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り，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．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既存研究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．

本研究では 2 つ代表的な質問曖昧化手法 [7, 10] の特許検索における安全性を実データを用いて評価する．そのため本研究では質問者の事前情報と新たな質問間の類似度で真の質問を選別する類似度攻撃 [8] と本研究で提案する事前情報を用いない攻撃手法を用いて既存の質問曖昧化手法を攻撃する．

また，上記の攻撃に対して既存研究より良い安全性を得られる新たな問曖昧化手法を提案する．

2 既存手法の流れと問題点

2.1 曖昧化手法

事前に質問をグループにする手法 (PDS)[6]：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)[3] 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．

PDS の問題点は真の質問ではなく標準質問を用いることによる再現率の低下である．

事前に単語をグループにする手法 (ETSQ)[7]：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[5] 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 [1] 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．

ETSQ の問題点はダミー単語が真の質問の単語のように 1 つのトピックに集中しないことである．

事前にトピックをグループにする手法 (HDGA)[10]：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)[2] 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[9] 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．

HDGA の問題点は真の質問が属するトピックという情報しか用いない点である．

2.2 攻撃手法

事前情報がある場合の類似攻撃 (SimAtt)[8]：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．

事前情報がないと利用できないことは SimAtt 問題点である．

3 本研究で提案する手法

3.1 攻撃手法

メイントピック攻撃 (MTA)：MTA は ETSQ の問題点を攻撃する手法である．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)：SimAtt2 は HDGA の問題点を攻撃する手法である．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．

3.2 単語ベクトルを用いた質問曖昧化

前節で挙げた既存手法の問題点を解決するため，本研究では単語ベクトルを用いる．トピック t の単語ベクトルとは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である．トピック t の属する質問の単語をその単語がトピック t の単語ベクトル内での順番にすることで質問を数字ベクトルにすることができる．また，数字ベクトル内の数字をトピック t の単語ベクトルに対応の順番の単語にすることでトピック t に属する質問を作ることができる．

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 は事前情報を持たない攻撃者に対応する質問曖昧化手法である．QOT では事前にトピックをグループにし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する．真のトピックの単語ベクトルを用いて真の質問を数字ベクトルにし，ダミートピックの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)：QOT は事前情報を持つ攻撃者に対応する質問曖昧化手法である．QOI では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．

4 評価実験

本研究では NTCIR-6[4] の無効資料調査タスクのデータセットを用いて評価実験をした結果が表 1 である．提案した事前情報を用いない攻撃手法は 60% 以上の確率で既存の質問曖昧化手法を見破った．事前情報を持たない攻撃者に対しては QOT-LSA が一番いい結果を得った．事前情報を持つ攻撃者に対しては QOI-LSA が一番いい結果を得った．

		攻撃者			
		MTA-LSA	MTA-LDA	SimAtt2	SimAtt
質問者	ETSQ	68.4	60.2	x	x
	HDGA	97.7	77.7	96.2	94.0
	QOT-LSA	12.2	23.8	49.4	70.4
	QOT-LDA	96.8	55.7	88.9	94.3
	QOI-LSA	17.2	26.2	56.9	50.3
	QOI-LDA	94.5	44.2	71.5	81.2

表 1. 検索曖昧化手法の比較

5 今後の課題

QOT と QOI は相互影響しないため， 両方同時に用いることが考えられる．また，どのような意味分析手法においても同じような強さを持つダミー質問を生成することも今後の課題として挙げられる．

参考文献

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

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[3] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), September 1990.

[4] Atsushi Fujii, Makoto Iwayama, and Noriko Kando. Overview of the Patent Retrieval Task at the NTCIR-6 Workshop. In *NTCIR*, 2007.

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

[6] 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.

[7] 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.

[8] 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.

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

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