3P-9

Web 検索エンジンを用いた分類手法の提案

金子 雅哉† 岡本 秀輔† 小花 聖輝†
成蹊大学 理工学研究科 理工学専攻†

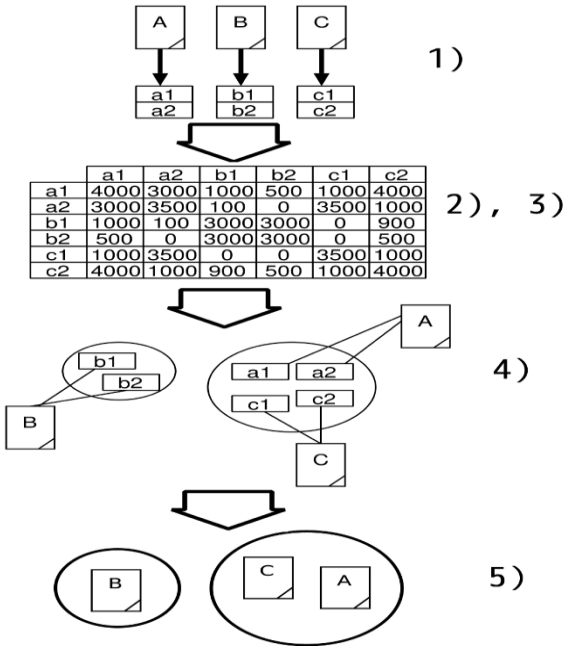
1. はじめに 近年、多くの文書を PDF ファイルとしてインターネット上から取得することが容易となった。本研究では多くの文書ファイルを自動的に分類することを試みる。そのために、各文章からその特徴を表す単語を複数個抽出する。そして検索エンジンにより、2 単語間の AND 検索のヒット数を取得する。そのヒット数を基に、データマイニングの分野で広く知られている K-平均法を用いたアルゴリズムを提案した。その提案したアルゴリズムの精度を評価するために、2 つの文書ファイルデータセットを用いて実験を行った。

2. アルゴリズムの概要 提案したアルゴリズムは、文書ファイルのデータセットを入力データとする。各文書ファイルからtfidf(またはBM25)[1]を用いて、代表単語を抽出する。文書ファイルは代表単語の分類結果を基に、グループ分けを行うことで分類される。図1にアルゴリズムの概要を示す。

提案したアルゴリズムは以下の5ステップで構成される。

1. tfidf(またはBM25)値を用いることで、文書ファイルから代表単語を抽出する。
2. 各文書から抽出した全ての代表単語をひとつにまとめる。
3. まとめた単語から2単語ずつ選択してweb検索におけるAND検索を行い、ヒット数を取得する。取得した全てのヒット数を基に、総当りの表を作成する。
4. ヒット数を用いて計算した2単語間のコサイン類似度をそれらの距離とみなし、K-平均法により代表単語をグループ分けする。
5. 代表単語の分類結果を基に入力データの文書ファイルを分類する。

ステップ3の際に各代表単語単独のヒット数も取得して、NGD(Normalized Google Distance)[2]



	a1	a2	b1	b2	c1	c2
a1	4000	3000	1000	500	1000	4000
a2	3000	3500	100	0	3500	1000
b1	1000	100	3000	3000	0	900
b2	500	0	3000	3000	0	500
c1	1000	3500	0	0	3500	1000
c2	4000	1000	900	500	1000	4000

図1: アルゴリズムの概要

表 1: 文書分類に使用するアルゴリズム

	tfidf	BM25	ヒット数重み	NGD 正規化
tfidf+weighting	○		○	
BM25+weighting		○	○	
tfidf+NGD	○			○
BM25+NGD		○		○

の計算に利用する。ステップ4ではヒット数を各単語の特徴とみなして、代表単語のグループ分けを行う。ステップ5ではグループ毎のtfidf(またはBM25)の合計値を、各ファイルの特徴とみなして入力データファイルのグループ分けを行う。

これらを踏まえて表1のような4つのアルゴリズムを作成する。2つのアルゴリズムは、tfidf(またはBM25)値の高い単語を各文章ファイルから代表単語として取り出す。また、その値でAND検索のヒット数に重み付けしたものを単語ベクトルとして、K-平均法により文書分類を行う。

残りの2つのアルゴリズムは前述の2つのアルゴリズムと同様に、tfidf(またはBM25)値の高い単語を各文章ファイルから代表単語として取り出す。その後、ヒット数の表をNGDを用いて正規化する。そして、正規化した表の値を単語ベク

Proposition for Clustering based on Web Search Engine
†「Masaya Kaneko ・*Seikei University*」

表 2: 文書ファイル 1 の tfidf 値

rank	word	tfidf
1	subtopic	.00954
2	query	.00831
…	…	…
9	compositional	.00395

表 3: 文章ファイル 2 の tfidf 値

rank	word	tfidf
1	computers	.0102
2	students	.00853
…	…	…
9	uptime	.00312

表 4: 文章ファイル 3 の tfidf 値

rank	word	tfidf
1	training	.00891
2	should	.00534
…	…	…
7	running	.00395

表 5: ヒット数の総当り表

	altitude	build	…	uptime
altitude	**783,000**	193,000	…	593
build	193,000	**2,620,000**	…	11,600
claim	33,500	20,700	**2,720,000**	6,900
…	…	…	…	…
uptime	593	11,600	…	**21,000**

表 6: 単語数

	K = 2	
Article ID	cluster 0	cluster 1
C_5_3-1.pdf	4	5
C_5_3-2.pdf	4	5
H_3_3-1.pdf	2	3

表 7: 各クラスタの tfidf 値

	K = 2	
Article ID	cluster 0	cluster 1
C_5_3-1.pdf	0.0371	**0.0467**
C_5_3-2.pdf	0.0158	**0.0188**
H_3_3-1.pdf	0.0071	**0.0253**

トルしてK-平均法により文書分類を行う。その際、単語ベクトルには重み付けを行わない。

3.分類結果 提案したアルゴリズムを用いたtfidf+weightingの分類例を示す。ACMの学術論文から3つのPDFファイルを選択する。

PDFファイルから単語を抽出する時は、Linuxオペレーティングシステム上でpdftotextコマンドを用いる。その後、それぞれの単語に関するtfidfの値を計算する。その結果を表2, 3, 4に示す。文章ファイル1はH.3.3、文章ファイル2、3はC.5.3に分類される論文である。表中のこれらの単語は文書ファイルセットの代表単語として扱われ、それらはひとつにまとめられる。

それらの単語中から2単語ずつ選択し、AND検索のヒット数を取得する。そして、代表単語すべての2単語間のヒット数を基に表を作成する(表5)。この表は各単語の特徴ベクトルを定義するために用いる。K-平均法よるグループ分けのために、この単語ベクトルが用いられる。表6はK=2(クラスタ数2個)の場合の代表単語の割り当て結果である。C.5.3に関する2つの論文の代表単語は共に、cluster0に4個、cluster1に5個のグループができる。H.3.3に関する論文の代表単語は、cluster0に2個、cluster1に3個のグループができている。

各クラスタのtfidfの合計値は表7のようになる。そして、表7の各クラスタのtfidfの合計値をそれぞれの文書ベクトルのとして、再度K-平均法を用いてグループ分けを行う。結果は以下のようになる。

```
cluster 0: C_5_3-1.pdf, C_5_3-2.pdf
cluster 1: H_3_3-1.pdf
```

C.5.3に関する文書ファイルは2つともcluster0に、H.3.3関する文書ファイルはcluster1 に割り当てられており、正しく分類されている。

提案したアルゴリズムを評価するために実験を行った。ヒット数を得るためにGoogle scholarを用いた。そして、20 newsgroups[3]と, ACM Computing Surveysの学術論文の2つのデータセットを2〜5種類までを用いた。

20 newsgroupsの文書ファイルデータセットではNGDを用いない場合はBM25の方がtfidfよりも精度が高いことがわかった。それは20 newsgroupsの各文書ファイルの総単語数に大きな違いがあるため、平均文書長（単語数）により正規化を行うBM25を用いた方が、tfidfよりも精度が高くなったと考えられる。またNGDを用いてヒット数を正規化した場合の方が、非正規化の表を用いた場合よりも分類精度が高いことがわかった。

ACM学術論文の文書ファイルデータセットを用いた文書分類の結果では、NGDによる表の正規化、非正規化またはtfidf、BM25に関わらず分類精度が低かった。それはACMの論文が全て情報に関する論文なので、カテゴリが異なっていても関連性があるからだと考えられる。

4. まとめ 本研究ではweb検索エンジンのAND検索によるヒット数を基にした、文書ファイルの自動分類アルゴリズムを提案した。2単語間のヒット数を取得する検索エンジンのAND検索を用いた。またNGDを用いて、ヒット数の表を正規化した。そしてコサイン類似度を距離として用いたK-平均法クラスタリングに利用した。これらの手法で文書ファイルの分類を行った。実験結果からNGDによりAND検索のヒット数の表を正規化することで分類精度が高くなり、また各文書ファイルの総単語数が大きく異なる場合は、BM25値を用いることで精度が上がることがわかった。

参考文献

[1] The BM25 Weighting Scheme. http://xapian.org/docs/bm25.html . [Online; accessed 27-December-2012].

[2] R.L.Cilibrasi and P. M. B. Vitanyi. The google similarity dis-tance. IEEE Transaction Knowledge and Data Engineering,19(3):370383, March 2007.

[3] J.Rennie. The 20 Newsgroups data set. http://qwone.com/jason/20Newsgroups/, 2008. [Online; accessed 17-Oct-2012].