

質問者のプライバシーを保護 する特許データベース検索 (研究紹介)

中川研 M2 胡 瀚林
指導教員：中川 裕志 教授

2016 年 7 月 1 日

① 背景紹介

② 既存研究

③ プライバシー分析

④ 参考文献

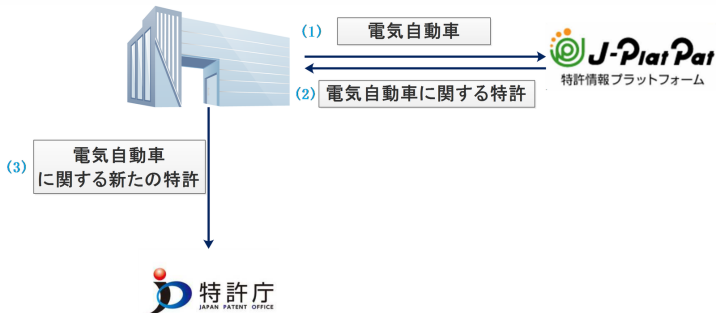
① 背景紹介

② 既存研究

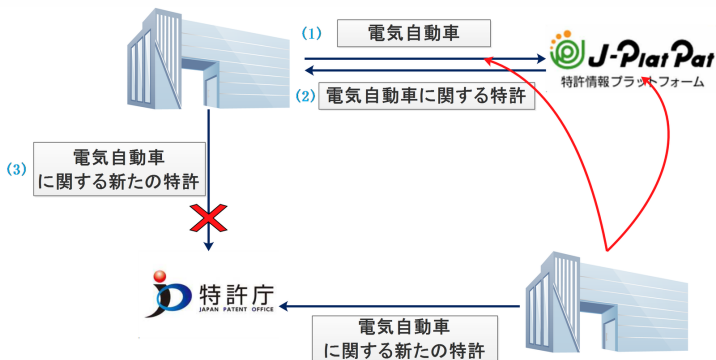
③ プライバシー分析

④ 参考文献

特許検索



特許検索



特許検索質問

メタノールを燃料とする車載用燃料電池システムおよび車

メタノール 水蒸気 反応 水素 透過 膜 自立 燃料 電池 システム 供給 ガス
アノード カソード 空気 排出

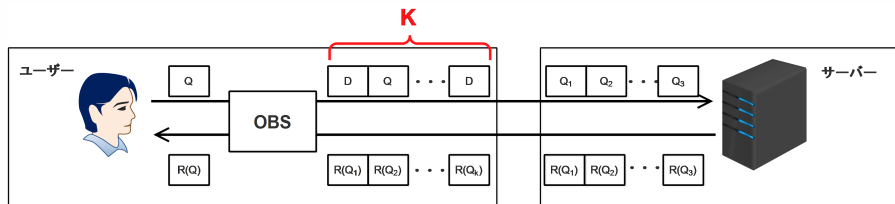
- 検索質問は単語 (名詞) の集合である
- 質問に含む単語数が多い
 - ウェブ検索:2.35 特許検索:20.1
- 専門用語が多い

テキスト検索



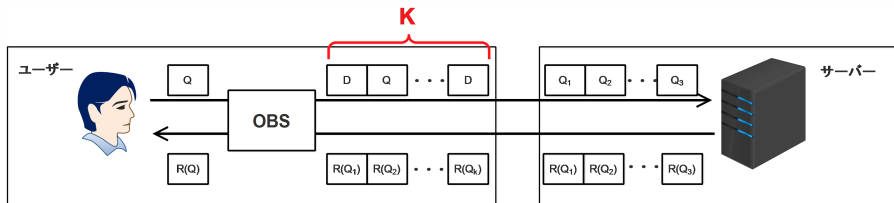
- 検索質問 Q : 単語の集合
- 質問 Q の検索結果 $R(Q)$: 文章の集合

Obfuscation Search



- 真の質問と $K - 1$ 個真の質問と区別できないダミー質問と同時に検索する
- サーバーが真の質問を見つける確率が $1/k$

Obfuscation Search:例



- 実践的には長い質問に対応できない
- 質問 q' を使うことより検索の精度と再現率が下がる

目標

- 長い質問に対応できる
- 専門用語が多いダミーを生成できる
- 検索の精度と再現率を維持できる

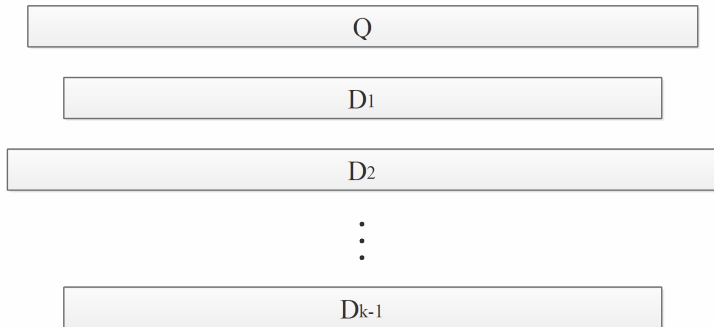
① 背景紹介

② 既存研究

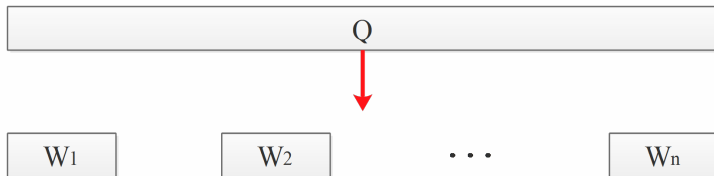
③ プライバシー分析

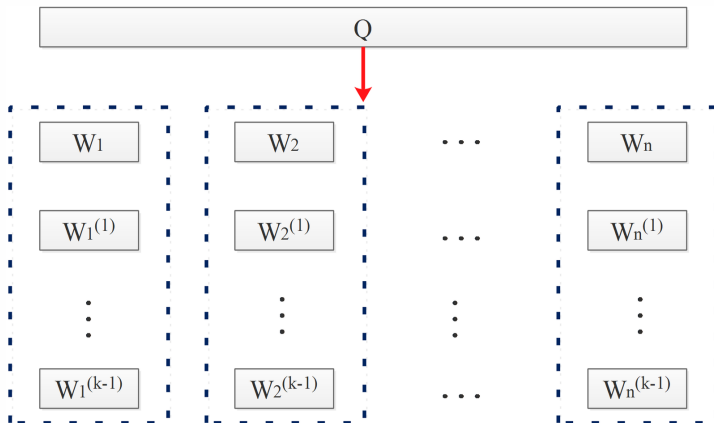
④ 参考文献

Embellishing Text Search Queries to Protect User Privacy (PDX10)



真の質問である可能性がある質問数: K





真の質問である可能性がある質問数: $K \rightarrow K^n$

テキスト検索



単語 W_i に対して文章 d_j のスコア: s_{ij}

質問 Q に対して文章 d_j のスコア: $s_j = \sum_{i \in Q} s_{ij}$

スコアが上位 m 個にある文章を質問 Q の検索結果として返す

準同型暗号

定義 (準同型暗号)

二つの暗号文 $Enc(m_1), Enc(m_2)$ が与えられた時に、平文や秘密鍵なしで $Enc(m_1 \circ m_2)$ を計算できる暗号

例 (加算ができる準同型暗号)

$Enc(\cdot)$: 暗号化 $Dec(\cdot)$: 復号

- ランダム性: $Enc(m) \neq Enc(m)$
- $Dec(Enc(m_1) \cdot Enc(m_2)) = m_1 + m_2$
- $Dec(Enc(m)^q) = m \cdot q, q \in \mathbb{Z}^+$

質問検索-ETS

$$Q \quad \begin{array}{|c|} \hline W_1^{(1)}, E(u_1^{(1)}) \\ W_1^{(2)}, E(u_1^{(2)}) \\ \vdots \\ W_1^{(k)}, E(u_1^{(k)}) \\ \hline \end{array} \quad \begin{array}{|c|} \hline W_2^{(1)}, E(u_2^{(1)}) \\ W_2^{(2)}, E(u_2^{(2)}) \\ \vdots \\ W_2^{(k)}, E(u_2^{(k)}) \\ \hline \end{array} \quad \dots \quad \begin{array}{|c|} \hline W_n^{(1)}, E(u_n^{(1)}) \\ W_n^{(2)}, E(u_n^{(2)}) \\ \vdots \\ W_n^{(k)}, E(u_n^{(k)}) \\ \hline \end{array} \quad u_i^{(k)} = \begin{cases} 0 & i, k \notin Q^* \\ 1 & i, k \in Q^* \end{cases}$$

単語 $W_i^{(k)}$ に対して文章 d_j のスコア: $s'_{ikj} = E(u_i^{(k)})(s_{ikj})$

質問 Q に対して文章 d_j のスコア: $s_j = \prod_{i,k \in Q} s'_{ikj}$

スコアが 0 ではない文章を全部返す

Wordnet

スクリーンショット

Synset 02068974-n ¹

Jpn: 海豚, ドルフィン, イルカ ²

Eng: *dolphin*

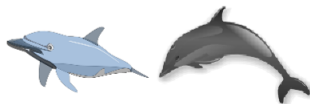
³ Jpn: くちばしのような鼻先を持つ様々な小型歯クジラ各種; ネズミイルカよりも大きい;
Eng: any of various small toothed whales with a beaklike snout; larger than porpoises;

Hype: [toothed whale](#)

Hypo: [delphinus](#) [delphis](#) [white whale](#) [grampus](#) [griseus](#) [bottlenose dolphin](#)
[pilot whale](#) [sea wolf](#) [river dolphin](#) [porpoise](#) ⁴

Hmem: [delphinidae](#)

SUMO: [c AquaticMammal](#) ⁵

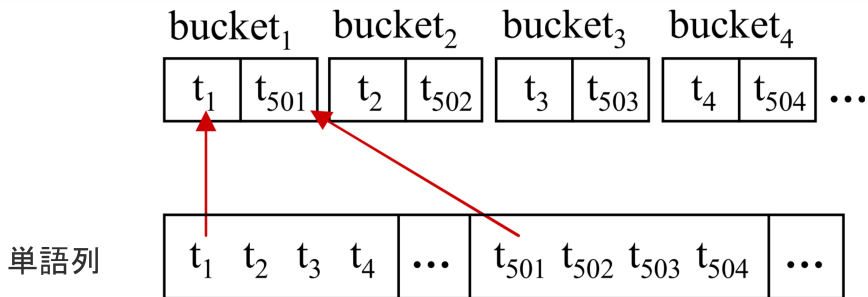


⁶

- ¹ synset番号(synset offset)
- ² 同義語(synonym)
- ³ 定義文・例文(gloss)
- ⁴ 関連synsetとのリンク
- ⁵ 他の言語資源とのリンク
- ⁶ 画像

単語を意味ごとに分類するデータベース
リンクを持つ synset を隣に並ぶことにより、すべての名詞を一列
に並ぶ: $t_1 t_2 t_3 \dots t_{1000}$

単語列



Wordnet

スクリーンショット

Synset 02068974-n ¹

Jpn: 海豚, ドルフィン, イルカ ²
Eng: dolphin

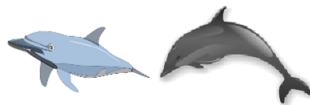
³ Jpn: くちばしのような鼻先を持つ様々な小型歯クジラ各種; ネズミイルカよりも大きい;
Eng: any of various small toothed whales with a beaklike snout; larger than porpoises;

Hype: [toothed whale](#)

Hypo: [delphinus](#) [delphis](#) [white whale](#) [grampus](#) [griseus](#) [bottlenose dolphin](#)
[pilot whale](#) [sea wolf](#) [river dolphin](#) [porpoise](#)

Hmem: [delphinidae](#)

SUMO: c [AquaticMammal](#) ⁵



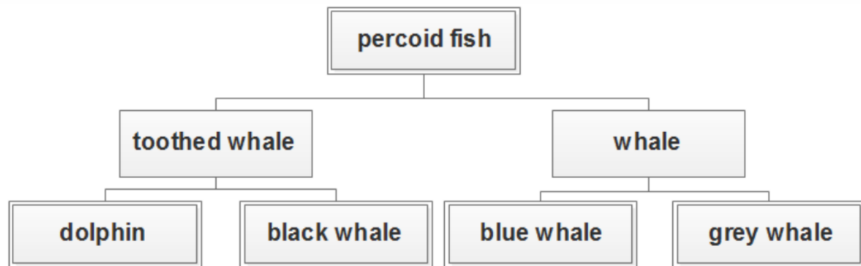
⁶

⁴

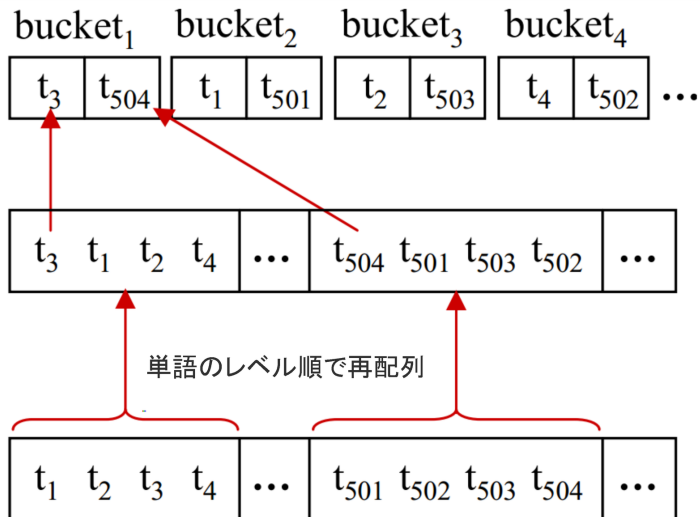
- ¹ synset番号(synset offset)
- ² 同義語(synonym)
- ³ 定義文・例文(gloss)
- ⁴ 関連synsetとのリンク
- ⁵ 他の言語資源とのリンク
- ⁶ 画像

実体/entity 以外全部の名詞の上位語が唯一に存在する
上下位関係を枝とすると、Wordnet 中の名詞が木の形になる

Wordnet



Wordnet



① 背景紹介

② 既存研究

③ プライバシー分析

④ 参考文献

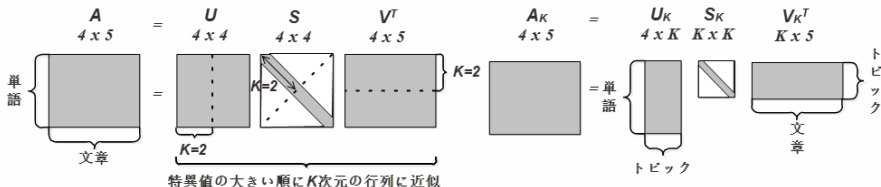
クエリ分析

メタノール	水蒸気	反応	水素	透過	膜	...	燃料
衡平	グンバイムシ	水力	上唇	ドアロック	沈殿	...	ベーキングパウダー
ルシタニア	ファースト	テアトル	水素	認知心理学	膜	...	運転者
メタノール	水蒸気	反応	長引かせること	透過	組織図	...	燃料
分限者	カランツ	意味合	発明品	イーサネットケーブル	原稿	...	黒泥土

真の質問の単語は全部燃料電池と関係あるが、ダミー単語の意味がバラバラである

もし単語が意味によって分類できるなら、燃料電池と関係がある単語が他のクラスに属する単語の数より多いことが考えられる

Latent Semantic Analysis



潜在的意味インデキシング

単語・文書行列 A を特異値分解 $A = USV^T$ し、 U 、 S 、 V の各列ベクトルを特異値が大きい順に K 個用いて A の低ランク近似 $A_K = U_K S_K V_K^T$ を得る。

このように低ランク分解によって、単語とトピックの関係を分析することができる

今回は同じ分類に属する全部の文章を 1 文章として LSA を行った

主意味攻撃

メタノール	水蒸気	反応	水素	透過	膜	...	燃料
衡平	ゲンバウムシ	水力	上唇	ドアロック	沈殿	...	ベーキングパウダー
ルシタニア	ファースト	テアトル	水素	認知心理学	膜	...	運転者
メタノール	水蒸気	反応	長引かせること	透過	組織図	...	燃料
分限者	カランツ	意味合	発明品	イーサネットケーブル	原稿	...	黒泥土

主意味攻撃

Input: 質問: $Q = \{t_i\}$, 単語のトピックベクトル集合 $L = \{\ell_i\}$

1: $R = \phi, \ell = 0$

2: $\ell = \sum_{t_i \in Q} \ell_{t_i}$

3: $\text{maintopic} = \text{argmax}_j \ell[j]$

4: **for all** $bk_k \in Q$:

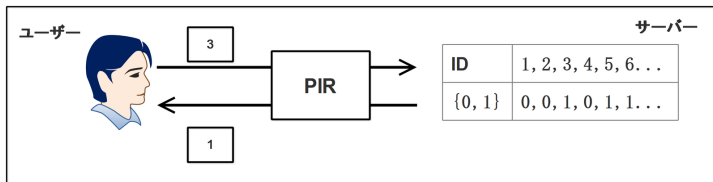
5: $R = R \cup \max_{t_i} q_{t_i}[\text{maintopic}]$

6: **return** R

プライバシー分析

重複を除いた単語数	2,973,096
文章数	3,496,253
質問数	2,908
質問平均単語数	21.0
主意味攻撃成功率	90.1%

Private Information Retrieval(PIR)



- 暗号などの手法を用いて質問の内容を完全に隠す

準同型暗号

ユーザー

質問生成

```
1: Input:  $i^*, n$ 
2: for  $i = 1, \dots, n$  :
3:   if  $i == i^*$  :
4:      $q_i = \text{Enc}(1)$ 
5:   else
6:      $q_i = \text{Enc}(0)$ 
7: return
    $Q = \{q_1, \dots, q_n\}$ 
```

復号

```
1: input:  $R$ 
2: return  $\text{Dec}(R)$ 
```

サーバー

結果計算

```
1: Input:  $Q, \{x_1, \dots, x_n\}$ 
2:  $R = 0$ 
3: for  $i = 1, \dots, n$  :
4:    $R = R \cdot q_i^{x_i}$ 
5: return  $R$ 
```

Note

$m_1 = m_2 \nRightarrow \text{Enc}(m_1) = \text{Enc}(m_2)$
 $\text{Dec}(R) = \sum_{x_i=1} \text{Dec}(q_i) = x_{i^*}$

性能

	OBS	OBS+PIR	PIR
サーバーの協力	不要	?	必要
安全性	弱い	?	強い
スピード	速い	?	遅い

安全性はOBSより強い、スピードはPIRより速い中間手法がほしい

国際特許分類

A61C 5/08A

セクション:A
サブセクション: 61
クラス: C
メイングループ:5
サブグループ:08

健康および娯楽
医学または獣医学:衛生学
歯科:口腔または歯科衛生
歯の充填または被覆
歯冠:その製造; 口中での歯冠固定

① 背景紹介

② 既存研究

③ プライバシー分析

④ 参考文献

Bibliography I

Rafail Ostrovsky and William E. Skeith Iii.

A Survey of Single-Database Private Information Retrieval: Techniques and Applications.
In Tatsuaki Okamoto and Xiaoyun Wang, editors, *Public Key Cryptography PKC 2007*,
number 4450 in Lecture Notes in Computer Science, pages 393–411. Springer Berlin
Heidelberg, April 2007.
DOI: 10.1007/978-3-540-71677-8_26.

HweeHwa Pang, Xuhua Ding, and Xiaokui Xiao.

Embellishing Text Search Queries to Protect User Privacy.
Proc. VLDB Endow., 3(1-2):598–607, September 2010.