D-027

特許の無効資料調査のための類似特許検索とリランキング

Patent Similarity Search and Reranking for Patent Invalidation

古川 修平† 関 洋平† 青野 雅樹†
Shuhei Furukawa Yohei Seki Masaki Aono

1. はじめに

日本の特許文献は，1 年間に約 40 万件が特許庁に出願され，そのうち約 15 万件が新たに登録される．特許の登録は先願主義で行うため，特許出願の際には，請求項で主張しようとする権利が先に出願されている他の特許文献に記載されていないかを検索，確認する必要がある．この無効資料調査は非常にコストが大きいため，機械で自動的にこの作業を行うシステムの実現が望まれている．

本研究では，無効資料調査を自動的に行うための前段階として，入力された文書と類似した特許文献を検索する手法を提案する．

提案手法では，入力文書と検索対象特許文献の文書ベクトルのコサイン類似度によってランキングした結果の上位 500 件に対して，特徴量空間が多様体構造を成すと仮定して解析を行う多様体ランキング[1]を用いて検索結果をリランキングする．

評価実験として，NTCIR-4 Patent Retrieval Task[4] で用いられたテストセットを用いて，提案手法を用いた類似文書検索と一般的なコサイン類似度による類似文書検索を DCG[3]を用いて比較，検討する．

2. 関連研究

特許の無効資料調査に関する研究はNTCIRワークショップを中心に数多く行われている．中でも，一度ランキング付けした検索結果に対してリランキングを行う手法として，似た特許には似たIPCが付けられるという特徴を利用した手法[2]などがある．

しかし，関連のある特許文献に別の IPC が付与される可能性があるため，IPC を用いると正解文献を取りこぼす恐れがある．本研究では，IPC など，特許から得られる固有の情報は用いず，文書の特徴ベクトルの構造のみを必要とする多様体ランキングを用いることで，データセットの分野に左右されない類似文書検索のためのリランキングを実現する．

3. 多様体ランキングを用いた類似文書検索

提案手法では，特許文書ベクトルをクエリ文書ベクトルとのコサイン類似度によってランキングした結果の上位 500 件に対して多様体ランキングを用いることで新たなランキング結果を得る．

多様体ランキングは，特徴量空間が多様体構造を成すと仮定し，特徴量間の類似度をもとに特徴量空間を解析する．本節では，多様体ランキングのアルゴリズムと，提案手法における適用について述べる．

3.1 多様体ランキングのアルゴリズム

以下のような m 次元の n 個のデータセットが与えられたとする．

$$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q, \boldsymbol{x}_{q+1}, \ldots, \boldsymbol{x}_n\} \subset \mathrm{R}_m$$

ここで，$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q$ は検索クエリのデータ，その他は検索対象データベースに含まれるデータである．また，$\boldsymbol{x}_i$ に対してランクキングスコア f_i を与えるランキング関数を f とする．この f は，$\boldsymbol{f} = [f_1, \ldots, f_n]^T$ というベクトルと見なす事ができる．最後に，$\boldsymbol{x}_i$ が検索クエリである場合に $y_i = 1$，それ以外では $y_i = 0$ となるベクトル $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ を定義する．これらを用いて，多様体ランキングのアルゴリズムは以下の手順になる．

i. データ同士の距離を昇順に並べ，連結グラフが作られるまで距離の近いデータ間を接続していく．
ii. データ $\boldsymbol{x}_i$ と $\boldsymbol{x}_j$ が接続している場合に，要素が $\boldsymbol{W}_{ij} = \mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j)$ で定義される類似行列 $\boldsymbol{W}$ を計算する．sim は 2 つのデータの類似度を計算する関数，対角要素は $W_{ii} = 0$ とする．
iii. (i, i) の要素が $\boldsymbol{W}$ の i 行の総和である対角行列 $\boldsymbol{D}$ を用いて，$\boldsymbol{W}$ を $\boldsymbol{S} = \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$ と対称正規化する．（Symmetrically normalize）
iv. $\boldsymbol{f}(t+1) = \alpha \boldsymbol{S} \boldsymbol{f}(t) + (1-\alpha)\boldsymbol{y}$ を収束するまで繰り返す．α は 0 以上 1 未満 の値をとるパラメータである．
v. $\boldsymbol{f}(t)$の収束値を $\boldsymbol{f}^*$とする．$\boldsymbol{f}^*$の値 f^*_iを降順にソートし，対応する文書 i の新たなランキング結果とする．

$\boldsymbol{f}(t)$は，$\boldsymbol{f}^* = \beta(\boldsymbol{I} - \alpha\boldsymbol{S})^{-1}\boldsymbol{y}$ に収束することが関数$\{f_i(t)\}$ に関する定理によって知られている．ただし，$\beta = 1-\alpha$ である．

3.2 類似特許検索への適用

本研究では，クエリ文書の文書ベクトルを $\boldsymbol{x}_1$とし，クエリ文書ベクトルに対してコサイン類似度で上位 i 番目の文書ベクトルを $\boldsymbol{x}_{i+1}$ とすることで，多様体ランキングを類似特許検索へ適用する．ただし，i は 500 以下である．

データセットが大きいため，連結グラフのリンクが少ないと，関連のあるデータ間が接続されない恐れがある．よって本研究では，連結グラフに k 近傍グラフ(k=10)を用いて，各データの 10 近傍のデータ群に対してリンクを持つグラフを作成する．グラフを作成するためのデータ間距離には，文書ベクトル間のコサイン類似度を 1 から引いた値を用いる．

類似行列 $\boldsymbol{W}_{ij}$ の要素の値となる $\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{x}_j)$にも，2 つのベクトル $\boldsymbol{x}_i, \boldsymbol{x}_j$のコサイン類似度を使用する．

† 豊橋技術科学大学
Toyohashi University of Technology

4. 実験

4.1 実験概要

提案手法の類似特許検索に対する効果を確認するために，多様体ランキングを利用したランキング（提案手法）と，文書ベクトルのコサイン類似度によって得たランキング（ベースライン）に対し，それぞれ上位 10 件の適合性を人手によって判定した．また，NTCIR-4 で用いられた無効資料調査タスクの正解セットによる適合性判定を行った．

検索課題として NTCIR-4 の PATENT Retrieval Task で用いられた正解文献が 1993 年から 1997 年の特許文献に含まれる Topic（010 から 040）を用いた．

検索対象集合として NTCIR-4 で使用されたデータセットの中から，1993 年から 1997 年までの日本語特許文献約 150 万件から Topic ごとの正解文献を含む 1 万件を抽出した．（正解文献以外はランダムに抽出）

4.2 実験条件

文書ベクトルは，特許文献全文を形態素解析器 MeCab[1] にかけた結果が「名詞」，「未知語」，または連続する「接頭辞，名詞，未知語，句読点以外の記号」で構成されるワードの TF 値を要素とした．このベクトル作成によって，1 万件の特許から約 45000 次元のベクトルが得られた．また，多様体ランキングに用いるパラメータ α は 0.44 とした．

4.3 適合性の判定方法

特許の類似性を人手で判定するには，要約の内容や請求項の内容が似ているかを比較することが自然である．

よって，検索課題として与えた文献の要約と請求項の内容を判定者に提示し，ランキング上位 10 件に出力された文献の内容と比較してもらい，適合性の判定を行った．

また，文献によって検索課題との類似性の度合いに違いが出ることを考慮し，適合性を以下の 3 段階で判定した．

- 適合：検索課題の要約と請求項の内容の 50%以上が出力文献中に現れる場合
- 部分適合：検索課題の要約と請求項の内容が 50%未満ではあるが出力文献中に現れる場合
- 不適合：検索課題の要約と請求項の内容が出力文献中に全く現れない場合

4.4 評価

ランキングの評価尺度として，多値の適合度に対応する DCG(Discounted Cumulative Gain)を用いた．

DCG は，次式で表される．

$$DCG_i = \begin{cases} G_1 & if\ i = 1 \\ DCG_{i-1} + \dfrac{G_i}{\log_b i} & otherwise \end{cases}$$

ここで，G_i は i 番目の文書の適合度の度合いを示す値である．本研究では，適合を 2，部分適合を 1，不適合を 0 とした．また，b は検索結果に対するユーザの寛容さを示すパラメータである．本研究では b=10 とした．

[1] MeCab URL: http://mecab.sourceforge.net/

提案手法とベースラインの全検索課題の検索結果上位 10 件までの DCG の平均値を示すグラフを図 1 に示す．図 1 より，提案手法がベースラインのランキングに比べ，類似特許検索において優位性を持つことが確認できる．

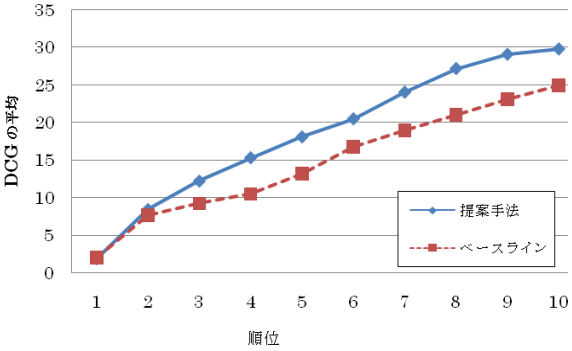

図 1　全検索課題の DCG の平均

4.5 考察

DCG の平均は提案手法を用いることで向上した．一方，Topic025（携帯電話装置）など 5 個のトピックでは，ベースラインの DCG が高くなった．これは，多様体ランキング適用前の検索結果の上位に不適合文書が数件存在し，多様体ランキングによってそれらの文書と関連する文書の順位が上がったためであると考えられる．

本研究の目的は，無効資料調査のための類似特許検索である．今回の手法を，NTCIR-4 の無効資料調査タスクの正解セットで評価したところ，ベースラインの検索とほぼ同等の精度となった．今回，ベクトル作成において，特許文献中のすべての項目を等しい重みで扱ったが，請求項部分の重みをあげるなどベクトル作成に工夫が必要である．

5. おわりに

特許の無効資料調査を機械的に行うために，多様体ランキングを用いた類似特許検索手法を提案した．その結果，類似特許文献の検索においては DCG の向上が見られた．一方，無効資料調査においては高い成果を上げることができなかった．多様体ランキングを適用する際の距離計算アルゴリズムと，無効資料調査に特化したベクトル作成が今後の課題である．

謝辞

NTCIR テストコレクションは国立情報学研究所の許諾を得て使用させていただきました．

参考文献

[1] D.Zhou, J.Weston, A.Gretton, O.Bousquet and B.Schölkopfk, “Ranking on Data Manifolds”, *Advances in Neural Information Processing Systems 16*, 169-176, MIT Press, Cambridge, MA, (2004).

[2] Kazuya Konishi, “Query Terms Extraction from Patent Document for Invalidity Search”, Proceedings of NTCIR-5 Workshop Meeting, December 6-9, (2005).

[3] K. Järvelin and J. Kekäläinen, “IR Evaluation Methods for Retrieving Highly Relevant Documents,” Proc. of the 23rd Annual International ACM SIGIR Conference and Development in Information Retrieval (SIGIR 2000), pp.41-48, Athens, Greece, July 2000.

[4] NTCIR(NII Test Collection for IR Systems) URL:http://research.nii.ac.jp/ntcir/index-ja.html