E-026

概念ベースを用いた新聞記事の分類

Classification of Newspaper Articles Using Concept-Base

若月 紀之 † **松田 全弘** † **渡部 広一** † **河岡 司** †
Noriyuki Wakatsuki Masahiro Matsuda Hirokazu Watabe Tsukasa Kawaoka

１．はじめに

多くの情報が行き交う現代においては効率よく情報を整理することが重要である．しかし，日々増大する電子化文書の整理には大きな労力を要する．そこで今回提案する手法は，分類の手掛かりとなる語を入力するだけで，大量の新聞記事の自動分類が可能になるものである．分類に際しては複数の辞書等から構築した概念ベース[1]を利用し，概念や記事の関連性を数値化する手法を用いた．

始めに分類の手掛かりとなる語と関連の深い記事を複数選び，サンプル記事とする．そしてサンプル記事と各々の記事の関連の深さを評価することによって各カテゴリへの仕分けを行う．この手法により文書分類におけるユーザの負担を大きく軽減できると考えられる．

2．概念ベースを用いた記事関連度計算

2.1 概念ベース

概念ベースは，日常的に使う語に関して複数の辞書等から機械的に構築された大規模で汎用的なデータベースである．

概念ベースにおいて概念 A は，その概念の意味を表す属性 a_i と，属性の重要性をあらわす重み w_i の対で表される．したがって概念 A の属性数を N 個とすると，概念 A は以下のように表すことができる．

$A=\{(a_1, w_1), (a_2, w_2), \cdots, (a_N, w_N)\}$

ここで，属性 a_i を概念 A の一次属性と呼ぶ．また，属性 a_i も概念ベースに登録されている 1 つの概念である．したがって，a_i からも同様に属性を導くことができる．a_i の属性 a_{ij} を概念 A の二次属性と呼ぶ．

概念ベースには約 9 万の概念を登録しており，重みは情報量を利用して付与されている．

2.2 概念間の関連度計算

2 つの概念がある時，概念ベースを利用することによってそれらの概念間の関連度[2]を求めることができる．関連度は概念間の意味的な関連の深さを 0～1 の実数値で表すものである．

2 つの概念の関連度は，一致度の和が最大になるような一次属性の組み合わせを作ることによって計算する．一致度は，それぞれの概念を二次属性まで展開し，一致する属性とその重みによって求める実数値である．

2.3 記事関連度計算

提案する手法では，概念間の関連度を計算する手法を用いて記事間の関連度を求める．つまり記事を概念ベース中の概念と同じように，属性とその重みの対の集合という形式で表すのである．そしてこの記事を 1 つの概念とみなすことで，概念間の関連度を計算する手法をそのまま利用することができる．記事間の関連度のことを記事関連度と呼ぶ．

記事の属性には，記事を構成している単語を用いる．まず，記事を形態素解析することによって得られる単語群の中から自立語を抜き出す．これは記事の意味内容を表す属性としては自立語が適切だと考えられるからである．

次に自立語の中から，概念ベースに収録されていない語を除く．これは関連度計算を可能にするためである．つまり記事関連度を求める際に記事の二次属性が必要となるのに対し，概念ベースに収録されていない語を記事の一次属性にしてしまうと二次属性を取得することが不可能となるからである．

記事の属性の重みには tf・idf を用いた．

3. 記事関連度を用いた記事分類

提案する手法ではユーザが入力する分類キーワードを基にサンプル記事を抽出し，各記事と関連度計算を行うことによって分類を行う．

3.1 分類キーワード

記事分類にあたってはまず，分類したいと考えるカテゴリごとにユーザがキーワードを入力する．この時キーワードはカテゴリの意味内容を端的に表す語とし，これを分類キーワードと呼ぶことにする．

多くのキーワードを与えることはユーザの負担増加になるため，1 つのカテゴリにつき 1 語だけの分類キーワードを与えることとする．ただし，分類キーワードは概念ベースで定義されている語とする．

3.2 サンプル記事の抽出

次に仕分け対象となる記事 1 つ 1 つについて，各カテゴリの分類キーワードとの関連度を計算する．この時，分類キーワードの一次属性は概念ベースに格納されている属性であり，記事の一次属性は 2.3 で述べた記事関連度計算で用いるものと同様である．

すべての組み合わせについて関連度を計算した後，カテゴリごとに分類キーワードと関連度の高い記事を複数抽出し，これをサンプル記事と呼ぶ．サンプル記事は分類キーワードと関連の深い記事であり，各カテゴリの記事を象徴するものと考えられる．

3.3 記事分類

最後に仕分け対象となる記事について，すべてのサンプル記事との関連度を計算する．そしてカテゴリごとにサンプル記事との関連度の平均値を求める．その中で最も大きな値となったカテゴリに記事を仕分けする．

4. 評価実験

テストデータとして，政治・社会・経済・スポーツのいずれかに分類されている新聞記事を用いた．実験では政治・社会・経済・スポーツの 4 カテゴリに分類することとし，それぞれ関係の深い分類キーワードを与えた．そのう

† 同志社大学大学院工学研究科
Graduate School of Engineering, Doshisha University

えで各記事が新聞中での元々の分類と同じカテゴリに仕分けされた場合を正解とした．今回は毎日新聞の記事 1000 件，朝日新聞の記事 208 件という 2 種類のテストデータについて実験を行い，それぞれの正答率を求めた．

4.1 実験 1－サンプル記事の件数

始めにサンプル記事の件数についての実験を行った．サンプル記事が多ければ計算コストが大きくなり，少なければ分類キーワードからの意味の増幅効果が小さくなると考えられる．そこで適切な件数を調べるため，各カテゴリのサンプル記事を 5 件～30 件とした場合の正答率を求めた．その結果，2 種類のテストデータのいずれの場合でもサンプル記事を 10 件とした場合の正答率が最も高いことがわかった．10 件という量は計算コストの面から見ても問題はないと考えられるため，以降の実験では各カテゴリのサンプル記事をそれぞれ 10 件とした．

4.2 実験 2－分類キーワードによる比較

ここでは 4 組の分類キーワードによって実験を行い，正答率の比較を行う．これは，同じ 4 カテゴリに分類する場合でもユーザによって与える分類キーワードが異なると想定されるため，キーワードの差異による正答率の変化を調べることを目的としている．提案手法について，表 1 に示す 4 パターンの分類キーワードによる実験を行った．

表 1：4 パターンの分類キーワード

カテゴリ	政治	社会	経済	スポーツ
分類キーワード A	政治	社会	経済	スポーツ
分類キーワード B	国会	事件	景気	試合
分類キーワード C	首相	事故	消費	選手
分類キーワード D	内閣	訴訟	株価	野球

実験の結果を図 1 に示す．2 種類のテストデータについてそれぞれ正答率を求めた．

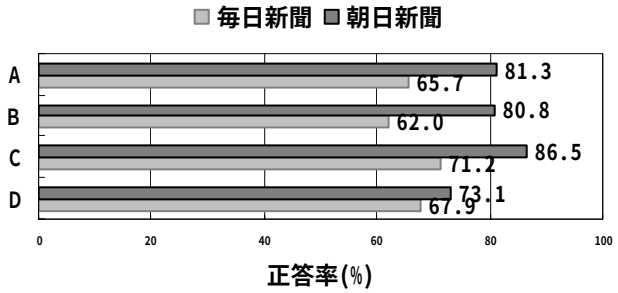

図 1：分類キーワードによる正答率の比較

4.3 実験 3－他の手法との比較

提案手法の有効性を調べるため，他の 2 種類の手法でも同様の実験を行い，結果を比較する．

(1) **キーワード一致**：概念ベースを用いず，分類キーワードのみから分類する手法である．分類キーワードそのものが記事に含まれているかどうかで仕分けを行う．

(2) **キーワード分類**：提案手法においてサンプル記事を抽出せず，分類キーワードとの関連度のみに基づいて仕分けを行う手法である．仕分け対象の記事について，分類キーワードそれぞれと関連度計算を行い，最も高い関連度を示したキーワードのカテゴリへ仕分けする．

ここでは朝日新聞の記事 208 件のテストデータを用いて実験を行った．使用した分類キーワードは表 2 の通りである．

表 2：分類キーワード

カテゴリ	政治	社会	経済	スポーツ
分類キーワード	政治	社会	経済	スポーツ

実験の結果を図 2 に示す．各手法について正答率を求めた．

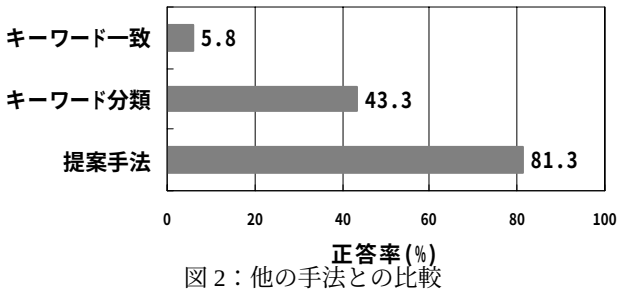

図 2：他の手法との比較

4.4 実験のまとめ

実験結果によると，分類キーワードによってある程度正答率に差があるものの提案手法が有効であることがわかる．一般に分類キーワードそのものが記事中に含まれていることは少ないと言える．しかし概念ベースの属性を用いて分類キーワードと記事の関連度を計算することで，キーワードの持つ意味に近い記事の抽出が可能になる．これによりわずか 1 語の分類キーワードを，記事で使われる多数の語の集合に拡げることができる．そのため記事の意味的内容にもとづく仕分けが，より精度の高いものになったと考えられる．

5. おわりに

本稿では，分類キーワードを入力するだけで新聞記事の分類が可能になる手法を提案し，それが有効であることを示した．しかしテストデータによって正答率に違いがあることがわかった．今後，新聞記事以外のテストデータについても調査を行うとともに，より一層の精度向上を図る必要がある．

本研究は文部科学省からの補助を受けた同志社大学の学術フロンティア研究プロジェクト「知能情報科学とその応用」における研究の一環として行った．

参考文献

[1]小島一秀，渡部広一，河岡司：“連想システムのための概念ベース構成法－属性信頼度の考え方に基づく属性重みの決定”，自然言語処理，Vol.9，No.5，pp.93-110，2002

[2]渡部広一，河岡司：“常識的判断のための概念間の関連度評価モデル”，自然言語処理，Vol.8，No.2，pp.39-54，2001