文書-法令条文グラフを利用した法律文書解析

吉川 克正
日本アイ・ビー・エム (株)
東京基礎研究所

村上 明子
日本アイ・ビー・エム (株)
ソフトウェア開発研究所

{katsuy, akikom}@jp.ibm.com

1 はじめに

法律文書解析は，裁判例，契約書，約款など，法的な情報を含む文書に対する解析研究である．法律文書は，内容の曖昧性を可能な限り排除することを求められる．ゆえに，その書式には特殊なルールが数多く存在し，新聞や Web データ等と比較して，一文当たりの文字数が多く，文書構造も複雑なことが多い．. 例えば,「又は」と「若しくは」では厳密な使い分けがされており，前者がより広い範囲の接続に利用されるなど，接続詞の優先度にも明確な序列がある．

類似文書検索は自然言語処理，情報検索の研究において，早くから数多くの研究があり，効果的な手法も確立されていると言える．しかしながら，法律文書データに対する類似文書検索は，一般的な文書検索手法では不十分な場合が考えられる．表 1 は，裁判例 Q「高知落石事件」の類似判例を検索した事例であるが，この中では主に 2 種類の検索誤りが示されている． 1 つ目は文書類似度は高いが，法的には類似性が低い文書を抽出してしまう問題で，C1 がその例にあたる．Q と C1 は，“道路”，“自動車”，“障害物”など，用語の重複は多いが，前者は既に起こった事故の国家賠償請求である．一方，後者は人格権的権利を主張した事例であって，訴訟類型がそもそも異なる． 2 つ目は文書類似度は低いものの，法的には類似度が高い文書を抽出できない問題で，C3 や C4 がその典型的な事例である．C3 と C4 はそれぞれ駅及び河川の機能的瑕疵を問うた国家賠償請求訴訟であり，Q の事例との法的類似性は高いと言える．表 1 中で，Q ともっとも法的に類似度が高いと考えられるのは C2 であり，ともに公道上の事故を対象として，賠償責任を問うた事例である．

これら 2 つの問題はいずれも，法的類似度という観点を明示的に導入することで解決できる．本研究ではこの法的類似度を,「共有する参照法令条文（法条）の数」と定義する．表 1 で裁判例 Q が参照している主要な法令条文は国家賠償法 2 条であり，事実，C2, C3, そして C4 はいずれもこの条文を参照しており，それを参照していない C1 だけが大きく異なることを明確に判別できるのである．

このように文書と参照法条間の関係を利用する有用性は直感的に理解できる．そこで本研究では 2 つのタスクに取り組むものとする．まず 1 つ目は，文書とそれが参照する法条を自動同定するタスク（参照法条同定）．もう 1 つはその自動同定された文書-参照法条間関係を用いて類似判例文書を検索するタスク（類似判例検索）である．いずれでも，文書と参照法条の 2 種類をノードとするグラフベースの手法を提案する．それぞれのタスクを異なる 2 種類のデータを利用して実験し，提案手法の効果を示す,

類似判例検索はそれそのものが，日々新たな法的問題に目を向け，情報を収集する必要のある法律家や法律学習者にとって価値がある技術だが，参照法条同定と組み合わせることで，既存の判例だけでなく，新規の案件に対しても利用できる技術になる．つまり，その案件に対しどの法律を適用し，どのようにその案件を解決するか，という情報を提供することができるようになると考えている．即ち，参照法条同定とは，類似判例検索のための付属技術ではなく，新たな案件を受けた法律家の思考プロセスを模倣した技術であると言える．

2 関連研究

法律情報処理技術は，これまでいくつかの評価型ワークショップ (Shared Task) を中心に，その研究が進められてきた．代表的なものとして，米国の e-Discovery のための情報検索を扱った，Text Retrieval Conference (TREC) Legal Track[1]，法律分野の構文解析を行う，Semantic Processing of Legal Texts (SPLeT) [3, 10]，司法試験の択一式問題を解析対象とした COLIEE-2015 [2] などがある．本研究でも，評価データの一つとして，COLIEE-2015 のデータを利用している．COLIEE-2015 では，IR タスク（Phase 1）と，QA タスク（Phase 2）

[1] http://trec-legal.umiacs.umd.edu/
[2] http://webdocs.cs.ualberta.ca/~miyoung2/COLIEE2015/

表 1: 文書類似度と法的類似度の違い

ID	判決日	名称	要旨
Q	[S45.8.20]	高知落石事件	国道上の障害物（落石）による自動車事故の賠償責任を問うた事例
C1	[H12.1.27]	車止め撤去請求事件	2 項道路上に設置された自動車の障害物に対する妨害排除請求
C2	[S50.7.25]	87 時間大型故障車放置事件	県道に放置された故障車の衝突事故の賠償責任
C3	[S61.3.25]	点字ブロック事件	旧国鉄の駅ホームに点字ブロックが未設置であったことによる瑕疵の有無
C4	[S50.6.26]	多摩川水害訴訟	国有河川の水害に対する国の賠償責任

があり，本研究では IR タスクのデータを利用する．

文書検索技術の研究は古くからあるが，その多くは少数のクエリ語から，求められる文書を検索するキーワードベースの手法である．しかし本研究で求められる文書検索は，ある文書そのものをクエリとして利用する類似文書検索 [2, 16, 5] であり，クエリ文書表現には Bag-of-Words ベクトル [14]，Generalized latent semantic analysis [9]，concept graph [1] など，様々な方法が考えられる．本研究では単純な TF-IDF 法により文書ベクトルを構築し，これを文書グラフに組み込むことで類似文書検索を実現している [13, 8].

3 グラフアルゴリズムによる文書検索法

この節では，本研究で利用するグラフベースの手法を，グラフ構造（3.1 節）及び検索アルゴリズム（3.2 節）に分けて述べる．このグラフ構造とアルゴリズムは，本研究で扱う 2 つのタスク，参照法条同定及び類似判例検索の双方で利用される．

3.1 文書-法条グラフ構造

本研究で構築するグラフの例を図 1 に示す．グラフには文書ノード（D ノード）と法条ノード (L ノード) という 2 種類のノードがあり，図 1 では D1-D4 が文書ノード，“国家賠償法 2 条 1 項”,“河川法 75 条”などが法条ノードとなっている．その 2 種類のノード間で，文書間エッジ (D-D)，文書-法条間 (D-L)，法条間 (L-L) の 3 種類のエッジが構築される．これらのノードとエッジの組み合わせにより，参照法条同定及び類似判例検索を行うものとする．文書間エッジの重みはそれぞれ TF-IDF を利用した文書ベクトルのコサイン類似度により推定しておくものとする．さらに法条も文書と同じくテキストで構成されるため，文書-法条間エッジと法条間エッジも文書間エッジと同様の方法により重みを推定できる．このようなグラフを利用することによる利点として，(1) 法的類似度を D-L エッジという直接的な形で組み込むことができること，(2) 直接的な類似性は低い場合でも，推移関係により間接的な類似性を考慮できること (A=B ∧ B=C ⇒ A=C) があげられる．

3.2 文書-法条グラフ上の検索アルゴリズム

グラフ上での近接（類似）ノード検索技術は数多いが [6, 15, 7, 4]，本研究では，大規模データでの高速検索

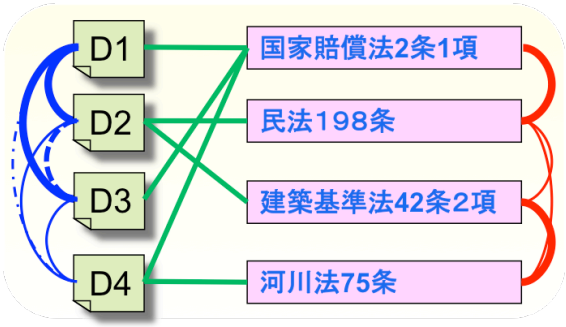

図 1: 文書-法条グラフの例

が可能な Random walk with restart (RWR) [12] を利用する．RWR は *OnTheFly* 法と *PreCompute* 法に分けられるが，本研究では実装の容易さと柔軟性の点から *OnTheFly* 法を採用する．*OnTheFly* 法は PageRank アルゴリズムに類似した以下の式によって表現できる．

$$\vec{r_i} = (1-c)\tilde{\mathbf{W}}\vec{r_i} + c\vec{e_i} \quad (1)$$

ここで，$\tilde{\mathbf{W}}$ は正規化済みの隣接行列，$\vec{r_i}$ は i 番目の文書をクエリとした時の n 次元ランクベクトル（n はノード数），$\vec{e_i}$ は i 番目の要素を 1 とする単位ベクトル，c はリスタート確率である．イテレーション毎に確率 $(1-c)$ でランダムステップを，確率 c でリスタートが選ばれる．即ち，c は PageRank の *dumping factor* に対応し，任意のノードに飛ぶ代わりに，初期ノード（クエリノード）に戻ることになる．このアルゴリズムの結果，クエリ文書を中心として，近接性の高いノードに対しより高いスコアがつくことになり，そのノードのスコアはそのままクエリ文書に対する類似度として扱うことができるのである．

4 実験と結果

この節では参照法条同定 (4.1) と法的類似度に基づく類似判例検索（4.2 節）の 2 つの実験を行い，それぞれその結果を示す．

4.1 参照法条同定

参照法条同定では，COLIEE-2015 Phase 1 のタスクと完全に対応するため，COLIEE-2015 のデータをそのまま利用した．COLIEE-2015 のデータは，司法試験の択一式問題の問題文をクエリ文書として，対応する法令条文を検索するタスクである．図 1 の例を利用するならば，文書 D1 をクエリとし，それと関連度の高い法

令条文 “国家賠償法 2 条 1 項”を検索するのがこのタスクになる．

COLIEE-2015 では学習データ (Train) 267 例及び評価データ (Test) 66 例が提供されているが，学習データには少ないため，学習は行わずに両方を評価データとして利用した．このタスクで検索対象となる法令は民法に限定されている．さらに民法の中から第 4 編の「親族」及び第 5 編の「相続」が除かれており，第 1 条から第 724 条までが解析対象となる．

この実験では英語に COLIEE-2015 で提供されているデータのうち，英語に翻訳済みのデータを利用しており，前処理では Stanford-CoreNLP [3] を利用して文分割及び，品詞タギングを行った．TF-IDF 法により，各文書の内容語に重みを付与し，文書ベクトルを生成，文書間のコサイン類似度により文書類似度を推定しておく．

文書-法条グラフには，文書類似度 0.1 以上の文書間エッジのみを用いた．グラフアルゴリズム（RWR）の疎行列計算には行列計算ライブラリ la4j[4] を利用した．また評価手法については情報検索の標準的な評価値である Mean Average Precision (MAP) 及び Mean Reciprocal Rank (MRR) を利用した．

表 2 に示すのが参照法条同定の結果である．参照法条同定では，文書をクエリとして法令条文を検索するので，文書-法条エッジ (D-L) は必ず必要となる．それがこのタスクを扱う上でのグラフの制約である．このタスクでは，文書間エッジ (D-D)，文書ー法条エッジ (D-L)，法条間エッジ (L-L) の組み合わせによって 4 つのモデルを構築した．その効果が際立つのは L-L エッジの付与で，D-L エッジのみのモデルと比較し，大幅に精度が向上していることが確認できる．一方，D-D エッジはほとんど効果がなく，(D-D+D-L+L-L) の組み合わせモデルでは，むしろ D-D+D-L モデルに比べて精度が落ちていることもわかる．

表 2: 参照法条同定結果

	Train		Test	
	MAP	MRR	MAP	MRR
D-L	0.101	0.269	0.132	0.375
D-D+D-L	0.104	0.276	0.132	0.376
D-L+L-L	0.153	0.454	0.168	0.523
D+D+D-L+L-L	0.156	0.465	0.167	0.519

4.2 類似判例検索

類似判例検索はある判例をクエリとして，それと類似度の高い判例を検索するタスクである．図 1 の例では，文書 D1 をクエリとし，それと類似度の高い文書 D2 を検索するのがこのタスクになる．

COLIEE-2015 では判例データを扱わないため，類似判例検索のためには，別の判例データ収集する必要がある．本研究では実験データとして，日本の最高裁判所 Web サイト [5] から最高裁判決文の 28778 文書を収集した．また総務省の法令データ提供システム [6] から 8142 法令中の 197509 条文を収集した．注意点として，これは，参照法条同定で利用した民法の 724 条までのデータと比較して極端に数が多いため，本実験では現実的な時間内で法条間エッジを構築することができなかった．

次にデータに必要な前処理として，PDF データの判決文を Poppler [7] を用いてテキストデータへ変換する．テキストデータに対しては，“.”, “。”, “?”などの主要な文末記号により単純な文分割を行った上で，MeCab [8] により形態素解析を行った．TF-IDF を利用した単語重み付け，文書ベクトル生成，文書間類似度によるグラフエッジ構築については 4.1 節と同様である．

評価データは行政試験用の判例を対象に，民法・行政法の 2 種類について，判例データベース [9] 及び重要判例集 [11] から 106 例を収集し，表 3 のように 20 カテゴリに人手で分類した．そして，ある判例をクエリ文書とした時に，同じカテゴリにある文書を正解の類似判例とみなすことにより，評価実験を行った．評価値についても，参照法条同定と共通の MAP 及び MRR を利用した．

類似判例検索の実験結果は表 4 に示した．この表では文書間エッジモデル (D-D)，文書-法条間エッジモデル (D-L)，組み合わせモデル (D-D+D-L) の 3 つ結果を，各々の法令に分けてまとめて評価した．いずれの法令においても，文書-法条エッジだけの利用では決して精度が高くはないが，文書間エッジと文書-法条間エッジを組み合わせることで大きく検索精度が向上していることを確認できる．

類似判例検索に対する定性的な分析も行う．1 の表 1 で示した「高知落石事件」をクエリ文書とした時，(D-D) モデルは 1 件の正解文書も出力できなかった．しかし，(D-D+D-L) モデルでは，表 5 のように 10-best までで 3 例，20-best まででは 5 例の正解を出力できた．法的類似度に基づく類似判例検索法が定性的にも意味のある結果を得られることがわかる．

[3] http://stanfordnlp.github.io/CoreNLP/index.html
[4] http://la4j.org/
[5] http://www.courts.go.jp/
[6] http://law.e-gov.go.jp/cgi-bin/idxsearch.cgi
[7] http://freedesktop.org/wiki/Software/poppler/
[8] http://taku910.github.io/mecab/
[9] http://www.gyosei-i.jp/page007.html

表 3: 類似判例検索評価データセット

法令	カテゴリ	事例数
民法	14	49
行政法	6	57
総数	20	106

表 4: 類似判例検索結果

	民法		行政法	
	MAP	MRR	MAP	MRR
D-D	0.038	0.103	0.057	0.152
D-L	0.016	0.046	0.017	0.049
D-D+D-L	0.051	0.124	0.067	0.188

表 5: (D-D+D-L) モデルの出力結果

ランク	事件番号	事件名
クエリ	昭和 42(オ)921	損害賠償請求 (高知落石事件)
1	平成 8(オ)1248	車止め撤去請求事件
2	昭和 34(オ)117	損害賠償請求
3	平成 8(オ)1361	通行妨害排除
4	平成 4(オ)1504	国道四三号・阪神高速道路騒音排気ガス規制等
5	平成 3(オ)1534	国家賠償
6	**昭和 63(オ)791**	**損害賠償（多摩川水害訴訟）**
7	**昭和 53(オ)492**	**損害賠償（大東水害訴訟）**
8	平成 1(オ)1628	損害賠償請求控訴、同附帯控訴事件
9	昭和 61(オ)256	損害賠償、仮執行の原状回復
10	昭和 61(オ)255	損害賠償、仮執行の原状回復
11	**昭和 47(オ)704**	**損害賠償請求（八七時間大型自動車放置事件）**
19	**昭和 46(オ)887**	**損害賠償請求（奈良赤色灯事件）**
20	**昭和 58(オ)1132**	**損害賠償（点字ブロック事件）**

5 おわりに

本稿では文書-法条間の関係を利用する一貫した技術により，法律文書解析を行った．その中では 2 つのタスクを扱っており，一つは判例を対象とした類似文書検索，もう一つは司法試験択一式問題を利用した参照法条同定である．類似文書検索では，文書-法条間関係を利用することで，大幅に検索精度を向上させられることを確認し，参照法条の同定では文書間エッジよりも法条間エッジの効果が高いことも確認できた．

今後の展開として，1 つは COLIEE-2015 の Phase 2 である QA タスクを扱うことを考えている．即ち，自動同定された参照法条を利用し，実際の司法試験択一式問題を解くことを試みる予定である．また，参照法条同定により内部的に構築されていると考えられるクラスタを法令中の編・章・節の構造と照らし合わせることで，どの程度マッチングが取れるかを分析するのも今後の大きな課題である．

参考文献

[1] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. Similarity search using concept graphs. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, CIKM '14, pp. 719–728, New York, NY, USA, 2014. ACM.

[2] Jeffrey Dean and Monika R. Henzinger. Finding related pages in the world wide web. In *Proceedings of the Eighth International Conference on World Wide Web*, WWW '99, pp. 1467–1479, New York, NY, USA, 1999. Elsevier North-Holland, Inc.

[3] Enrico Francesconi, Simonetta Montemagni, Wim Peters, and Daniela Tiscornia, editors. *Semantic Processing of Legal Texts: Where the Language of Law Meets the Law of Language*, Vol. 6036 of *Lecture Notes in Computer Science*. Springer, 2010.

[4] Yasuhiro Fujiwara, Makoto Nakatsuji, Makoto Onizuka, and Masaru Kitsuregawa. Fast and exact top-k search for random walk with restart. *Proc. VLDB Endow.*, Vol. 5, No. 5, pp. 442–453, January 2012.

[5] Qixia Jiang and Maosong Sun. Semi-supervised simhash for efficient document similarity search. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 93–101, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[6] Yehuda Koren, Stephen C. North, and Chris Volinsky. Measuring and extracting proximity graphs in networks. *ACM Trans. Knowl. Discov. Data*, Vol. 1, No. 3, December 2007.

[7] Dmitry Lizorkin, Pavel Velikhov, Maxim Grinev, and Denis Turdakov. Accuracy estimate and optimization techniques for simrank computation. *The VLDB Journal*, Vol. 19, No. 1, pp. 45–66, February 2010.

[8] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[9] Irina Matveeva. Document representation and multilevel measures of document similarity. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Doctoral Consortium*, pp. 235–238, New York City, USA, June 2006. Association for Computational Linguistics.

[10] Simonetta Montemagni, Wim Peters, and Daniela Tiscornia. *Semantic Processing of Legal Texts*. Springer, 2010.

[11] Kazuhiko Nishimura. *Perfect Collection of Important Cases*. Jyutaku Shimposha, 2015.

[12] Jia-Yu Pan, Hyung-Jeong Yang, Christos Faloutsos, and Pinar Duygulu. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pp. 653–658, New York, NY, USA, 2004. ACM.

[13] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, Vol. 18, No. 11, pp. 613–620, November 1975.

[14] Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986.

[15] Hanghang Tong, Christos Faloutsos, and Yehuda Koren. Fast direction-aware proximity for graph mining. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pp. 747–756, New York, NY, USA, 2007. ACM.

[16] Xiaojun Wan, Jianwu Yang, and Jianguo Xiao. Towards a unified approach to document similarity search using manifold-ranking of blocks. *Inf. Process. Manage.*, Vol. 44, No. 3, pp. 1032–1048, May 2008.