【VLDB2011勉強会】

Session 13: Human-Computer Interaction

担当：田島 敬史（京都大学）

Query Expansion Based on Clustered Results

by Ziyang Liu (Arizona State) et al.

目的

文書検索結果をクラスタリングしクラスタ毎に追加検索語を出す

- 従来の「クラスタラベリング」手法：各クラスタの代表的な語
- 提案手法：ちょうどそのクラスタが解となるような語を求める

問題設定

各クラスタに対し最適(F-measure)な語集合(and検索)を求める

課題

- NP-hard (APX-hard, 定数 bounded な近似さえできない)
- 効率よく求める近似手法を二つ提案
  - 単語を一つづつ greedy に選んでいく．精度は良いが遅い．
  - 成績の良い候補解の間を探索して解空間を効率よく探索
    早いが精度は悪い．

手法1

C: クラスタ, q: 現時点での C を表す検索式の候補

各語 k について，q に追加・削除した場合の利益と損失を定義

- 追加の利益 $|(R(q) - C - R(k)|$
  k を追加したら解から除外できる不適切解の数 (or スコアの和)
- 追加の損失 $|(R(q) \cap C - R(k)|$
  k を追加したら解から除外される適切解の数 (or スコアの和)
- 削除の利益 $|(R(q - \{k\}) \cap C - R(k)|$
  k を削除したら解に追加できる適切解の数 (or スコアの和)
- 削除の損失 $|(R(q - \{k\}) - C - R(k)|$
  k を削除したら解に追加される不適切解の数 (or スコアの和)

1. 利益/損失の比が最大のものを追加または削除
2. 各語の追加・削除の利益・損失を更新
3. 最大の利益/損失比が1以下になったら終了

手法2

F-measure(精度と再現率の調和平均)の最適化なので，結局，精度と再現率のバランス

例：

1. 精度 10%，30%，50%，70%，90% の各々について，再現率を最大化する検索式，$q_{10}, q_{30}, \ldots, q_{90}$ を見つける．
2. 隣接ペアの中で，q_{50} と q_{70} が最も成績の平均が良いとする．
3. 精度 50%，55%，60%，65%，70% の各々について，再現率を最大化する検索式，$q_{50}, q_{55}, \ldots, q_{70}$ を見つける．

「精度がほぼ30%で再現率が最大の検索式」の見つけ方？

1. 現在の検索式の解中の不適切解をランダムに選ぶ
2. その解を除外できる語で，利益/損失比が最大のものを追加
3. 精度がほぼ 30% になるまで繰り返す