プライバシーを保護する特許検索

中川研究室 修士 2 年 胡 瀚林
指導教員： 中川 裕志 教授

2016 年 7 月 1 日

概要

1 INTRODUCTION

1.1 Patent Search

特許文章の特徴
特許検索の目的と方法
ーー新規性調査 (Novelty Search)

1.2 Patent Versus Non-patent Literature

特許文章と普通の文章の区別

2 PRIVATE INFORMATION RETRIEVAL

PIR の背景紹介

2.1 Private Information Retrieval

2.2 Obfuscation-Based Private Search

既存手法とその手法が特許検索に適用できない理由

3 LATENT SEMANTIC MODELS

3.1 tf-idf

3.2 Latent Semantic Indexing

長所:計算簡単
短所:トピックベクトルが直交である

3.3 Probabilistic Latent Semantic Indexing

長所:確率的モデル
短所:トレーニングセットに含まれていない文章 (質問) の分析が困難である

3.4 Latent Dirichlet Allocation

長所:確率的モデルトレーニングセットに含まれていない文章 (質問) の分析が簡単
短所:学習するときは単語数 x トピック数の行列を用いて反復するので学習するには時間がかかる (30 トピック、1000 反復は 3 日かかる)

4 privacy-protecting patent search

提案手法
評価 (攻撃) 方法

5 EXPERIMENT

実験
1 質問者:tfidf 攻撃者 LSA
2 質問者:LSA 攻撃者 LSA
3 質問者:LDA 攻撃者 LSA
4 質問者:LSA 攻撃者 LDA

6 CONCLUSIONS

7 FUTURE WORKS

符号	意味
N	辞書中の単語の数
$T = 1, 2, 3, \ldots, N$	単語集合
M	コーパス中の文書の数
$D = 1, 2, 3, \ldots, M$	文章集合
K	トピック数
$\ell_i = t_1, t_2, \ldots, K$	単語 i のトピックベクトル
ℓ	質問のトピックベクトル

表 1　表記法

Algorithm 1 メイントピック攻撃

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

$R = \phi$, $\ell = 0$
$\ell = \sum_{t_i \in Q} \ell_{t_i}$
$maintopic = argmax_j \ell[j]$
for all $bk_k \in q$ **do**
    $R = R \cup max_{t_i} l_{t_i}[maintopic]$
end for
return R

Algorithm 2 類似攻撃

Input: 質問集合:$Q = \{q_i^r \,|i \in \{1,2,3,4\}, r \in \{1,2,\ldots,R\}\}$, 単語のトピックベクトル集合 $L = \{\ell_i\}$

$p_i = q_i^1 \;\; i \in \{1,2,3,4\}, result = \phi$
for $r = 2, 3 \ldots, R$ **do**
    for $i = 1, 2, 3, 4$ **do**
        $j = argmax_j \frac{p_i \cdot q_j^r}{|p_i||q_j^r|}$
        $d_i = \frac{p_i \cdot q_j^r}{|p_i||q_j^r|}$
        $temp_i = \frac{1}{r}(p_i(r-1) + q_j)$
    end for
    for $i = 1, 2, 3, 4$ **do**
        $p_i = temp_i$
    end for
    $result = result \cup argmax_i d_i$
end for
return $result$