E-016

クラメールの連関係数を援用した類似文書検索システムの提案

A Framework for a Similar Documents Retrieval System Using Cramer's Coefficient of Association

樗松理樹†
Masaki Kurematsu

1．はじめに

現在、コンピュータを利用して数多くの文書に容易にアクセスできる環境が整ってきている。それとともに、それらの文書を効率良く処理する技術の開発が活発化[1]している。この分野の課題の一つとして、類似文書検索 [2]がある。類似文書検索の基本的な方法は、①元となる文書を何らかのモデルに変換する。②検索対象となる文書集合の各文書を同じモデルに変換する。③モデル間の類似度を計算し、類似度の高いものを抽出する。というものである。モデルとしては、文書中に出現する語の TF*IDF からなる文書ベクトルや、文書中に出現する語の出現確率に基づく確率モデルなどがある。これらの考えに基づく商用システムなども開発されているが、まだ精度には課題が残っている。そのため、さらなる手法についての研究が進められているのが現状である。

本研究では、このような類似文書検索に対し、カテゴリデータ間の関係を示すクラメールの連関係数[3]を援用するアプローチを検討した。本稿では、本手法を示すとともに、プロトタイプを用いた評価実験結果について報告する。

2．クラメールの連関係数を援用した類似文書検索システム

2.1 本研究で対象とする類似文書検索

本研究における類似文書検索は、特定の文書と文書集合中の全文書とを比較し、類似している文書を検索するというものである。端的にいえば、文書をクエリとした文書検索となる。これを実現するためには、任意の二つの文書の類似度を求める必要がある。この点に対し、クラメールの連関係数を援用する。

2.2 クラメールの連関係数

クラメールの連関係数は、カテゴリデータ間の関連の程度を表す指標の一つであり、二つのカテゴリの連関（独立性）を測る指標である。k 個の要素からなるカテゴリデータ A と l 個の要素からなるカテゴリデータ B 間のクラメールの連関係数 $C_{A,B}$ は、式(1)によって求めることができる。

$$C_{A,B}=\sqrt{\frac{\chi^2}{n\times\min\{k-1,\ l-1\}}} \qquad \chi^2=n\left(\sum_{i=1}^{k}\sum_{j=1}^{l}\frac{f_{i,j}^2}{f_i f_j}-1\right) \quad \text{..式(1)}$$

ここで、n はデータの総数、$f_{i,j}$、は A の i 番目の要素 A_i と B の j 番目の要素 B_j が一緒に出現したデータ数、f_i は A_i が出現したデータ数、f_j は B_j が出現したデータ数を示す。

†岩手県立大学ソフトウェア情報学部

またクラメールの連関係数は $0 \leqq C_{A,B} \leqq 1$ の値をとり、1 の時に完全に連関となり、二つのカテゴリデータ間には強い相関があると言える。

2.3 クラメールの連関係数の援用方法

本研究では、クラメールの連関係数を文書の類似度と見立て、援用する。以下、その算出方法を説明する。

① 類似度を求めるために、文書を文書ベクトルに変換する。文書ベクトルの各要素は、形態素解析を用いて抽出した名詞および名詞列とその出現回数である。名詞および名詞列の順番は、出現回数（降順）、辞書順（昇順）でソートする。

② ①で作成した文書ベクトル A および B をカテゴリデータとみなし、式(1)によってクラメールの連関係数を求める。ここで式(1)中の各値は以下のように求める。

k =文書 A 中の名詞および名詞列の個数
l ＝文書 B 中の名詞および名詞列の個数、
fi,j ＝文書 A の i 番目の語句 $W_{A,i}$ の個数
　×文書 B の j 番目の語句 $W_{B,i}$ の個数
　×語の類似度($W_{A,i}$, $W_{B,j}$)
語の類似度(A,B)= 2×(語句 A と語句 B の共通意味数
　÷（語句 A の意味数＋語句 B の意味数）
なお意味数は、計算機可読型辞書(MRD)から求める。
$f_i = f_{i,1}+f_{i,2}+\cdots+f_{i,k}$
$f_j = f_{1,j}+f_{2,j}+\cdots+f_{l,j}$
n＝ $f_{i,j}$ の総和

③ ②で求めた類似度の高い順に結果をユーザに提示する。

2.4 特許検索システムの構築

以上の提案内容に基づき、処理する文書を特許公報に限定した検索システムを構築した。そのスクリーンショットを図 1 に示す。

図 1 ：特許検索システム

今回特許公報に限定した理由としては、特許は文書構造が明確であるとともに、類似文書の評価が行われていることから、本手法の評価に適切なタスクであると考えたためである。また、①特許公報の内容把握、分類、情報蓄積は人が行っており工数がかかること、②内容把握の結果や分類が個人に依存し多様化する傾向にあること、③多様化のため、共有が困難になっていること、といった課題が特許公報処理にはあり、その解決も視野に入れている。

システムとしては、類似文書検索の他に、語彙抽出、検索式による検索などの機能も用意したが、本稿とは直接関係しないため、割愛する。

3. 評価実験

3.1 実験概要

本提案手法の有用性を評価するために、2.4 で示したシステムを用い、評価実験を行った。評価実験では、X 社にて実際に特許公報の業務に携わる A 氏に協力を依頼し、A 氏による評価と比較した。

実験においては、A 氏が所属する X 社の製品に関する X 社の特許公報 1 件を特定の文書、この文書との類似度を求める文書として、X 社の同製品の別の特許公報 1 件及び公開済みの特許公報 26 件を用いた。特許公報は、A 氏により、「かなり近い」「近い」「あまり近くない」「近くない」の 4 段階のカテゴリ分けがされている。このカテゴリわけと類似度の傾向が一致すれば、本システム、提案手法の有用性が高いと評価する。また比較においては、特許公報全体ではなく、特許公報において注目すべき、「請求項の範囲」「解決すべき課題」「解決手段」にかかる部分ごとに行った。

また本システムにおいては形態素解析としては京都大学、黒橋・河原研究室で公開されている JUMAN[4]を、MRD としては日本語 WordNet[5]を用いている。

3.2 実験結果

実験結果を図 2 に示す。請求については、特許公報による差異はあまり見られなかった。課題および手段に関しては、「かなり近い」が高くなる傾向が見受けられた。その一方で、「あまり近くない」「近くない」において類似度の高い特許公報が出てくる場合も見られた。また X 社の公報については、請求、課題、手段もいずれも 0.6 以上の類似度を得ている。

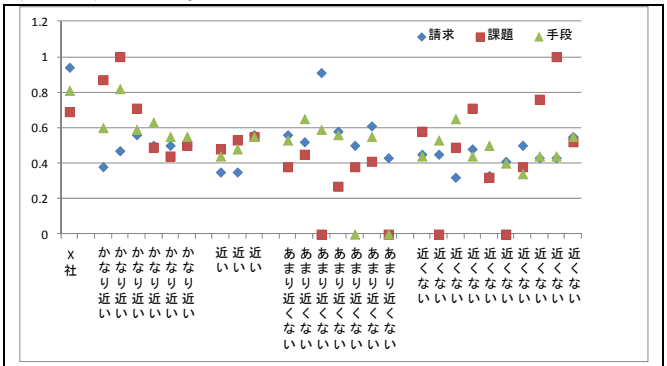

図 2 ：実験結果（全体）

3.3 評価

特許の部分ごとの類似度の平均値を表 1 に示す。「あまり近くない」と「近くない」において値の逆転が起きているが、全体として、「かなり近い」＞「近い」＞「あまり近くない」「近くない」という傾向を見ることができる。また、「請求」「課題」「手段」の 3 つの類似度の算術平均に対し、同様の処理を行った場合、「かなり近い」（0.59）、「近い」（0.48）、「あまり近くない」（0.42）「近くない」（0.47）と同様の傾向が見られた。一方、「X 社」に 2^4「かなり近い」に 2^3、「近い」に 2^2、「あまり近くない」に 2^1、「近くない」に 2^0の評価値を与えた場合の相関係数は、請求が 0.41、課題が 0.38、手段は 0.50 であった。この結果から、類似度とカテゴリに弱い相関があると考えられる。

以上のことから、本手法は、人の評価結果と一致しているとは言えないが、全体としては、妥当な結果を得ていると考えられる。このことから、本手法が有用である可能性を示すことができた。

表 1:**実験結果**（特許の部分ごとの類似度の平均値）

		範囲		
		請求	課題	手段
カテゴリ	かなり近い	0.49	0.67	0.62
	近い	0.42	0.52	0.49
	あまり近くない	0.59	0.27	0.41
	近くない	0.44	0.48	0.47

4. おわりに

本稿では、特定の文書に対する類似文書検索に対し、文書ベクトルとクラメールの連関係数を用いた手法を提案した。特許公報を用いた評価結果から、本手法が有用に働く可能性を示せた。しかし、精度に改善の余地を残すことから、語句の位置などの情報も考慮した連関係数の計算方法の改善、より多くの文書による評価などが今後の課題として挙げられる。

謝辞

評価実験にご協力いただいた X 社 A 氏に感謝の意を表します。また本研究の一部は、科研費・基盤 C（課題番号 24500121）の助成を受けております。

参考文献

[1] 亀井真一郎，田邊栄一，和泉憲明：“自然言語処理の高度化による知的生産性の向上：1.知の共創のための自然言語処理技術 -情報マネジメント技術を俯瞰する-”，情報処理学会誌 Vol.44 No.10，pp. 1007-1011（2003）

[2] 徳永 健伸，“情報検索と言語処理 (言語と計算) ”，東京大学出版会 (1999)

[3] 武藤眞介，“統計解析ハンドブック”，朝倉書店（1995）

[4] JUMAN：http://nlp.ist.i.kyoto-u.ac.jp/index.php?JUMAN

[5] 日本語 WordNet：http://nlpwww.nict.go.jp/wn-ja/