Session 10:
In-RDBMS inverted indexes revisited

担当: 波多野 賢治 (同志社大学)

論文の概要

- 全文検索エンジン
  - inverted index が不可欠
- RDBMS
  - RDBMS 内に inverted index を実装するには細工が必要
    - 列指向クエリエンジン
    - 行指向クエリエンジン

使い分けたら
パフォーマンスが上がったよ！

行指向 (通常の RDBMS)

- 行をヒトカタマリのデータ
  - データの追加，削除，更新は行単位
  - ディスクへの読み書きも行単位

オンライントランザクション機能に優位性

研究者番号	研究者名	機関コード	所属	特徴
20362832	渡辺知恵美	12102	筑波大学	ほんわか
70314531	天笠俊之	12102	筑波大学	物静か
80263440	石川佳治	13901	名古屋大学	聡明
80314532	波多野賢治	34310	同志社大学	デカい

列指向 (カラム型 RDBMS)

- 列をヒトカタマリのデータ
  - 列ごとにまとめて処理するのに特化

列ごとの集計処理，データ圧縮，並列処理に優位性
ただ，オンライントランザクション機能は苦手

研究者番号	研究者名	機関コード	所属	特徴
20362832	渡辺知恵美	12102	筑波大学	ほんわか
70314531	天笠俊之	12102	筑波大学	おっとり
80263440	石川佳治	13901	名古屋大学	聡明
80314532	波多野賢治	34310	同志社大学	デカい

inverted index (転置インデックス)

文書基準

Doc ID	words
0	{渡辺,知恵美,筑波,…}
1	{天笠,俊之,筑波,大学,…}
2	{石川,佳治,名古屋,大学,…}
	. . .

語基準 (転置インデックス)

word	Doc IDs
渡辺	{0}
筑波	{0,1}
大学	{0,1,2}
	. . .

- 検索ワード {筑波,大学}
  - {0,1} ∩ {0,1,2} = {0,1}
- RDBMS での inverted index の格納
  - <word, Doc ID, tf, df, doclen, offset>

ZigZag Merge JOIN

- inverted index から行指向/列指向データの作成，マージ結合
  - 行指向 (赤字の属性ごとにクラスタ化，B-tree)
    - <word, Doc ID, tf, df, doclen>
    - <word, Doc ID, offset>
  - 列指向
    - word, Doc ID, offset をソートして格納
    - B-tree 等は用いずフラットファイルで扱う

語基準と文書基準のデータをマージするため
結合の様子が ZigZag に見える

ソートされた列指向データを持ちいてマージ結合するので高効率
列指向データが圧縮されているのでさらに高効率