NBT EVO FTS Tokenizer Investigation

Mar. 19， 2014

Bi-gram and Pre-segment Tokenizer(1)

Name	Pre-segment	Bi-gram
延安高架路 (Yan’an Elevated Road)	延安 (Yan’an) 高架路 (Elevated Road)	延安 (Yan’an) 安高 (No meaning) 高架 (Elevated) 架路 (No meaning)

北京首都国际机场(Beijing Capital International Airport)	北京(Beijing)
	首都(Capital)
	国际(International)
	机场(Airport)
		北京(Beijing)
		京首(No meaning)
		首都(Capital)
		都国(No meaning)
		国际(International)
		际机(No meaning)
		机场(Airport)

Bi-gram and Pre-segment Tokenizer(2)

	Pre-segment	Bi-gram
Raw Data Team’s Job	Provide the pre-segment tokenizing result. 延安|高架路 北京|首都|国际|机场	No job
Compiler Team’s Job	1) Convert pre-segment tokenizing result to NDS FTS table. 2) FTS4AUX table is needed	1) Using bi-gram tokenizer to tokenize the names and fill the NDS FTS table. 2) FTS4AUX table is not needed.
Application Team’s Job	Using FTS4AUX to tokenize input string.(not realized by NDS association so far)	No additional job needed.

Comparison (1)

- Data

  Beijing POI names (total count: 30278)

- Method

  (1) bi-gram

  ```
  CREATE VIRTUAL TABLE fts USING fts4(name, tokenize=ndsunicode61);
  INSERT INTO fts SELECT name FROM BJ.poi
  SELECT name FROM fts WHERE fts MATCH ‘XXXXX’
  ```

  (2) pre-segment

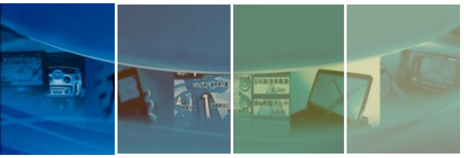
  ```
  CREATE VIRTUAL TABLE fts USING fts4(name);
  CREATE VIRTUAL TABLE fts_terms USING fts4aux(ft);
  INSERT INTO fts SELECT pre_segment_name FROM BJ.poi
  SELECT name FROM fts WHERE fts MATCH ‘XXXXX’
  ```

- Testing Environment

  PC(32 bit, CUP 1.8G, Memory 4G), SqliteSpy

Comparison (2)

	Bi-Gram	Pre-segment
Size	25,476KB	20,960KB
Query ('首都‘)	7.9-42ms (5 times)	2.45-9.8ms(5 times)
Query ('首都机场‘)	2.85-16.30ms (5 times)	1.54-9.73ms(5 time) 北京首都国际机场 was not found because tokenizing input string was still not realized.
Query (‘都国‘)	2.02-8.64ms (5 times) 北京首都国际机场 was found, but is it reasonable?	1.20-4.47ms(5 time) 北京首都国际机场 was not found because 都国 is not a word.



Mapping your way

NAVINFO 四维图新

mapscape

Vielen Dank

Merci

谢谢

Dank u wel

Thank you

ありがとうございます