Information Retrieval

Weike Pan

Exercise 3.1

In the permuterm index, each permuterm vocabulary term points to the original vocabulary term(s) from which it was derived. How many original vocabulary terms can there be in the postings list of a permuterm vocabulary term?

ANSWER: One.

If there is no terminal symbol $, in the postings list of a permuterm vocabulary term, there can be more than one original vocabulary terms.

For example, for two original vocabulary terms leaf and flea, we can have a same permuterm vocabulary term.

Exercise 3.2

Write down the entries in the permuterm index dictionary that are generated by the term mama.

ANSWER: mama$, ama$m, mama, amam, $mama.

Exercise 3.3

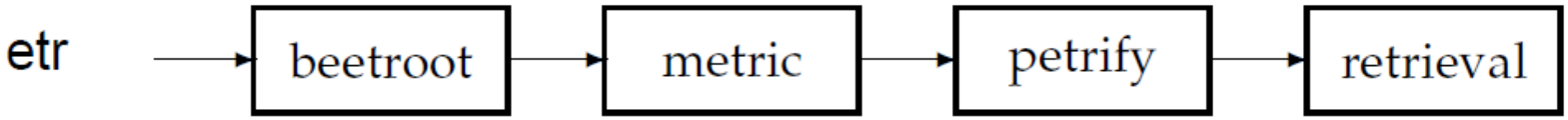
If you wanted to search for s*ng in a permuterm wildcard index, what key(s) would one do the lookup on?

ANSWER: ng$s*.

Exercise 3.4

Refer to Figure 3.4; it is pointed out in the caption that the vocabulary terms in the postings are lexicographically ordered (按字典顺序). Why is this ordering useful?

ANSWER: A lexicographic ordering will make the merging of the two k-grams lists efficient, i.e. O(x+y) steps, where x and y are the sizes of the two lists.



► **Figure 3.4** Example of a postings list in a 3-gram index. Here the 3-gram etr is illustrated. Matching vocabulary terms are lexicographically ordered in the postings.

Exercise 3.5

Consider again the query fi*mo*er from Section 3.2.1. What Boolean query on a bigram index would be generated for this query? Can you think of a term that matches the permuterm query er$fi* in Section 3.2.1, but does not satisfy this Boolean query?

ANSWER: The Boolean query is $f AND fi AND mo AND er AND r$.

The term filibuster will match the permuterm query er$fi* in Section 3.2.1, but does not satisfy this Boolean query.

Exercise 3.6

Give an example of a sentence that falsely matches the wildcard query mon*h if the search were to simply use a conjunction of bigrams.

ANSWER: His personality is *moonish*.

Exercise 3.7

If |S| denotes the length of string S, show that the edit distance between s1 and s2 is never more than max{|s1|, |s2|}.

Exercise 3.8

Compute the edit distance between paris and alice. Write down the 5*5 array of distances between all prefixes (前缀) as computed by the algorithm in Figure 3.5.