Übung: Tolerant Retieval, Index Construction

November 14, 2009

K-gram Index AND-Query

- Consider the query "fi*mo*er". What Boolean query on a bigram index would be generated for this query?
- Benutzen Sie als Start- bzw. End-Symbol $

Jaccard Koeffizient

- Berechnen Sie die Jaccard Koeffizient zwischen dem Frageterm "bord" und den Termen:
  - "border", "borderline", "lord", "morbid" und "sordid"
- Der Jaccard Koeffizient (zwischen den Mengen A und B) ist definiert durch $|A \cap B|/|A \cup B|$
- Die Mengen seien hier jeweils die "2-grams" der Terme
- Was ist der Vorteil des Jaccard Koeffizient als Maß der String-Ähnlichkeit für eine "unscharfe Suche" bzw. "Rechtschreibkorrektur" gegenüber der Anzahl der übereinstimmenden k-grams?

Naive Phrasenkorrektur

Consider the four-term query "catched in the rye" and suppose that each of the query terms has five alternative terms suggested by isolated-term correction. How many possible corrected phrases must we consider if we do not trim the space of corrected phrases, but instead try all six variants for each of the terms?

Soundex

- Berechnen Sie den Soundex Code von:
  - "Chebyshev", "Tchebycheff"
  - "Mary", "Mira"
- Was folgern Sie daraus?

Indexaufbauzeit für (naives) Sortieren auf der Festplatte

If we need n log2 n comparisons (where n is the number of termID-docID pairs) and 2 disk seeks for each comparison, how much time would index construction for Reuters-RCV1 take if we used disk instead of memory for storage and an unopti- mized sorting algorithm (i.e., not an external sorting algorithm)? Use the system parameters of page 5 of the lecture slides.

Prinzip BSBI

Gegeben sind (beispielhaft) folgende termID-docID Paare als Output der Dokumentenverarbeitung:
(1,1), (2,1), (3,1), (4,1), (5,1), (6,2), (7,2), (2,2), (3,2), (1,3), (3,3), (5,3), (7,3), (3,4), (2,4), (1,4), (7,4)
Simulieren Sie die Index-Konstruktionschritte des BSBI unter der Annahme das maximal sechs termID-docID Paare in den Speicher passen.

- Sortieren innerhalb der Blöcke
- Merge der Blöcke