特許検索における質問意図の曖昧化

中川研 M2 胡 瀚林
指導教員：中川 裕志 教授

2016 年月日

特許

特許とは？

- 特許法第１条には、「この法律は、発明の保護及び利用を図ることにより、発明を奨励し、もつて産業の発達に寄与することを目的とする」とある。
- 特許制度は、発明者には一定期間、一定の条件のもとに特許権という独占的な権利を与えて発明の保護を図る一方、その発明を公開して利用を図ることにより新しい技術を人類共通の財産としていくことを定めて、これにより技術の進歩を促進し、産業の発達に寄与しようというものである。

特許

特許請求の範囲

【請求項１】植物の種子をパルプ繊維の水懸濁液に混合して抄紙する播種シートの製造方法。
【請求項２】水懸濁液にさらに水溶性接着剤を添加する請求項１記載の播種シートの製造方法。
【請求項３】あらかじめ種子を低粘度多価アルコールで被覆する請求項１記載の播種シートの製造方法。

特許請求の範囲の作成方法

８技術用語は、学術用語を用いる。
９用語は、その有する普通の意味で使用し、かつ、明細書及び特許請求の範囲全体を通じて統一して使用する

国際特許分類

A61C 5/08A

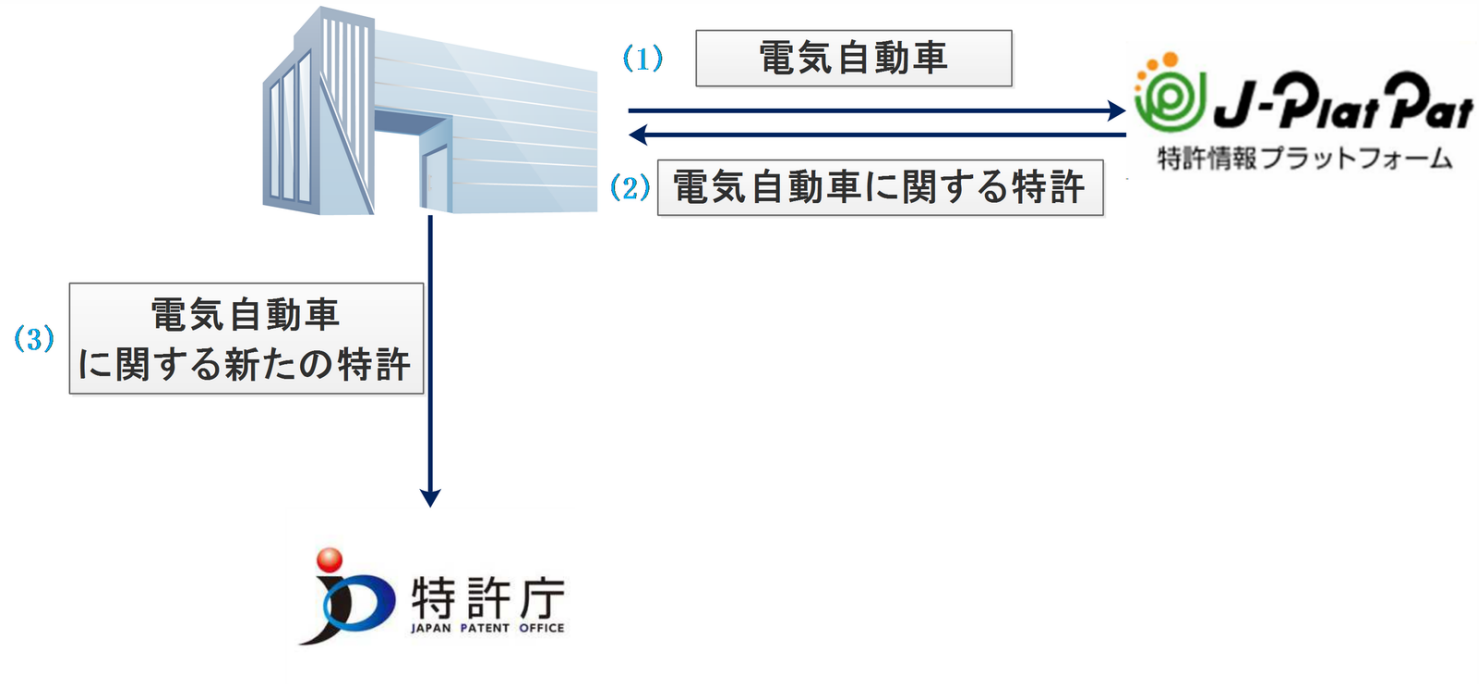
セクション:A	健康および娯楽
サブセクション : 61	医学または獣医学:衛生学
クラス: C	歯科:口腔または歯科衛生
メイングループ:5	歯の充填または被覆
サブグループ:08	歯冠:その製造; 口中での歯冠固定

全ての特許が人の手によって分類されている

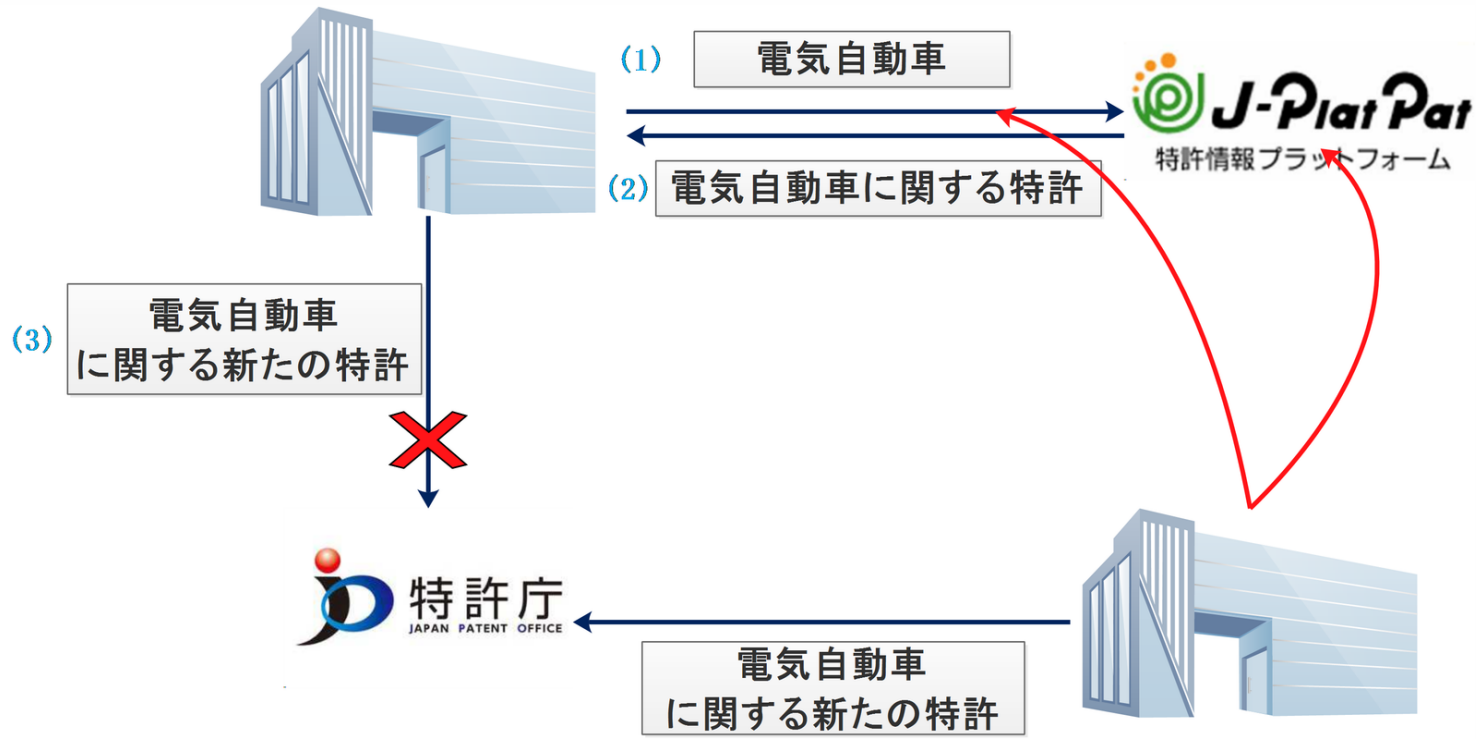
特許検索

検索タイプー	検索対象 (specification)	検索目的
技術水準調査 (State of the Art Search)	**イデア**	**自分の発明に関連する背景知識を得る**
新規性調査 (Novelty Search)	**特許文章**	**特許登録の可能性を判断する**
侵害調査 (Infringement Search)	**商品と** **商品に関連する技術**	**権利侵害とならないかを判断する**

新規性調査


(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

新規性調査


(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

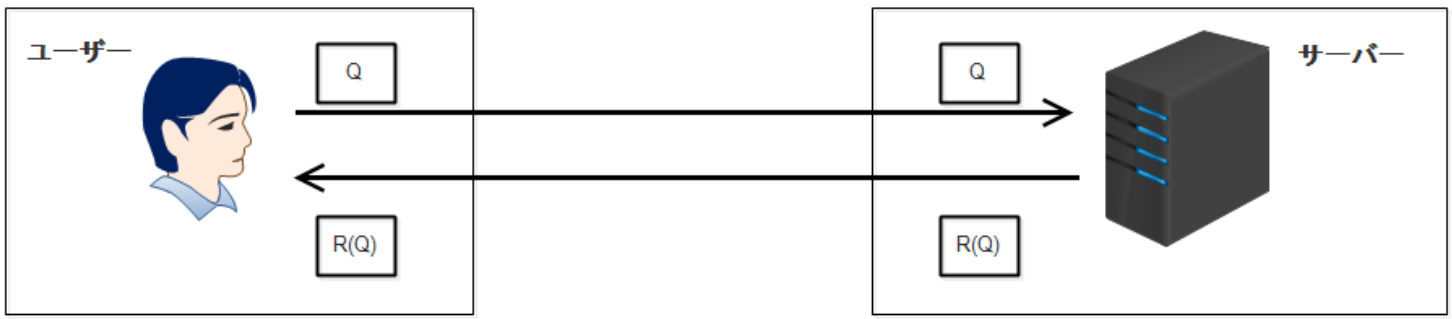
特許検索質問

播種シートの製造方法

植物 種子 パルプ 繊維 水 液 混合 抄 紙 播種 シート 製造 方法 水溶 性 接着 剤 添加 記載 度 価 アルコール 被覆

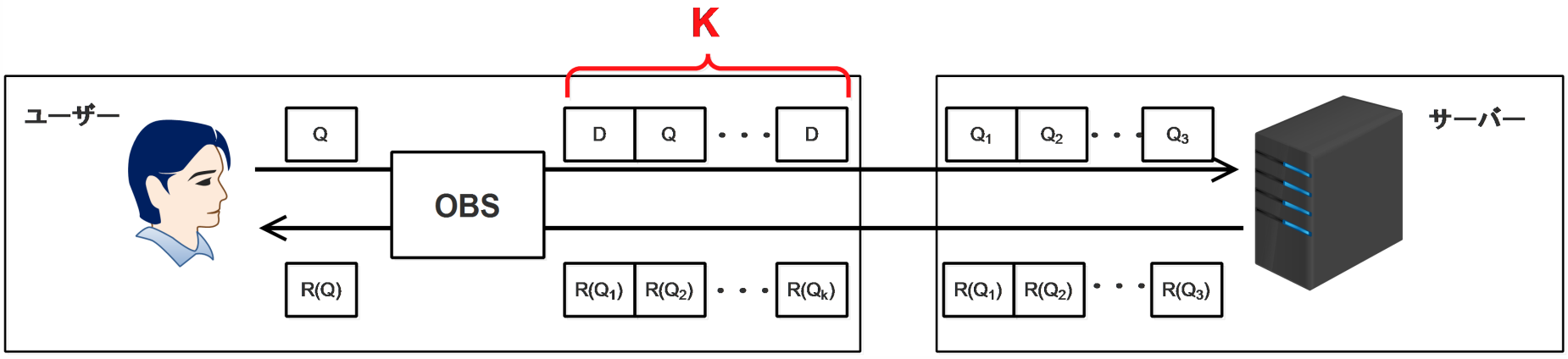
- 請求項から名詞を取り出し，検索質問とする
- 検索質問は単語 (名詞) の集合である
- 質問に含む単語数が多い
  - ウェブ検索:2.35 特許検索:20.1
- 専門用語が多い

テキスト検索



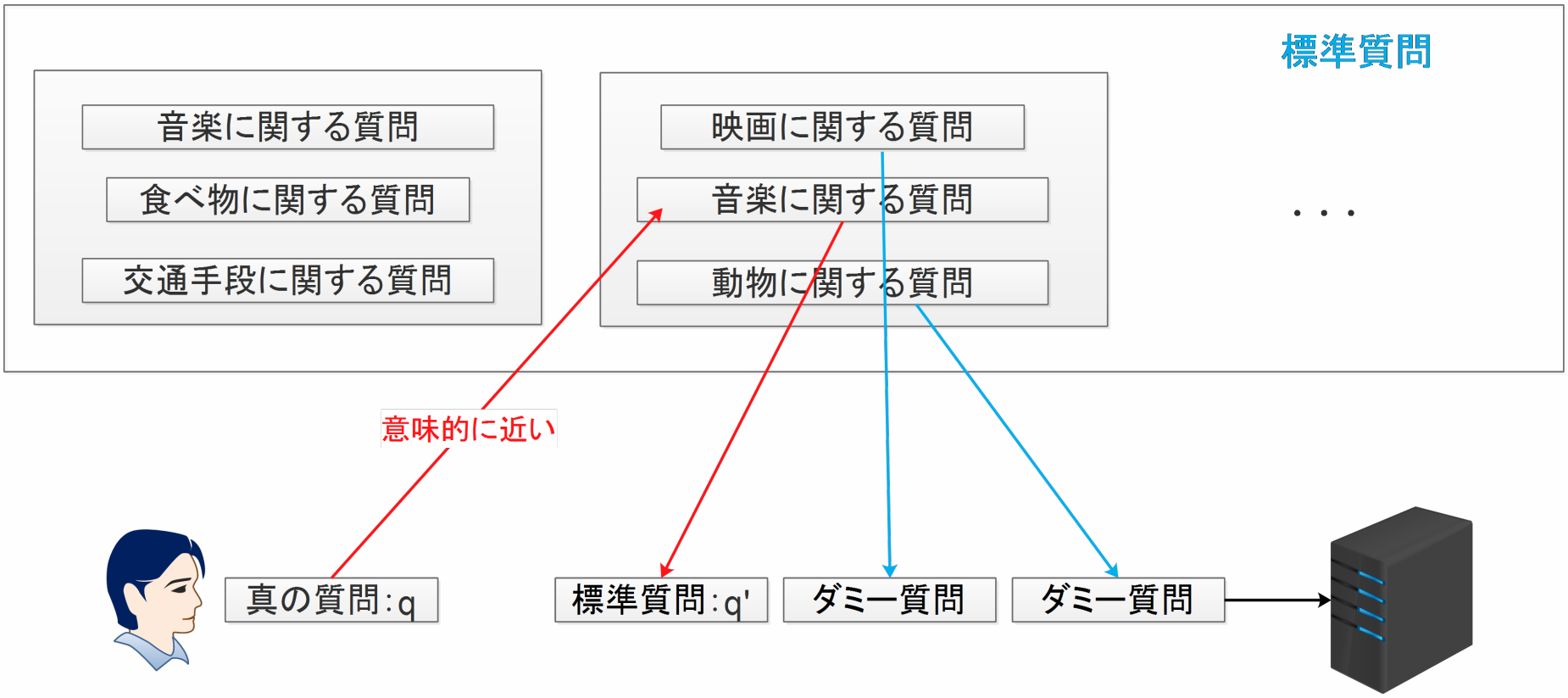
- **検索質問 q:単語の集合**
- **質問 q の検索結果 $R(Q)$:文章の集合**

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

Obfuscation Search

ユニバーサル質問集合:Q

W を全ての単語の集合とする．ユニバーサル質問集合 Q とは W の冪集合である，つまり

$$Q = P(W) = \{X | X \subset W\} \tag{1}$$

質問-トピックスコア関数:*rscore*

T を全ての可能なトピックの集合とする．質問 q とトピック t の関係を表す関数とは

$$rscore : Q \times T \rightarrow \mathbb{R} \tag{2}$$

質問間距離関数:*dist*

質問 q_1 と質問 q_2 間の距離を表す関数とは

$$dist : Q \times Q \rightarrow \mathbb{R} \tag{3}$$

目標

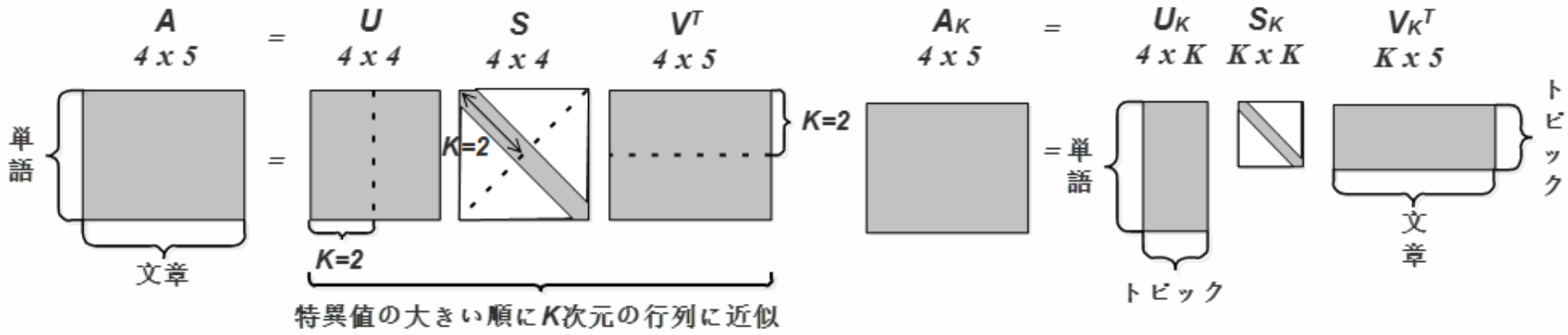
- 長い質問に対応できる
- 専門用語が多いダミーを生成できる
- 検索の精度と再現率を維持できる

Providing Privacy through Plausibly Deniable Search (MC09)

質問 q をユーザーが入力した質問とする．ダミー質問生成システム D が k 個の質問を含んでいる質問集合 $D(q_u) = \{q_1, \ldots, q_k\}$ を出力しサーバーに提出する．
$D(q_u)$ が以下の性質を持つなら，$D(q_u)$ を PD-質問集合といい，D を k ー否認可能検索という

1. $\exists q_i \in D(q_u), q_i$ と q_u が意味的に近い
2. $\forall q_j \in D(q_u), D(q_j) = D(q_u)$
3. $\forall q_j \in D(q_u), q_j$ が違うトピックに含まれる
4. $\forall q_j \in D(q_u), q_j$ が同じような尤もらしさを持つ

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

Providing Privacy through Plausibly Deniable Search (MC09)

質問-トピックスコア関数:$rscore_{LSI}$

S_K を単語 · 文書行列 A の低ランク近似の結果とし，$S_K(i,j)$ を S_K の (i,j) 番目の要素とする．LSI による質問 q とトピック t の関係を表す関数とは

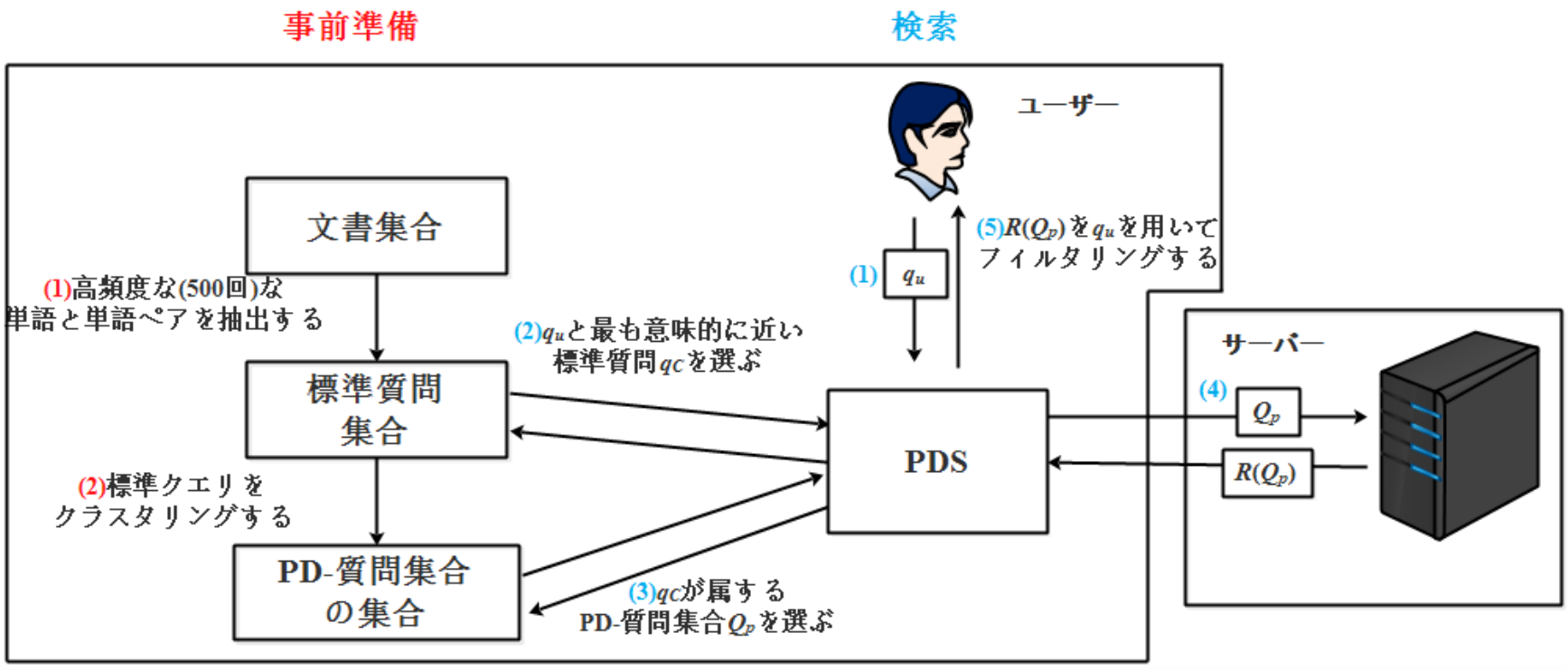
$$rscore_{LSI}(q,t) = \sum_{w \in q} S_K(w,t) \tag{4}$$

質問間距離関数:$dist_{LSI}$

LSI による質問 q_1 と質問 q_2 の距離を表す関数とは

$$dist_{LSI}(q_1,q_2) = 1 - \frac{\sum_{t \in T} rscore_{LSI}(q_1,t) \cdot rscore_{LSI}(q_2,t)}{\sum_{t \in T}(rscore_{LSI}(q_1,t^2))^{1/2} + \sum_{t \in T}(rscore_{LSI}(q_2,t)^2)^{1/} \tag{5}$$

Providing Privacy through Plausibly Deniable Search (MC09)



Providing Privacy through Plausibly Deniable Search (MC09)

問題点

- 質問の長さの増加に伴って標準質問の数が指数的に増加するため，長い質問対応できない
- 真の質問ではなく，真の質問に意味的に近い標準質問を用いるため，精度と再現率が低い
- (MC09) ではPD-質問集合を作るときは質問間で距離しか配慮していないため、同じトピックについて複数検索すると真の質問が属するトピック出現回数がほかのトピックより多くなる．したがって出現回数が一番多いトピックに属する質問が真の質問となる可能性が大きい．

Embellishing Text Search Queries to Protect User Privacy (PDX10)

- 質問の全体ではなく単語ごとにダミー単語を混ぜる.
- 単語バケットを事前に作り，真の質問の単語と同じバケットにある他の単語をダミー単語とする.

バケット作り

- 同じバケットにある単語の特殊さは近いが，意味的には大きい違いがある.
- 任意の2つのバケットの全ての単語間の意味的距離の差が近い

Wordnet

スクリーンショット

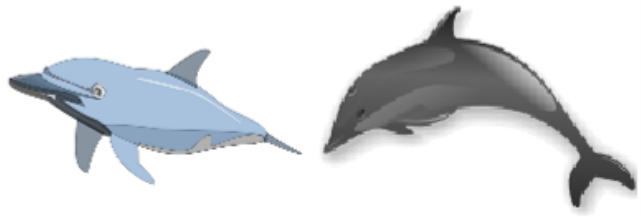
Synset 02068974-n 1

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

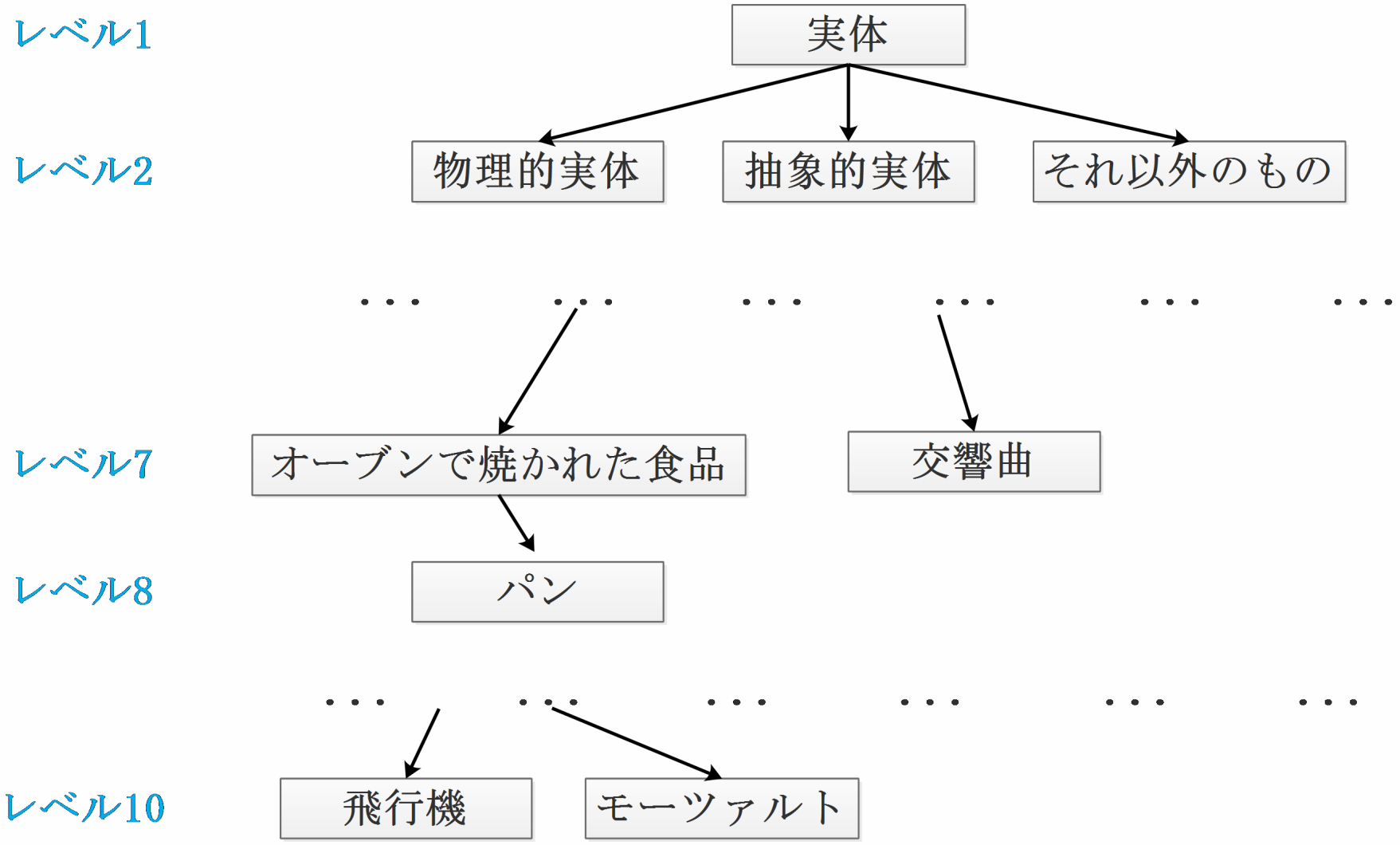


6

- 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**

Wordnet


レベル1
実体
レベル2
物理的実体
抽象的実体
それ以外のもの
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

Embellishing Text Search Queries to Protect User Privacy (PDX10)

単語間距離関数:$dist_{WordNet}$

単語 a の $synset$ を A とし単語 b の $synset$ を B とする．$synset\ A$ と $synset\ B$ 間の最短パスを $PATH(A, B)$ とする． WordNet **による単語間距離関数** $dist_{WordNet}$ **とは**

$$dist_{WordNet}(a, b) = |PATH(A, B)| \tag{6}$$

Embellishing Text Search Queries to Protect User Privacy (PDX10)

問題点

- WordNet に含まれていない専門用語が多い
- 単語バケットを作るときは単語間で距離しか配慮していない．

On masking topical intent in keyword search (WR14)

- Hash 関数を用いてダミー質問が属するトピックを決定する
- LDA を用いてトピック t に単語 w の出現確率 $Pr(w|t)$ を計算し、$Pr(w|t)$ からランダムで取り出した単語の集合をダミー質問にする

latent Dirichlet allocation (BNJ03)

On masking topical intent in keyword search (WR14)

問題点

- 単語数が多いデーターベースの LDA 計算
- 真の質問が同じトピックに属するとき対応するダミー質問も同じトピックに属するが，これだけで安全だと言えるか

SimAttack (PCB$^+$16)

類似度:*sim*

Input: 質問 q, ユーザープロフィール P_u, スムージングパラメータ:α

$coef[i] \leftarrow 2 \cdot |q \cap q_i| \cdot \frac{1}{|q|+|q_i|}$

$coef \leftarrow sort(coef)$

$sim \leftarrow coef[0]$

for $i \in [1, |P_u|]$:

$sim \leftarrow \alpha \cdot coef[i] + (1 - \alpha) \cdot sim$

Output: sim

SimAttack

Input: 質問集合 Q, ユーザープロフィール Pu, スムージングパラメータ:α

$q^* = \text{argmax}_{q \in Q}\, sim_{q,Pu}$

Output: q^*

既存研究

	潜在意味分析手法	質問列の対応	長い質問の対応
(MC09)	LSI	X	X
(PDX10)	WordNet	X	O
(WR14)	LDA	O	O

単語ベクトル

単語ベクトル l_t

T を全てのトピックの集合とし W を全て単語の集合とする．トピック t の単語ベクトル l_t とは

$$
\begin{aligned}
& l_t = \{w_1, w_2, \ldots, w_{|W|}\}, \\
& \forall w \in l_t, w \in W \\
& \forall 1 \leq i \neq j \leq, w_i \neq w_j \\
& \forall 1 \leq i < j \leq |W|, rscore(w_i, t) \geq rscore(w_j, t)
\end{aligned}
\tag{7}
$$

トピック間の距離

トピック間の cos 距離関数 $dist_{cos}$

l_{t_1}, l_{t_2} をトピック t_1, t_2 の単語ベクトルとする．トピック t_1, t_2 の cos 距離関数 $dist_{cos}$ とは

$$dist_{cos}(t_1, t_2) = dist_{LSA}(l_{t_1}[1:1000], l_{t_2}[1:1000]) \quad (8)$$

トピック間の $coef$ 距離関数 $dist_{coef}$

l_{t_1}, l_{t_2} をトピック t_1, t_2 の単語ベクトルとする．トピック t_1, t_2 の $coef$ 距離関数 $dist_{coef}$ とは

$$dist_{coef}(t_1, t_2) = 1 - \frac{l_{t_1}[1:1000] \cap l_{t_2}[1:1000])}{1000} \quad (9)$$

流れ

		文章単語行列		
	tfidf	LSI	LDA	
		トピック単語ベクトル (辞書)		
cos-**近い**	cos-**遠い**	**データベース分割**	coef-**近い**	coef-**遠い**
		トピック分類		
	同じ位置	**同じ位置+ノイズ**	zipf **分布**	
		ダミー質問		

SimAttack New

SimAttack$_{New}$

Input: **質問集合列** $\hat{Q} = \{Q_1, Q_2, \ldots, Q_n\}$, **スムージングパラメータ**:$\alpha$

1: **for** $j \in |Q_1|$:
2: $\quad \hat{Pu}[j] = Q_1[j]$
3: $\quad \hat{Put}[j] = \Phi$
4: $\quad d[j] = 0$
5: **for** $i \in [2, n]$:
6: $\quad$ **for** $j \in |Q_i|$:
7: $\quad\quad \hat{Put}[j] = \text{argmax}_{Pu \in \hat{Put}}\ sim_{Q_i[j], \hat{Put}[j]}$
8: $\quad q_i^* = \text{argmin}_{Q_i[j] \in Q_i}\ sim_{Q_i[j], \hat{Put}[j]}$
9: $\quad$ **for** $j \in |Q_i|$:
10: $\quad\quad \hat{Pu}[j] = \hat{Put}[j] \cap Q_i[j]$

Output: q^*

メイントピック攻撃

MainTopicAttack

Input: 質問集合 Q

1: **if** $Q = \{q_1, q_2, \ldots, q_{|Q|}\}$:
2: $\quad q^* = \mathrm{argmax}_q \max_t rscore_{LSA}(q, t)$
3: **else if** $Q = \{w_1^1, w_1^2, \ldots, w_1^k, \ldots, w_n^k\}$:
4: $\quad t^* = \mathrm{argmax}_t rscore_{LSA}(Q, t)$
5: $\quad$ **for** $i \in \{1, 2, \ldots, n\}$:
6: $\quad\quad w_i^* = \mathrm{argmax}_{w_i^j} rscore_{LSA}(w_i^j, t^*)$
7: $\quad q^* = \{w_1^*, w_2^*, \ldots, w_n^*\}$

Output: q^*

実験

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
国際特許分類数	623

実験

LSA ランダムトピック

評価方法:ダミー質問数	3	4	5	6	7i
SimAtt New		78.2	78.5	77.3	63.8
SimAtt LSA	65.5	55.6	50.8	48.2	50.7
Maintopic	75.2	72.3	67.6	64.2	58.5
100 番までの検索結果重複率	6.0	1.9	2.0	1.9	1.9

同トピック

評価方法	SimAtt New	SimAtt LSA	Maintopic	重複率
tfidf	25.1	25.0	25.1	28.5
LSA	27.6	27.0	25.8	20.7
LDA	25.8	24.6	24.8	32.3

データベース分割 (8)

評価方法	SimAtt New	SimAtt LSA	Maintopic	重複率
データベース分割 (8)	52.3	11.6	42.1	1.5

実験

LSA トピック分類方法

評価方法	SimAtt LSA	Maintopic	重複率
coef-far	88.7	69.2	5.2
coef-near	35.5	28.2	8.5

まとめ

Bibliography I

David M. Blei, Andrew Y. Ng, and Michael I. Jordan.
Latent dirichlet allocation.
Journal of machine Learning research, 3(Jan):993–1022, 2003.

M. Murugesan and C. Clifton.
Providing Privacy through Plausibly Deniable Search.
In *Proceedings of the 2009 SIAM International Conference on Data Mining*, Proceedings, pages 768–779. Society for Industrial and Applied Mathematics, April 2009.

Albin Petit, Thomas Cerqueus, Antoine Boutet, Sonia Ben Mokhtar, David Coquil, Lionel Brunie, and Harald Kosch.
SimAttack: private web search under fire.
Journal of Internet Services and Applications, 7(1):1, 2016.

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.

Peng Wang and Chinya V. Ravishankar.
On masking topical intent in keyword search.
In *2014 IEEE 30th International Conference on Data Engineering*, pages 256–267. IEEE, 2014.