CS848 Paper Submission Site

Go To CS848 Site

Submit or Update A Review For Paper #25

Go To CS848 Submissions

It is currently Thursday 28th of October 2004 10:34:44 PM EDT

Paper # 25 (Download paper of type application/pdf, 724744 bytes)	
Title:	CORDS: Correlation Detection by Sampling
Abstract:	

You have already finalized your review for this paper.
You can no longer modify it, but you may view it.

If you made a mistake in your review and you want it "unfinalized",
you may send mail to the program chair asking them to unfinalize it

Send yourself this review by email

Attribute	Value
Are you finished with this review?	Finalize, I am done editing
Provide a short summary of the paper	This paper presents CORDS, a method for detecting correlations between column data in a DBMS. The authors begin by presenting some introductory information about CORDS, as well as a discussion of related work in the field. The authors begin their discussion of CORDS by presenting the method used to find dependencies. The algorithm begins by generating candidate pairs of columns for key columns. Some heuristics are applied to prune the potentially large search space. The authors then present their method of statistically sampling for a Chi Square test. The authors then present their algorithm for performing this process. The authors present ways that CORDS can be used to improve query optimization. CORDS data can be used to avoid plans that are very expensive (by orders of magnitude) that result from incorrect independence assumptions, among other things. The authors also present a method to determine which column-group statistics should be recommended for use by the optimizer as maintaining all possible correlations between columns is usually prohibitively expensive. The article next moves to a discussion of experiments using the CORDS implementation. CORDS properly detected the correlations that were present in manufactured data, and did not report any incorrect correlations. Overall, CORDS reduces the worst-case execution time by otders of magnitude, and the average case execution time slightly. Only a few queries experienced increases in execution time, and those increases were small when present. The authors wind down their paper with a discussion of how CORDS performed on some real-world data (i.e. census and car information). CORDS accurately detected many correlations between data items in both databases. This information could be used to save many orders of magnitude of incorrect execution time estimates.

	The authors conclude with some closing remarks.
What is the strength of the paper? (1-3 sentences)	This paper presents a novel way to use statistical sampling to detect statistical correlations among data in database columns.
What is the weakness of the paper? (1-3 sentences)	This paper fails to present some performance information (see below). Although the idea is novel, it is slow and may not work well in practice with tables of smaller number of rows. Also, it is difficult to read and understand due to the highly technical presentation.
Your qualifications to review this paper	I know the material, but am not an expert
Writing Quality	Average
Relevance to query processing?	The paper is relevant to query processing
Experimental Methodology	**Good**
Novelty of paper	This is a new contribution to an established area
Overall paper merit	**A novel or new contribution to this area with good methodology, or an incremental contribution paper that has excellent methodology. A must read for anyone in the area.**
In your opinion, will this paper be important over time?	Good
Provide additional detailed comments to the author	You have presented a great way to determine correlations between dependent columns. Your method will surely benefit many existing and forthcoming systems. However, I have some issues with your presentation. -The algorithm appears to be slow (i.e. taking n^2 time to enumerate all the possible combinations. -The algorithm requires a large sample size. Would it be more appropriate to use the statistical relevance formula for a representative sample size? What about tableswith smaller numbers of tuples? -Although you provide the number of correlations and the amount of time saved by the CORDS system with census and car databases, you do not discuss the accuracy of your system in detecting these correlations. This is important information that should have been provided. -Your paper is very technical in nature and is thus difficult to read in certain parts. This reader got lost in the statistical theory discussions. Greater care should be taken to ensure your discussions are easy to understand.
Additional comments to PC (not seen by author)	**Although this paper has a few major flaws, I believe the concept is important enough to publish.**

Goto Main Index Close Window