1A3-03

判例の構造を利用した判例文書検索の試み

The trial of the judicial precedent document retrieval focusing on the structure

江越 裕紀　安村 禎明　片上 大輔　新田 克己
Hiroki Egoshi　Yoshiaki Yasumura　Daisuke Katagami　Katsumi Nitta

東京工業大学大学院総合理工学研究科
Interdisciplinary Graduate School of Science and Engineering,
Tokyo Institute of Technology

Recently, keyword search is used to search for a judicial precedent document mainly. However, the accuracy is not satisfactory, because it is hard for users to choose effective queries. In this paper, we propose a retrieval method for judicial precedent documents using the structure of a judicial precedent and TF*IDF. The approach discussed here uses TF*IDF for the each part of document, paying attention to the deviation of words which appears in a paragraph peculiar to judicial precedent documents. As the first step, we divided documents into several parts and added XML tags in which the role of these parts are shown. We determined keyword vectors by calculating TF*IDF for each part. We calculated the similarity between the key document and documents in the database using the vectors for each part.

1. はじめに

判例文書とは，過去の裁判の内容を文書にまとめたものである．判例データベースの検索は，一般にはキーワード検索が使われている[判例マスター 03][判例体系 01]が，現状のキーワード検索では絞込みに手間がかかる問題がある．また，ある記述やある判例に関して，関連の判例を検索する場合は，適切なキーワードを見つけることが困難であるため，それらの記述や判例に類似する判例を直接検索する手法が求められている．そのため，事件の記述や判例を入力とし，TF*IDF 等を利用して判例の検索を行う試みがなされている．例えば，[督永 03]の場合，判例データベースのアブストラクトを用いて，その類似性に着目した類似検索に一定の効果があることを報告している．しかし，文書の類似性という場合に，事件の事実レベルの類似性を重視することもある一方，法的レベルの類似性を重視する場合もある．このような目的別の類似判断をするためには，文書全体に対してではなく，文書の構造に着目して，事実が多く含まれている部分や，法律概念が多く含まれている部分に対して類似判断をすることが効果的であると考えられる．

本稿では，判例文書の構造情報と，語の頻度情報を扱う手法と組み合わせて検索精度を向上させる方法について述べ，このような考えの妥当性を確かめるために予備的な実験を行った．

2. カテゴリを分けた検索システム構築

2.1 判例の構造

判例文書の多くは，図 1 のように，内容によって段落を分けた階層構造になっている．この構造のうち，「争いのない事実」では事件の概要が記述され，「原告の主張」，「被告の主張」，及び，「裁判官の判断」には，法的概念での記述を含んでいる．事実の記述と法的概念での記述の文章量の比率は 1 対 9 であり，判断部分がさらに長くなるものもある．各内容に従い，出現する語に偏りが見られるため，内容ごとに段落にまとめられているこの構造を利用し，各部分に対して検索を行うことで精度向上を図る．

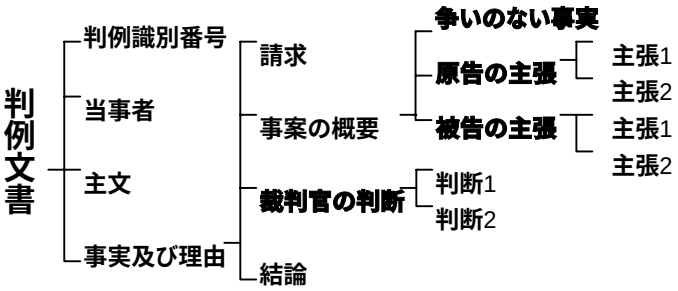

図 1:判例文書構造の例

2.2 文書検索の流れ

各段落に対し，段落の内容を示す XML タグ（原告の主張等）を付加する．それら各段落，またはいくつかの段落のまとまりに対し，TF*IDF 値が高い語でキーワードベクトルを作成する．このベクトルを用い，実験対象に選んだ文書との類似度を計算し，各部分のベクトルを用いた場合の類似度の比較評価を行う．

2.3 段落ごとの内容に着目した構造解析

前述のように判例文書の階層構造を利用して，自然言語で記述された本文を段落別に分割する．各箇条書きの内容を示す表題，もしくは最初の一文からセンテンスの内容を決定し，内容を示す<原告の主張>，<争いのない事実>といったタグを挿入する．このタグを用いて，各内容について比較した検索を行うことができる．

```
<項目 Depth:2 Key= 争点に対する原告 Speaker=原告 Contents=主張>
三　争点に対する原告の主張
<項目 Depth:3 Key= 本件写真について>
1　本件写真について
<項目 Depth:4 Key= 肖像権の侵害>
(1)　肖像権の侵害
　被告は、本件写真の撮影及び掲載について、原告に対して一切の許諾を求めずに行ったのであるから、本件写真の掲載及び頒布は、原告の
```

図 2:タグ付け例

連絡先：江越裕紀，東京工業大学大学院総合理工学研究科
Email：egoshi@ntt.dis.titech.ac.jp

2.4 判例構造を用いた類似事例検索

キーワード検索における有効なクエリ選択が難解であるという問題に対して, 本研究では単語出現頻度を用いる TF*IDF による検索を試みる.

IDF には以下の式を用いる. 全文書数を N, 索引語 t が 1 回以上生起する文書の数を df とする.

$$idf(t) = \log \frac{N}{df(t)} + 1$$

前述の, 内容による構造解析で事実概要記述部分と主張･判断記述部分を大別した場合, 事実概要を記述した段落には, 会社名, 雑誌名など固有名詞が多く見られ, 法律用語はあまり使用されないという特徴があり, 主張･判断を記述した段落には, 一般用語に加えて,「プライバシー権」,「肖像権」といったような法律用語を多く用いているという特徴がある. この出現する単語の部分による偏りを利用し, 各部分に対して別々に TF*IDF による類似度計算を行うことで, 着目している事例に関しての詳細な検索が可能であり, 単純に全文に対して行った場合よりも検索精度が向上すると考えられる.

判例文書の構造を用いた TF*IDF による類似事例検索には以下の 3 通りの方法と応用が考えられる.

(1) 判例全文に対して類似検索を行う方法. 類似度計算には全文中で TF*IDF 値の高い語を用いる.

(2) 事件概要部分と主張･判断部分を分け, それぞれの部分に対して類似検索を行う方法. 各記述部分に出現する語の偏りに着目している. 類似度計算には各部分中で TF*IDF 値の高い語を用いる.

(3) (2)と同様であるが, 原告及び被告の主張に共通して現れる語が, 争点を表す重要語であるとして類似度計算に用いる方法.

さらに, 類似度計算に用いる語をあらかじめ用意しておき, この出現頻度を用いる方法も考えられる. 用いる語は, 文書をよく特徴付けると思われる法律用語から人手で選び, さらにそれらから精度向上に関して妥当であると思われるものを, 実験を繰り返し評価することで選び出していく必要がある.

これらの方法は, 応用になるほど精度向上が期待できる反面, 人手で有効な法律用語を選定し, リストを作成することなど, システム構築にも手間がかかるようになる. 本稿では, これらのシステム構築の第一歩として(1), 及び, (2)の方法を比較実験し, 検索精度に関する評価を行う.

3. 実験及び評価

評価実験に用いる判例は, 刑事事件よりも裁判の判断が難しく, 過去の判例文書の参照が多いという理由から, 民事事件のものを使う. さらに事実関係記述部分を省略せずに記述した地方裁判所での事件のうち, 特に「損害賠償」についての判例とする. これらの判例から 200 件の文書を用意し, これと別にサンプルに選んだ 1 件の判例との類似検索を行い, 評価する.

200 件の内容の内訳は, 交通事故 80 件, 株式トラブル 50 件, 名誉毀損 10 件, 医療トラブル 2 件, その他詐欺等である. 本実験では「刑事被告人の護送の途中を撮影, 雑誌掲載した週刊誌」に対する訴訟の事件をサンプルとし, システムが名誉毀損の判例を類似事件として評価することを見込んでいる.

これら 200 件の判例文書に対し, 以下の 3 つの部分に分け, それぞれに関して類似検索を行い, 比較検討する.

- 事実概要記述部分
- 主張･判断記述部分
- 判例全文

ここで, 文書の形態素解析には茶筌[松本 99]を用い, 名詞, 固有名詞のみを抽出している. 類似度計算に用いるベクトル長は 100 とする.

3.1 実験結果

200 件中, 名誉毀損に関する判例文書 10 件を目的文書として類似文書検索を行った. 主張･判断部分に関する検索結果を以下に示す.

検索元ファイル ID:27816711

判例 ID	類似度	概要
28050101	0.86389	小説記述による肖像権侵害
28050104	0.85646	新聞掲載による名誉毀損
28041977	0.82013	雑誌掲載による名誉毀損
28050852	0.77835	ネット掲示板による名誉毀損

図 3:類似検索結果

上記 4 件はプライバシー侵害, 名誉毀損に関する判例であり, 基準とした判例に関係の高い内容である. 類似度上位 10 件中に上記含む 6 件の類似判例が確認された. 全文に対しての検索結果はこれに類似したものであるが, 事実に関する検索では検索結果上位に異なる結果をみることができた.

3.2 考察

主張･判断部分, 及び, 判例全文に対しての結果が類似しているのは事実概要の記述量の比率が, 判例全文に対して小さいためであると考えられ, 事実部分での検索結果に目的文書が得られなかったことも, 適当な TF 値を得るには記述量が少なかったためと考えられる. しかし, その中でも「交通」,「取引」等事件内容を示唆すると思われる語を多くベクトルに含む特徴が見られ, 人手でのベクトル語の抽出に利用できると思われる. また, 主張･判断部分のキーワードベクトルには「～権」,「～的」といった, 事実部分では TF*IDF 値の低い語が上位にランクしている特徴を利用し,「肖像権」,「プライバシー権」といった判例の内容を特徴付ける語を人手で選ぶことで精度を改善できる可能性がある.

4. おわりに

本稿では, 判例文書が内容ごとに段落に分かれている構造に着目し, 事実関係と法的概念を記述した部分に大別して TF*IDF を用いた類似検索を行い, 評価, 考察を行った. 以後の課題は, 主張･判断部分の全文に占めるが大きいことから, 2.4(3)に述べた, 争点に関する語の抽出, 及び, 本稿の実験で得られた主張･判断部分で TF*IDF 値の高い語を考慮し, 人手で有効なキーワードベクトルの選定を行っていくことである.

参考文献

[判例マスター 03] 判例マスター. 新日本法規出版. 2003.

[判例体系 01] 判例体系. 第一法規出版. 2001.

[松本 99] 松本, 北内, 山下, 平野, 松田, 浅原:日本語形態素解析システム「茶筌」Version 2.0 使用説明書, NAIST Technical Report, NAIST-IS-TR99012, 奈良先端科学技術大学院大学, 1999.

[督永 03] 督永, 樋口, 若木, 新田:判例データベースからの類似文書検索システムの開発と評価, 電子情報通信学会, 2003.