修士論文

特許検索における質問意図の曖昧化

48-156229　**胡瀚林**

指導教員　中川裕志 教授

2017 **年** 1 **月**

東京大学大学院情報理工学系研究科数理情報学専攻

概要

企業が特許を取る前に，類似な特許が既に存在するかを確かめるために特許データベースを検索する必要がある．しかし，検索の質問から企業秘密が漏洩する可能性がある．ウェブテキスト検索の質問からユーザーの検索意図を守る手法が多数存在している．その中真の質問と同時にダミー質問を提出する質問曖昧化手法が一番効率的、現実的である．本論文では特許検索における既存な質問曖昧化手法 [**?**, **?**, **?**] を実装し，類似度攻撃 [**?**] で既存手法の安全性を評価した．

また、類似度攻撃 [**?**] を含め、多くの既存な質問曖昧化に対する攻撃手法は攻撃者が質問者の事前情報を持つと仮定する．本論文では事前情報なしの攻撃手法を提案し，その攻撃手法に対応できる既存な質問曖昧化の改良と新たな質問曖昧化手法を提案する．

目次

第 1 章

はじめに

テキスト検索をするとき，検索質問をサーバー側に渡さなければならない．しかし，検索質問から質問者の情報が漏洩する危険があることが AOL 事件 [?] より証明された．特許検索の場合は検索質問が研究開発動向など企業秘密を含んでいるため，一般的なウェブ検索の質問者より質問のプライバシー問題を重視している．そのような問題を解く様々な手法が存在している．[] や [] などの IP アドレスの匿名化メカニズムは登録情報が必要な検索サーバーに対応できない．また検索質問のみから質問者を一意に特定されてしまう可能性がある [?]．プライベート情報検索 (Private Information Retrieval)[] は計算量的安全性を持つが，サーバー側で大量の計算が必要であるため実用するのは難しい．曖昧化検索 (Obfuscation Search)[] は真の質問を分析し適切な $K - 1$ 個のダミー質問を生成し真の質問と同時に検索する．安全性が弱いが，効率よく質問者の検索意図を守ることができる．

本論文の構成は次の通りである．第二章では特許文章と特許検索の特徴を述べる．第三章では既存な質問曖昧化メカニズム [?, ?, ?] を述べる．第四章では曖昧化メカニズムがよく用いる意味分析手法を述べる．第五章では既存な攻撃手法 [?] を述べ，[?] の改良と新たな攻撃手法を提案する．第六，七章では新たな質問曖昧化手法を提案する．最後に, 第八章で評価実験を述べ，第九章で全体をまとめる．

第 2 章

特許

特許検索質問のプライバシーを保護する手法を説明する前に特許検索と特許そのものを簡単に紹介する必要がある．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．[?] 特許を取るには以下の条件を満たさなければならない: 新規性:公知の発明と同様の発明は特許を受けることができない; 進歩性:先行技術に基づいて容易に発明をすることができる発明は特許を受けることができない．単一性:発明の単一性の要件を満たさない二以上の発明は一つの願書で出願することができない.

特許を受けようとする発明を特定するために特許請求の範囲を記載する必要がある.

図 2.1 で表した例のように，特許の請求項は特定の書き方がある．誤解を招かないように技術用語は、学術用語を用いる．また，一般的な文章は単語をなるべく重複しないようにする一方，特許文章は単語を全体を通じて統一して使用する.

2.1 特許分類

特許の一つ特徴は全ての特許が人の手によって分類されている．特許分類を用いることより検索する特許文章が減り，似たようなキーワードを含むが分類が違う特許文章を排除することができる．今最も使われている特許分類が世界知的所有権機関 (WIPO) による管理されている国際特許分類 (IPC) である．国際特許分類は階層構造であり，一番上の階層は A から H までの 8 個のセクションである．セクション以下は**??**に表したように四つの階層に分類されている.

2.2 特許検索

(19)日本国特許庁(JP)　(12)公開特許公報(A)　(11)特許出願公開番号
特開2016-208844
(P2016-208844A)
(43)公開日　平成28年12月15日(2016.12.15)

(51)Int.Cl.　F I　テーマコード（参考）
A01B 35/04 (2006.01)　A01B 35/04 E　2B034
A01B 35/04 B

審査請求 未請求　請求項の数 4　OL　（全 13 頁）

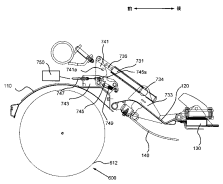
(21)出願番号　特願2015-92186(P2015-92186)
(22)出願日　平成27年4月28日(2015.4.28)

(71)出願人　390010836
小橋工業株式会社
岡山県岡山市南区中畦６８４番地
(74)代理人　110000408
特許業務法人高橋・林アンドパートナーズ
(72)発明者　河原　文雄
岡山県岡山市南区中畦６８４番地　小橋工業株式会社内

Ｆターム(参考)　2B034 AA03 BA06 BB01 BB02 EA02 EB06 EB33 JA06

(54)【発明の名称】農作業機

(57)【要約】
【課題】代かき作業機を昇降させる必要がない場面において、オート装置が代かき作業機を昇降させることを防止する。
【解決手段】本発明の一実施形態に係る農作業機は、耕耘作業を行うロータリ作業部を回転自在に支持する機体と、機体に設けられ、ロータリ作業部の上部を覆うカバー部と、カバー部の後端部に回動可能に支持されたエプロンと、エプロンの背面に取り付けられた支持部材と、カバー部に取り付けられ、エプロンのロック状態とフリー状態を切り替え可能なエプロン回動制御部と、を備え、エプロン回動制御部は、後端部が支持部材に対し取り付けられたロッド部と、ロッド部の前端部を回動自在に支持し、被係合部を有する第１アーム部と、カバー部に回動自在に支持され、係合部を有する第２アーム部と、第２アーム部を回動させる駆動部とを有し、係合部は、第１アーム部が回動するときに、被係合部の回動を規制するように構成されてもよい。
【選択図】図１



【特許請求の範囲】
【請求項１】
耕耘作業を行うロータリ作業部を回転自在に支持する機体と、
前記機体に設けられ、前記ロータリ作業部の上部を覆うカバー部と、
前記カバー部の後端部に回動可能に支持されたエプロンと、
前記エプロンの背面に取り付けられた支持部材と、
前記カバー部に取り付けられ、前記エプロンが自由に回動できない状態であるロック状態と前記エプロンが自由に回動できる状態であるフリー状態とに切り替えることができるエプロン回動制御部と、を備え、
前記エプロン回動制御部は、後端部が前記支持部材に対して、摺動可能に取り付けられたロッド部と、
前記ロッド部の前端部を回動自在に支持し、前記カバー部に対して回動自在に支持されて、被係合部を有する第１アーム部と、
前記カバー部に回動自在に支持され、係合部を有する第２アーム部と、
前記第２アーム部を回動させる駆動部とを有し、
前記係合部は、前記ロッド部が前方に移動するのに伴い前記第１アーム部が回動するときに、前記被係合部の回動を規制することを特徴とする農作業機。
【請求項２】
前記被係合部は、ピン部材であることを特徴とする請求項１に記載の農作業機。
【請求項３】
前記駆動部は、ワイヤとワイヤ制御部を含むことを特徴とする請求項１又は請求項２に記載の農作業機。
【請求項４】
耕耘作業を行うロータリ作業部を回転自在に支持する機体と、
前記機体に設けられ、前記ロータリ作業部の上部を覆うカバー部と、
前記カバー部の後端部に回動可能に支持されたエプロンと、
前記エプロンの背面に取り付けられた支持部材と、
前記カバー部に取り付けられ、前記エプロンが自由に回動できない状態であるロック状態と前記エプロンが自由に回動できる状態であるフリー状態とに切り替えることができるエプロン回動制御部と、を備え、
前記エプロン回動制御部は、後端部が前記支持部材に対して、摺動可能に取り付けられたロッド部と、
前記ロッド部の前端部を回動自在に支持し、前記カバー部に対して回動自在に支持されて、係合部を有する第１アーム部と、
前記カバー部に回動自在に支持され、被係合部を有する第２アーム部と、
前記第２アーム部を回動させる駆動部とを有し、
前記被係合部は、前記ロッド部が前方に移動するのに伴い前記第１アーム部が回動するときに、前記係合部の回動を規制することを特徴とする農作業機。
【発明の詳細な説明】
【技術分野】
【０００１】
本発明は、農作業機に関する。特に、本発明は、エプロンが自由に回動できない状態であるロック状態とエプロンが自由に回動できる状態であるフリー状態とに切り替えることができるエプロン回動制御部を備える農作業機に関する。
【背景技術】
【０００２】
耕耘ロータにより耕耘された耕土を整地するエプロン（第１整地板）とエプロンの後部に上下方向に回動自在に設けられて耕土表面を均平にするレベラ（第２整地板）を備える農作業機、例えば、代かき作業機は、一般に、走行可能な走行機体の後部に三点リンク連結機構を介して昇降可能に連結されて、走行機体の前進走行とともに進行しながら代かき

図 2.1. 特許文章例

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

表 2.1. 国際特許分類例:A61C 5/08

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

表 2.2. 特許検索タイプー

符号	意味
N	辞書中の単語の数
$W = \{1, 2, 3, \ldots, N\}$	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
$T = \{1, 2, 3, \ldots, K\}$	トピック集合
$\ell_i = \{t_1, t_2, \ldots, K\}$	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 2.3. 表記法

第 3 章

曖昧化検索

曖昧化検索は質問者が検索したい真の質問と質問者側で生成したダミー質問を一緒に検索サーバーに提出し，真の質問がどれかを曖昧化するものである．本論文では以下のモデル [] を用いて既存な曖昧化検索メカニズムを分析する．質問者 Alice がとある検索サーパーに質問を出して手に入れたい情報を検索し，検索サーバーが semi-honest な攻撃者であることを仮定する．

質問が単語の集合であり，質問の定義域を単語集合の冪集合にする．

定義 1. 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

$$Q = P(W) = \{A|A \subset W\} \tag{3.1}$$

Alice のプロフィールを多項分布と仮定し，Alice が持つ真のプロフィールを X とする．

定義 2. 質問者のプロフィール X．T を全てのトピックの集合とする．質問者のプロフィール X とは

$$X = \{x_i|i \in T\} \tag{3.2}$$

x_i は質問者がトピック i に対して持つ興味の強さを表す．

曖昧化検索メカニズムは Alice のコンピュータで実行する．曖昧化検索メカニズムが意味分析ツール S を用いて真の質問 q_R を分析しダミー質問 q_D を生成し，検索サーバーに提出する．質問 q とトピック t の関係を表す関数と質問間の距離は以下のように定義する，

定義 3. 質問-トピックスコア関数:$rscore_S$．T を全ての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

$$rscore_S : Q \times T \to \mathbb{R} \tag{3.3}$$

定義 4. 質問間距離関数:$dist_S$．質問 q_1 と質問 q_2 間の距離を表す関数とは

$$dist_S : Q \times Q \to \mathbb{R} \tag{3.4}$$

検索サーバーが Alice からもらった質問をすべて記録し，その質問たちを分析し得るプロフィールを Y にする．

曖昧化検索は 3 つ違うレベルな目的がある．まずは質問そのものの曖昧化である．質問者が検索した真の質問 q_R はどの質問であるかをわからないようにする．2 つ目は質問意図の曖昧化である．質問者が検索したいものは何であるかをわからないようにする．最後は質問者のプロフィール X の曖昧化である．Y から質問者が興味を持つトピックは何であるかをわからないようにする．

質問の曖昧化ができたとしても質問意図の曖昧化ができると限れない．林檎とリンゴの 2 つ質問から真の質問を確定することができないが，質問者が林檎について検索したいことが確定できる．同じように林檎と梨の 2 つ質問から質問者が検索したいを確定することができないが，質問者が果物に興味を持つことが確定できる．本論文では質問意図の曖昧化をメインにする．

3.1　否認可能検索を利用したプライバシー保護 [?]

否認可能検索という概念を提出したのは [] である．

定義 5. k − 否認可能検索質問 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 −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

3.2　質問者のプライバシーを保護する質問加工法 [?]

3.3　質問意図を曖昧化するキーワード検索 [?]

Algorithm 1 HDGA(On Masking Topical Intent in Keyword Search)

Input: 質問:q_1

$Q = \{q_1\}\delta_{q_1} = \underset{t \in T}{\operatorname{argmax}} Pr[t|q_1]$
for all $t \in T \setminus \{\delta_{q_1}\}$ **do**
    $e_t = h(\delta_{q_1}||t||s)$
end for
$T_D = \{t^1_{q1}, t^2_{q1}, \ldots, t^2_{q1} | \forall t_1 \ \in T_D, \forall t_2 \ \in T \setminus T_D, e_{t_1} > e_{t_2}\}$
for all $t \in T_D$ **do**
    while $\underset{t \in T}{\operatorname{argmax}} Pr[t|q'] \neq t$ **do**
        randomly select $|q_1|$ keywords for t based on $Pr[w|t]$, to form a dummy query q'
    end while
    $Q = Q \cup \{q'\}$
end for
Shuffle queries in Q **return** Q

第 4 章

意味分析

4.1 tf-idf

4.2 **潜在意味解析**

4.3 **潜在的ディリクレ配分法**

第 5 章

プライバシー分析 (攻撃手法)

5.1 メイントピック攻撃

Algorithm 2 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 $L = \{\ell_i\}$

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

5.2 類似度攻撃 [?](事前情報あり)

Algorithm 3 類似度計算

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

for $q_i \in P_u$ **do**
    $coef[i] \leftarrow 2 \cdot |q \cap q_i| \cdot \frac{1}{|q|+|q_i|}$
end for
$coef \leftarrow sort(coef)$
$sim \leftarrow coef[0]$
for $i \in [1, |P_u|]$ **do**
    $sim \leftarrow \alpha \cdot coef[i] + (1 - \alpha) \cdot sim$
end for

Output: sim

Algorithm 4 類似度攻撃

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

$q^* = \underset{q \in Q}{\operatorname{argmax}}\, sim_{q,Pu}$

Output: q^*

5.3　類似度攻撃 2(事前情報なし)

Algorithm 5 類似度攻撃

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:$\alpha$

for $j \in |Q_1|$ **do**
　$\hat{Pu}[j] = Q_1[j]$
　$\hat{Put}[j] = \Phi$
　$d[j] = 0$
end for
for $i \in [2, n]$ **do**
　for $j \in |Q_i|$ **do**　State $\hat{Put}[j] = \underset{Pu \in \hat{Put}}{\operatorname{argmax}}\, sim_{Q_i[j], \hat{Put}[j]}$
　end for
　$q_i^* = \underset{Q_i[j] \in Q_i}{\operatorname{argmin}}\, sim_{Q_i[j], \hat{Put}[j]}$
　for $j \in |Q_i|$ **do**
　　$\hat{Pu}[j] = \hat{Put}[j] \cap Q_i[j]$
　end for
end for

Output: q^*

第 6 章

質問曖昧化 (提案手法)

事前に単語をグループにする [?]，質問をグループにする [?] と同じようにトピックをグループにすることよりトピック出現頻度で質問者が興味あるトピックを特定することが防ぐと考えられる．

[?] ではハッシュ関数でトピックをグループにしているが，各トピック間の関係を配慮していない．また，[?] では各ダミートピックから単語をランダムに選ぶため，真の質問に含まれている単語は違っても属するトピックは同じならダミー質問が同じような性質を持つ。一方，同じ真の質問に対して同じダミー質問を生成することができない．真の質問に含まれている単語という情報を用いてないことが [?] に提案した手法が Simattack に弱い原因だと考えられる．simattack から真の質問を守るために真の質問が同じ単語を含むとき，ダミー質問も同様に同じ単語を含んでほしい．それを実現するため提案手法では単語ベクトルを用いた．

6.1 単語ベクトル

定義 6. 単語ベクトル 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

$$
\begin{aligned}
&l_t = \{w_1, w_2, \ldots, w_{|W|}\}, \\
&\forall w \in l_t, w \in W \\
&\forall 1 \leq i \neq j \leq, w_i \neq w_j \\
&\forall 1 \leq i < j \leq |W|, rscore(w_i, t) \geq rscore(w_j, t)
\end{aligned}
\tag{6.1}
$$

質問のメイントピックを計算し，質問に含まれている単語をその単語が質問のメイントピックの単語ベクトルにいる順番にすれば，質問を数字ベクトルで表わすことができる．同様にトピックが決めれば，そのトピックの単語ベクトルを用いて数字ベクトルを質問に翻訳することができる．

単語ベクトル内の単語が単語とそのトピックの関連値の大きい方から小さい方までに並ぶため，単語ベクトルに同じ順番を持つ単語がその単語ベクトルを持つトピックに対して同じ様な関連性を持つと考えられる．また，同じ数字ベクトルで表わせる質問もその質問が属するトピックに対して同じ様な関連性を持つと考えられる．

したがって，単語ベクトルを通じて違うトピックに属するが似たような特徴を持つ質問を作るこたができる．

6.2　質問曖昧化

第 7 章

データベース分割

特許分類を用いることにより特許データベースを分割することができる．分割したデータベース各々に対して同じような信憑性を持つ質問を提出すると真に検索したいデータベースを隠すことができると考えられる．

第 8 章

評価実験

8.1 **データベース**

8.2 tfidf vs lda vs lsa

8.3 **データベース分割**

8.4 **検索結果分析 (真の質問が当たられる確率 vs ダミー質問と真の質問の検索結果の類似度)**

第 9 章

おわりに

16

謝辞

付録 A