Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

Week 7: Web Search

Assignment 1: Web Retrieval vs. Classical IR

a) What are the main differences between Web Retrieval and Classical IR?
The challenges of IR in regards to the web lie in the following areas:

- The sheer **size** of the internet (with multiple trillion sites currently indexed) makes information retrieval in this data set a non-trivial task.
- Websites are connected through **links**, which not only enables crawling for discovery of new content, but also has meaning for document relevance etc.
- Due to the large amount of money that can be earned online, there are lots of shady activities in order to fool search engines and users into visiting certain websites (**spam**, black-hat SEO etc.)
- Data in the internet is **heterogeneous**. This creates many problems, as information retrieval systems have to deal with different genres, languages, document types (text, images, videos etc.), but also a very diverse set of information needs that is to be satisfied.

b) What are the types of queries you find in the Web? Give an example for a query type.
There are several types of queries to be found on the internet, the most important of which are navigational queries (users searching for a website, such as "youtube"), informational queries ("how tall is the tv tower", "best audio software") as well as transactional queries ("sparta movie download").

c) What are reasonable business models for Web Search? Describe two scenarios where different business models would make sense.

1. Traditionally, web search engines tend to make money through **advertisements**. For money, they display links next to corresponding search queries. Due to the relevance of these ads to the user queries, this tends to be a rather successful approach.
2. **Subscriptions** work well for topical search engines that focus on doing one thing really well, and have a well-defined target userbase.

Assignment 2: (Programming) Web Search

Print the ranking of Wikipedia titles and snippets in Wikipedia together with its NDCG value for the queries:
a) "Anschluss Luhmann"
- -
b) „Actinium"
- -

Assignments Information Retrieval (SS 2014)

Stefan Bunk 756006
Franz Liedke 757081
Cornelius Bock 754432

b) "Information Retrieval"

- -

[The creation of the index took so long that we were unable to test these queries. We have not yet found a way to speed things up significantly. We will try to provide our results as part of the next assignment.]