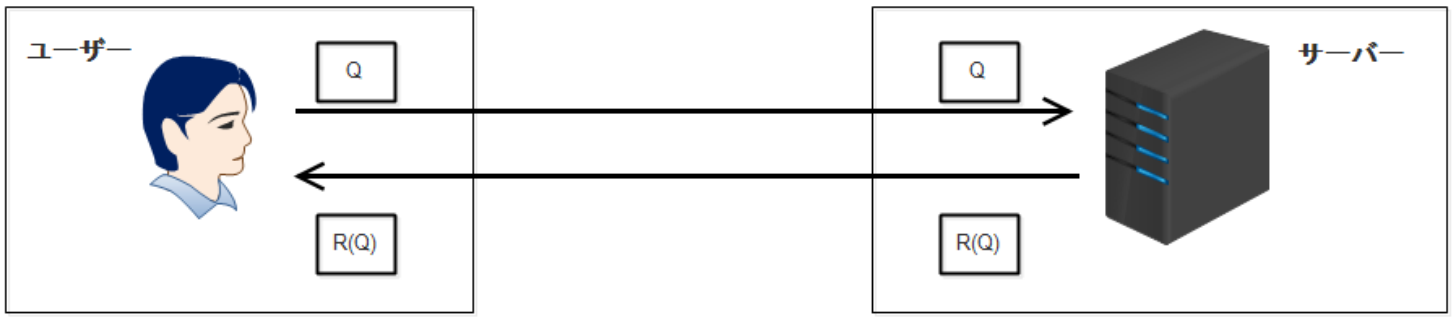
Private Information Retrieval

胡瀚林

May 18, 2016

1. 背景
2. 守り方
3. 攻撃方
4. 参考文献

Private Information Retrieval



- Q:**検索質問**
- R(Q):**質問Qの検索結果**

Location vs Keyword

- Location
  - **地図**
  - **乗換案内**
  - **近くのラストラン**
- Keyword
  - **ウェブ検索**
  - **データベース検索**
  - **クラウドストア検索**

AOL 事件

AOL 質問ログ

AnonID	Query	QueryTime	ItemRank	ClickURL
4417749	care packages	2006-03-02 09:19:32	10	http://booksforsoldiers.com
4417749	care packages	2006-03-02 09:19:32	9	http://www.brandonblog.com
4417749	movies for dogs	2006-03-02 09:24:14		
4417749	blue book	2006-03-03 11:48:52	1	http://www.kbb.com
4417749	best dog for older owner	2006-03-06 11:48:24	1	http://www.canismajor.com
4417749	best dog for older owner	2006-03-06 11:48:24	5	http://dogs.about.com

- 2006 年 8 月 4 日、AOL(American OnLine) が 650,000 人以上のユーザーの匿名化された検索質問ログを研究目的でリリースした。

AOL 事件

AOL 質問ログ

AnonID	Query	QueryTime	ItemRank	ClickURL
4417749	care packages	2006-03-02 09:19:32	10	http://booksforsoldiers.com
4417749	care packages	2006-03-02 09:19:32	9	http://www.brandonblog.com
4417749	movies for dogs	2006-03-02 09:24:14		
4417749	blue book	2006-03-03 11:48:52	1	http://www.kbb.com
4417749	best dog for older owner	2006-03-06 11:48:24	1	http://www.canismajor.com
4417749	best dog for older owner	2006-03-06 11:48:24	5	http://dogs.about.com

- 2006 年 8 月 4 日、AOL(American OnLine) が 650,000 人以上のユーザーの匿名化された検索質問ログを研究目的でリリースした。
- 2006 年 8 月 9 日、ID 4417749 の名前、年齢、住所などが特定された。(Bar)

Location vs Keyword

猫　？　犬

- 位置間の距離は簡単に計算できるが、
  単語間の距離は計算しにくい
- 単語の次元数が高い

猫？

- ノイズを加えにくい

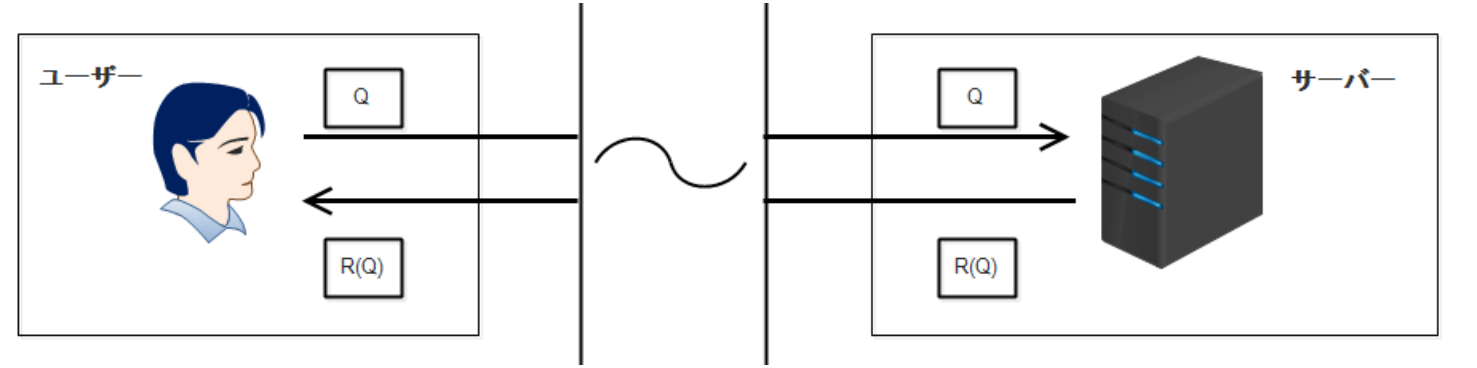
1 背景

2 守り方

3 攻撃方

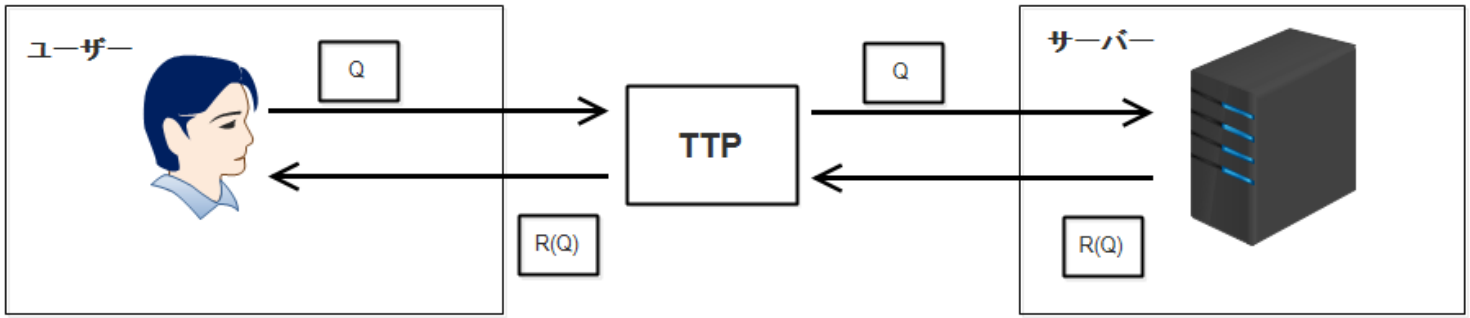
4 参考文献

Anonymity



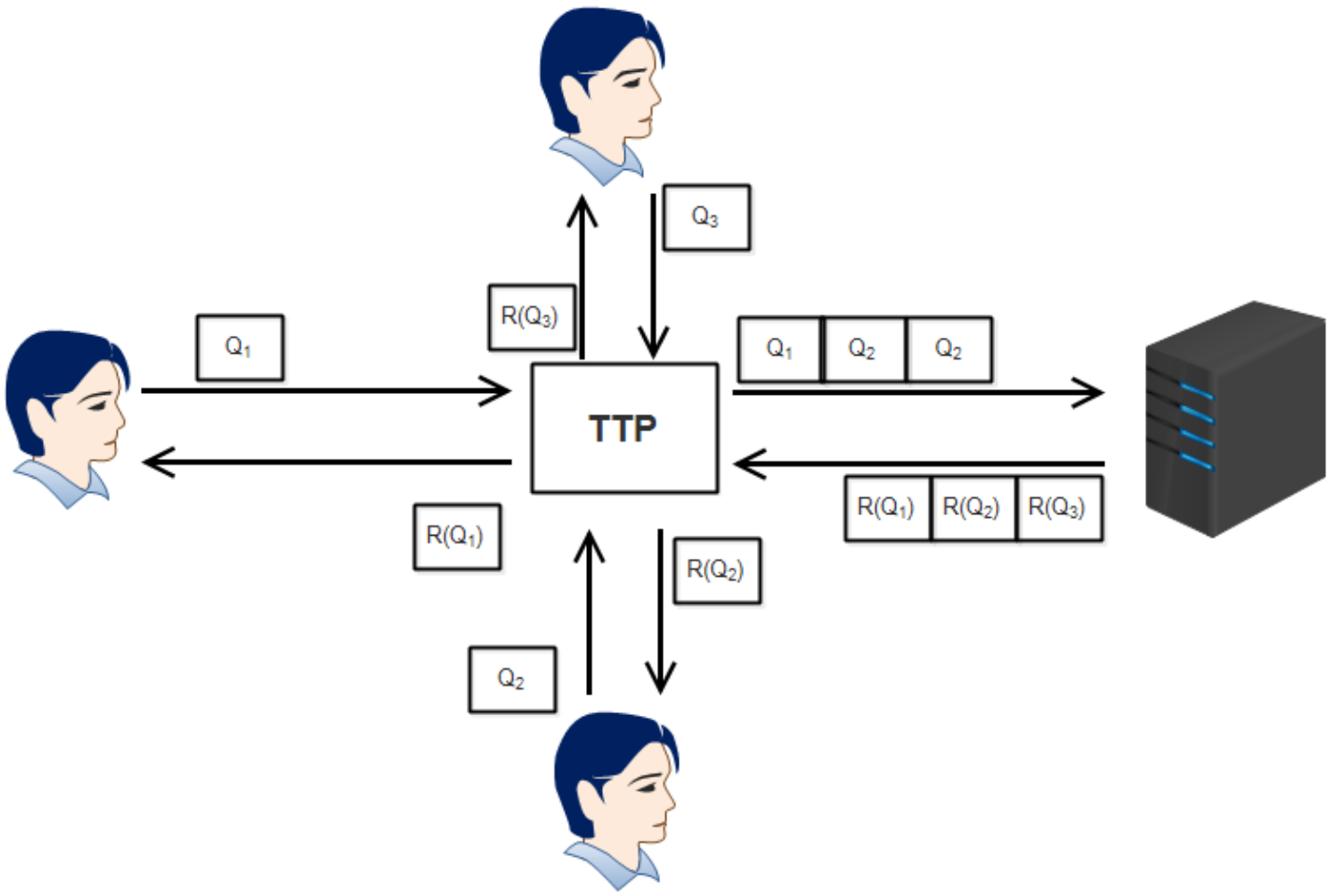
- 質問者を隠す

Tursted Third Party



- 質問者の IP アドレスなどを隠す

Tursted Third Party



- 複数の質問者を混ぜて検索する

Perturbation

Location

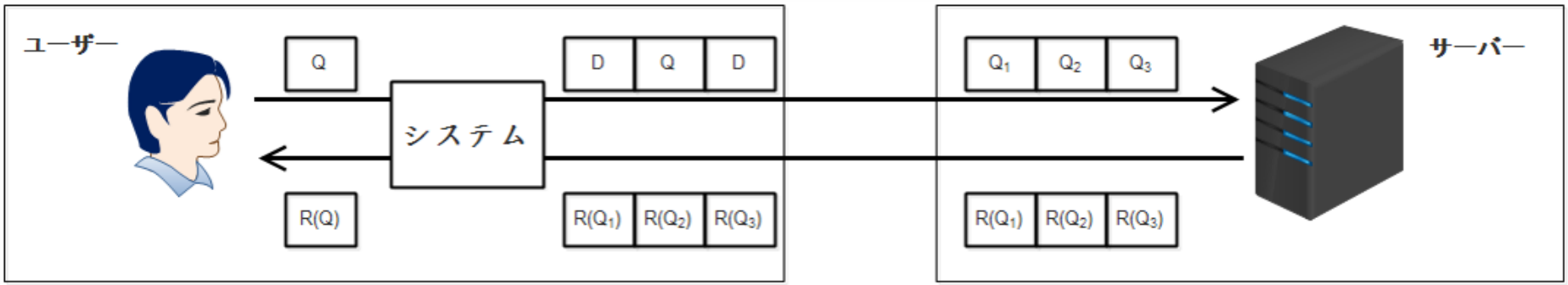
- Geo-indistinguishability (ABCP13)

Keyword

- **質問を一般化して検索する** (AED12)

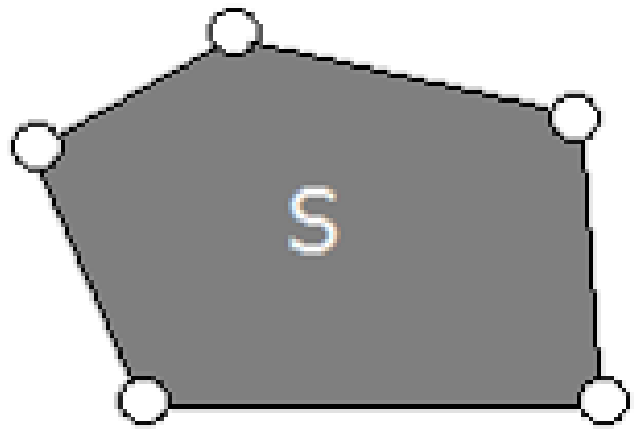
  リンゴ ⇒ 赤 果物
- **事前に標準質問を作って、本当の質問の代わりに使う** (MC09)

Obfuscation



- 複数の質問を混ぜて検索する

Obfuscation-Location



定義 ((k, s) – *privacy* (LJY08))

本当の位置と $k-1$ 個のダミー位置に囲まれた図形の面積が S 以上ある

Obfuscation-Keyword (BTD12)

問題

どうのようなダミー質問がいいダミー質問

Obfuscation-Keyword (BTD12)

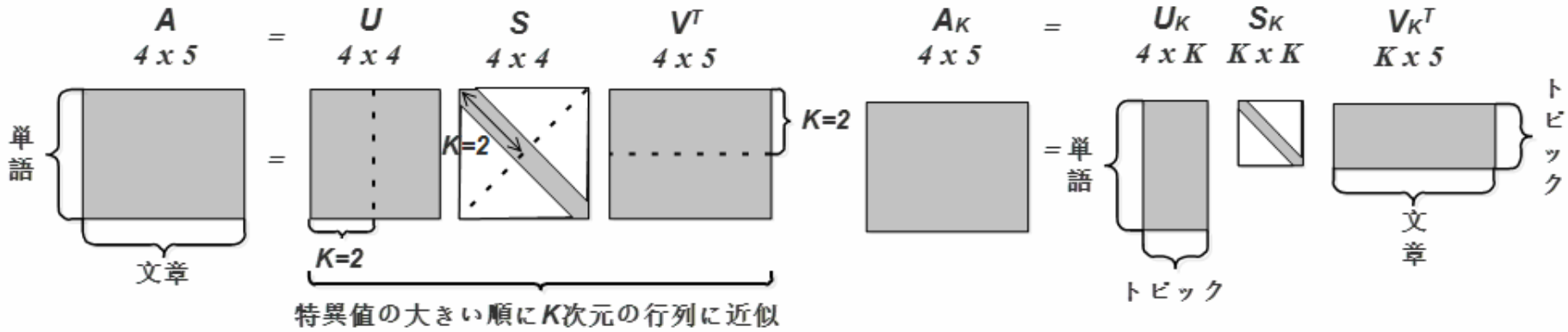
問題

どうのようなダミー質問がいいダミー質問

Plausibly Deniable Search (MC09)

- 本当の質問との“ 距離 ”が遠い
- 本当の質問と似たような“ 確率 ”で提出される

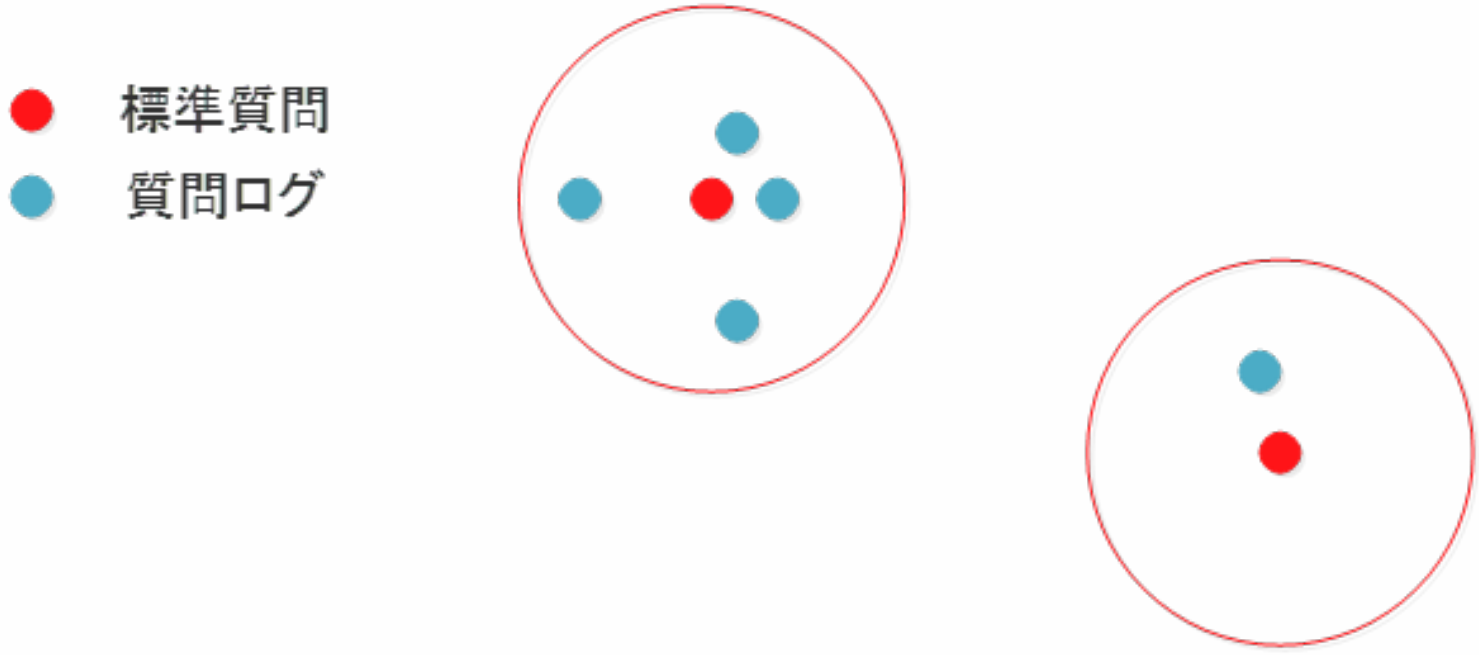
Latent Semantic Analysis



潜在的意味インデキシング

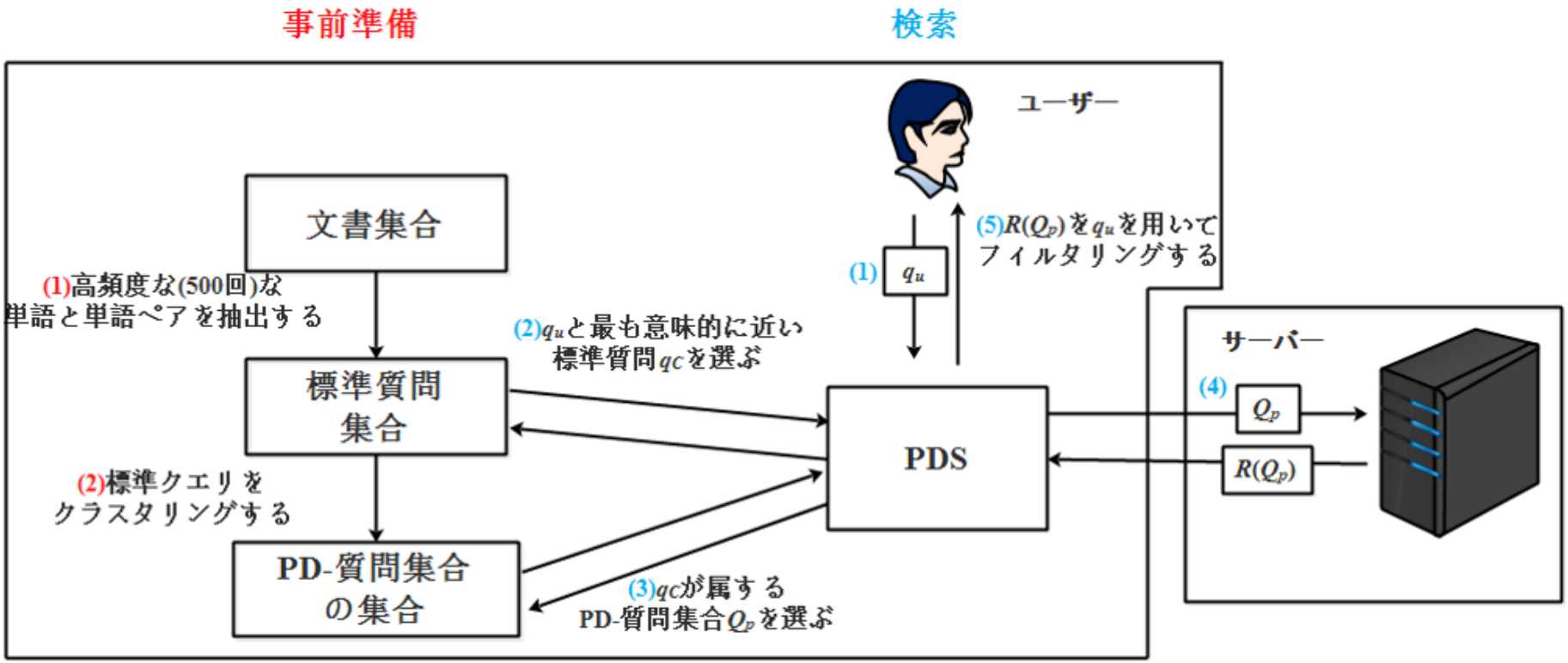
単語・文書行列 A を特異値分解 $A = USV^T$ し、U、S、V の各列ベクトルを特異値が大きい順に K 個用いて A の低ランク近似 $A_K = U_K S_K V_K^T$ を得る。
このように低ランク分解によって、単語とトピックの関係を分析することができる

Plausibly

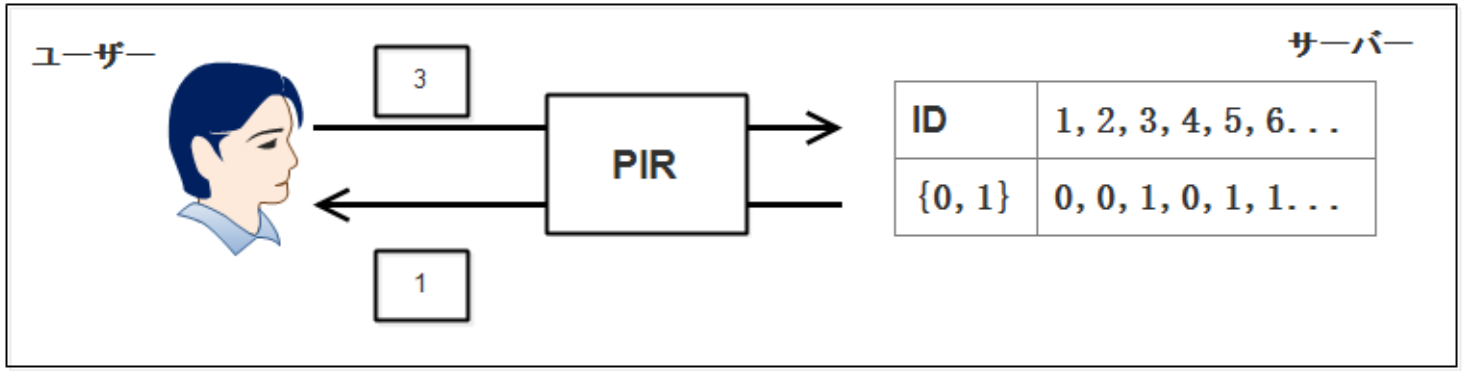


- 質問の近傍の中の質問数で“ 確率 ”、あるいは尤もらしさを計算する
- 質問数が多いほど“ 確率 ”が高いとする

Plausibly Deniable Search



PIR (OI07)



- 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

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**

$Dec(Enc(m_1) \cdot Enc(m_2)) = m_1 + m_2$

準同型暗号

ユーザー

質問生成

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

復号

1: input:R
2: **return** $Dec(R)$

サーバー

結果計算

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

Note

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

PIR I

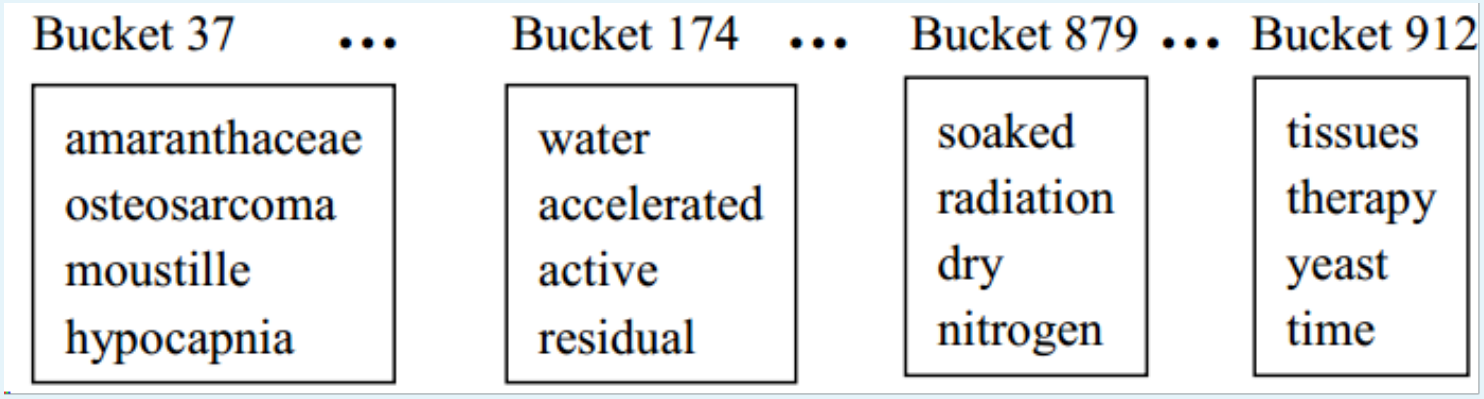
- 1995 Chor et al. : Multiserver PIR
  - **情報理論から見ると** single-database PIR **ができない**
- 1997 Kushilevitz and Ostrovsky : computational single-database PIR
  - *quadratic residuosity computational assumption*
  - **通信量**: $O(2^{\sqrt{lognloglogN}})$
- 1999 Cachin et al. : s-PIR
  - Φ – *hiding number* – *theoretic assumption*
  - **通信量**: $O(log^8 n)$
- 2000 Kushilevitz and Ostrovsky : Private Block Retrieval
  - *Naor* – *Yung one* – *way* 2 – *to* – 1 *trapdoor permutations*
  - *Goldreich* – *Levinhard* – *corepredicates*
  - **通信量**: $n - cn/2k + O(k^2)$

PIR II

- 2005 Gentry and Ramzan : Multiserver PIR
  - Φ – *hiding number* – *theoretic assumption*
  - **通信量**: $O(log^2 n)$
- 2007 Aguilar-Melchor and Gaborit : computationally-efficient PIR
  - *lattice* – *based*
  - a few thousand bit-operations per bit in the database
  - 2010 Olumofin and Goldberg:**応答時間は普通の方法の千分の一くらい**
- 2013 Yi et al. : PBR
  - *Fully homomorphic encryption*
  - **通信量**: $(\gamma + \gamma/)$
  - **計算量**: $(\gamma^2 + \gamma/2)$
  - **計算時間**:2*min*
  - **通信時間**: 4.5*s*(100 – *Mb*/*second*)

Obfuscation + PIR

Embellishing Text Search Queries to Protect User Privacy (PDX10)



- 本当の質問との“ 距離 ”が遠い
- 本当の質問と似たような“ 確率 ”で提出される
- 質問ではなく単語ごとにダミーを加える

Wordnet

スクリーンショット

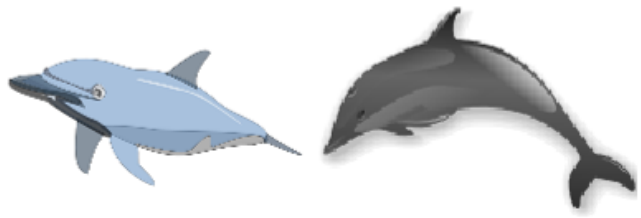
Synset 02068974-n 1

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

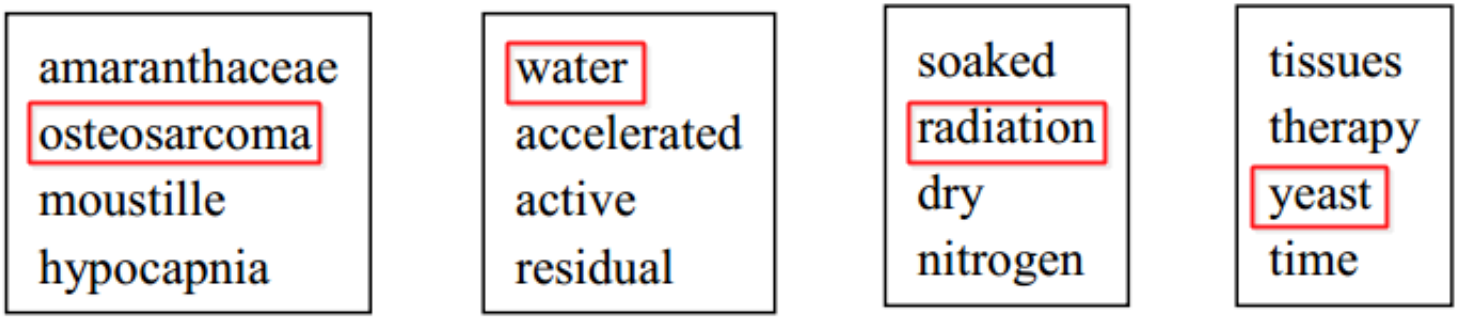


6

- 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

- “ 距離 ”:単語間の最小リンク数
- “ 確率 ”:単語が属する synset のレベル

Obfuscation + PIR



$q = \{\langle amaranthaceae, E(0)\rangle, \langle osteosarcoma, E(1)\rangle, \dots\}$ t_i : **単語** i
d_j : **文章** j L_i : t_i **の検索結果** p_{ij} : d_j **に対して** t_i **の** *score*

Query processing by the search engin

Input:Embellished query q
Let $R = \phi$
for all $\langle t_i, E(u_i)\rangle \in q$:
  for all $\langle d_j, p_{ij}\rangle \in L_i$:
    if $\exists\langle d_j, E(score_j)\rangle \in R$:
      $E(score_j) = E(score_j) \cdot E(u_i)^{p_{ij}}$
    else
      Insert $\langle d_j, E(u_i)^{p_{ij}}\rangle$ into R
return R

Snapshot vs Sequence

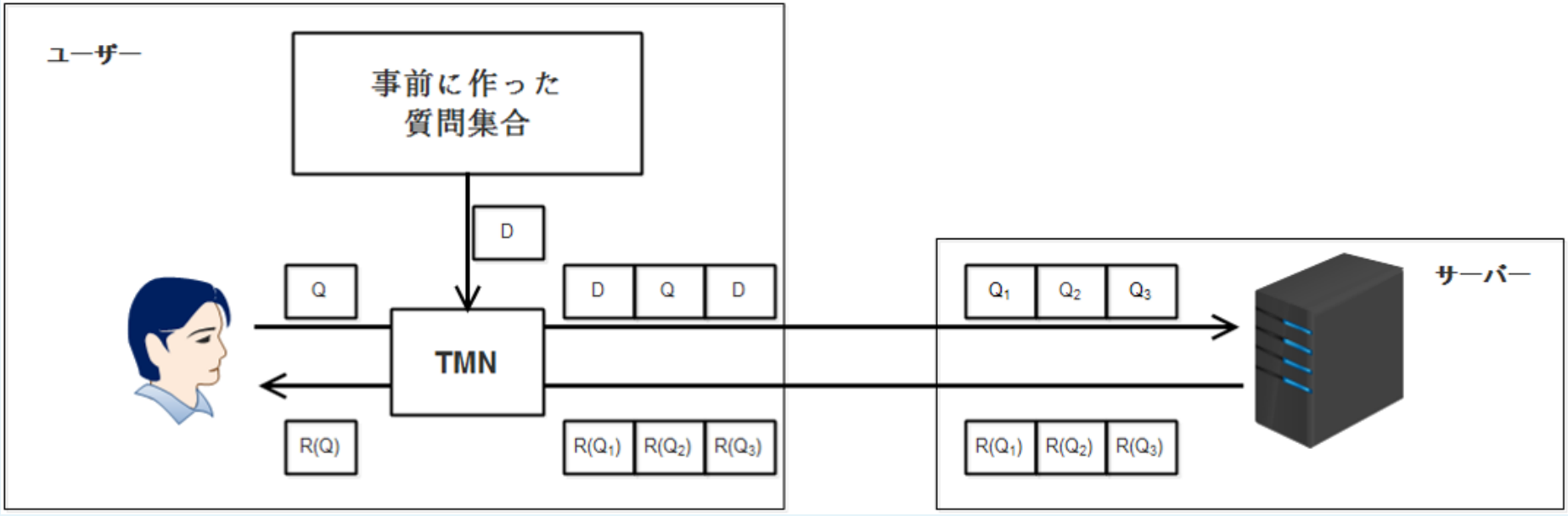
Location

MaskIt (GNG12)

- ユーザーの行動パタンを Markov chain で表現する
- 攻撃者がユーザーの Markov chain を持っていると仮定する
- Markov chain を用いて公開していけない情報を特定する

Snapshot vs Sequence

Keyword



TrackMeNot (HN09)

一つ一つの質問ではなく、ユーザーの質問ログにノイズを加える

Others

クラウドストア検索

- CryptDB (PRZB11)
  - **ユーザーが自分が暗号化した、クラウド上のデータを暗号化したままで検索する方法**

データベースの情報を守る

- Simulatable Auditing (KMN05)
  - **ユーザーの差分質問からデータベース上の情報を守る**

Linkage attack

Websearch

SimAttack (PCB+16)
Quantifying Web-Search Privacy (GSS+14)
TrackMeNot-so-good-after-all (ARJP12)

匿名化

Robust De-anonymization of Large Sparse Datasets (NS08)

差分攻撃

The Mastermind Attack on Genomic Data (Goo)

Bibliography I

Miguel E. Andrs, Nicols E. Bordenabe, Konstantinos Chatzikokolakis, and Catuscia Palamidessi.
Geo-indistinguishability: Differential Privacy for Location-based Systems.
In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, CCS '13, pages 901–914, New York, NY, USA, 2013. ACM.

Avi Arampatzis, Pavlos S. Efraimidis, and George Drosatos.
A query scrambler for search privacy on the internet.
Information Retrieval, 16(6):657–679, October 2012.

Rami Al-Rfou', William Jannen, and Nikhil Patwardhan.
TrackMeNot-so-good-after-all.
arXiv:1211.0320 [cs], November 2012.
arXiv: 1211.0320.

Zeller Barbaro.
A Face Is Exposed for AOL Searcher No. 4417749 - New York Times.

E. Balsa, C. Troncoso, and C. Diaz.
OB-PWS: Obfuscation-Based Private Web Search.
In *2012 IEEE Symposium on Security and Privacy*, pages 491–505, May 2012.

Michaela Gtz, Suman Nath, and Johannes Gehrke.
MaskIt: Privately Releasing User Context Streams for Personalized Mobile Applications.
In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, SIGMOD '12, pages 289–300, New York, NY, USA, 2012. ACM.

Bibliography II

Michael T. Goodrich.
IEEE Xplore Abstract - The Mastermind Attack on Genomic Data.

Arthur Gervais, Reza Shokri, Adish Singla, Srdjan Capkun, and Vincent Lenders.
Quantifying Web-Search Privacy.
In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 966–977, New York, NY, USA, 2014. ACM.

Daniel C. Howe and Helen Nissenbaum.
TrackMeNot: Resisting surveillance in web search.
Lessons from the Identity Trail: Anonymity, Privacy, and Identity in a Networked Society, 23:417–436, 2009.

Krishnaram Kenthapadi, Nina Mishra, and Kobbi Nissim.
Simulatable Auditing.
In *Proceedings of the Twenty-fourth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '05, pages 118–127, New York, NY, USA, 2005. ACM.

Hua Lu, Christian S. Jensen, and Man Lung Yiu.
PAD: Privacy-area Aware, Dummy-based Location Privacy in Mobile Services.
In *Proceedings of the Seventh ACM International Workshop on Data Engineering for Wireless and Mobile Access*, MobiDE '08, pages 16–23, New York, NY, USA, 2008. ACM.

Bibliography III

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.

A. Narayanan and V. Shmatikov.
Robust De-anonymization of Large Sparse Datasets.
In *2008 IEEE Symposium on Security and Privacy (sp 2008)*, pages 111–125, May 2008.

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.

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.

Bibliography IV

Raluca Ada Popa, Catherine M. S. Redfield, Nickolai Zeldovich, and Hari Balakrishnan.
CryptDB: Protecting Confidentiality with Encrypted Query Processing.
In *Proceedings of the Twenty-Third ACM Symposium on Operating Systems Principles*, SOSP '11, pages 85–100, New York, NY, USA, 2011. ACM.