Space-Constrained Gram-Based Indexing for Efficient Approximate String Search

Tech ID: 20771 / UC Case 2007-322-3

BRIEF DESCRIPTION

Researchers at the University of California, Irvine have developed two new compression techniques called “DiscardLists” and “CombineLists” to optimize gram-based indexes.

FULL DESCRIPTION

Many information systems need to support approximate string queries. For example, given a collection of textual strings, such as person names, telephone numbers, and addresses, we want to find the strings in the collection that are similar to a given query string. Many existing algorithms use gram-based inverted indexes to answer approximate string queries. These indexes can be notoriously large compared to the size of their original string collection and the large index size causes problems for applications. Nonetheless, high efficiency of applications often requires these indexing structures to reside in main memory. Therefore, methods to reduce the index size while retaining high query efficiency are desired.

ADVANTAGES

Researchers at the University of California, Irvine have developed two techniques called “DiscardLists” and “CombineLists” to optimize gram-based indexes. The DiscardLists technique removes some of the lists in the index. A cost-based algorithm is proposed to iteratively select lists to discard based on its effect on the average query performance for a given query workload if it is discarded. In each iteration, the algorithm needs to evaluate the effect of discarding any of the remaining lists while considering all previously discarded lists. The effect includes the number of new false positives, the cost of processing the reduced set of lists for each query, and the reduction of the total index size. The CombineLists technique combines some of the correlated lists to reduce the index size.

The experimental results of the techniques show that while the index size can be safely reduced up to 60%, query performance even improves for lower compression ratios (10%-40%) as compared to uncompressed indexes.

STATE OF DEVELOPMENT

Software implementing algorithms have been written and tested on real data.

INTRODUCTION

PATENT STATUS

Country	Type	Number	Dated	Case
United States Of America	Issued Patent	7,996,369	08/09/2011	2007-322

CONTACT

Ben Chu
ben.chu@uci.edu
tel: .

OTHER INFORMATION

CATEGORIZED AS

- » **Communications**
  - » Internet
  - » Other
- » **Computer**
  - » Software

RELATED CASES

2007-322-3, 2007-322-2

UCI Beall
Applied Innovation

5270 California Avenue / Irvine,CA
92697-7700 / Tel: 949.824.2683