文書の特徴語抽出に関する技術の調査と実験

井内 寛 (0710951)

北陸先端科学技術大学院大学 情報科学研究科

2011 年 11 月

キーワード: 特徴語抽出, 法律文, 索引付け.

現代において、世の中には大量の情報が溢れかえっている。人口の増加と通信手段の発達に伴い、一日あたりにやり取りされる文書の情報量は、近年増加の一途をたどっており、効率的な情報収集に努めなければいけない。

ユーザが情報の収集を行おうとしたとき、大量の文書から必要かどうかを判断することは現実的には困難である。したがって、その時ユーザは検索という方法を取り、必要な情報に関係する情報を探し出すことになる。

検索により得られた文書は、ランク付けもされているため、ユーザは順に目を通していき、目的の情報を手に入れることができる。

検索によって必要な情報へと素早くたどりつくためには、事前に整理された情報の手引きの存在や、必要な情報が既にまとめられている必要がある。だが、それらの作成にはコストがかかり、また、真に必要な情報であるかどうかを事前に知ることは困難である。

機械的に特徴語を抽出することができれば、事前での情報の収集に大きく役立つことになり、機械的な特徴語抽出の手法に関して調査・考察することは重要である。

また、このような大量の文書があふれかえっている分野として、法律制定の分野が挙げられる。法律文の電子データ・検索システムが公開されているが、法律単位でのインデックスが不十分であり、現状では、検索に対して十分とはいえない。また、法律文の特徴語の辞書化やタグ付けなどの資源の整備も進んでいない。

そこで、各特徴語抽出における手法を用いて法律文に適用し、問題点を洗い出し評価することは、大いに重要であると考えられる。

よって、本研究において特徴語抽出の先行研究の調査を行い、その手法を実際の法律文に適用し、考察することを目的とする。

特徴語抽出における手法は、すでに多くの先行研究がなされいる。そこで、特徴語を抽出する目的による分類を行った。また、分類した特徴語抽出の手法に関連する先行研究の調査を行った。

(1) 情報検索

文書における情報検索を目的として、文書を特徴付けるための特徴語抽出の手法が挙げられる。この手法に関して、「TF-IDF」、「SVM を利用する手法」、「KeyGraph」、「χ^2 値を用いる手法」の調査を行った。

(2) 情報抽出

あらかじめ指定された情報を文書中から抽出することを目的とした、情報抽出に関する特徴語抽出の手法が挙げられる。この手法に関して、「固有表現抽出の手法」の調査を行った。

(3) 専門用語抽出

専門分野のコーパスから専門用語を自動的に抽出することを目的とした、特徴語抽出の手法が挙げられる。この手法に関して、「出現頻度と連接頻度に基づく手法」の調査を行った。

また、これらの手法について実用的な特徴語抽出システムの構築を念頭において考察を行った。また、特に法律文の検索システムへの利用を想定して、情報検索を目的とした特徴語抽出の手法を中心に検討を行った。

調査と考察により、実際の法律文の検索システムへの利用にあたり、「カイ 2 乗値を用いる手法」が事前にコーパスや大量の文書に関する統計データが必要ではなく、人手やコストが掛らず、スケーラビリティの観点からも有効であると考えた。この手法を用いて、実際の法律文である「国民年金法」、「労働基準法」、「建築基準法」に対して検証を行った。

この結果、この手法は、頻出語を基準とするが、出現頻度による頻出語で、すでに十分よい特徴語になっている場合には、文書の一部分の概念が特定された特徴的な語になっているケースが多く、逆に頻出語が「月」や「例」など一般的な語が多く、特徴語としての情報量が少ない場合には、適切な特徴語となっているケースが多くなる傾向が窺えた。

実験の結果、本実験で用いた「χ^2 値を用いる手法」は実際の法律文の検索システムに適用できるほどの精度は得られなかった。

しかしながら、本実験の手法は他の TF-IDF などの他の手法に置き換わるような手法ではないが、他の手法と組み合わせて用いることで、より適切な特徴語の抽出ができると考えられる。

また、事前にコーパスや大量の文書に関する統計データを用意することなく、手元にある文書だけで処理できるという手軽さが大きな特徴であり、この特徴を活かした改良を行うことが今後の課題といえる。