情 報 学 基 礎 56－4
（1999. 11. 30）

文書間の関係に基づくキーワード自動抽出の検討

岡本 東　児玉英一郎　菅原 光政　宮崎 正俊
岩手県立大学 ソフトウェア情報学部
{lfo,kodama,sugawara,miyazaki}@soft.iwate-pu.ac.jp

大量の文書に対して効率的な検索を行なうためには、個々の文書の特徴を表すキーワードを適切に自動抽出することが必要である。しかしながら従来の統計的手法では、単一の文書と文書空間全体から得られる情報のみを利用しているため、文書空間内における、対象文書に関係した部分集合の特徴を示すキーワードを得難いといった問題があった。そこで本研究では、文書間の参照や分類といった文書間の関係を利用することにより、より精度の高いキーワードを抽出する方法について提案する。また本稿では、我々の提案手法に基づき $tf \cdot idf$ を拡張し、その有効性を検証する。

Automatic Keyword Extraction based on Relation of Documents

Azuma OKAMOTO　Eiichiro KODAMA
Mitsumasa SUGAWARA　Masatoshi MIYAZAKI
Faculty of Software and Information Science
Iwate Prefectural University
{lfo,kodama,sugawara,miyazaki}@soft.iwate-pu.ac.jp

In order to retrieve a target document efficiently from the number of document space, it is necessary to extract the keywords automatically, which indicate a characteristic of each document. Though there exist several statistical methods to extract keywords, these methods analyze only a target document and whole document space, and it is difficult to extract keywords in a related document group. So, we propose to use the relation between documents such as references and classifications. We also extend $tf \cdot idf$ method based on our proposal and evaluate it.

1 はじめに

電子化された文書の利点の一つとして、大量の文書集合の中から、全文検索を行うことによって目的とする文書を探し出すことができるということがあげられる。

しかしながら、コンピュータやネットワークの普及および発達に伴って、現在では膨大な量の電子化された文書が流通・蓄積されているため、単純な全文検索では検索結果として不適切ものを含んだ大量の文書が示されたり、検索に時間がかかるなどの問題点がある。

効率的な文書検索を行うためには、それぞれの文書のキーワード (重要語) を予め用意しておき、そのキーワードに対して検索を行ったり、該当する文書におけるキーワードの重要さを考慮して検索結果に優先順位をつけるなどの工夫が必要とされるが、文書数が膨大であるため、これを手作業で行うのは非現実的である。

このようなキーワードを統計処理等によって文書から自動的に抽出するための研究は広く行われている [1][2]。しかしながら、対象としている文書にあって容易に利用できる情報 (語の出現数、文書構造等) および文書空間全体における情報 (語の文書頻度等) のみから、適切なキーワードを抽出することは困難である。そこで、こういった容易に利用できる情報以外の情報をいかにして取り出し利用するかがポイントとなる。たとえば、対象となる文書中の、文書の構造から得られる情報を用いて、より精度の高いキーワードを抽出する試みがある [6][7]。

一方で、世の中に存在する文書の多くは、その著者・書かれた時期・扱う分野・対象・その他によって予め体系づけて分類されていたり、そうでない場合にも後から分類することが可能である。また、明確な分類が困難な場合にも、文書間の関係を明らかにできるものは多い。

本研究では、このような既に与えられている文書の分類や文書間の関係を利用して、より精度の高いキーワードを抽出する方法について検討を行う。

2 文書間の関連性とキーワード

大量の文書の集合からキーワード検索によって目的する文書を取り出す際、その文書の集合がどのようなものであるかによって、有効となるキーワードが変わってくる [4]。

例えば、自然言語処理の分野の文献を探す際、情報科学全般を扱う文献データベースの検索システムに対して「自然言語処理」というキーワードを用いて絞り込みを行い、目的の文書に近づくことは可能である。しかし、自然言語処理の分野のみを扱う文献データベースの検索システムに対して「自然言語処理」というキーワードを与えることにはあまり意味がない。

検索システムの側においても同様で、自然言語処理の分野のみが集められている文献に「自然言語処理」というキーワードを用意しておいても意味がないばかりか、不要なキーワードは検索速度の低下などの弊害をもたらす可能性がある。つまり、検索システムの側で用意するべきキーワードも、文書の集合の特徴に応じて変化させるべきであると言える。

また、全文書空間の違いばかりでなく、同じ文書空間中にあっても、その文書の見方を変えることによってキーワードが変化する可能性がある。

ある文書 i_a に注目しつつ別の文書 i_b の特徴を表すキーワードを決定する場合、文書 i_a, i_b に共通するものは重要ではなく、文書 i_b に固有のものが重要なキーワードとなるであろう。ここで、注目する文書が別の文書に変わるとキーワードも変化する。

また、文書集合に含まれる文書をいくつかのグループに分類した場合、グループの特徴を表す語がそのグループに含まれる文書のキーワードとなる可能性がある。このキーワードも分類の仕方によって変化する。例えば論文が分野別に分類されているというように、文書の内容に密接した分類がなされていれば、その分野の特徴をよく表したキーワードが得られると考えられる。また、論文を著者の出身地で分類するなど、文書の内容と無関係な分類がなされていてもキーワードは得られないであろう。

そこで本研究では、内容の類似関係が既に与えられている文書集合を考察対象とする。

旧来からある紙に印刷された図鑑や専門書などはこのような文書集合にあたり、文書間の類似関係が明確となるように、体系づけられた分類がなされている。現在、これらの書物は電子化が進められており、このような文書は今後増えていくことが予想される。

本研究における文書間の関連性を考慮したキー

ワードとは、従来手法である文書内での出現回数や全文書空間内での文書頻度のみで決定されたキーワードではなく、それらに加えて、関係のある文書に多く現れる単語は重要性の高い語とし、関係のない文書に多く現れる単語は重要性の低い語であるものとして決定されたキーワードとする。

3 文書間の関連性を考慮した文書空間のモデル

ここで、文書間の関連性を考慮した文書空間のモデルを提案する。以後の議論はこのモデルに基づいて行う。

3.1 最小モデル

文書間の関係を表現するには様々な方法があるが、ここでは、複数の文書の間には互いに関係が「ある」「ない」の2値のみが存在するものとする。互いに関係がある文書のみが含まれる文書空間を文書グループと呼ぶ。

文書グループを含む文書空間を表現する上で最小限必要とされるのは、キーワード抽出の対象とする文書を i_0 と、i_0 の含まれている文書グループ I に含まれる文書 i_1、および文書グループ I には含まれない文書 i_2 の3文書である。これが文書グループを構成する文書空間の最小モデルとなる。全文書空間を A とすると以下のように表すことができる。

$$\begin{aligned} i_0, i_1, i_2 &\in A \\ i_0, i_1 &\in I \subset A \\ i_2 &\notin I \end{aligned}$$

ここで、文書 i_0 に含まれる単語 j が文書 i_1 及び文書 i_2 にそれぞれ含まれるかどうかで分類すると、表1で示されるように単語 $j_0, \cdots j_3$ の4通りに分類される。

3.2 最小モデルにおける文書間の関連性を考慮したキーワード抽出

ここで、文書間の関連性を考慮したキーワードを抽出するため、各単語の重要度を考える。第2節で

表 1: 最小モデル

	i_0	i_1	i_2
j_0	○	×	×
j_1	○	○	×
j_2	○	×	○
j_3	○	○	○

○:含まれる, ×:含まれない

述べた通り、文書間の関連性を考慮したキーワードは、出現回数や文書頻度が等しい語であっても、関係のある文書に多く現れる単語ほど重要性が高いとし、関係のない文書に多く現れる単語ほど重要性が低いとする。

これを、最小モデルにおける文書頻度が等しい単語 j_1, j_2 に対してあてはめると以下の式が成り立つ。

$$w_{model}(i_0, j_1) > w_{model}(i_0, j_2)$$

ただし、文書 i_0 において j_1, j_2 の出現回数は等しいとし、$w_{model}(i_0, j)$ は、文書 i_0 における単語 j の重要性を表すものとする。

3.3 最小モデルに基づく文書例

最小モデルに基づいた文書の例を以下に示す。

文書 (a):

サケ科のサケは、最も良く知られた降海型の魚で、産卵のために北日本の河川に遡上する。
日本のほか、カラフト、アラスカなどにも分布する。

文書 (b):

サケ科の魚であるサクラマスは、ヤマメの降海型といわれ、同種のサツキマスによく似るが、産卵のために東日本の河川に遡上する。

文書 (c):

コイ科の魚であるフナは、全国の河川、湖沼に広く見られる。繁殖力が強く、日本のほか、アジアやヨーロッパにも分布する。
鑑賞魚としての金魚は、フナを改良したものである。

文書 (a) と文書 (b) には「サケ科の降海型の魚」について述べられているという関係があり、同一の文書グループに含まれるものとする。また、「コイ科の魚」について述べている文書 (c) はこの文書グループには含まれない。このようにすると、これらの文書 (a),(b),(c) は先に述べた最小モデルの i_0, i_1, i_2 にそのままあてはめることができる。

ここで、文書 (a) に含まれる単語を最小モデルにおける単語 $j_0, \cdots j_3$ にあてはめると、表 2 のようになる。

表 2: 出現語と最小モデルとの対応

j	文書 (a) に現れる単語
j_0	サケ, 北日本, カラフト, アラスカ
j_1	サケ科, 降海型, 産卵, 遡上
j_2	日本, 分布
j_3	魚, 河川

この文書例では文書グループの内外を問わず「日本の河川に分布する魚」について述べられており、「日本」「河川」「分布」「魚」のように j_2, j_3 に分類される単語は文書 (a) の特徴を表すキーワードとして適切ではない。一方、文書 (a) を含む文書グループは「サケ科の降海型の魚」について述べている文書の集合であり、「サケ科」「降海型」のように j_1 に分類される単語は文書 (a) の特徴を表すキーワードになり得ると考えられる。

これは、文書間の関連性を考慮したキーワードを最小モデルに基づいて表現した式、

$$w_{model}(i_0, j_1) > w_{model}(i_0, j_2)$$

と一致する。

4 $tf \cdot idf$ への文書間の関連性の導入

4.1 $tf \cdot idf$ とその問題点

全文書空間を A、対象とする文書を i_0 ($i_0 \in A$) とし、i_0 における単語 j の重要性を表す $tf \cdot idf$ は、一般的に以下の式で表される。

$$w(i_0, j) = tf(i_0, j) \times \log \frac{N_A}{df_A(j)}$$

各項の意味は以下の通り。

$w(i_0, j)$: 文書 i_0 における単語 j の重要性
$tf(i_0, j)$: 文書 i_0 における単語 j の出現回数
N_A: 文書空間 A における文書数
$df_A(j)$: 文書空間 A における単語 j の文書頻度

ここで、文書グループ内でのみ文書頻度の高い語について考えると、文書グループ内での文書頻度が高い文書は、その文書グループの特徴を表した重要語である可能性が高い。しかし、この式で計算される重要性は、文書頻度が高ければ高いほど、$idf(j) = \log \frac{N_A}{df_A(j)}$ が小さくなり、結果として $w(i, j)$ も低い値となって、重要性は低いとされてしまう。つまり、文書グループ内でのみ文書頻度が高い単語においては、文書グループを考慮した重要性と $tf \cdot idf$ が示す重要性とが無関係であったり逆の傾向を示すという問題が発生する。

第 3.1 節で述べた最小モデルに対して $idf(j)$ を計算すると表 3 のようになる。

表 3: 最小モデルにおける idf

j	df_A	$idf(j)$
j_0	1	1.585
j_1	2	0.585
j_2	2	0.585
j_3	3	0.000

この結果によれば、$idf(j_1) = idf(j_2)$ であり $tf(i_0, j_1) = tf(i_0, j_2)$ の場合においては、

$$w(i_0, j_1) = w(i_0, j_2)$$

となってしまう。これが

$$w_{model}(i_0, j_1) > w_{model}(i_0, j_2)$$

とは一致しないことからも上記問題の存在が確認できる。

4.2 $tf \cdot idf$ の拡張

ここでは、全文書空間についての文書頻度のかわりに文書グループ外における文書頻度のみに着目するという単純な手法で、上記問題の解決を試みる。

全文書空間を A、対象とする文書 i を含む文書グループを I $(i \in I \subset A)$ とし、idf における全文書空間 (A) のかわりに文書グループ外の文書空間 $(A-I)$ を用いて変形したものが以下の式である。

$$idf'(j) = \log \frac{N_A - N_I + 1}{df_A(j) - df_I(j) + 1}$$

ここで、$N_I = 1$、すなわち文書グループが形成されておらず i 以外に文書グループ I に含まれる文書が存在しない場合、$idf'(j) = idf(j)$ である。

本稿では、この idf' と従来の idf を包含する以下の式を提案する。

$$eidf(j,k) = \log \frac{N_A - k(N_I - 1)}{df_A(j) - k(df_I(j) - 1)}$$

ただし $0 \leq k \leq 1$ であり、境界値では

$$\begin{aligned} eidf(j,0) &= idf(j) \\ eidf(j,1) &= idf'(j) \end{aligned}$$

がそれぞれ成り立つ。

この式を利用して、文書グループ I に含まれる文書 i における単語 j の重要性を、

$$ew(i,j,k) = tf(i,j) \times eidf(j,k)$$

で計算するものとする。

5 実験

5.1 最小モデルへの適用

最小モデルにおける $j_0, \cdots j_3$ に対して、第4節で提案した $eidf$ を計算した結果を表4に示す。

この結果から、単語 j_1 と j_2 のように、従来の idf(今回の $eidf$ において $k=0$ の場合) では重要性が等しいと計算されていたものが、今回提案する $eidf$ を用いることによって文書グループ内外での文書頻度を考慮した重要性を計算することができると言える。

また、文書例 (a),(b),(c) の (a) に出現する単語の重要度を計算した結果を表5に示す。この文書例では、単語 $j_0, \cdots j_3$ についてすべて $tf(i_0,j) = 1$ であるため、$ew(i_0,j,k) = eidf(j,k)$ である。

表 4: 最小モデルにおける $eidf$

j	df_A	df_I	$eidf(j,\mathbf{k})$		
			0	0.5	1
j_0	1	1	1.585	1.322	1.000
j_1	2	2	0.585	0.737	1.000
j_2	2	1	0.585	0.322	0.000
j_3	3	2	0.000	0.000	0.000

表 5: 最小の場合の ew

j	$ew(i_0,j,\mathbf{k})$		
	0	0.5	1
サケ, 北日本, カラフト, アラスカ	1.585	1.322	1.000
サケ科, 降海型, 産卵, 遡上	0.585	0.737	1.000
日本, 分布	0.585	0.322	0.000
魚, 河川	0.000	0.000	0.000

この結果 $ew(i_0,j_1,k) > ew(i_0,j_2,k)$ (ただし $k > 0$) からも同様に、文書グループ内外での文書頻度を考慮した重要性を計算することができると言え、また、それがこの文書例におけるキーワード抽出にも有効であることが確認できる。

5.2 4文書モデルへの適用

最小モデルに対して、文書 i_0 と同じ文書グループに含まれない文書をもう1文書 (i_3 とする) 用意したものを4文書モデルとする。

$$\begin{aligned} i_0, i_1, i_2, i_3 &\in A \\ i_0, i_1 &\in I \subset A \\ i_2, i_3 &\notin I \end{aligned}$$

この場合、$N_I = N_{A-I}$ となる。4文書モデルにおいて、文書 i_0 に含まれる単語を分類すると、表6で示されるように単語 $j_0, \cdots j_5, j_2', j_3'$ の8通りとな

る。ここで、文書 i_2, i_3 は入れ替えても df_I の値には影響がないため、単語 j'_2, j'_3 の場合はそれぞれ j_2, j_3 と同じとみなし、以下は $j_0, \cdots j_5$ の 6 通りで考える。

表 6: 4 文書モデル

	i_0	i_1	i_2	i_3
j_0	○	×	×	×
j_1	○	○	×	×
j_2	○	×	○	×
j_3	○	○	○	×
j_4	○	×	○	○
j_5	○	○	○	○
j'_2	○	×	×	○
j'_3	○	○	×	○

○:含まれる, ×:含まれない

これらの単語について $eidf$ を計算した結果を表 7 に示す。この文書モデルにおいても最小モデルと同様に、文書グループを考慮した重要性を計算できると言える。

表 7: 4 文書モデルにおける $eidf$

j	df_A	df_I	$eidf(j, \mathbf{k})$		
			0	0.5	1
j_0	1	1	2.000	1.807	1.585
j_1	2	2	1.000	1.222	1.585
j_2	2	1	1.000	0.807	0.585
j_3	3	2	0.415	0.485	0.585
j_4	3	1	0.415	0.222	0.000
j_5	4	2	0.000	0.000	0.000

ここで、最小モデルの文書例であった文書 (a), (b), (c) に、以下の文書 (d) を追加する。

文書 (d):

> コイ科のタナゴは、二枚貝のエラの間に産卵するという特徴を持つ。かつては全国の河川や湖沼に広く分布していたが、近年では山間部の水のきれいな、ごくわずかの場所にのみ生息する。

文書 (d) は「サケ科の降海型の魚」について述べたものではないため、文書 (a) と同じ文書グループには含まれない。よって、文書 (a),(b),(c),(d) は 4 文書モデルの i_0, i_1, i_2, i_3 にそのままあてはめることができる。

最小モデルの場合と同様に、文書 (a) に含まれる単語を 4 文書モデルにおける単語 $j_0, \cdots j_5$ にあてはめると、表 8 のようになる。

表 8: 出現語と 4 文書モデルとの対応

j	文書 (a) に現れる単語
j_0	サケ, 北日本, カラフト, アラスカ
j_1	サケ科, 降海型, 遡上
j_2	日本
j_3	魚, 産卵
j_4	分布
j_5	河川

この文書例の場合も、単語 $j_0, \cdots j_5$ についてすべて $tf(i_0, j) = 1$ であるため、$ew(i_0, j, k) = eidf(j, k)$ である。この結果を表 9 に示す。

表 9: 4 文書の場合の ew

j	$ew(i_0, j, \mathbf{k})$		
	0	0.5	1
サケ, 北日本, カラフト, アラスカ	2.000	1.807	1.585
サケ科, 降海型, 遡上	1.000	1.222	1.585
日本	1.000	0.807	0.585
魚, 産卵	0.415	0.485	0.585
分布	0.415	0.222	0.000
河川	0.000	0.000	0.000

この結果からも同様に、文書グループを考慮した重要性を計算することができ、それがこの文書例においても有効であるると言える。

5.3 実験内容と結果

この実験では第3節で述べた文書空間のモデルを用いて、我々の提案する式によって文書間の関連性を考慮したキーワード抽出ができるかどうかを検討した。また、モデルに対応する実際の文書を利用して、モデル上での文書間の関連性を考慮したキーワードと、実際の文書におけるキーワードとが一致することを確認した。

最小モデルの文書例に対して $tf \cdot idf$ を用いた場合、「サケ科」や「降海型」の重要性 $w(i_0, j_1)$ と「日本」「分布」の重要性 $w(i_0, j_2)$ との重要性が等価であると計算されてしまっていた。今回提案した手法で計算した場合には、$ew(i_0, j_1, k) > ew(i_0, j_2, k)$ (ただし $k > 0$) となるため、例えば重みなしのキーワードを抽出するだけならば、$ew(i_0, j_1, k) \geq t \geq ew(i_0, j_2, k)$(ただし $k > 0$) となる閾値 t を設ければよい。このように、キーワードとして採用する語と捨てる語に切り分けることが可能になっている。この差は k を増大させるほど顕著になるが、適切な k の値については今後の検討課題とする。

6 むすび

本研究では、キーワード自動抽出において文書間の関連性を考慮した手法を用いることについて考察を行った。そして、従来からキーワード自動抽出にて使用されている $tf \cdot idf$ を用いた手法をベースに、$tf \cdot idf$ の式に文書間の関連性を考慮するよう変更を加えた。また、関係のある文書の為す空間のモデルを提案し、 これを用いて我々の提案した式の正当性、有効性について理論的な検証をする為の実験を行った。その結果、我々の提案した式が文書間の関連性を考慮して、うまくキーワードを抽出できることを確認した。

今後の研究の方向性として、今回我々が行った考察を更に発展させ、文書間の関係の深さに応じたキーワード抽出や文書に含まれないキーワードの付与などを挙げる。今回の研究では、文書間の関連性を関係が「ある」「ない」の2値で表現することで単純化して考察を進めたが、文書間の関係をファジーに定義する方が関連性をより明確に反映できるとの考えから、現在、文書間の関係の深さに応じたキーワード抽出についても考察を進めている。また、一般的傾向として、“キーワードで検索すると、適合率は高く再現率は低い。”、“全文検索だと、再現率は高く適合率は低い。”といったことが挙げられる[3]が、我々はこれを解決する為、文書に含まれないキーワードの付与についても考察を進めており、これによって、適合率は下がるかもしれないが、再現率を高めて行こうと考えている。

この2点について十分研究を進めた後、我々の提案する手法を用いて、プロトタイプシステムを作成し、情報検索システムの評価指標である再現率と適合率を用いた F-measure ($F = \frac{(a^2+1) \times P \times R}{a^2 \times P + R}$) などを使用して評価する予定である。

謝辞

本研究当初から参画し共に議論していただいた、三石大・岩手県立大学ソフトウェア情報学部助手に感謝いたします。

参考文献

[1] 長尾真, 水谷幹男, 池田浩之: 日本語文献における重要語の自動抽出, 情報処理, Vol.17, No. 2, pp.110-117 (1976)

[2] 諸橋正幸: 自動索引付け研究の動向, 情報処理, Vol.25, No. 9, pp.918-925 (1984)

[3] 木谷強, 高木徹, 木原誠, 関根道隆: フルテキストと抽出キーワードを利用した情報検索, 情報処理学会研究報告, NL-115-18, pp.129-134 (1996)

[4] 小泉敦延, 奥田敬, 伊藤秀一: 文書集合における重要語の抽出, 電子情報通信学会技術研究報告, DE98-1, pp.1-6 (1998)

[5] 中渡瀬秀一: 統計的手法によるテキストからのキーワード抽出法, 電子情報通信学会技術研究報告, DE95-2, pp.9-16 (1995)

[6] 原正巳, 中島浩之, 木谷強: テキストのフォーマットと単語の範囲内重要度を利用したキーワード抽出, 情報処理学会論文誌, Vol. 38, No. 2, pp.299-309 (1997)

[7] 仲尾由雄: 文書の話題構成に基づく重要語の抽出, 情報処理学会研究報告, FI-50-1, pp.1-8 (1998)

[8] 金沢輝一, 高須淳宏, 安達淳: 文書関連性を考慮した検索方式, 情報処理学会研究報告, DBS-116-48, pp.165-172 (1998)

[9] 渡辺日出雄: 文章内容を反映したキーワードの重要度付け, 情報処理学会第 52 回 (平成 8 年前期) 全国大会講演論文集 (4), pp.193-194 (1996)

[10] 畑本直樹, 岩瀬成人: キーワード抽出方式についての検討, 情報処理学会第 56 回 (平成 10 年前期) 全国大会講演論文集 (3), pp.91-92 (1998)

[11] 別所礼子, 広瀬雅子, 小川泰嗣, 西村美苗: テキストデータベースのためのキーワード抽出法, 情報処理学会第 45 回 (平成 4 年後期) 全国大会講演論文集 (4), pp.219-220 (1992)