Information Retrieval

Tutorial

1. Recommend a query processing order for the query:
   (tangerine OR trees) AND (marmalade OR skies) AND (kaleidoscope OR eyes)
   given the following postings list sizes:

	Term	Postings size
a.	eyes	213312
b.	kaleidoscope	87009
c.	marmalade	107913
d.	skies	271658
e.	tangerine	46653
f.	trees	316812

2. Are the following statements true or false?

   a. In a Boolean retrieval system, stemming never lowers precision.
   b. In a Boolean retrieval system, stemming never lowers recall.
   c. Stemming increases the size of the vocabulary.
   d. Stemming should be invoked at indexing time but not while processing a query.

3. We have a two-word query. For one term the postings list consists of the following 16 entries:

   [4,6,10,12,14,16,18,20,22,32,47,81,120,122,157,180]

   and for the other it is the one entry postings list:

   [47]

   How many comparisons would be done to intersect the two postings lists with the following two strategies.

   a. Using standard postings lists
   b. Using postings lists stored with skip pointers, with a skip length of $\sqrt{P}$

4. Compute the Jaccard coefficients between the query 'bord' and
   a. border
   b. lord
   c. morbid
   d. sordid

Solutions

1. Using the conservative estimate of the length of unioned postings lists, the recommended order is: (kaleidoscope OR eyes) (300,321) AND (tangerine OR trees) (363,465) AND (marmalade OR skies) (379,571) However, depending on the actual distribution of postings, (tangerine OR trees) may well be longer than (marmalade OR skies) because the two components of the former are more asymmetric. For example, the union of 11 and 9990 is expected to be longer than the union of 5000 and 5000 even though the conservative estimate predicts otherwise.
   Time for processing :
   (i) (tangerine OR trees) = O(46653+316812) =O(363465)
   (ii) (marmalade OR skies) = O(107913+271658) = O(379571)
   (iii)(kaleidoscope OR eyes) = O(46653+87009) = O(300321)
   Order of processing:
   a. Process (i), (ii), (iii) in any order as first 3 steps (total time for these steps is O(363465+379571+300321) in any case)
   b. Merge (i) AND (iii) = (iv): In case of AND operator, the complexity of merging postings list depends on the length of the shorter postings list. Therefore, the more short the smaller postings list, the lesser the time spent.
   The reason for choosing (i) instead of (ii) is that the output list (iv) is more probable to be shorter if (i) is chosen. c. Merge (iv) AND (ii): This is the only merging operation left.

2. a. False. Stemming can increase the retrieved set without increasing the number of relevant documents.
   b. True. Stemming can only increase the retrieved set, which means increased or unchanged recall.
   c. False. Stemming decreases the size of the vocabulary.
   d. False. The same processing should be applied to documents and queries to ensure matching terms

3. a. Applying MERGE on the standard postings list, comparisons will be made
   unless either of the postings list end i.e. till we reach 47 in the upper postings list, after which the lower list ends and no more processing needs to be done. Number of comparisons = 11
   b. Using skip pointers of length 4 for the longer list and of length 1 for the shorter list, the following comparisons will be made: 1. 4 & 47 2. 14 & 47 3.22 & 47 4. 120 & 47 5. 81 & 47 6. 47 & 47 Number of comparisons = 6

4. Jaccard co-efficients between the following terms:
   a. bord and border : 3/5
   b. bord and lord : 2/4
   c. bord and morbid : 1/7
   d. bord and sordid : 2/6