1 Web IR

1. What are the most important differences between Web IR and traditional IR systems? Explain the differences with respect to links, queries, users, documents, and context.

2. What are the types of search queries in Web IR? Give examples for each type. Which of these types apply to traditional IR systems?

3. Explain the following “Bow Tie” diagram. Where should a crawler start to catch as many pages as possible?

![Figure 1: Bow tie diagram of interlinked web pages](image)

4. How can we estimate the size of the Web using existing Web search engines?

5. How can we detect duplicate and near duplicate document on the Web?

6. Search engines $A$ and $B$ each crawl a random subset of the same size of the Web. Some of the pages in a subset are duplicates (same content, different URL). Assume that duplicates are distributed uniformly amongst the crawled pages. Further, assume that a duplicate is a page that has exactly two copies (no page has more copies). $n$ performs duplicate elimination and indexes only one copy of each duplicate page whereas $A$ does not. If 45% of $A$’s indexed URLs are present in $B$’s index, while 50% of $B$’s indexed URLs are present in $A$’s index, what fraction of the web consists of pages that do not have a duplicate?