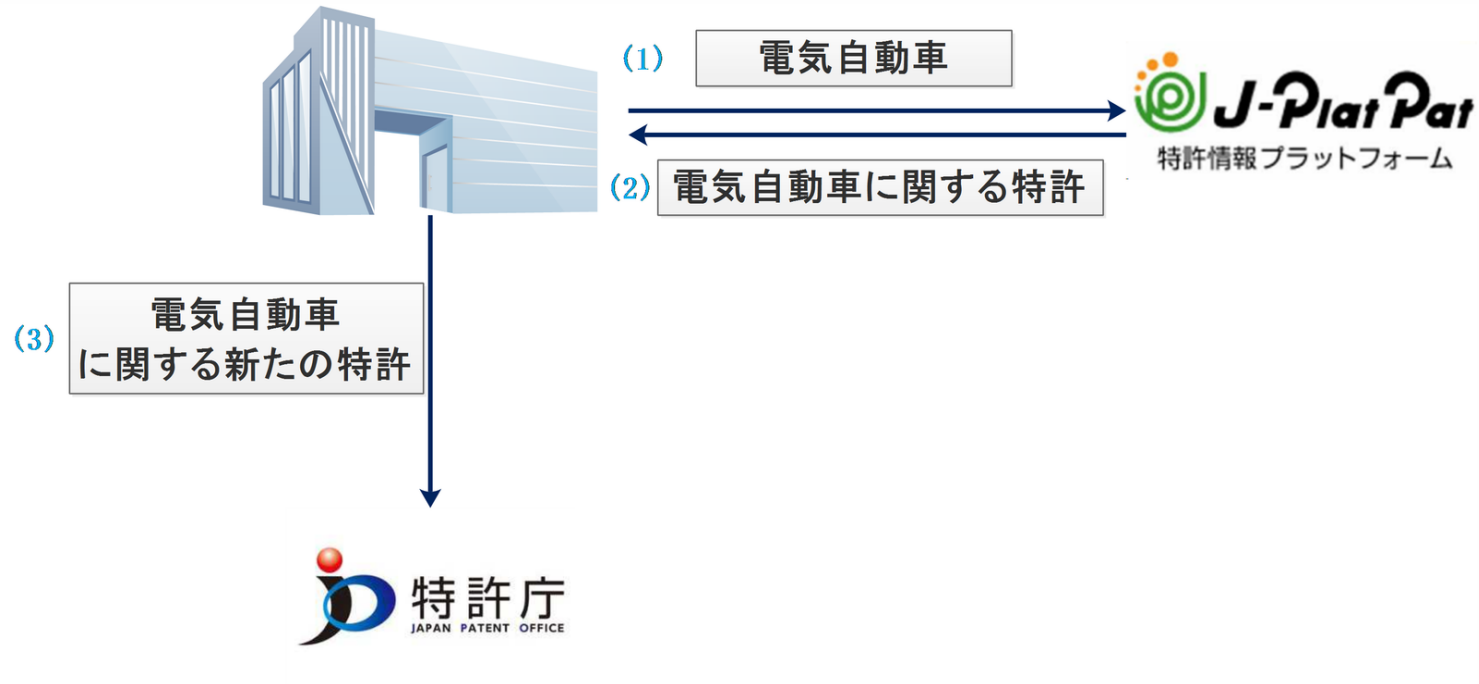
質問者のプライバシーを保護する特許データベース検索 (研究紹介)

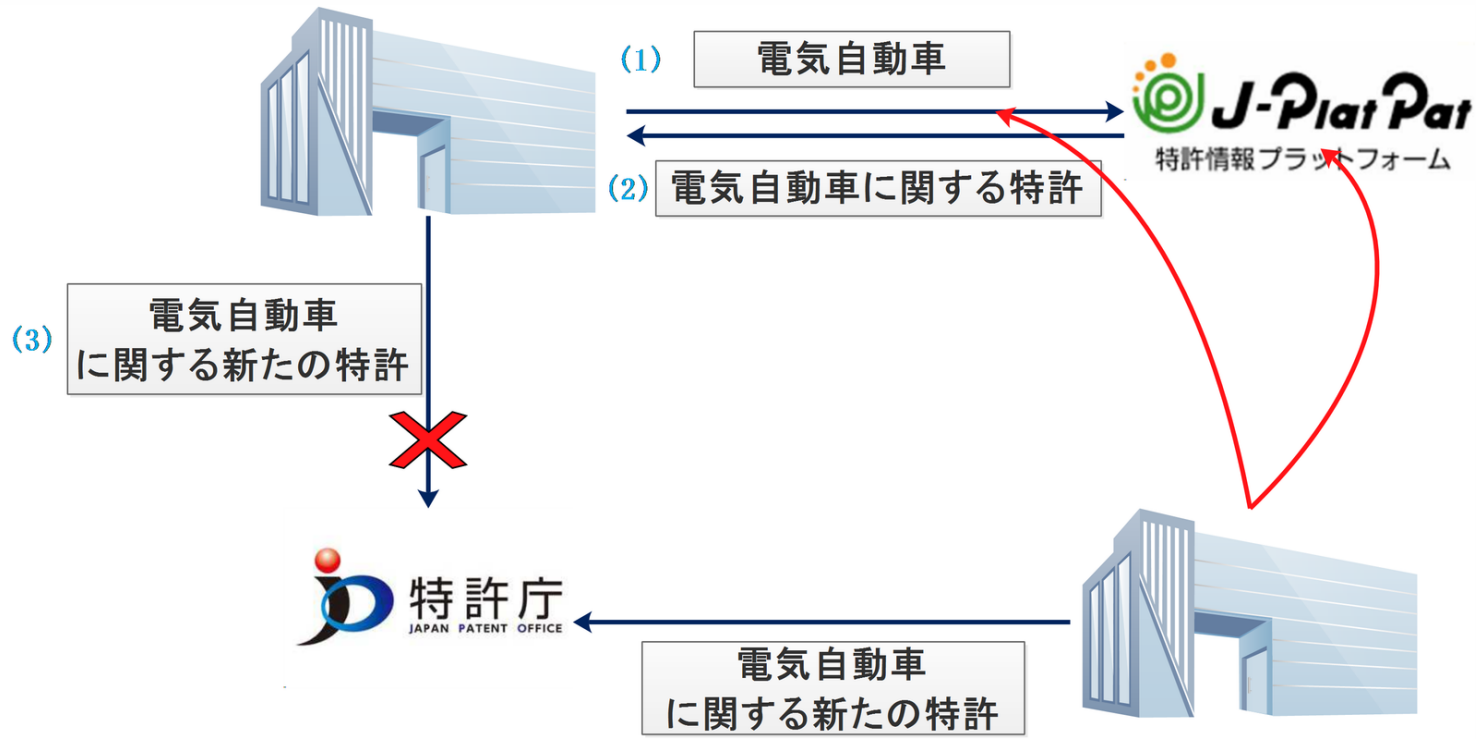
中川研 M2 胡 瀚林
指導教員：中川 裕志 教授

2016 年 7 月 1 日

特許検索


(1) 電気自動車
J-PlatPat
特許情報プラットフォーム
(2) 電気自動車に関する特許
(3) 電気自動車
に関する新たの特許
特許庁
JAPAN PATENT OFFICE

特許検索


(1) 電気自動車
J-PlatPat
特許情報プラットフォーム
(2) 電気自動車に関する特許
(3) 電気自動車
に関する新たの特許
特許庁
JAPAN PATENT OFFICE
電気自動車
に関する新たの特許

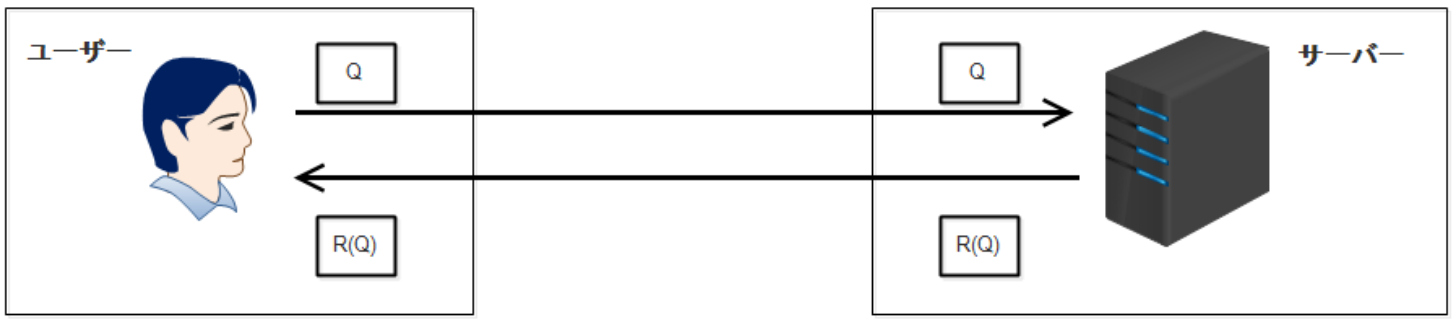
特許検索質問

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

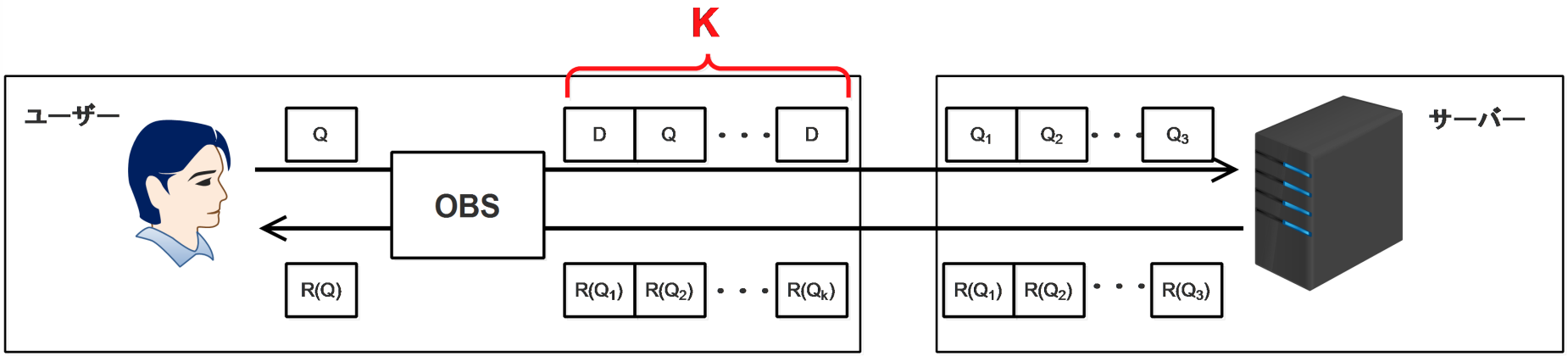
- 検索質問は単語 (名詞) の集合である
- 質問に含む単語数が多い
  - ウェブ検索:2.35 特許検索:20.1
- 専門用語が多い

テキスト検索



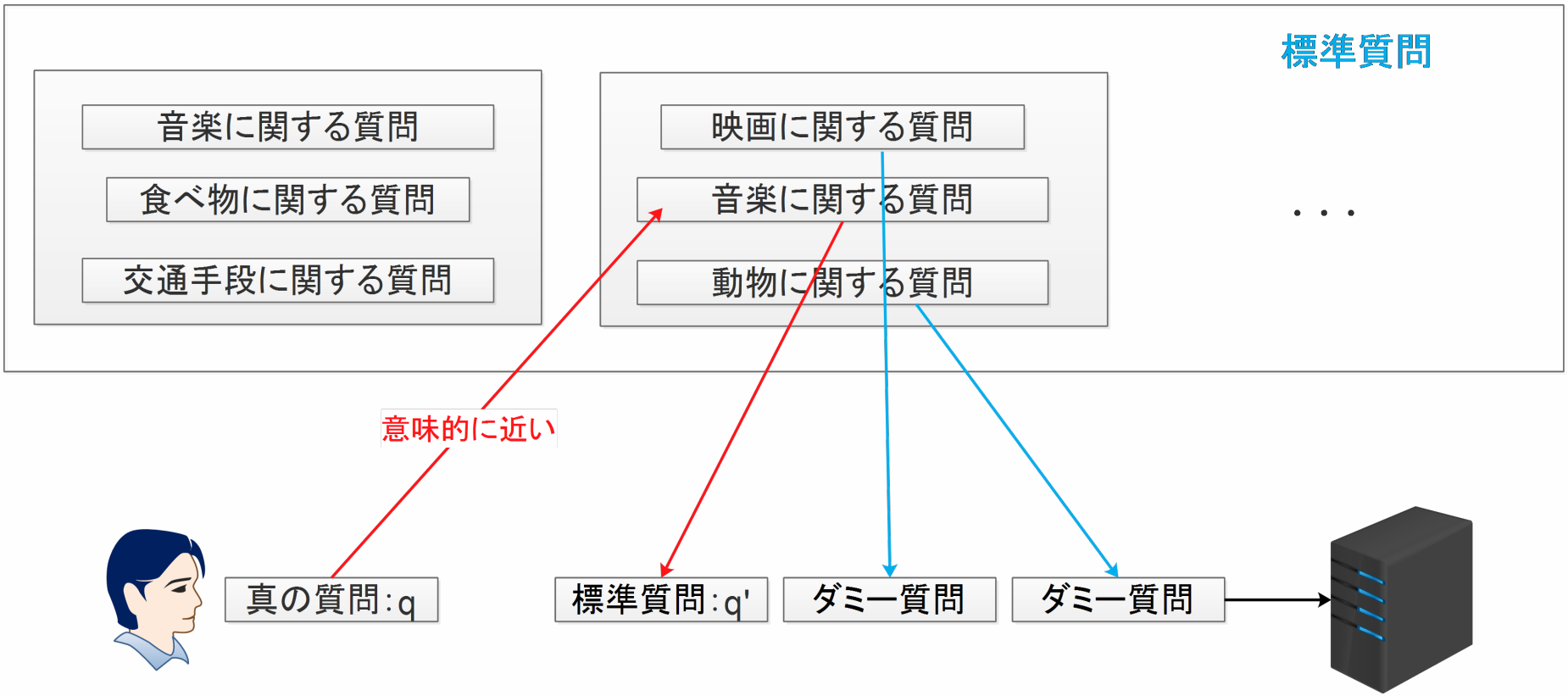
- **検索質問 Q:単語の集合**
- **質問 Q の検索結果 $R(Q)$:文章の集合**

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

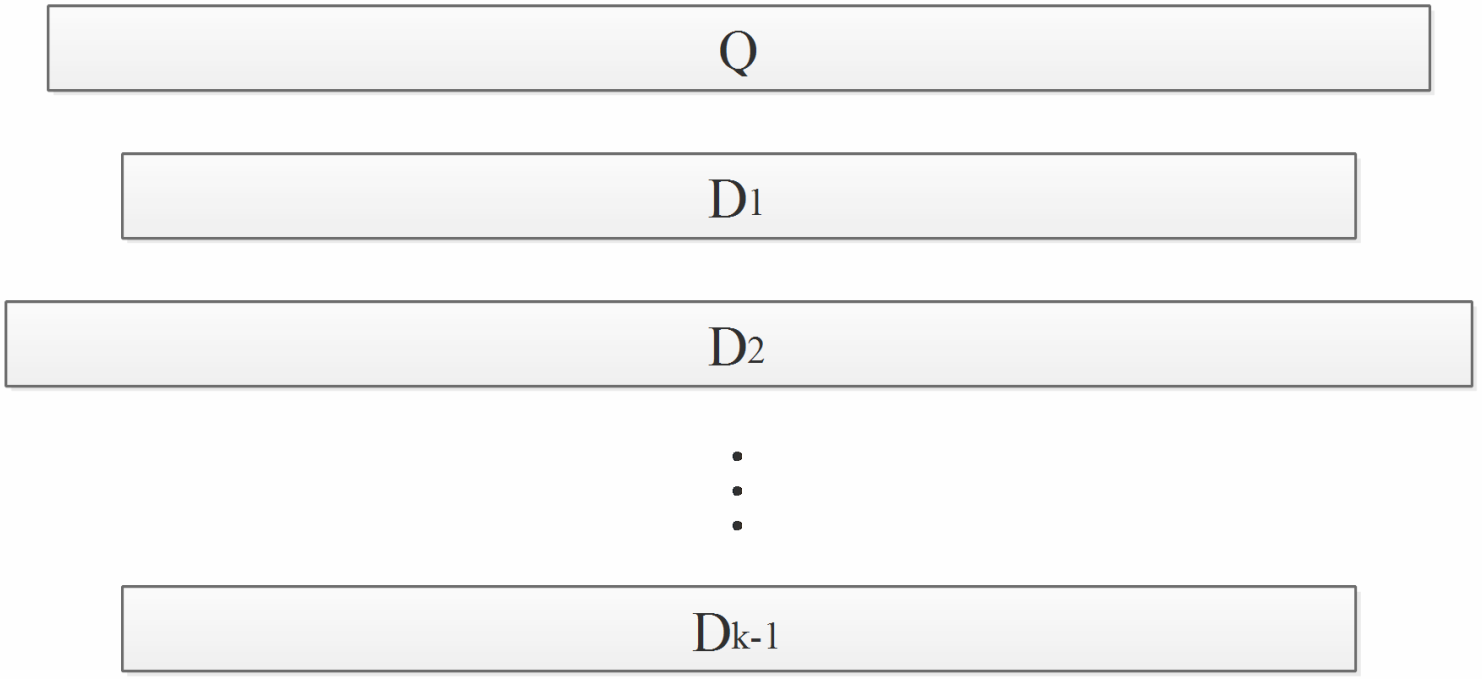


- 実践的には長い質問に対応できない
- 質問 q' を使うことより検索の精度と再現率が下がる

目標

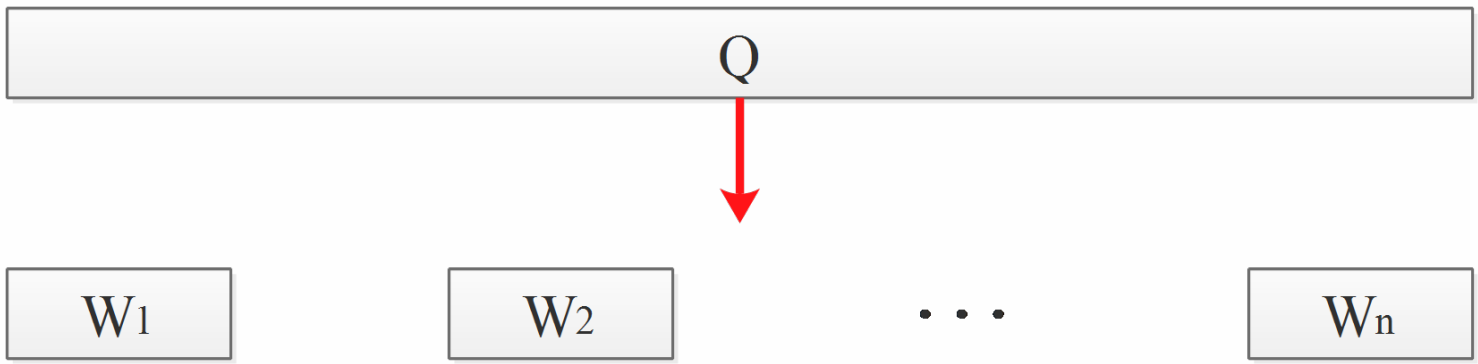
- 長い質問に対応できる
- 専門用語が多いダミーを生成できる
- 検索の精度と再現率を維持できる

Embellishing Text Search Queries to Protect User Privacy (PDX10)

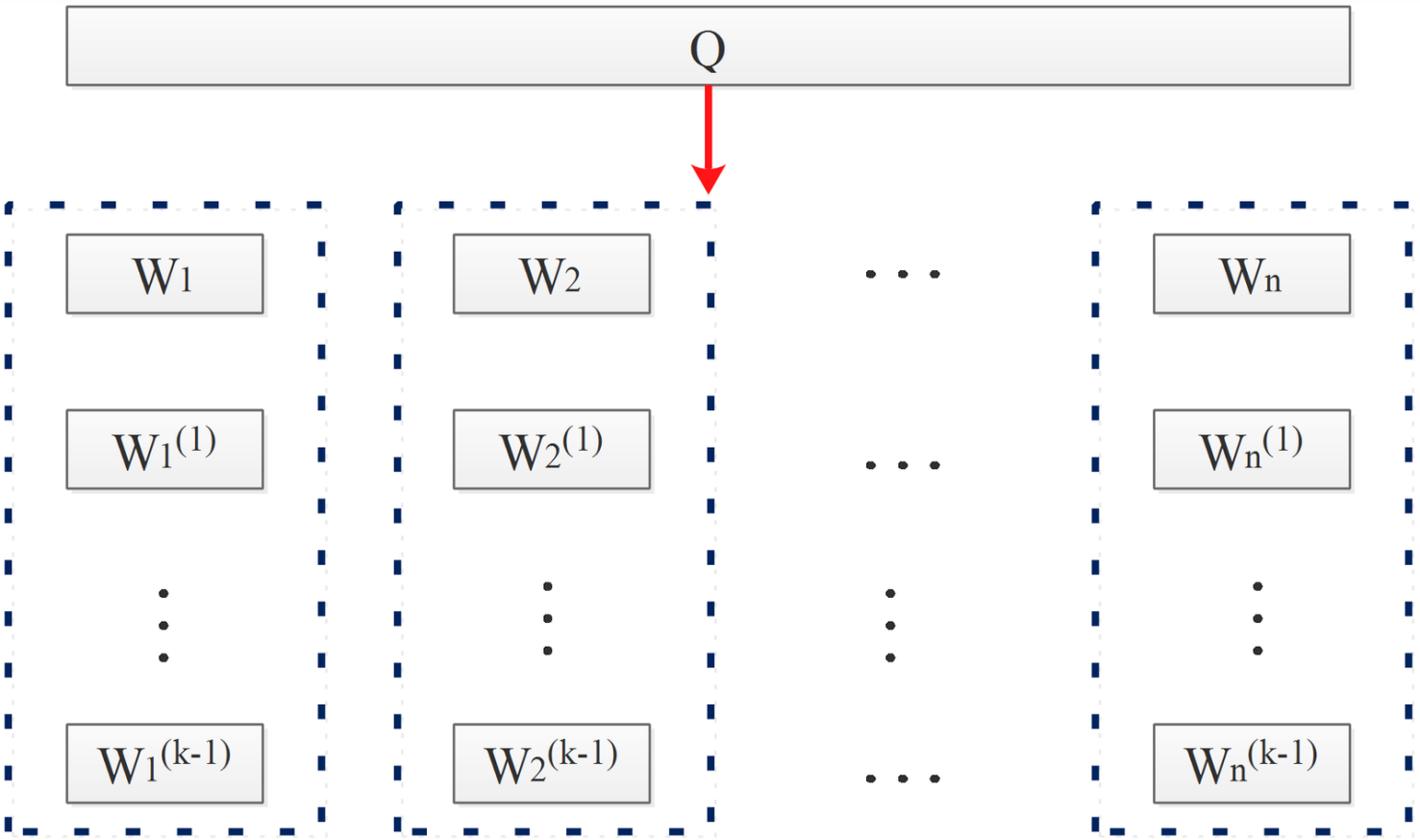


真の質問である可能性がある質問数:*K*

ETS



ETS



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)

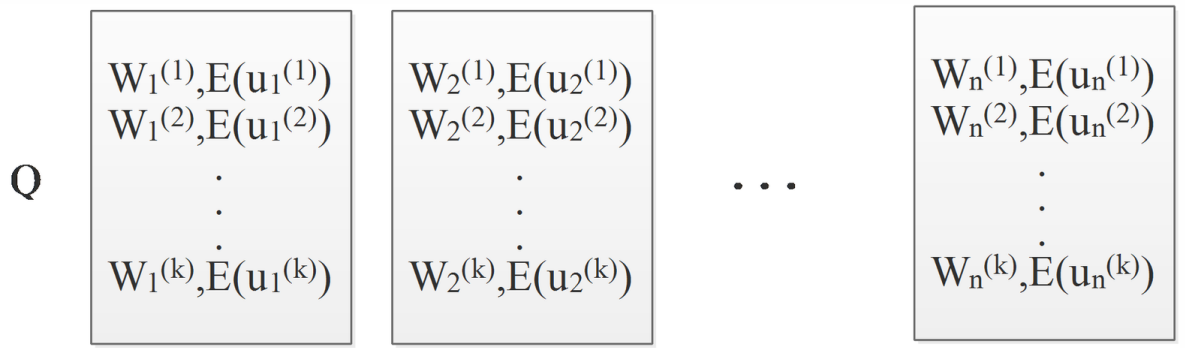
二つの暗号文 $Enc(m_1), Enc(m_2)$ が与えられた時に、平文や秘密鍵なしで $Enc(m_1 \circ m_2)$ を計算できる暗号

例 (加算ができる準同型暗号)

$E(\cdot)$: **暗号化**　$D(\cdot)$: **復号**

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

質問検索-ETS



$$u_i^{(k)} = \begin{cases} 0 & i,k \notin Q^* \\ 1 & i,k \in Q^* \end{cases}$$

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

質問 Q に対して文章 d_j のスコア: $s_j = \prod_{i,k \in Q} s'_{ikj} = E(\sum_{i,k \in Q^*} s_{ikj})$

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

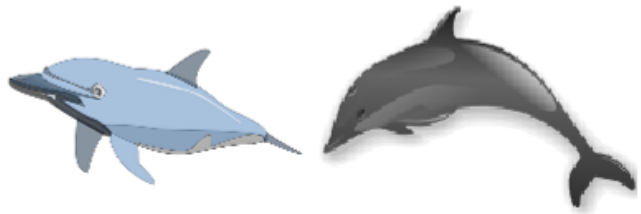
Synset 02068974-n 1

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: ⊂ AquaticMammal 5



6

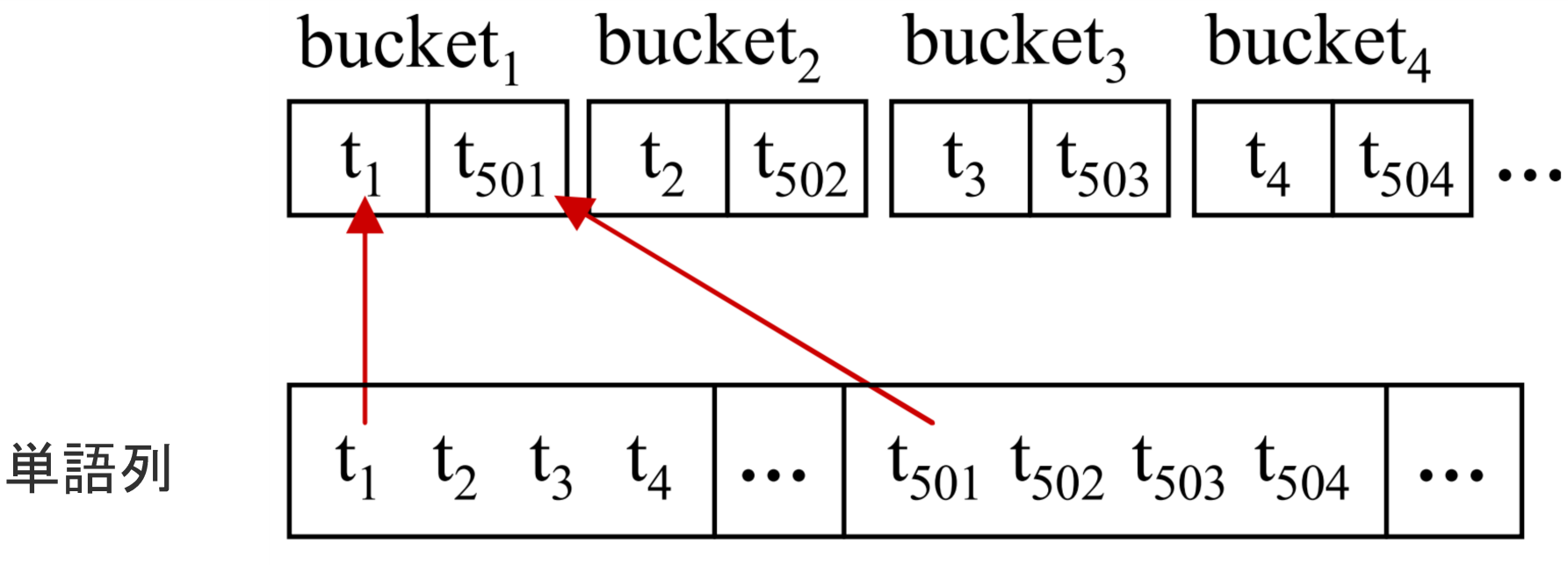
- 1 synset番号(synset offset)
- 2 同義語(synonym)
- 3 定義文・例文(gloss)
- 4 関連synsetとのリンク
- 5 他の言語資源とのリンク
- 6 画像

単語を類義関係のセット (synset) **でグループ化し、一つの** synset **が一つの概念に対応する**
各 synset **は上位下位関係などの関係で結ばれている**

バケツ作り

- 全ての synset を関係数が多いから小さい順で処理する
- 同じ単語を持つ synset を隣で並ぶ
- 反意関係，上位下位関係，構成被構成関係を synset を隣で並ぶ

単語列


bucket1
bucket2
bucket3
bucket4
t1
t501
t2
t502
t3
t503
t4
t504
…
単語列
t1 t2 t3 t4 … t501 t502 t503 t504 …

Wordnet

スクリーンショット

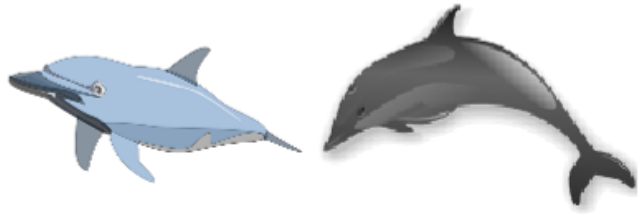
Synset 02068974-n 1

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: ⊂ AquaticMammal 5

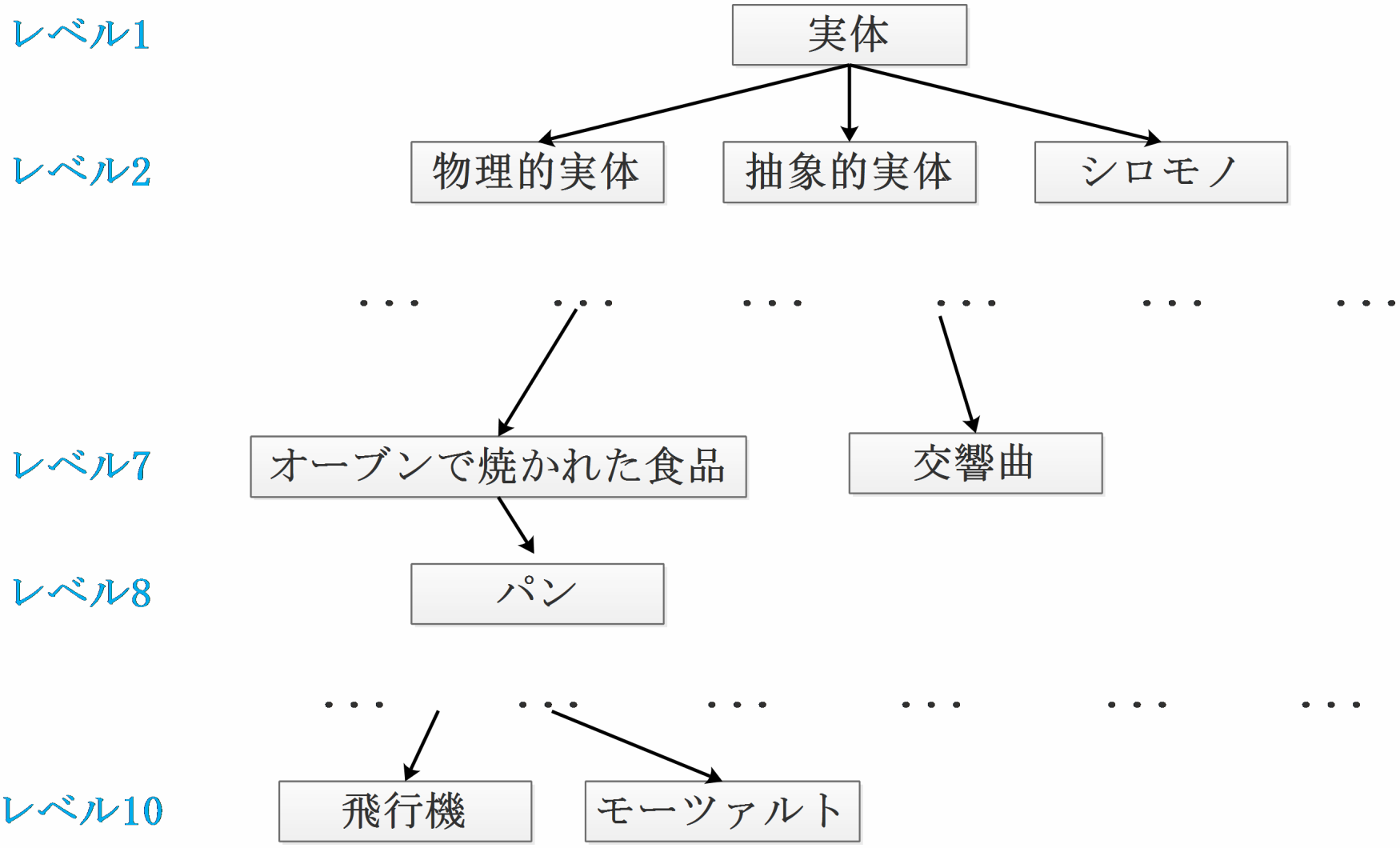


6

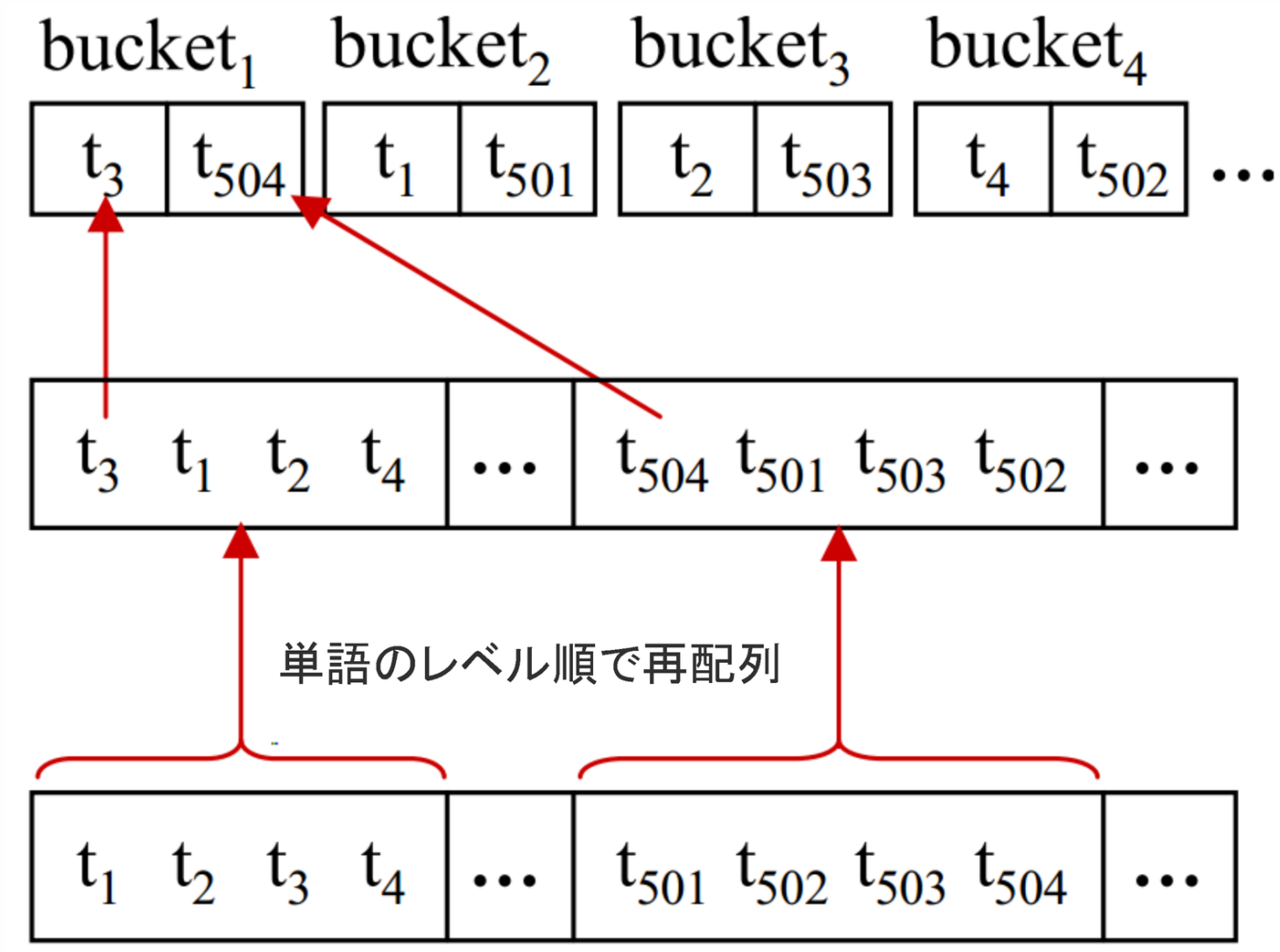
- 1 synset番号(synset offset)
- 2 同義語(synonym)
- 3 定義文・例文(gloss)
- 4 関連synsetとのリンク
- 5 他の言語資源とのリンク
- 6 画像

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


レベル1
実体
レベル2
物理的実体
抽象的実体
シロモノ
レベル7
オーブンで焼かれた食品
交響曲
レベル8
パン
レベル10
飛行機
モーツァルト

Wordnet

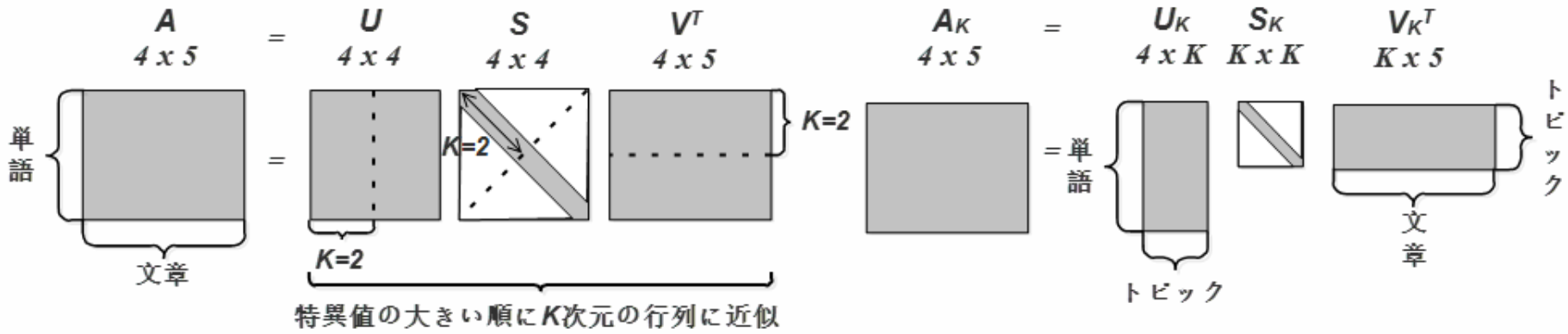

bucket1
bucket2
bucket3
bucket4
t3 t504
t1 t501
t2 t503
t4 t502
...
t3 t1 t2 t4 ... t504 t501 t503 t502 ...
単語のレベル順で再配列
t1 t2 t3 t4 ... t501 t502 t503 t504 ...

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 Indexing



潜在的意味インデキシング

単語 · 文書行列 A の (i, j) 番目の要素は i 番目の単語が j 番目の文章に出現した回数である
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
A_K の (i, j) 番目の要素は i 番目の単語と j 番目のトピックの関係を表す

国際特許分類

A61C 5/08A

セクション:A	健康および娯楽
サブセクション : 61	医学または獣医学:衛生学
クラス: C	歯科:口腔または歯科衛生
メイングループ:5	歯の充填または被覆
サブグループ:08	歯冠:その製造; 口中での歯冠固定

今回は同じ分類に属する全部の文章を 1 文章として LSI を行った

メイントピック攻撃

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

メイントピック攻撃

- ダミーを含んでいる質問のメイントピックを確定する
- 各単語バケツの中，メイントピックと一番関係強い単語を真の質問単語にする

主意味攻撃:例

モーツァルト

飛行機

パン

交響曲

	t_1(食べ物)	t_2(音楽)	t_3(交通手段)
w_1(モーツァルト)	0	1	0
w_2(交響曲)	0	1.5	0
w_3(パン)	1.5	0	0
w_4(飛行機)	0	0	1

ユーザー質問：モーツァルト 交響曲

$$\ell_Q = \ell_{w_1} + \ell_{w_2} + \ell_{w_3} + \ell_{w_4} = (1.5, 2.5, 1)$$
$$Maintopic = argmax_t \ell_Q[t] = t_2$$

主意味攻撃:例

モーツァルト

飛行機

パン

交響曲

	t_1(食べ物)	t_2(音楽)	t_3(交通手段)
w_1(モーツァルト)	0	1	0
w_2(交響曲)	0	1.5	0
w_3(パン)	1.5	0	0
w_4(飛行機)	0	0	1

ユーザー質問：モーツァルト 交響曲

$$\ell_{w_1}[t_2] = 1 > \ell_{w_4}[t_2] = 0$$
$$\ell_{w_3}[t_2] = 0 < \ell_{w_2}[t_2] = 1.5$$
$$Q^* = \{\textbf{モーツァルト}, \textbf{交響曲}\}$$

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%

まとめ

- 質問を単語ごとに分割し，暗号と組み合わせする手法
- 質問のメイントピックを保護するのは難しい
- Wordnet ではなく他のダミー単語を生成するツールが欲しい

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.