【VLDB 2013勉強会】

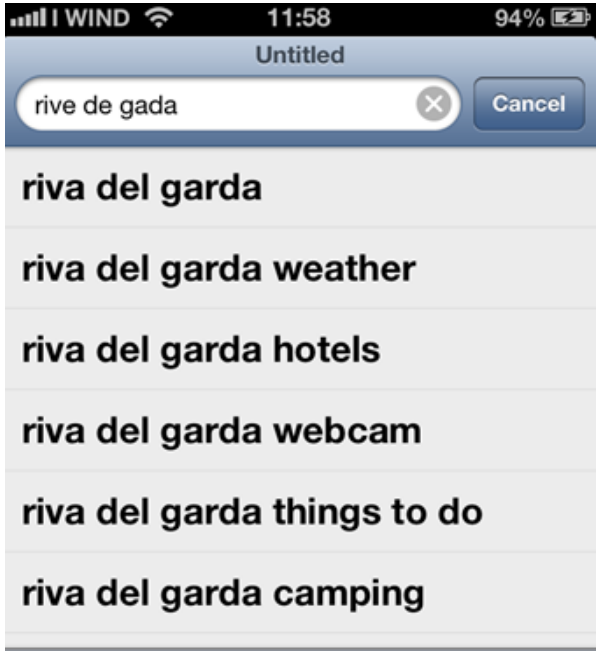
Session 12: Spatial and Text

担当: 石川佳治

Some figures are copied from VLDB 2013 proceedings.

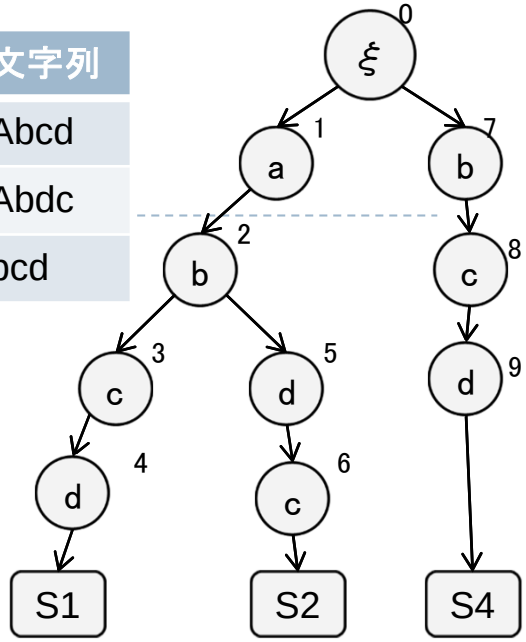
Efficient Error-tolerant Query Autocompletion

- C. Xiao, J. Qin, W. Wang, Y. Ishikawa, K. Tsuda, K. Sadakane (Nagoya U./UNSW/NII/AIST)
  - Chuan Xiao（肖 川）：石川研ポスドク
- サーチ処理におけるキーワード補完
  - 入力文字列の誤りも許す
- 編集距離に基づく接頭辞探索 (Edit Prefix Search)
  - 例：編集距離の閾値 $\tau = 1$
  - 検索文字列 q = “abc”
  - 文字列集合 S = {“acdefg”, “cda”, …}
  - “acdefg”は条件にマッチ：接頭辞の“ac”についてED(“abc”, “ac”) = 1が成立



アプローチ

ID	文字列
S1	Abcd
S2	Abdc
S4	bcd



- 既存手法
  - 文字列集合 S についてトライ(trie)を構築
  - 検索時には編集距離を考慮しながら探索
  - 問題点：多数のノードへのアクセス $O(|\Sigma|^{\tau})$
- 本論文の貢献
  - 新たなデータ構造(Deletion Variants Trie)の提案
  - Edit Prefix Search問題をExact Prefix Search問題に変換
  - 1,000倍の高速化(ただし20倍の索引サイズ)
- Deletion Variantの概念を利用：s = “abcd” の場合
  - 0-variants：abcd {}
  - 1-variants：bcd {1}, acd {2}, abd {3}, abc {4}
  - 2-variants: cd {1,1}, bd {1,2}, bc {1,3}, ad {2,2}, ac {2,3}, ab {3,3}

4番目の文字を削除

アプローチ（続き）

- Deletion Variantを用いた判定
  - 例：s = “abcd”, q = “abxd”のとき，両者とも (abd, {3}) というDeletion Variantを持ち，τ = 1の編集距離
  - 性質：二つのDelation Variants (x, Dx), (y, Dy)があるとき，x = yかつ $|Dx \cup Dy| \leq \tau$ ならED(s, q) $\leq \tau$
- Deletion Variants Trie：s = “abcd”でτ = 2のとき，{abcd, #bcd, a#cd, ab#d, abc#, ##cd, #b#d, ･･･} をトライに挿入
- 検索処理：q = “abc”，τ = 2のとき
  - τ = 2を考慮し，abc, #bc, a#c, ab#, ##c, #b#, a## と展開
  - 検索文字列の列挙（それぞれを実際に探す文字列に展開）
    - a#cの場合，ac, a#c, #a#c, a##c, a#c#
  - 実際には，検索文字入力に応じて適応的に処理 $O(\tau \cdot (|q| + \tau)^{\tau})$