修士論文 要旨

特許検索における質問意図の曖昧化

数理情報学専攻 48-156229 胡 瀚林

指導教員 中川 裕志 教授

1 はじめに

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．テキスト検索をするとき，検索質問をサーバ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．ウェブテキスト検索の質問から質問者の検索意図を守る手法が多数存在している．その中では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の質問曖昧化手法 [?, ?, ?] を実装し，類似度攻撃 [?] で特許データベースにおける既存手法の安全性を評価した．また，類似度攻撃を含め，多くの既存の質問曖昧化に対する攻撃手法は攻撃者が質問者に関する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する既存の質問曖昧化の改良と新たな質問曖昧化手法を提案し，特許データベースにおける評価実験を行う．

2 特許の概要

特許文書は発明を正確に規定するために普段に使わない学術用語を用い，単語を全体を通じて統一して使用して単語を曖昧性を無くす．また特許文書は世界標準である国際特許分類コードが付いている．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．

3 既存研究

事前に質問をグループにする：否認可能検索 (PDS) は文書集合から高頻度な単語と単語ペアをシード質問として抽出し，潜在意味分析 (LSA)[?] を用いてシード質問をトピック空間にマップし，トピック空間に距離が近いシード質問をクラスタリングして標準質問にし，トピック空間に距離が遠い標準質問で PD-質問集合を構築する．検索する場合は，質問者が検索したい質問の代わりに事前に用意した標準質問集合からトピック空間において質問者が検索したい真の質問と最も近い標準質問が属する PD-質問集合をサーバに提出し，サーバから検索結果を得，質問者側で真の質問を用いて検索結果をフィルタリングする．

事前に単語をグループにする：質問者のプライバシーを保護する質問加工法 (ETSQ) は単語を類義関係のセット (synset) でグループ化する WordNet[?] を用いて意味的に遠い単語を１つ単語バケットにし，真の質問単語が属するバケットの中の他の単語を全てダミー単語として質問に加え，１つの加工した質問として検索サーバに提出する．暗号したままの暗号文を加算できる加算可能な準同型暗号 [?] を用いることにより真の質問の単語だけ検索することができる．

事前にトピックをグループにする：質問意図を曖昧化するキーワード検索 (HDGA) は潜在的ディリクレ配分法 (LDA)[?] を用いてコーパスにおける各トピック t における単語 w の出現率 $Pr(w|t)$ を計算する．検索する場合はハッシュ関数 HRW[?] を用いてダミートピック t' を選び，$Pr[w|t']$ に基づいて単語をランダムに選び，真の質問と同じ長さのダミー質問を作る．

4　本研究で提案するアルゴリズム

以上の背景を踏まえ，本研究では以下のアルゴリズムを提案した．

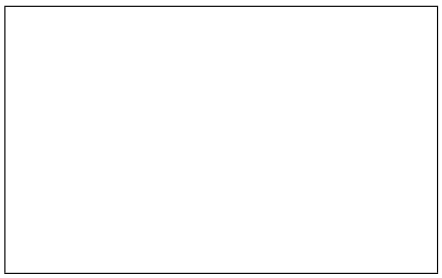

図 1. 提案アルゴリズム．

参考文献