Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Marjan Celikik

Search Engines WS 09/10

http://ad.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 5

complete until Tuesday, November 24th

Exercise 1

Write a program for intersecting two inverted lists with the basic linear-time algorithm. You may take your code from Exercise Sheet 1, but go over it again and try to make it as fast as possible. Describe at least one improvement you did, and say how much it improved the running time. (See Exercise 3 for which measurements should be done.)

Exercise 2

Write a program for intersecting two inverted lists with the asymptotically optimal exponential/ binary-search algorithm from the lecture. You only get points for this exercise, if in your measurements from Exercise 3 you check that this algorithm gives exactly the same result as your algorithm from Exercise 1.

Exercise 3

For both of your algorithms, the one from Exercise 1 and the one from Exercise 2, measure the following: running time, the ratio number of comparisons / sum of number of elements in the two lists, million elements processed per second, MB processed per second. Use randomly generated lists, that is, to generate a list of (expected) size m with doc ids from $1..n$, pick each doc id with probability m/n. Pick $n = 10^8$ and for m pick all combinations of $10^3, 10^4, 10^5, 10^6$ for your two lists, that is, for each algorithm and for each of the four measures above there should be 16 measurements. Repeat each measurement ten times and form the average. For each algorithm print your 16 averages in a 4×4 table.

Exercise 4

Compare the two tables from Exercise 3. That is, identify the (significant) differences and similarities and then try to explain them. When you cannot make sense of something, keep in mind that there might be bugs in your code due to which you get weird results. The ideal goal of any discussion of a series of experiments is that the results make perfect sense and one can explain that perfect sense. Try to achieve that goal.

Exercise 5* (Optional / Bonus)

Prove that for positive integers k and m with $k \leq m$, it holds that $k \cdot \ln(1 + m/k) \leq k + m$. That is, the exponential/binary-search algorithm is never worse than the straightforward linear algorithm, at least asymptotically.