E-014

クラメールの連関係数を援用した類似文書検索の評価

An Evaluation of Similar Documents Retrieval with Cramer's Coefficient of Association

樗松理樹†
Masaki Kurematsu

1．はじめに

社会の情報化とともに、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。このような分野の課題の一つとして、文書の自動分類[2]がある。自動分類の基本的な方法は、①事前に用意した各カテゴリをあらかじめモデルに変換する。②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、最も類似したカテゴリに文書を割り振る。というものである。モデルとしては、各文書を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。しかし、これらの方法では事前にカテゴリや訓練データを用意する必要があり、与えられた文書集合内で文書分類することは難しい。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、このような文書の自動分類（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。

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

†岩手県立大学ソフトウェア情報学部

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。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 それぞれの文書から、形態素解析を用いて名詞および名詞列を抽出する。これらを句と呼ぶ。
② 各句に対し、その出現回数を求める。
③出現回数が事前に決めた回数以上の句のみを利用し、表 1 のクロス表を作成する。

表 1　クロス表

		文書 B			
		句 1	…	句 N	小計
文書 A	句 1	V[1][1]	…	V[1][N]	C[1]＝ V[1][1]＋…+ V[1][N]
		…		…	
	句 M	V[M][1]	…	V[M][N]	C[M]= V[M][1] ＋…+ V[M][N]
	小計	R[1]= V[1][1] ＋…＋ V[1][M]	…	R[N]= V[N][1] ＋…＋ V[N][M]	ALL= V[1][1]＋…＋ V[1][N]＋ V[2][1]＋…＋ V[M][N]

表 1 において、V[i][j]は、以下の式で求める。

V[i][j] ＝SimX（文書A 句i，文書B 句j）
×文書A 句i の出現数
×重み関数（文書A 句i，重要文フィルタ）
×文書B 句j の出現数
×重み関数（文書B 句j，重要文フィルタ）

ここで SimX（文書 A 句 i，文書 B 句 j）は句 i と句 j の辞書に基づく類似度である。辞書としては、専門家が作成した専門用語辞書 1、専門用語辞書 2 および汎用辞書を用いる。汎用辞書としては、日本語 WordNet[4]を用いる。専門辞書 1 には、語句とそれに対する代表語が記載されている。ここで代表語とは、その語句の概念を示す代表的な言葉である。専門用語辞書 2 には、語句とその語句の代表語が記載されている。ただし、代表語に対して、Equal（同義）と Nearly（類似）の関係を与える。これらの辞書毎の類似度を調べ、その最大値を句と句の類似度とする。

SimX（句i，句j）=Max{Sim（句i，句j，辞書k）}

辞書＝{専門用語辞書 1，専門用語辞書 2，日本語 WordNet}

専門用語辞書 1 に対する計算方法は次のとおりである。

- 句 i の代表語と句 j の代表語が等しい場合は 1
- 上記以外は 0

専門用語辞書 2 に対する計算方法は次の通りである。

- 句 i の類義語と句 j の代表語が一致する場合
  - どちらの代表語も同義なら 1
  - 句 i または句 j の代表語が同義なら、0.6
  - どちらの代表語も類義なら、0.3
  - 上記以外は 0

日本語 WordNet では、各単語に英単語が関連付けられている。これに着目し、以下の式で計算する。

2×句 i と句 j の共通する英単語数
÷（句 i の英単語数＋句 j の英単語数）

重み関数は、句が出現した文が重要文か否かで重みを変えるものである。一つでも重要文に出現した場合は、2 を、それ以外は 1 を与える。重要文か否かの判定については、重要文フィルタを用いる。この部分の詳細については、2.4 で述べる。

④ 上記で求まった V[i][j]に対して、クラメールの連関係数を 1 におさめるため、以下の処理を行う。
文書 A 句 i と一致する文書 B 句 k があるとき、V[i][k]以外はすべて 0 とする。また、複数ある場合は、最大値を選択する。

⑤ クラメールの連関係数を援用した次の手順により類似度を求める。

$$Sim = Kai \div \{ALL \times (min\{N, M\} - 1)\}$$
$$Kai = \Sigma\Sigma \{(V[i][j] - E[i][j])^2 \div E[i][j]\}$$
$$E[i][j] = (C[i] \times R[j]) \div ALL$$

2.4 重要文抽出手法の検討

特許の内容把握において、その特許の内容を特徴づける重要文の抽出は有意義である。重要文を提示するだけでも、特許業務にかかる人の負担を軽減することが期待できる。本研究では、この点に対し、以下の流れで検討を行った。

①人手による重用部分抽出

複数の特許に対し、専門家に重要部分と判断する箇所にチェックをいれてもらった。重要部分は、1 文ではなく、文の一部も含まれる。

②重要文の解析

①でチェックをいれた重要部分を含む文を抜き出し、N-gram の抽出を行った。N-gram とは N 個の文字の並びを、先頭から 1 文字ずつずらして取り出すものである。今回は N の数を 2 から 4 とした。この N-gram のうち、重要文に頻出するものがあれば、重要文特有の表現になっている可能性が高い。今回は、ここで得られた N-gram のうち、出現回数の高いもの、語句として意味が取れるもの、含まれる語句が一般的なものを注目し、抽出した。また特許は、要約、請求項などいくつかのブロックに分かれている。このブロック単位で傾向が違うことも考えられる。そのため、注目すべき N-gram をブロックごとに抽出した。

③重要文抽出機能の開発

②で得た N-gram の結果をもとに複数のフィルタを作成した。このフィルタは、ブロックを示すタグ、N-gram パターン列で構成される。N-gram パターン列は出現順にも意味を持ち、例えば、発明.*特許、と特許.*発明とでは働きが異なる。さらに、このフィルタと一つでもマッチする文を抽出する機能を開発した。

④重要文抽出機能の評価

④で構築した機能が有用であるかを、専門家に評価を依頼した。結果として抽出した 1025 文中 766 文（約 75%）が有用と評価された。また、こちらが提案したものと異なるフィルタの提案を受けた。この結果から、本機能は有用である可能性が示された。

本フィルタにより分類した結果を文書類似度計算の重み関数において利用する。

フィルタにより上記機能の有用性が変わること、および人手で行うのは負荷が高いと考えられることから、今後は、フィルタ構築支援機能の開発を行う必要がある。

2.5 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、JAVA を用いて構築した。そのスクリーンショットを図 1 に示す。

今回特許公報に限定した理由としては、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、といった特許公報にかかる課題を解決することにある。また、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。

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、2.5 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、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、「文書を非常に高い」「高い」「中程度」「低い」の 4 段階でランク付けされている。このランク付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。

上記の専門家による評価結果と、提案手法および従来手法との比較検証を行った。以下、その内容を説明する。

利用する句の切り出し方法として、形態素解析の他、N-gram、辞書にある語句との最長一致、文字種区切りの 4 つの方法を用いた。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[5]を用いた。N-gram としては 2-Gram を用いた。辞書にある句との最長一致では、2.3 であげた日本語 WordNet および、専門家が作成した専門用語辞書内にある

語と一致する文字列を句として取り出す。文字種区切りにおいては、JAVA で用いられている文字種を用い、文字種が変わったところまでを句として取り出す。

また、形態素解析においては、【要約】【課題の解決手段】などの特許文書におけるブロックのタグを考慮する場合と考慮しない場合も考慮した。

文書間の類似度計算には、クラメールの連関係数のほか、文書ベクトル法[6]による方法も用いた。文書ベクトル法では、次のように計算を行った。

出現回数を要素の値とする文書ベクトルとして以下の式で Cos 類似度を取り出す。

$$\frac{\sum x[i]\times y[i]}{\sqrt{\left(\sum x[i]*x[i]\right)}\sqrt{\left(\sum y[i]*y[i]\right)}}$$

ここで、x[i]…文書 x 句 i の出現回数。y[i]…文書 y 句 i の出現回数。i の範囲は、x と y の語句の和集合の数となる。一方しかない場合は、値を 0 とする。

また、文書においても 2 章で述べた重要文で抽出した文書のみを用いる場合、重要文以外を用いる場合、両方を用いる場合についても考慮した。

これらの組み合わせにより文書類似度を、同一の文書集合に対して算出し、専門家の評価と比較した。3.1 で述べたように人による評価は、「文書を非常に高い」「高い」「中程度」「低い」の 4 段階でランク付けされている。そのため、各ランクの文書に対する類似度の平均を求め、その平均の変動を評価する。各ランクの文書に対する類似度の平均が右上がりになれば、専門家の評価に近いと判断する。

3.2 実験結果

結果の概要を図 2 に示す。

図 2 において、XG は文字種区切り、ZG は辞書にある語句を、Cos は文書ベクトルによる類似度算出を意味する。

また、Each はブロックタグを考慮した場合、ALL は考慮しない場合を意味する。

横軸の A、A’、B、C は人による評価結果であり、それぞれ、「非常に高い」「高い」「中程度」「低い」を意味する。

3.3 評価・考察

傾向として右上がりとなったのは、全文から得た形態素と文書ベクトルとの組み合わせ、全文から得た 2-gram と文書ベクトルとの組み合わせであった。

クラメールの連関係数に対しては、専門家の評価が「非常に高い（A）」の類似度が最も高くなる傾向が見られたが、また、「高い（A’）」や「中程度（B）」の類似度が低くなる、V 字傾向になった。これは句の切り出し方や対象となる文章を変えた場合にも同様の傾向が出ている。本システムの目的は類似性の高い文書を見つけることである。そのことから考えれば、「非常に高い（A）」を見つけることは出来るが、「低い（C）」が「高い（A’）」よりも評価が高く評価されることとなり、不十分と言える。また SD も値が小さい。

このような傾向がでる理由の一つとしては、句の類似度の扱いが考えられる。現在、一致する句がない場合、類似度を加算している。クロス表中の値が分散することになり、結果として値が低くなる。「非常に高い」は一致する句が多いため、この問題は解消されると考えられるが、「高い」や「中程度」はこの傾向が強いと予想される。そのため、クラメールの連関係数を用いるには、類似度の扱いを再度検討する必要がある。

また、句の切り出し方で見た場合、形態素を用いた場合、類似度は低い値になりやすい。辞書を用いた方が高い傾向がある半面、「高い」が低くなる傾向がでている。文字種区切りは比較的高いが、差が出にくい。一方 2-gram はやや値は低いが、差が出やすい。

今回の実験結果では、類似度が低くなりやすく、SD も狭い、また、「高い」「中程度」の類似度が下がる傾向が見られた。これは、クラメールの連関係数の計算式の性質上、類義語が多い場合、ノイズになるのが原因と考えられる。また計算量は文書ベクトルに対し大きい。これらの点から、提案したクラメールの連関係数を用いた文書類似度の計算手法は現時点では利用価値が低い。そのため、クラメールの連関係数を用いるには、その性質に基づき、相違度で利用するなど再度の検討検証が必要である。

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、従来手法に比べ、明確な有用性を示すことができなかった。しかし、精度に改善の余地を残すことから、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 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] 日本語 WordNet：http://nlpwww.nict.go.jp/wn-ja/

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

[6] 北 研二，津田和彦，獅々堀正幹：“情報検索アルゴリズム”， 共立出版 (2002)

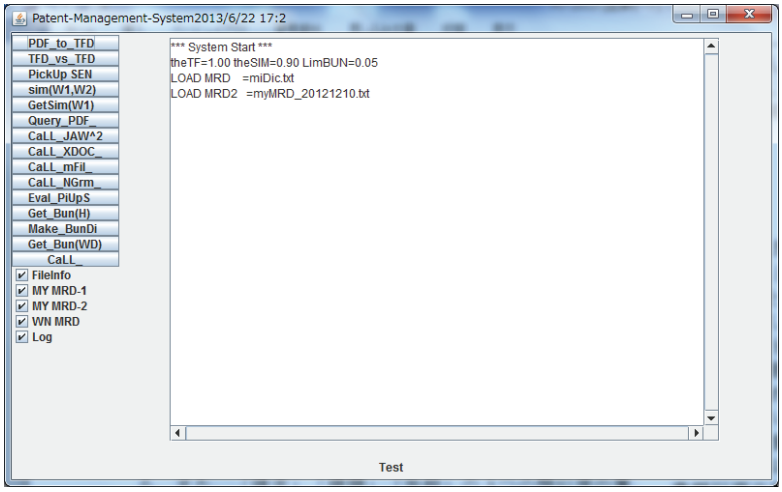

図 1　システムのスクリーンショット

	人の評価ごとの平均値＋SD			
計算方法	C	B	A'	A
辞書(Cos) 重用文以外	0.81 ±0.06	0.76 ±0.07	0.64 ±0.21	0.82 ±0.08
辞書(Cos) 重用文のみ	0.79 ±0.14	0.60 ±0.28	0.54 ±022	0.79 ±0.13
辞書(Cos) 全文	0.81 ±0.07	0.74 ±0.08	0.61 ±0.23	0.82 ±0.10
形態素(クラメール) Each	0.30 ±0.06	0.25 ±0.05	0.28 ±0.04	0.38 ±0.21
形態素(クラメール) ALL	0.50 ±0.03	0.50 ±0.03	0.47 ±0.01	0.54 ±0.15
形態素(Cos) Each&ALL	0.27 ±0.07	0.28 ±0.08	0.31 ±0.11	0.40 ±0.20
2G(クラメール) 全文・・・MRD なし	0.60 ±0.02	0.60 ±0.04	0.59 ±0.01	0.63 ±0.12
2G(Cos) 全文	0.45 ±0.07	0.55 ±0.10	0.58 ±0.12	0.61 ±0.15
XG(クラメール) 全文	0.59 ±0.03	0.57 ±0.04	0.58 ±0.00	0.61 ±0.12
XG(Cos) 全文	0.70 ±0.07	0.70 ±0.09	0.65 ±0.03	0.73 ±0.11
ZG(クラメール) 全文・・・MRD なし	0.68 ±0.02	0.66 ±0.03	0.66 ±0.01	0.70 ±0.10
ZG(Cos) 全文	0.38 ±0.09	0.39 ±0.08	0.36 ±0.05	0.45 ±0.22

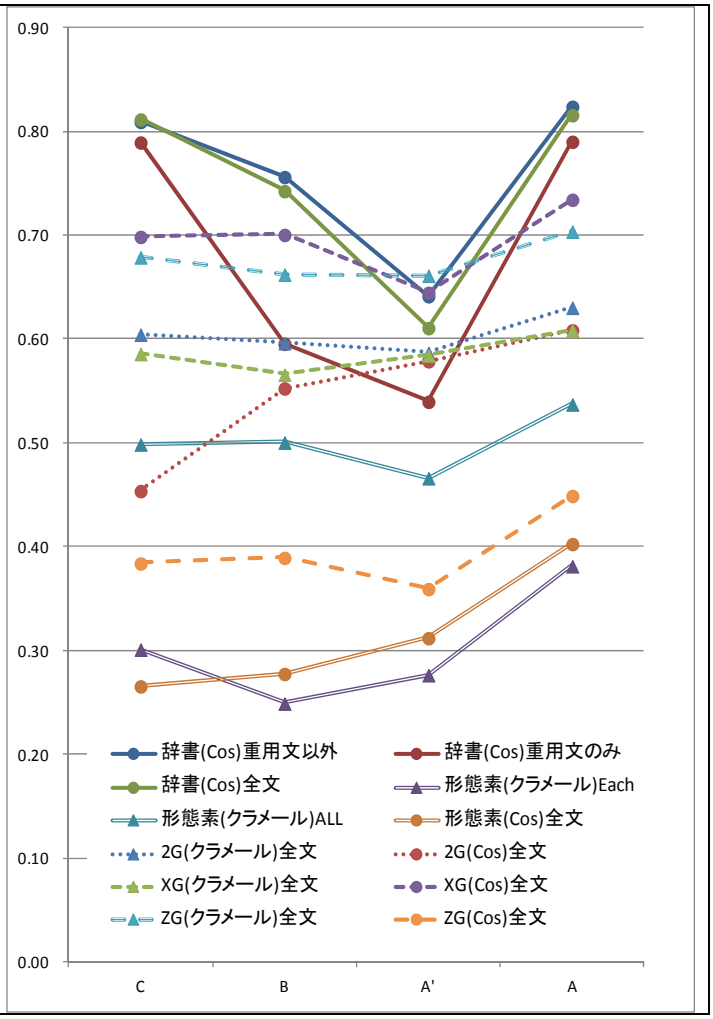

図 2　人による評価単位における、各手法の類似度平均値の比較