

【ICDE2014勉強会】

Session 8:  
Keyword Search II:  
Privacy and Social Data

担当：吉國，柿澤（お茶大），岡田（筑波大）

\* 図は論文より引用

# On Masking Topical Intent in Keyword Search

Peng Wang and Chinva V. Ravishankar,  
(University of California-Riverside)

担当 : 柿澤(お茶大)

# On Masking Topical Intent in Keyword Search

(Peng Wang and China V. Ravishankar, University of California-Riverside)

---

- ▶ キーワード検索におけるユーザのプライバシー保護
  - ▶ キーワードを検索エンジンに与えると、ユーザがどんなことに興味を持っているかが分かってしまう
- ▶ この問題の解決策となるTIOスキーマ
  - ▶ 本物のクエリを、ダミークエリの中に混ぜる
  - ▶ ダミークエリを生成することで、ユーザの意図を隠す
- ▶ 既存のスキーマに対し、3つの攻撃モデルで効率とプライバシー保護度合いを比較 (割愛)
- ▶ ダミークエリ生成スキーマDGAの改良スキーマHDGAを提案
  - ▶ HDGAが効果的かつ効率の良い手法であることを示す

# On Masking Topical Intent in Keyword Search

(Peng Wang and China V. Ravishankar, University of California-Riverside)

## ▶ モデルとなるスキーマ

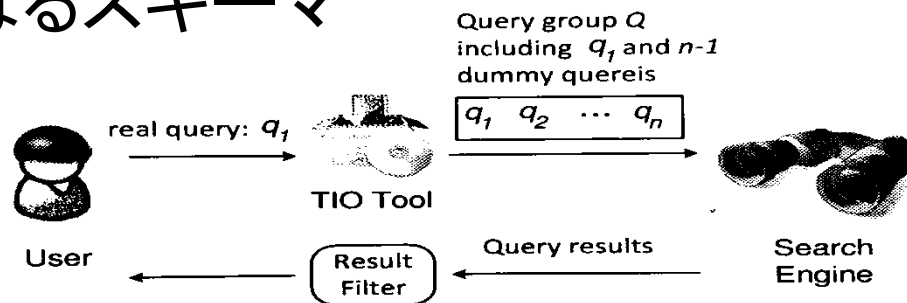


Fig. 1: Scheme model.

## ▶ Keyword-Based Obfuscation Scheme(KBO)

- ▶ ユーザは、クエリを検索エンジンに提供する
- ▶ ダミークエリが生成され、本物のクエリとともに一纏めにされる
- ▶ 検索エンジンは、ダミーを含む全てのクエリに対して結果を返す
- ▶ ダミークエリに対する結果が取り除かれる

## ▶ Topical Intent Obfuscation(TIO)

- ▶ KBOの要領で、さらにユーザの意図を隠すことができる

テキストベースの検索において、ユーザのプライバシーを保護した手法

# On Masking Topical Intent in Keyword Search

(Peng Wang and China V. Ravishankar, University of California-Riverside)

- ▶ Topical Intent Obfuscation(TIO)
  - ▶ 意味論的分類アルゴリズム(SCA)
    - ▶ LDA(Latent Dirichlet Allocation)によってドキュメントをトピックに分類
  - ▶ ダミークエリ生成アルゴリズム(DGA)
    - ▶ 本物のクエリとマッチするトピックを見つける
    - ▶ 別のトピックとマッチしたダミークエリを作って、本物のクエリを紛らわせる
    - ▶ 各ダミークエリ中のキーワードは、同じトピックにマッチする
    - ▶ このため、ダミークエリのキーワードは意味に一貫性を持つ



# On Masking Topical Intent in Keyword Search

(Peng Wang and China V. Ravishankar, University of California-Riverside)

## ▶ HDGA

### ▶ TIOのDGAの脆弱性を改良

- ▶ 意味の一貫したダミークエリを一度に生成する
- ▶ ダミークエリをグループ内で完全に独立に生成する
- ▶ ある2つのデータセットに対し、
  - ▶ クエリのグループサイズ
  - ▶ クエリ中のキーワード数

を変化させた時の、クエリグループ生成時間の推移

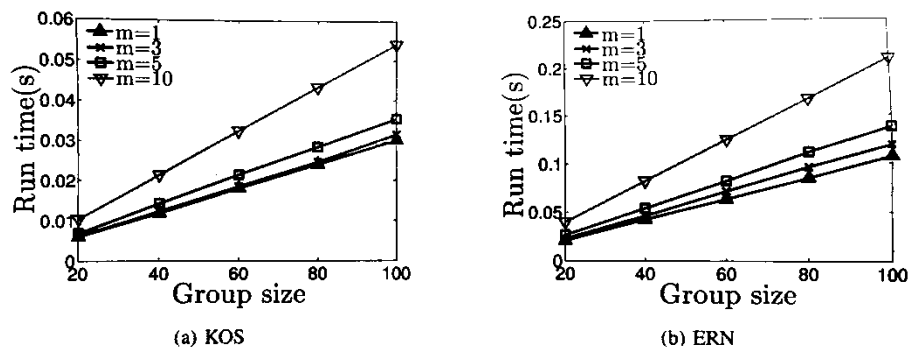


Fig. 13: Time used in query group generation.

非常に短い実行時間で  
HDGAを稼働できる

# Head, Modifier, and Constraint Detection in Short Texts

Zhongyuan Wang y, Haixun Wang , Zhirui Hu

担当 : 吉國(お茶大)

# 概要

- ▶ 検索クエリなどの短い文における“主要語”，“修飾語”，“制約語”を検出する手法の提案
  - ▶ クエリなどは文法無視が多い
  - ▶ 検索意図を正確にとるためにも必要な処理

popular    iphone 5s    smart cover

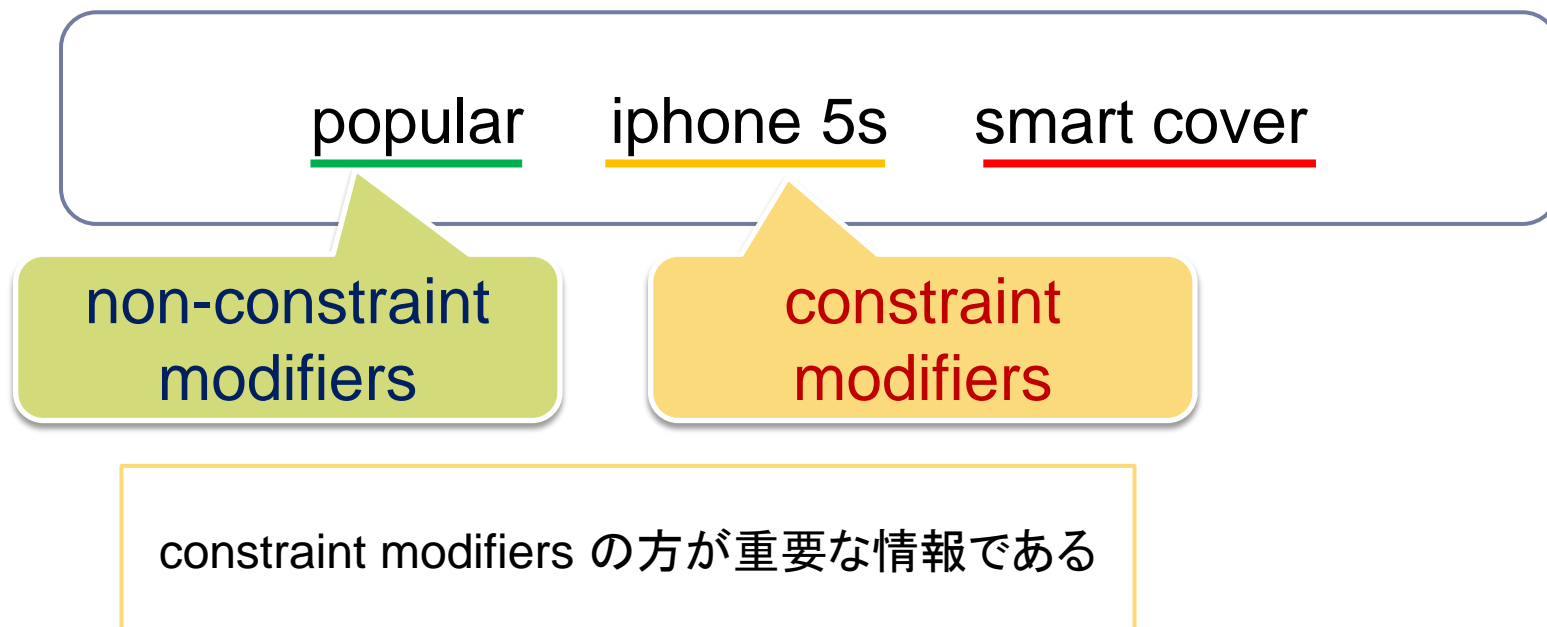
修飾語

主要語



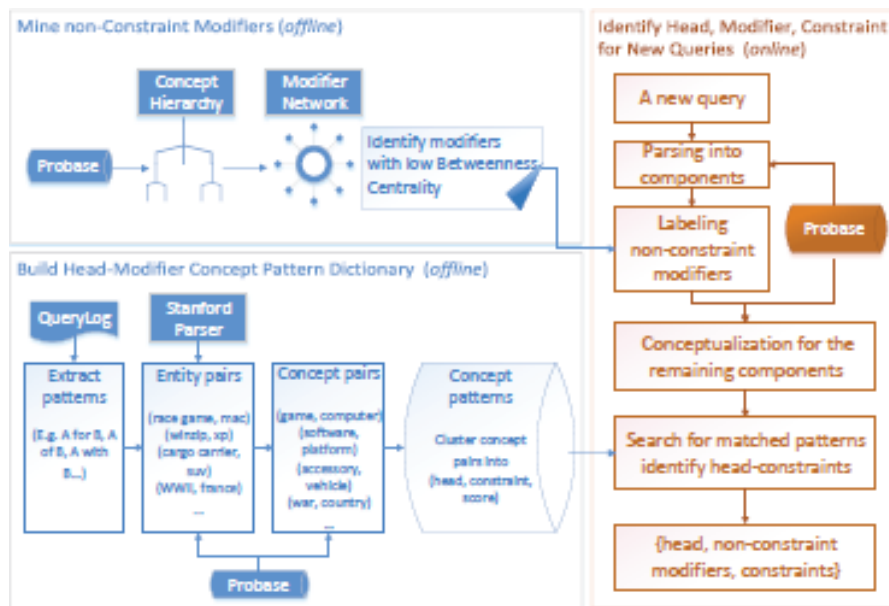
# 概要

- ▶ 検索クエリなどの短い文における“主要語”，“修飾語”，“制約語”を検出する手法の提案
  - ▶ クエリなどは文法無視が多い
  - ▶ 検索意図を正確にとるためにも必要な処理



# フレームワーク

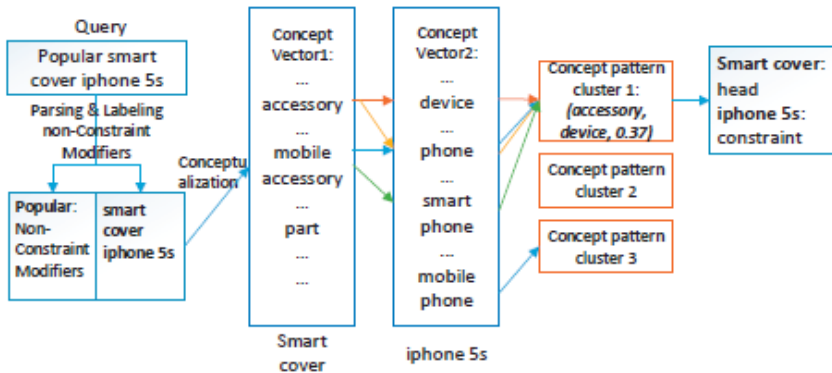
- ▶ head – modifier の関係を利用して検出していく
  - ▶  $(concept_{[head]}, concept_{[modifier]}, score)$
  - ▶ head – modifier パターン取得のために Probase を利用
    - ▶ entities と concepts のタグをもつ大きい単語ネットワーク



Framework for head, modifier, and constraint detection

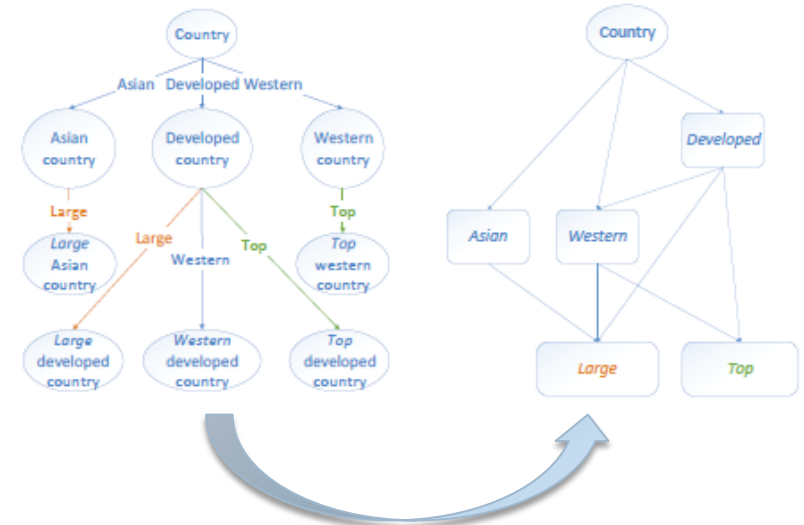
# Head and Modifier Detection

- ▶ 2つ以上の要素を持つクエリに対して検出する場合



- ▶  $(concept_{[head]}, concept_{[modifier]}, score)$
- ▶  $score$  の高い組合せパターンを検出
  - ▶ 数式は論文を参照してください

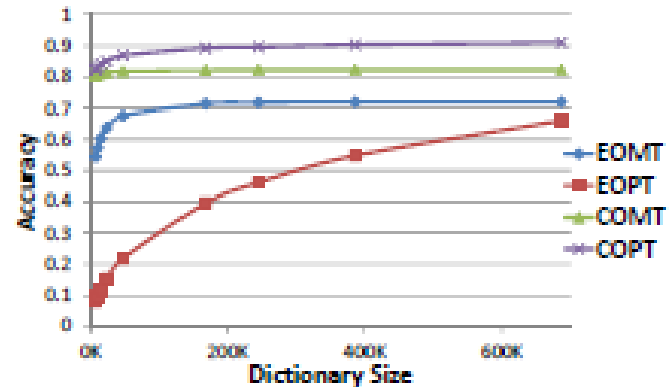
- ▶ non-constraint modifiers は Modifiers network において中心性の高いもととする



Concept hierarchy treeから  
Modifier Networkを生成

# Experiments

- ▶ BINGのクエリログ6か月分を利用して実験
  - ▶ さまざまな実験を行っているが一つだけ紹介



- ▶ 提案手法が一番精度が良くなった
- ▶ 辞書のサイズが小さくても精度が高い