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 [Michael and Tom, 2006] 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 [Petit et al., 2016] 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．

特許検索の質問は一般なウェブ検索質問より長い．

3 既存研究

3.1 曖昧化手法

事前に質問をグループにする手法 (PDS)[Murugesan and Clifton, 2009]：否認可能検索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)[Deerwester et al., 1990] 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手法 (ETSQ)[Pang et al., 2010]：質問者のプライバシーを保護する質問加工法は単語を類義関係のセット (synset) でグループ化する WordNet[Miller, 1995] 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パに提出する．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 [Benaloh, 1994] 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手法 (HDGA)[Wang and Ravishankar, 2014]：質問意図を曖昧化するキーワード検索は潜在的ディリクレ配分法 (LDA)[Blei et al., 2003] 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[Thaler and Ravishankar, 1998] 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．

3.2 攻撃手法

事前情報がある場合の類似攻撃 (SimAtt)[Petit et al., 2016]：SimAtt は質問者が提出した質問と攻撃者が事前に得た質問者の質問ログ間の類似度を計算し，同じ質問グループの中の質問ログとの類似度が一番高い質問を真の質問とする．

4 本研究で提案するアルゴリズム

4.1 攻撃手法

メイントピック攻撃 (MTA)：ダミー質問が真の質問と同様に全ての単語が 1 つのトピックに集中することが失敗したら，真の質問と真の質問のメイントピックの関連値がダミー質問とダミー質問のメイントピック

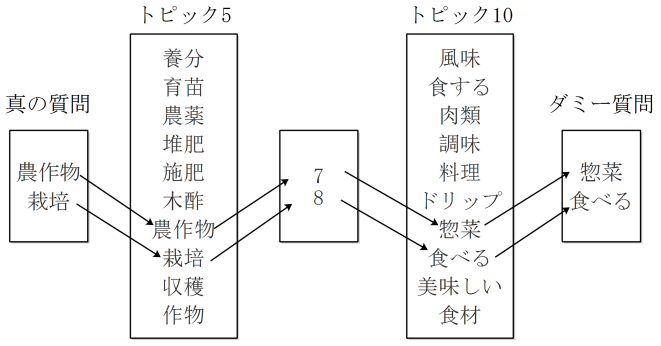
の関連値より強いと考えられる．MTA は 1 つの質問グループの中で自分のメイントピックとの関連値が一番高い質問を真の質問とする．

事前情報がない場合の類似度攻撃 (SimAtt2)：攻撃者が事前的に質問者の真の質問ログを持たないと SimAtt を用いることができない．SimAtt2 は意味的に近い一連の質問が真の質問の列であると考える．SimAtt2 は 1 つの質問グループに属する質問と同じ数の質問列を可能な真の質問列として保存し，次にきた質問グループの各質問に対して各可能な真の質問列との類似度を計算し，一番類似度が高い質問列に加える．類似度が一番高い質問と質問列のペアを真の質問の列とする．

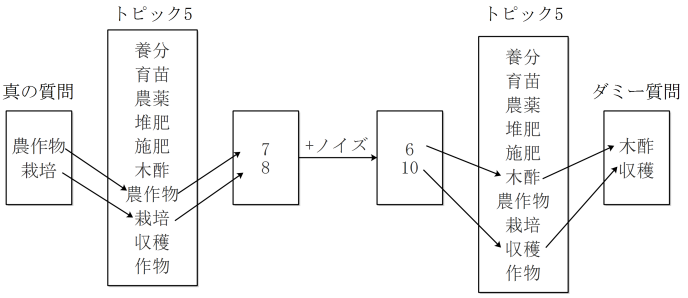
4.2 単語ベクトルを用いた質問曖昧化

単語ベクトル：全ての単語を単語とトピック t の関連値を大きい順に並べるベクトルをトピック t の単語ベクトルという．

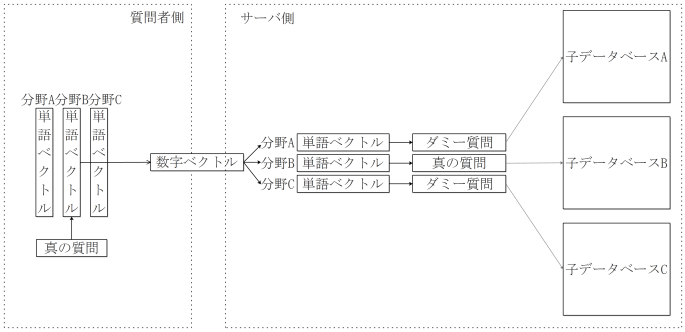
質問者が検索したいトピックを曖昧化する質問曖昧化 (QOT)：QOT は事前にトピックをグループにする．検索する場合は真のトピックが属するトピックグループの中の他のトピックをダミートピックとし，以下の図のように単語ベクトルを用いてダミー質問を作る．



質問者が検索したいトピックにおける質問曖昧化 (QOI)：QOI は以下の図のように単語ベクトルを用いて真の質問を数字ベクトルにし，数字ベクトルの各要素に対して雑音を加え，ダミー質問にする．



データベース分割：データベース分割では他の曖昧化と違って全ての子データベース，るいはトピックについて以下の図のように質問を提出する．



参考文献

[Benaloh, 1994] Benaloh, J. (1994). Dense probabilistic encryption. In *Proceedings of the workshop on selected areas of cryptography*, pages 120–128.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Deerwester et al., 1990] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6).

[Michael and Tom, 2006] Michael, B. and Tom, Jeller, J. (2006). A Face Is Exposed for AOL Searcher No. 4417749 - New York Times.

[Miller, 1995] Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

[Murugesan and Clifton, 2009] Murugesan, M. and Clifton, C. (2009). 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.

[Pang et al., 2010] Pang, H., Ding, X., and Xiao, X. (2010). Embellishing Text Search Queries to Protect User Privacy. *Proc. VLDB Endow.*, 3(1-2):598–607.

[Petit et al., 2016] Petit, A., Cerqueus, T., Boutet, A., Mokhtar, S. B., Coquil, D., Brunie, L., and Kosch, H. (2016). SimAttack: private web search under fire. *Journal of Internet Services and Applications*, 7(1):1.

[Thaler and Ravishankar, 1998] Thaler, D. G. and Ravishankar, C. V. (1998). Using name-based mappings to increase hit rates. *IEEE/ACM Transactions on Networking (TON)*, 6(1):1–14.

[Wang and Ravishankar, 2014] Wang, P. and Ravishankar, C. V. (2014). On masking topical intent in keyword search. In *2014 IEEE 30th International Conference on Data Engineering*, pages 256–267. IEEE.