データベースシステム 123－11
（２００１．１．２３）

ラフ集合理論による分類に関する研究

青山 聡　鮑 永广　石井 直宏

名古屋工業大学

概 要

ドキュメント分類は情報検索や情報フィルタリング技術など様々な分野で利用されている。分類手法として数多くの研究、提案がなされている。その中でデータマイニングの手法の一つであるラフ集合理論に基づいた分類に注目した。本稿では従来のラフ集合理論の分類システムにさらにドキュメント中の単語間の関係を考慮することを提案した。従来の方法ではそれぞれの単語は全く別々に扱っていた。そのため分類の精度はドキュメントに依存していた。しかし、単語間の関係を考慮して関係のある単語をグループ化して扱うことで、ドキュメントに依存しないより精度の高い分類が可能であると考えた。

A Document Classification Based on The Rough Set Theory

Satoshi Aoyama Yongguang Bao and Naohiro Ishii

Nagoya Institute of Technology

Abstract:

A document classification is attractive in information retrieval and information filtering (IR/IF). Many classification methods have been proposed. We discuss the Rough Set - Based Classification Approach. In this paper,we introduced the classification system,which take account of a relationship between words in documents. The previous system depended on documents because it separately handled each words. However,by grouping the related words,our system performs good classification without depending on documents.

keyword:

document classification,rough set theory,feature grouping

1 はじめに

近年、WWW (World Wide Web) の急速な発達により、インターネット上に提供される情報量は膨大になっている。この様な状況下で、ユーザが必要とする情報を見つけ出すことは、非常に困難である。そのため、情報検索や情報フィルタリングの技術への重要性が高まっている。情報フィルタリングとはユーザが必要とする情報だけをふるい分けユーザに提供する技術である。こういった情報フィルタリングを利用することでユーザの情報管理を支援することができる。

そのなかでも特に, 自然言語で記述された膨大なドキュメントをその内容に基づいて適切なカテゴリに分類するドキュメント分類技術は重要である。情報フィルタリングだけでなく, テキスト検索をはじめとする幅広い分野において情報の組織化などに利用されている。

情報検索や情報フィルタリングの分野では扱うドキュメントを単語やその単語の出現頻度から成るベクトルと見なすことがある。[2] このベクトルはそれぞれのベクトル自身が持つ傾向により、どのカテゴリに分類するかを決定するルールベースの分類に使われる。しかし、ドキュメントを表すベクトルは単語を要素としているため、数千から数万の次元を持っている。そのため、実際のルール作成にはかなりの計算量が必要となる。

本研究では、ラフ集合を用いた分類手法に注目した。[3] この手法ではあらかじめ分類されたトレーニングドキュメントからカテゴリ間の違いを表す最小限の単語 (キーワード) 集合を見つけ出す。つまり、高次元の単語空間の次元を大幅に削減する。この導き出された単語集合により膨大な未分類のドキュメントを容易に分類することができる。この手法は Alexios Chouchoulas らによって提案されたシステムである。[1] 本研究ではこの手法に単語間の関係を導入した。単語間の関係としては、関連性と類似性の 2 つの関係に注目した。このような単語間の関係を考慮して分類することでより精度の高い分類が可能であると考えた。まか、この他にも従来より様々な改善をすることができる。

2 背景

ラフ集合の理論はデータ集合の次元を削減するのに有効な手法である。[1][3] この章ではデータマイニングの手法の一つであるラフ集合理論の基礎について紹介する。

2.1 情報システム (information system)

$S = (U, Q, V, f)$ という 4 つのタプルで表現されるものを情報システムと呼ぶ。U はオブジェクトの有限集合、Q は属性の有限集合、V は属性が取りうる値の集合、そして、f は情報関数と呼ばれるれ、オブジェクトと属性からその属性の値を割り当てる:

$$f(x, a) \in V \mid x \in U, a \in Q$$

例えば、情報システムとして次の表1を考える。

表 1: 情報システムの例

U	a_1	a_2	a_3	d
x_1	0	0	1	0
x_2	1	0	1	0
x_3	2	2	0	1
x_4	2	1	2	2
x_5	2	1	1	2

この場合情報システム $S = (U, Q, V, f)$ は、オブジェクトの集合 $U = \{x_1, x_2, x_3, x_4, x_5\}$、属性の集合 $Q = \{a_1, a_2, a_3, d\}$、属性の値 $V = \{0, 1, 2\}$ となる。ここで一般的に属性集合 Q は状態 (condition) 属性 C と決定 (decision) 属性 D に分けることができる。状態属性とはオブジェクトの特性を示す属性である。一方、決定属性とは例えば分類処理などでそのオブジェクトが含まれるカテゴリ、クラスを示す属性である。Q を $C = \{a_1, a_2, a_3\}$、$D = \{d\}$ と状態

属性と決定属性に分けるとき、この情報システムは決定表 (decision table) とも呼ばれる。

2.2 識別不能関係 (indiscernibility relation)

P を属性集合 Q の部分集合とする。そのとき、$IND(P)$ を識別不能関係といい、次に定義する等値関係を表す。

$$IND(P) = \left\{ \begin{array}{l} (x,y) \in U \times U : \\ f(x,a) = f(y,a) \quad for\ all\ a \in P \end{array} \right\}$$

もし二つのオブジェクト $(x,y) \in IND(P)$ であるならば、x、y は P に関して識別不能であるという。また、$U/IND(P)$ は $IND(P)$ においてすべて等しい (識別できない) とされるクラスの集まりを示す。つまり、$U/IND(P)$ のそれぞれの要素は P において識別不能なオブジェクトの集合である。

$X \subseteq U$ と $P \subseteq Q$ に対して次の二つの集合を定義する。

$$\underline{P}X = \bigcup \{Y \in U/IND(P) : Y \subseteq X\}$$
$$\overline{P}X = \bigcup \{Y \in U/IND(P) : Y \cap X \neq \emptyset\}$$

$\underline{P}X$、$\overline{P}X$ をそれぞれ X の $P-$ 下近似 (lower approximation)、$P-$ 上近似 (upper $-$) と呼ぶ。$\underline{P}X$ とは属性の集合 P を用いてある X の要素として確実に識別されるオブジェクトの集合を表す。一方 $\overline{P}X$ は P を用いてある X の要素として識別される可能性を持つすべてオブジェクトの集合である。

【例】表 1 において、$P = \{a_1\}$、$X = \{x_1, x_3, x_4\}$ とした場合を考える。その時 $U/IND(P)$ は次のようになる。

$$U/IND(P) = \{\{x_1\}, \{x_2\}, \{x_3, x_4, x_5\}\}$$

次に X の $P-$ 下近似 $\underline{P}X$、$P-$ 上近似 $\overline{P}X$ を求める。

$$\underline{P}X = \{x_1\}$$
$$\overline{P}X = \{x_1, x_3, x_4, x_5\}$$

情報システムにおいて属性同士は互いに依存して影響し合っている場合がある。その度合いを示すためにラフ集合理論は依存度という尺度を定義している。属性集合 R 上の属性集合 P の依存度を $\gamma_R(P)$ で表す。$\gamma_R(P)$ は次の式によって定義される。

$$\gamma_R(P) = \frac{\| POS_R(P) \|}{\| U \|}$$

なお、$\gamma_R(P)$ は $0 \leq \gamma_R(P) \leq 1$ である。ここで、$POS_R(P)$ は正領域 (positive region) といい、次の式で表される。

$$POS_R(P) = \bigcup_{X \in U/IND(P)} \underline{R}X$$

$\| \sim \|$ は集合の要素数を示す。

$POS_R(P)$ は U の中から R の属性だけを使って $IND(P)$ により求めた識別不能なクラスから分類することできるオブジェクトを集めたものである。

2.3 核 (core) とリダクト (reduct)

核とリダクトはラフ集合において重要な概念である。リダクトとは余分な属性を取り除き、本来の情報システムで識別可能なすべてのオブジェクトを同様に識別できる属性だけを選び出し集めたものである。核はすべてのリダクトの共通部分である。もし、$a \in P \subseteq Q$ において、$IND(P) = IND(P - \{a\})$ ならば、属性 a は P に対して不要 (dispensable) であるという。そうでなければ、属性 a は不可欠 (indispensable) であるという。不可欠な属性は情報システムのオブジェクトについてなんらかの重要な情報を持っている。そのため本来識別できる能力が損なわれるため取り除くことはできない。

集合 $P \subseteq Q$ においてそのすべての属性が不可欠であるなら直交 (orthogonal) しているという。部分集合 $E \subset P$ が P のリダクトであるとは E が直交していて、かつ、P で識別できるオブジェクトが識別できることをいう。P

のリダクトは$RED(P)$と表し、次のように定義する。

$$E = RED(P) \Leftrightarrow \left(\begin{array}{l} E \subset P, \quad E \text{は直交}, \\ IND(E) = IND(P) \end{array} \right)$$

Pの核は、すべてのPのリダクトの共通部分なので、次のように定義する。

$$CORE(P) = \bigcup_{R \in RED(P)} R$$

3 リダクトによる分類

Chouchoulasらは、ラフ集合理論を用いた分類システムを提案した。このシステムはキーワード抽出部とリダクトを求める部分の2つから構成されている。キーワード抽出部ではまずそれぞれのカテゴリからトレーニングデータのドキュメントを読み込み、ドキュメント中の単語を取り出す。取り出された単語は後で述べるような計算式から重要度を評価される。その中からキーワードとなる重要な単語だけが選びだされる。それぞれのドキュメントはそれらのキーワードを要素に持つ高次元のデータ集合として表現される。キーワード抽出の際のキーワードkの重みは次の式で求めていた。

$$w(k) = -log\left(\frac{N_k}{N}\right) f_k \, w_f$$

N_k: kを含むメッセージ数
N: 全メッセージ数
f_k: 現在のメッセージ中のkの出現頻度
w_f: ユーザに依存した単語の重要度

この評価式から得られた値を調べて、すべてのドキュメントに共通して出現する単語やほんの少ししか出現していない単語はキーワードとしてふさわしくないとして取り除かれる。

次にドキュメントの集合からリダクトを求める。余分な属性を取り除き、最小限のキーワードを求める。それを元にして分類するためのルールを作成する。

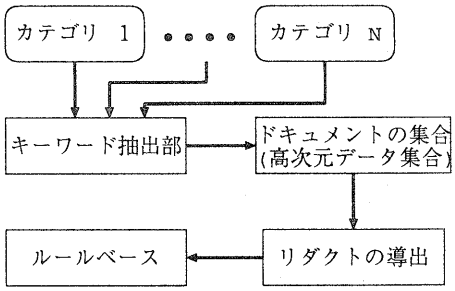

図 1: システムの流れ

システムの概要を図1に示す。

このシステムではEメールを扱っている。それぞれのフォルダには良く似たEメールが保存されていると想定して、Eメールの分類を行っていた。

我々はこのシステムを参考にして、Eメールよりもより一般的なドキュメントとしてインターネット上のニュース分類に応用した。そして、より精度の高い分類を行うためにこのシステムに次の章で述べる" 単語間の関係" を考慮した分類を試みた。

4 単語間関係による単語のグループ化

Eメールは本文以外にもヘッダなどの特殊な情報を持っている。この点でEメールは一般的なドキュメントとは呼べない。本研究ではラフ集合理論を用いて一般的なドキュメントを分類するために" 単語間の関係" を考慮して、それに基づいて単語のグループ化をして分類を行うことにした。従来のシステムはリダクトとして最小のキーワードの集合を求めていた。しかし、そのままニュースなどの一般的なドキュメントの分類に利用した場合にドキュメントによってかなり分類の精度が左右されるのではないかと考えた。ドキュメントの依存度を減らすためには、このグループ化は有効である。従来のキーワードとしては一つ一つの単語を別々に扱っていた。しかし、

単語のグループ化を行うことで、もし誤ったキーワードが含まれていた場合でもその影響を小さく押さえることができる。

”単語間の関係”として本研究では単語間の類似性と関連性に注目した。類似性とは、単語自体が持つ意味により単語同士がどれくらい似ているかを表す。また、関連性とは単語のドキュメント中の出現傾向から単語同士がどれくらい関連があるかを表している。このような関係を調べて関係があると思われる幾つかの単語をまとめてグループ化を行う。このグループ化された単語集合をキーワードとして改めてリダクトを求める。

単語のグループ化を行うことで他にも様々な利点が考えられる。グループ化することでキーワードの総数は従来に比べかなり削減される。膨大なドキュメントを扱う場合でも従来の方法ではキーワードは増え続ける一方であった。キーワード数の削減はリダクトを求める処理に与える影響を小さくすることができる。次に、キーワードは幾つかの単語がグループ化されたものなので、キーワードの情報量、重要度が従来に比べて高い。そのため、最小のキーワードの集合であるリダクトを求めたとき、そのリダクトからは従来以上の情報が得られる。このことはより精度の高い分類へとつながる。また、求められたリダクトをユーザが解析する場合、キーワードにはなんらかの関係のある単語の集まりなので、どのように分類するのかがユーザにとって理解しやすいはずである。

実際にどのように単語間の類似性、関連性を求めたか述べる。

4.1 類似性:意味的アプローチ

ここでは単語間の類似性に関する類似度を測ることのできる”シソーラスを用いた単語間類似度の判定法”について述べる。[8]

4.1.1 シソーラスを用いた類似度判定法

Word-Netは、プリンストン大学において開発されたオンラインの語彙参照システムである。[9] Word-Netは、英語の名詞、動詞などを言語心理学に基づいて分類し有向非循環グラフ、つまりセマンティックネットで手動的に表現したシソーラスである。シソーラスとは、語句を意味によって分類・配列した語彙集のことを言う。

一般的に単語間の類似度を判定する際には、記号としての単語を比較するためにその単語の意味を表す概念の比較を行う。そのため、単語がどのような概念に対応するかという「単語と概念の関係」と「概念間の関係」についての情報が必要となる。Word-Netはこの2つの情報のデータベースを持つシステムである。

Word-Netにおける概念とグラフの対応は以下の通りである。

- ノード　　：概念
- エッジ　　：上下関係
- 道のりの長さ：概念間の距離
- 先祖ノード　：上位概念
- 子孫ノード　：下位概念
- ルート　　：最も抽象的な概念
- リーフ　　：最も具体的な概念

Word-Netを用いた類似度判定法には様々なものがあるが、本研究では情報量に基づいた類似度判定法を使用する。

単語ペア $\{w_1\}$、$\{w_2\}$ 間の類似度を次のような手順で判定する。

1. 判定を行なう単語のペア $\{w_1\},\{w_2\}$ をWord-Netシステムのテーブルを用いて概念の集合 C_1, C_2 に各々変換する。($|C_1| = m, |C_2| = n$)
2. 概念集合 C_1, C_2 の各々の概念ペア ($m \times n$ 個) を比較し、概念間類似度 $Sim(c_i^1, c_j^2)$ を計算する。($c_i^1 \in C_1, c_j^2 \in C_2$　ただし $1 \leq i \leq m, 1 \leq j \leq n$)
3. 概念間類似度 $Sim(c_i^1, c_j^2)$ の正規化を行なう。

4. 正規化した概念間類似度の中で最大のものを単語間類似度 $Sim(w_1, w_2)$ とする。

概念間類似度の計算には概念ペアが共有している上位概念 c の情報量 $info(c)$ を用いる。$info(c)$ は、子孫数の多い抽象的な概念ほどその情報量は少ないとみなして、低い値を取るように定めている。概念ペア $[c_i^1]$、$[c_j^2]$ が共有する上位概念の中で最も概念ペアに近い位置にあるものを最短共有上位概念 $[c_{ij}^{12}]$ とする。従って、単語ペア $\{w_1\}$、$\{w_2\}$ 間の単語間類似度 Sim_{wn} は以下の式で表すことができる。

$$Sim_{wn}(w_1, w_2) = \max_{i,j}[Sim_{wn}(c_i^1, c_j^2)]$$

ただし、

$$Sim_{wn}(c_i^1, c_j^2) = info(c_{ij}^{12}) \times adj(c_i^1, c_j^2)$$

$$adj(c_i^1, c_j^2) = \frac{MaxInfo}{max[info(c_1), info(c_2)]}$$

$MaxInfo$: 基準値 (リーフ概念の情報量 = 16.010725)

この判定法により単語間の意味的な類似度を求めることができる。しかし、概念ペアの意味的なつながりが無い場合や Word-Net に登録されていない単語との類似度は計算できないということある。

4.2 関連性:統計的アプローチ

潜在的意味分析 (LSA: Latent Semantic Analysis) という手法を用いて単語間の関係を統計的に調べる。LSA はドキュメントの集まりから単語の出現頻度などを用いて統計な手法で潜在的な単語の関連性を抽出することができる。[4][5][6] ここでは LSA の概要について述べていく。

LAS を実行するにはまず単語とドキュメントの行列を用意する。その単語 × ドキュメントの行列 A の要素は、個々のドキュメントにおけるそれぞれの単語の出現頻度である。

$$A = [a_{ij}]$$

a_{ij} はドキュメント i における単語 j の出現頻度である。また、単なる出現頻度とはせず、ドキュメントや単語の重要度で重みづけされている場合もある。

次に LSA では行列 A に対して特異値分解 (singular value decomposition) を行う。この分解により行列 A は 3 つの行列に分解すされる。

$$A = U\Sigma V^T$$

U、V はそれぞれ AA^T と A^TA の固有ベクトルからなる直交行列である。Σ は AA^T の固有値を要素に持つ対角行列である。単語 × ドキュメントの行列を $m \times n$ 行列として、A の階数を $r(\leq min(m,n))$ としたとき、U、Σ、V は図 2 のように表せる。

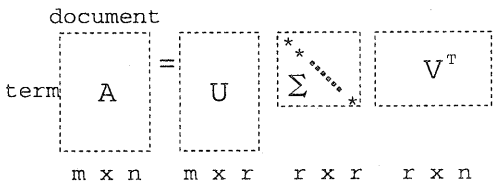

図 2: 3 つの行列への分解

SVD の特徴として U、Σ、V より小さい行列を使って A の近似を導き出せることにある。例えば Σ 中の固有値が昇順に並んでいるとする。そして、大きいものから k 個選び出した $k \times k$ 行列を Σ_k とする。U、V については、選ばれた k 個の固有値に対応する k 個の固有ベクトルからなる行列をとして、それぞれを U_k、V_k とする。そして、行列 $\hat{A}$ を次のようにおく。

$$\hat{A} = U_k\Sigma_k V_k^T$$

$\hat{A}$ を図で表すと、図 3 のように表せる。

このとき $\hat{A}$ は階数 k の A における (最小二乗法で求めうる) 最適な近似である。

A の近似 $\hat{A}$ を解析することで、単語間、ドキュメント間そして単語 – ドキュメント間の関連性を調べることができる。ここでは単語

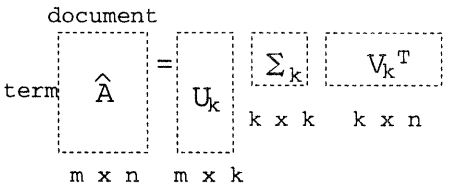

図 3: A の近似 $\hat{A}$

間の関連性に注目する。単語間の関係を調べるため $\hat{A}$ の行ベクトル間の内積を調べる。それぞれの行ベクトルはドキュメントの集合からそれぞれの単語の出現パターンを表している。その出現パターンのベクトルの内積を調べることで単語間の関連性がわかる。ベクトル間の内積は $\hat{A}\hat{A}'$ と計算することによりその要素にはすべての単語間の内積を求めることができる。U_k は直交行列であり、S_k は対角行列なので、$\hat{A}\hat{A}'$ は次にようになる。

$$\hat{A}\hat{A}' = U_k \Sigma_k^2 U_k^T$$

行列 (i,j) 要素は i 番目と j 番目の単語間の内積を表している。

この方法では、固有名詞などの特殊な単語含めてすべての単語間の関係を求めることができる。

4.3 単語のグループ化

従来のキーワードを $k_i(i = 0, 1..n)$ としたとき、グループ化されたキーワード $K_j(j = 0, 1..m(< n))$ は次のように表せる。

$$K_j = \bigcup_{s \leq r(k_i, k_l)} k_l$$

$r(k_i, k_l)$ は、k_i, k_l の先に述べた方法により得られた類似度または関連度を示す。s はそれぞれの関係があると見なすことができる閾値とする。このようにして求められたキーワード K_j に対して、ドキュメントのベクトルを構成し直す。そして、そのドキュメントの集合からリダクトを求める。我々のシステムの流れを図 4 に示す

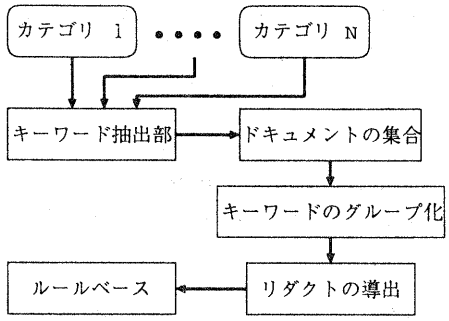

図 4: システムの流れ

5 関連研究

今回は最小のリダクトを求めるアルゴリズムを用いた。最小のリダクトだけではなく複数のリダクトを知識として知識ベースで分類を行う手法が研究されている。すべてのリダクトを求めることは NP 問題である。そのため、効率的にリダクトを求める様々なアルゴリズムが提案されている。また、比較的簡単に複数のリダクトを求めるアルゴリズムもある。この知識ベースによる分類に今回の手法を組み合わせることでさらなる展開が期待できる。

参考文献

[1] Alexios Chouchoulas and Qiang Shen *A Rough Set-Based Approach to Text Classification* In 7th International Workshop, RSFDGrC'99, Yamaguchi,Japan,November 1999, pp.118-129

[2] C.J. van Rijsbergen *Information Retrieval* Butterworths,United Kingdom (1990)

[3] Z.Pawlak *Rough Set: Theoretical Aspects of Reasoning·About Data* Kluwer Academic Publishers,Dordrecht (1991)

[4] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman,

R. *Indexing by latent semantic analysis* Journal of the American Society for Information Science, 41, 391-407.(1990)

[5] Landauer, T. K., Foltz, P. W., & Laham, D. *Introduction to Latent Semantic Analysis.* Discourse Processes, 25, 259-284.(1998).

[6] Foltz, P. W. *Using Latent Semantic Indexing for Information Filtering.* In R. B. Allen (Ed.) Proceedings of the Conference on Office Information Systems, Cambridge, MA,40-47.(1990)

[7] 情報科学技術協会編::「情報検索の基礎」, 日外アソシエーツ株式会社 (1995)

[8] Masanobu Kobayashi, Xiaoyong Du, Naohiro Ishii *A New Measure of Word Similarity Based on Information Content* , Proc. of International Symposium on Database, Web, and Cooperative Systems (DWACOS'99),pp. 85-90, Baden-Baden,Germany (Aug. 1999).

[9] Word-Net のホームページ:: http://www.cogsci.princeton.edu/~wn/