3U－6

文書間のリンク関係と類似度を利用した特徴ベクトル生成方式によるWWW文書自動分類

加藤英晴　手塚祐一
NTTソフトウェア研究所 ソフトウェア技術研究部

1．はじめに

多種多様な情報が溢れるインターネット上において、NTTDirectoryなどのメニュー検索ページは、ユーザーが必要な情報を探し出す有効な手段になっている。しかし、これらの分類体系は人手によって作成されているのが現状であり、運用、管理には多大な労力が必要とされる。そこで、WWW上の文書を自動的に分類する技術が必要であり、文書間のリンク関係と類似度を加味した文書の特徴ベクトル生成方式を提案する。

2．WWW上の文書の特徴

WWW上の文書（NTTDirectoryに登録されている文書と、それらの文書からリンクされている文書、約9000文書）に対して、文書内単語情報量、文書間のリンクと類似度の関係の2点について調査した。

2.1.文書内単語情報量

文書自動分類で一般的に用いられているベクトル空間モデルでは、文書内に出現する単語の統計情報を利用しており、文書内出現単語情報量が十分にある事が望ましい。

そこで、WWW文書に対して、単位文書当たりの名詞単語出現数を調査した。図1の名詞単語出現数の分布を示す結果により、文書内名詞単語出現数が100個以下である文書が全体の約6割を占めており、WWW上の文書は出現単語数の少ない文書が多いと言える。

2.2.リンク先文書との関連性

二つの文書間の類似度を示す指標を、単純に二つの文書に共通して出現する名詞単語の数として、リンク関係のある文書とない文書での違いを調査した。表1に示す結果により、リンク関係のある文書間のほうが共通して出現する名詞単語数が多く、WWW上のリンクで関連付けられた文書は類似性・関連性が高いことが予想できる。

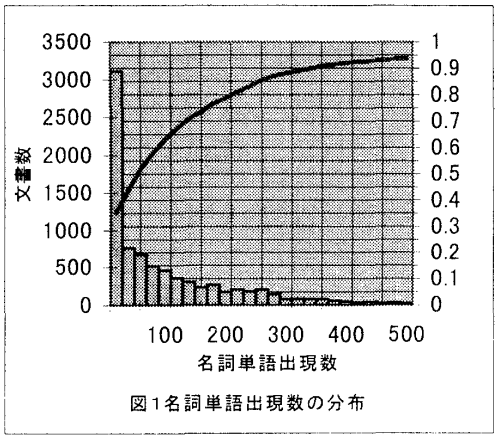

図1名詞単語出現数の分布

表1　共通出現名詞単語数とリンクの関係

	リンク関係あり	リンク関係なし
S	13.20%	1.90%

S= 2文書共通名詞単語数/2文書名詞単語総数

3．問題点

2.1.で示したように、WWW上の文書は出現単語数が少ないため、より適切に文書の特徴を表現する特徴ベクトルを生成するためには、特徴ベクトルを生成するための情報量を増やす事が望ましいと考える。そこで、2.2.で示したリンクで関連付けられた文書間の類似性・関連性を利用する事が有効である。

文書の特徴ベクトルを生成する際に、その文書のリンク先文書の情報を利用する方法はこれまでに行われており、リンク先文書を利用することにより分類精度が向上することが報告されている[1]。しかし、全てのリンク先文書を闇雲に利用すれば関連性のない文書も利用することになり、不適切な情報も特徴ベクトルの生成に利用するため効率的でない。

4．提案する特徴ベクトル生成方式

リンク先文書を全て利用するのではなく、リンク先文書集合から不適切な特徴をもつ文書を除外した文書のみを利用して特徴ベクトルを生成する方式を提案する。リンク先文書集合に対してクラスタリングを行うことにより、他の文書と類似していない文書を除外する。以下にそのアルゴリズムの概要を示す。クラスタリングの対象は、特徴ベクトルを求めようとする文書と、その文書のリンク先文書とする。

i）各文書に対して、その文書一つからなるクラスタiをつくる。

C_i：クラスタiの特徴ベクトル

ii）全ての２クラスタ間の類似度S_{ij}を求め、最大類似度を与える組(a, b)を得る。

$$S_{ij} = C_i \cdot C_j / (|C_i| \cdot |C_j|)$$

iii)最大類似度を与える組(a, b)のクラスタを融合して新しくクラスタｃを生成する。

$$C_c = (q_a \cdot C_a + q_b \cdot C_b) / (q_a + q_b)$$

q_i:クラスタｉを構成する文書数

iv)全てのクラスタ間の類似度S_{ij}が閾値α以下であり、かつ、最大文書数のクラスタが唯一であれば終了する。

v）クラスタリング対象から、クラスタａ, ｂを取り除きクラスタｃを加え、再びii）へ戻る。

クラスタリングによって形成されるクラスタの概念図を図２に示す。最大文書数のクラスタの特徴ベクトルを、求める文書の特徴ベクトルとする。

これによりリンク先文書集合中の不適切な文書の影響を抑えられ、より適切な特徴ベクトルを生成できると考えられる。

５．WWW文書自動分類システムと検証方法

提案した特徴ベクトル生成方式を用いたWWW文書自動分類システムを試作した。

5.1. システム概要

このシステムは、既存の分類体系における各分野の特徴を、その分野に既に登録されている文書を利用して生成する「学習フェーズ」と、未分類の文書をこの分類体系に自動的に分類する「分類フェーズ」に分かれる。文書の特徴ベクトルを構成するキーワードには分野ごとの出現頻度に偏りのある特徴素を選出し、各キーワードに対する特徴ベクトルは文書内キーワード共起関係を利用して生成する方法[2]を用いる。以下に自動分類の流れを示す。

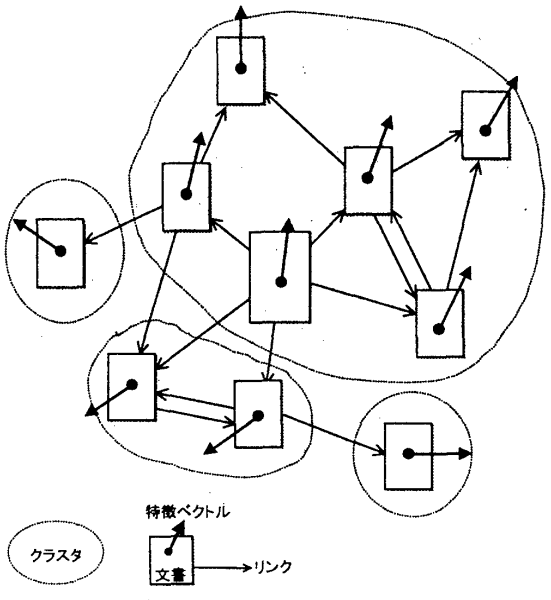

図2 クラスタの概念図

・学習フェーズ

ｉ）文書収集、ＨＴＭＬタグ除去、形態素解析

ii）特徴素（名詞bigram）抽出

iii）キーワード選出

iv）キーワードの特徴ベクトル生成

ｖ）文書の特徴ベクトル

vi）クラスタリングにより文書の特徴ベクトル生成

vii）分野の特徴ベクトル生成

・分類フェーズ

ｉ）文書の特徴ベクトル生成

ii）クラスタリングにより文書の特徴ベクトル生成

iii）分野の特徴ベクトルとの類似度により分類

5.2. 提案する方式の検証方法

全てのリンク先文書の集合を利用した特徴ベクトル生成方式と、クラスタリングを行ったリンク先文書集合を利用した特徴ベクトル生成方式による分類精度の違いを調べる。分類精度は、人手によって既存の分類体系に既に分類されている文書に対し、人手による分類を正解とすることにより、再現率、適合率を計算し評価する。

６．まとめ

本稿では、リンク先文書に対するクラスタリングを用いたWWW文書の特徴ベクトル生成方式を提案した。また、この方式を用いたWWW文書自動分類システムを試作した。

今後は、試作したシステムにより、大量の文書を用いて、提案した特徴ベクトル生成方式による効果を検証する予定である。

参考文献

[1] 落谷亮:「WWWページの分類におけるテキストの特徴分析法」情報処理学会自然言語処理研究報告 118–14, p. 85–90 (1997)

[2] 湯浅夏樹、上田徹、外川文雄:「大量文書データ中の単語間共起を利用した文書分類」情報処理学会論文誌 Vol. 36, No. 8, p. 1819–1827 (1995)