do not match the initial intent. Moreover, the more specific the query gets the more erratic the suggested topics become. I would therefore assume that such sites can help *human* users to extend their query. Using them automatically would require a deep understanding of the information need.

b)

Consider the characteristics of the patent data that you either discovered so far in the U.S. patent xml files or in patent-related material on the Web. Can you think of a user-friendly way to present your search engine results clustered? For example, based on the names of the authors, or the classification of each patent or the content of the titles? (hint: there are many xml elements apart from the title and the abstract that could be useful for the patent clustering)

The most obvious candidate for clustering is the *category* tag. Besides those, US patents have a *kind code*[1] that describes how different patent relate to each other. Author and the authors affiliation (i.e. the company/institution he/she works for) might be relevant categories as well. A relevant piece of information might also be the expiry date of a patent.

[1] `http://www.uspto.gov/learning-and-resources/support-centers/electronic-business-center/kind-codes-included-uspto-patent`

Assignment 2

Consider some patent-related queries. Compare the result pages of Google and Bing for the above queries.

a)

What elements are similar, different, the same?

Both engines rank the Wikipedia article first, thus giving a generic description of the topic. They both provide a summary extracted from the Wiki article in a info box outside the scope of the usual search results. Google provides some images relevant to the query. Bing on the other hand displays a list of related queries (for instance ”Access Control System”, ”User Access Control Windows 7” for the query ”access control”). Bing allows more explicit control of language and location with its ”Narrow by language” and ”Narrow by region” functions.

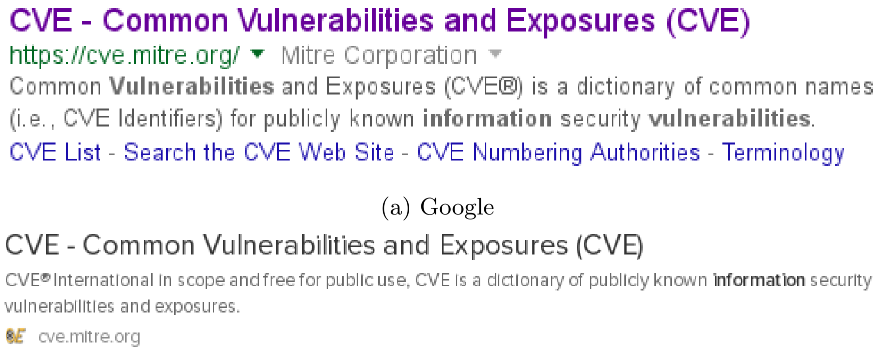
For the query ”computer-readable media” both search engines ranked an article from USPTO first. After that however Google used synonyms and different grammatical versions for a term to yield the Wikipedia article on ”Machine-readable medium”, whereas Bing listed a couple of more patent specific articles before altering the query slightly.

One thing I noticed was that Bing ranks results from Microsoft, MSDN or about Microsoft products very high, even if they are hardly related.

b)

Find two different result snippets for the same page and say which one you prefer and why.

Looking for ”vulnerability information” I found the 2 snippets depicted in Figure 1. Both show the first sentence on the website, Google however extended



CVE - Common Vulnerabilities and Exposures (CVE)

https://cve.mitre.org/ ▾ Mitre Corporation ▾

Common **Vulnerabilities** and Exposures (CVE®) is a dictionary of common names (i.e., CVE Identifiers) for publicly known **information** security **vulnerabilities**.

CVE List - Search the CVE Web Site - CVE Numbering Authorities - Terminology

(a) Google

CVE - Common Vulnerabilities and Exposures (CVE)

CVE® International in scope and free for public use, CVE is a dictionary of publicly known **information** security vulnerabilities and exposures.

cve.mitre.org

(b) Bing

Figure 1: Results for ”vulnerability information”

the content by substituting abbreviations. Moreover Google ads direct links to useful functions or sub pages such as the CVE list or an internal search. Though this clutters the display it makes Google more efficient to use which I would prefer.

Assignment 3

a)

Explain in a few words how your snippet generation method works in your pdf file.

We generate snippets by looking for the query terms in the search result documents. We fetch 5 words before and 5 words after the query terms in the document and merge overlapping text ranges. After that we shorten these text ranges to fit in sentence boundaries.

b)

See results.txt