CS 276B: Sample midterm questions

1. Web size estimation:

(a) When estimating the size of the web using a random walk on web pages, we do not use the "teleport to a random web page" operation used in the pagerank computation. Why not?

(b) Consider the capture/recapture strategy for estimating the size of the web. We draw a random number of pages from one search engine and check whether they are indexed in a second search engine. Consider the test for whether the chosen page is present in another engine. List at least two sources of bias in this test.

(c) Two web search engines A and B each generate a large number of pages uniformly at random from their indexes. 30% of A's pages are present in B's index, while 50% of B's pages are present in A's index. What is the number of pages in A's index relative to B's?

(d) Now let us consider a scenario in which we use two crawls to estimate the frequency of duplicates on the web. Web search engines A and B each crawl a random subset of the web of the same size. Some of the pages crawled will be duplicates – exact textual copies of each other at different URLs. Assume that duplicates are distributed uniformly amongst the pages crawled by A and B. Further, for this exercise we will define a duplicate as a page that has exactly two copies. No pages have more than two copies. A indexes pages without duplicate elimination whereas B indexes only one copy of each duplicate page. [Note: the two random subsets have the same size *before* duplicate elimination.] 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?

2. Web graph

The in-degree of web pages follows a distribution similar to Zipf's law: the kth most "popular" page has in-degree proportional to 1/(k^2.1). As the largest in-degree value goes to infinity, does the fraction of pages with in-degree k grow, stay the same, or diminish?

Solutions

1. Web size estimation

a. Pagerank is run on a set of known web pages that are already crawled. We can teleport since we can randomly select pages from this set. [Estimating the size of this subset of the web is trivial: just count.] The point of web size estimation is that we don't already have an exhaustive list of the pages of interest. So we can't teleport.

b. 1) We assume that the population search engines draw from is the whole web. It is not: for example, nonconnected, non-submitted pages are not part of the population.

2) We assume that each page has the same probability to be drawn. That is not correct: popular/important pages (yahoo.com) are more likely to be drawn than obscure pages.

3) duplicates

c. |A| = 5/3 |B|

d. The size of A is pN if p is the sampling rate and N the size of the web. Let 1-2f be the proportion of non-duplicates.

Let B be the crawl with dups.

Let B' be the dup free version of B.

Deterministically, we know

(1) |A| = |B| = pN

(2) |A^B' | = .45 |A| = .45 |B| = .50 |B' |

so that

(3) |B' | / |B| = 0.90

We expect:

(4) We expect that |A^B' | = p * 0.9 |B| = p * 0.9 * pN = 0.9p^2N

Using (1) & (2), we have that 0.9p^2N = 0.45 pN, so that

(5) p = 0.5

We now have p, but still need to find f:

We expect:

B to have 2fp^2N dups with both copies in B, so that fp^2N pages get dropped. But using (3), the number dropped must be 0.10|B|, so that fp^2N = 0.10|B| = 0.1pN, so that f = 0.1/p = 0.1 / 0.5 = 0.2

So f = 0.2. 1-2*f = 60% of the web has no duplicates.

2. It stays the same. Here the frequency at k is $\frac{\frac{1}{k^{2.1}}}{\sum_{i=1}^{n}\frac{1}{i^{2.1}}}$.

The denominator converges to a constant for large n, hence the frequency at k stays unchanged as n approaches infinity.