1G3-5

Web 情報から抽出したカテゴリ属性と記事関連度計算を用いた記事分類方式

Article classification method using the calculation of the degree of association between articles and category attributes extracted from Web information

藤江 悠五[*1] Yugo Fujie
渡部 広一[*1] Hirokazu Watabe
河岡 司[*1] Tsukasa Kawaoka

[*1] 同志社大学大学院工学研究科
Department of Knowledge Engineering and Computer Sciences, Graduate School of Engineering, Doshisha University

Abstract: In this research, it aims to achieve the intelligent processing that classifies a lot of documents according to the content with the computer as one of the applications of intelligent information processing. It is proposed the new article classification method by applying the category attribute extracted from Web information, the Concept-base that registers the semantic attributes with weights and the degree of association, and the calculation of the degree of association between articles.

1. はじめに

本研究では，知的な情報処理の応用の一つとして，多数の文書をその内容によって分類するという知的処理をコンピュータで実現することを目指す．Web 情報から抽出したカテゴリ属性，汎用知識ベースである概念ベースと関連度計算手法[荒木 2006]にもとづく記事関連度計算手法[倉田 2006]を用いて，記事分類方式を提案する．

2. 関連技術

2.1 TF･IDF 重み付け

本研究では，構文解析ソフト茶筌[Chasen 奈良先端科学技術大学院大学 2003]を利用し，「名詞」，「形容詞」，「動詞」などを索引語として採用する．索引語に対する重み付けは，情報検索の分野で広く用いられている TF･IDF 重み付け[Salton 1988]を採用する．TF･IDF による重み付けとは，語の出現頻度と特定性を表す尺度に基づいた重み付け手法である．文書 における語 の重みは，以下の式で定義される．

$$w_t^d = tf_d(t) \cdot idf(t) \tag{1}$$

$tf_d(t)$ とは，文書 d 内での語 t の出現頻度 $tf(t,d)$ と文書 d 内の全ての語数から求められる相対頻度である．以下の式で定義される．

$$tf_d(t) = \frac{tf(t,d)}{\sum_{s \in d} tf(s,d)} \tag{2}$$

また，$idf(t)$ は語 t が出現する文書数によって決まり，以下の式で定義される．

$$idf(t) = \log\left(\frac{N}{df(t)}\right) \tag{3}$$

ここで，N は検索対象群での全文書数であり，$df(t)$ とは，検索対象群で語 t が出現する文書数である．以下，全ての手法において，この重み付け手法を用いる．

2.2 F 尺度

F 尺度とは，システムの有効性を評価するものとして，再現率（必要なものをどれだけ漏れることなく分類できるか）と精度（分類したものの中に必要なものがどれだけあるか）の両方を考慮に入れた評価尺度である．F 値とは評価尺度の値で，0～1 の数値で表される．

2.3 表記一致手法

本節では，単純な単語の表記のみを活用した文章間の類似度計算方式について述べる．

この方式は，単純に単語の表記が一致した割合から類似度を求めるものであり，以下の式で定義する．

$$Score(X,Y) = \frac{m/x + m/y}{2} \tag{4}$$

ここで，x は文章 X の単語数，y は文章 Y の単語数，m は文章 X と Y の両方に出現する単語数である．

また，TF･IDF による重み付けを行い，その重みの割合から，類似度を求める方法を重み付き表記一致方式と呼ぶ．

3. 概念ベースと関連度計算

概念ベースとは，電子辞書などから自動構築された知識ベースである．ある一つの概念 A を属性 a_i と重み w_i によって，次のように定義する．

$$A = \{(a_1, w_1), \cdots, (a_N, w_N)\} \tag{5}$$

一つの語（概念 A）は概念 A の意味特徴を表す単語（属性と呼ぶ）とその属性の重要度（重み）の対の集合で表される．概念数は，約 9 万語で一つの概念につき平均 30 個の属性が存在する（図 1）．

概念：約9万語

概念	属性／重み
医者	医師／0.3, ･･･, 治療／0.1
太陽	日光／0.2, ･･･, 月／0.05
美術	芸術／0.4, ･･･, 鑑賞／0.07
：	：

図 1 概念ベース

関連度とは，二つの概念 A と B の関連の強さを定量化した相対的な値である．

関連度は 0 から 1 までの連続値をとり，関連の強い概念同士では高い値となり，関連の弱い概念同士では低い値となる．例えば，概念「医者」と「病院」の関連度は 0.72，概念「医者」と「太

連絡先：同志社大学工学研究科 知識情報処理研究室，〒610-0394 京都府京田辺市多々羅都谷 1-3, Tel: 0774-65-6944

陽」の関連度は 0.04 となる．このように概念同士の関連の強さを定量化すれば，数値の大小比較によって，曖昧である概念間の関連性の強弱をコンピュータに判断させることができるようになる．この概念ベースを利用して概念と概念の関連の強さを定量化する手法が関連度計算である．本研究では，重み比率付き関連度計算を利用する．

3.1 重み比率付き一致度

2 つの概念 A, B でその一次属性を a_i, b_j, 重みを u_i, v_j とし，属性がそれぞれ L 個，M 個($L \leqq M$)とすると

$$A = \{(a_i, u_i) | i = 1 \sim L\}$$
$$B = \{(b_j, v_j) | j = 1 \sim M\} \quad (6)$$

と表現でき，概念 A, B の一致度 MatchWR(A,B)は以下のようになる．

$$\mathrm{MatchWR}(A, B) = \sum_{a_i = b_j} \min(u_i, v_j)$$
$$\min(\alpha, \beta) = \begin{cases} \alpha (\beta \geq \alpha) \\ \beta (\alpha > \beta) \end{cases} \quad (7)$$

重み比率付き一致度は一致する属性のうち小さい方の重みの和となるが，これは両方の属性に共通して存在する重み分は有効だと考えるためである．

3.2 重み比率付き関連度

概念 A, B のうち属性数の少ない概念を A ($L \leqq M$)とし，概念 A の一次属性の並びを固定する．

$$A=\{(a_1, u_1),\ (a_2, u_2),\ \cdots,\ (a_L, u_L)\} \quad (8)$$

概念 B の各一次属性を対応する概念 A の各一次属性との一致度(MatchWR)の合計が最大になるように並べ替える．

$$B_x=\{(b_{x1}, v_{x1}),\ (b_{x2}, v_{x2}),\ \cdots,\ (b_{xL}, v_{xL}) \quad (9)$$

概念 A と概念 B_x との関連度 ChainWR(A,B)は，

$$\mathrm{ChainWR}(A, B_x) = \sum_{i=1}^{L} \mathrm{MatchWR}(a_i, b_{xi}) \times \frac{(u_i + v_{xi})}{2} \times \frac{\min(u_i, v_{xi})}{\max(u_i, v_{xi})}$$

$$\min(\alpha, \beta) = \begin{cases} \alpha (\beta \geq \alpha) \\ \beta (\alpha > \beta) \end{cases} \quad \max(\alpha, \beta) = \begin{cases} \beta (\beta \geq \alpha) \\ \alpha (\alpha > \beta) \end{cases} \quad (10)$$

となる．すなわち，重み比率付き関連度は対応する一次属性の一致度と，それらの属性の重みの平均および重みの比に比例する．

4. 記事関連度計算方式

記事間の関連の強弱をコンピュータに判断させるために，記事間の関連の度合いを定量化した「記事関連度」を定義する．記事は複数の単語から構成されているため，単語間の関連度を求める考え方を拡張し，記事間の類似度を求める手法が提案されている．記事を 1 つの概念，記事の索引語を概念の一次属性と見なし，索引語の意味を考慮するために，概念ベースを用いて索引語からその属性を取得し，単語間の関連度と同様の方法で関連度を計算する．その計算方式を単語拡張方式と呼ぶ．記事中のすべての索引語を同等に扱うと，その記事を特徴的に表す索引語の意味が薄れてしまう．そこで，索引語に TF･IDF 重み付けを行う．

5. 記事分類方式

本研究においては，分類の基準となりやすいスポーツ面･経済面･社会面･政治面の記事のみを分類する．記事の内容によって分類カテゴリを決定するので，記事は本文のみを用い，見出し部分は用いない．

5.1 Web から抽出したカテゴリ属性での分類

カテゴリの情報のみを与えて分類を行ってみる．カテゴリと記事との間の関連の強さを定量的に表すため，カテゴリの属性を Web から抽出し，そのカテゴリ属性を一次属性とし，カテゴリと記事の関連度を計算する．スポーツ面･経済面･社会面･政治面の記事を，表1に示す4組のカテゴリを入力して記事群を分類した．

表 1 入力するカテゴリ

カテゴリ	カテゴリ 1	カテゴリ 2	カテゴリ 3	カテゴリ 4
スポーツ	スポーツ	試合	選手	野球
経済	経済	景気	消費	株価
社会	社会	事件	事故	訴訟
政治	政治	国会	首相	内閣

分類対象となる記事群は，前述の 4 カテゴリの記事を合計して 1000 件集めたものを用いた．

正解の判定方法は，各記事が分類されたカテゴリが，その記事が本来新聞紙面上で属しているカテゴリと同じならば正解とする．カテゴリの情報のみで記事を分類した結果を図 2 に示す．

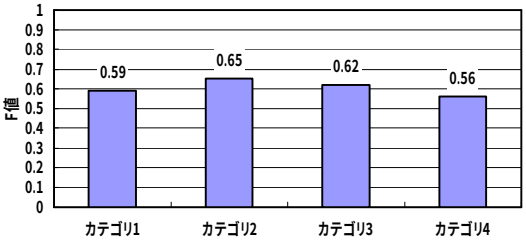

図 2 カテゴリの情報で分類した結果

カテゴリに用いる語によって属性が変わるため，図 1 より F 値が変化している．そこで属性の改良が精度向上につながると考えられる．

5.2 精度向上手法

本節では前節で述べた属性の改良を考慮に入れた精度向上のための手法を述べる．

一つ目は，Web 情報からの属性抽出手法は“＆検索”が可能である特徴を利用し，例えば「＆ニュース，＆新聞」など，分類する対象の総称を分類の手がかりとして加える．この方法で属性の精錬を図る．

二つ目は抽出した属性の中から親カテゴリとの関連が高い属性を選別する．その属性を親とし，さらに Web から属性を抽出し，カテゴリ属性の拡張を図る．

5.3 精度向上手法の実験と評価

前節で述べた属性の精錬と拡張を行い，F 値で評価した．分類結果の平均値をカテゴリ毎に図 3 に示す．

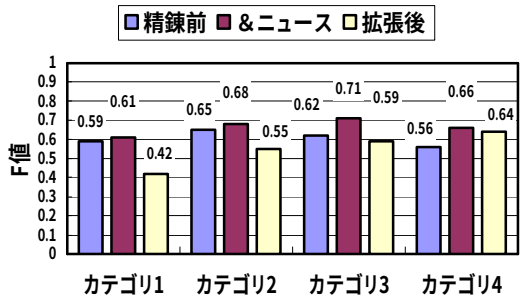

図 3 属性変更による分類結果の比較

図 3 の結果を見ると，属性の精錬は効果を示しているが，精錬後に単純に拡張を行っても評価は上がらなかった．しかし，カテゴリ 1 において，経済に分類されて適合している記事が 86 件あった．このうち拡張前に経済に適合した記事は 17 件であり，69 件の記事が拡張することによって新たに適合した．このことから属性の拡張は一度拡張前に分類した後に用いると有用である．

5.4 本研究で提案する記事分類方式

ここまでの結果をふまえ，分類の手がかりとしてコンピュータにカテゴリと分類対象の情報を与え，それを手がかりとして Web からカテゴリ毎に属性を抽出する．その属性をもとに記事群を分類するが，そこで閾値を設定し，関連度の割合が閾値より大きい記事のみ分類する．閾値以下の記事は，拡張した属性を用い分類する．

提案システムの処理の流れを以下の図 4 に示す．

① 分類記事群を読み込み，形態素解析し，TF･IDF 重み付けを行う．
② 各カテゴリの分類の手がかりであるカテゴリ･分類対象の情報を，ユーザーが入力して与える．
③ Web 情報からカテゴリの属性を抽出し，カテゴリ毎の属性群と全分類記事群と関連度計算を行う．関連度の割合が閾値より大きいものを分類する．
④ 関連度の割合が閾値以下の記事群をカテゴリの属性を拡張した属性で分類する．

図 4 提案する記事分類システムの処理の流れ

6. 提案方式による実験と評価

6.1 実験

提案した記事分類手法で実際に分類を行う．分類するカテゴリは，スポーツ面･経済面･社会面･政治面の 4 カテゴリである．テストデータとする分類対象の記事群は以下の 2 種類である．

･毎日新聞 94 年 CD-ROM の記事より，各カテゴリの記事をほぼ同数ずつ集め，4 カテゴリ合計して 1000 件の記事を集めたもの．

･朝日新聞 Web サイト上の記事より，各カテゴリの記事を各カテゴリについてほぼ同数ずつ集め，4 カテゴリ合計して 208 件の記事を集めたもの．

人間が手がかりとして与えるカテゴリは，表 1 に示す 4 組のカテゴリとし，分類対象はニュースとする．人間がこれらのカテゴリの情報を入力して与え，システムは分類結果を出力する．正解判定として，各記事に分類されたカテゴリと，その記事が本来新聞紙面上で属しているカテゴリと同じならば正解とする．

また本システムの有効性を評価するため，語の表記一致によって分類した結果，既存手法である記事分類方式[若月 2005]をあわせて示し，本システムによる分類結果と比較する．本システムでは，分類対象の記事群以外に人間が入力するものはカテゴリ･分類対象の情報のみ，システムによる出力は分類結果のみである．

6.2 評価

閾値は前述の 2 種類のテストデータを用いて，検討した結果，閾値 0.6 で二種類のテストデータとも最適値を得た．システムの評価として，表 1 の 4 組のカテゴリの情報，分類対象ニュースを入力し，閾値 0.6 で評価を行った．提案手法による毎日新聞･朝日新聞それぞれの記事の分類結果の F 値の平均を図 5 に示す．

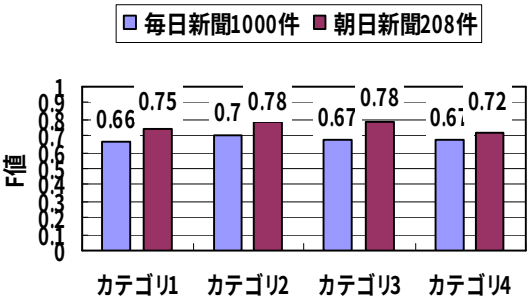

図 5 提案方式による分類結果

また，表記一致手法，既存手法，提案手法の各分類結果の F 値の平均を並べて比較したものを図 6 に示す．

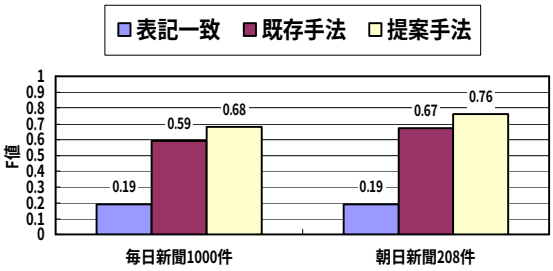

図 6 提案方式と表記一致，既存手法との比較

以上のことから，提案する記事分類手法が，新聞記事の分類において有効であることがわかった．

7. 考察

前章までの結果から，人間が与える分類の手がかりをもとに Web から情報を取得し，汎用知識ベースである概念ベースと関連度計算を用いて，多数の新聞記事を内容によって分類する知的な記事分類方式が実現できた．このシステムでは人間は分類記事群と一組の手がかりとなるカテゴリ･分類対象をコンピュータに渡すだけでよく，手がかりの詳細を述べるといった人間の負担も軽減することができた．

8. おわりに

本研究では，概念ベースと関連度計算を用いて新聞記事を分類するための手法を提案した．まず任意の未分類の記事群を分類する場合を想定し，記事群の内容によって臨機応変に分類するにはその場で手がかりを示す必要があることを指摘した．また実際にその手がかりとなるもの考え，実験を行ってそれを検討した．そして効果的な手がかりの与え方を提案し，実際にそれをもとに分類を行った．また，手がかりを人間が与える過

程でのいくつかの問題点を指摘し，提案方式によりその問題点も解決した．

謝辞

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

[荒木 2006] 荒木孝允，渡部広一，河岡司：共通・類似属性を考慮した概念間関連度計算方式，情報処理学会第 68 回全国大会講演論文集 4N-2，2006．

[倉田 2006] 倉田篤史，渡部広一，河岡司：概念ベースと関連度計算を用いた記事関連度計算方式，情報処理学会研究報告 2006-NL-171 pp.19-24，2006．

[Chasen 奈良先端科学技術大学院大学 2003] 奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座 松本研究室，http://chasen.aist-nara.ac.jp/hiki/ChaSen/，2003．

[Salton 1988] Salton, G. and Buckley, C.: Term-weighting approaches in automatic text retrieval , Information Processing & Management Vol.41 No.4 pp.513-523 , 1988.

[若月 2005] 若月紀之，松田全弘，渡部広一，河岡司：概念ベースと関連度計算を用いた新聞記事の分類，情報処理学会研究報告 2005-NL-165 pp.67-72，2005．