分野連想語を利用した未知語に対する分野の自動推定
A Field Estimation Method for Unknown Words by Field Association Words

安原寛之 森田和宏 泓田正雄 青江順一
Hiroyuki Yasuhara Kazuhiro Morita Masao Fuketa Junnichi Aoe

徳島大学大学院 先端技術科学教育部 システム創生工学専攻
Graduate School of of Advanced Technology and Science,,the University of Tokushima

Recently, the research on the automated classification of the document is paid to attention because of an increase of the Web document. In the classification of the document, there is a technique that uses the field association word that associates a specific topic field. However, there is a problem that the classification of the document becomes difficult when the unknown word to which the field has not adhered exists. In this paper we proposes the technique for automatically presuming the field of the unknown word by collecting the document to the analysis by the Web retrieval that makes the unknown word that query and using the co-occurrence relation between the unknown word and the field association words

1. はじめに

現在，電子化された文書は増大の一途を辿っている．そのため，ユーザにとって不必要な情報が増え必要な情報の取得が難しくなってきている．そこで，ユーザが必要とする情報を効率的に取得するために，文書を話題ごとに自動分類する技術の必要性が高まっている．

人間は，文書全体を読まなくても代表的な単語を見るだけで，政治やスポーツなどの分野を認知できる．このことから，文書断片内の数少ない単語情報から分野を決定するために文書の話題分野を連想する単語である分野連想語[1]の構築は重要な研究課題となっている．

しかし，分野が付いていない未知語が文書内に多数存在すれば，文書を自動的に分類することが困難である．これは，文書の自動分類には，単語の分野が必要であるからである．そこで，本稿では，文書分類の精度向上の前段階として，未知語に対して自動的に分野を推定する手法を提案する．

2. 分野連想語

2.1 分野連想語

分野連想語とは，特定の分野を連想することのできる単語をいう．例えば，野球の分野連想語としては，「ホームラン」や「ピッチャー」がある．一方，「時間」や「場所」は，特定の分野を連想できる単語ではないので，分野連想語として扱わない．以降，＜野球＞のように分野名は＜＞内に表記する．

2.2 分野連想語の構築手法

分野連想語の構築手法に関しては，様々な研究がある．．

辻らの研究[1]では，予め人手で分野分類した文書から抽出した単語がどの分野に多く含まれているか，各分野に対する出現比率によって決定している．

佃らの研究[2]では，分野名を検索キーワードとして，各分野の文書を収集する．収集した文書から単語を切り出し，単語ごとの出現比率によって連想語候補を決定する．さらに，連想語候補が正しいかどうかを，Web を用いて検証することで連想語を構築している．

連絡先：氏名 安原 寛之，所属 徳島大学大学院，住所 徳島市中常三島町1丁目 25 番地ドルフ日野 403，Tel 090-1015-3495，E-mail hiya268@is.tokushima-u.ac.jp

これらの研究では，収集したコーパスにない単語には連想語を取得できないという問題点がある．そこで，Hashimoto[3]らの研究では，未知語に注目し，未知語を検索キーワードとすることで文書を収集し，既存の辞書を用いて分野を決定している．しかし，精度は 67.5%ほどと改善の余地がある．本稿では，未知語に対して更なる精度向上を目指して分野を推定することを目的とする．

3. 提案手法

本システムは，文書収集部，分野候補決定部，分野候補検証部の 3 つにより構成される．システムの概要を図 1 に示す．システムの詳細について以降の節で述べる．

3.1 文書収集部

文書収集部では，分野連想語となる単語を抽出するために必要となる文書を，Web 検索エンジン[4]を利用して収集する．分野連想語の対象となる未知語を検索キーワードとして，文書を収集する．ここでの文書とは，サマリといい，検索結果ページにおいて，タイトル以下にある文書の要約のことである．サマリを用いる理由としては，Web ページ全体と比べて，広告など未知語と関連しない文書が少ないためである．収集する文書数は 100 文書とする．

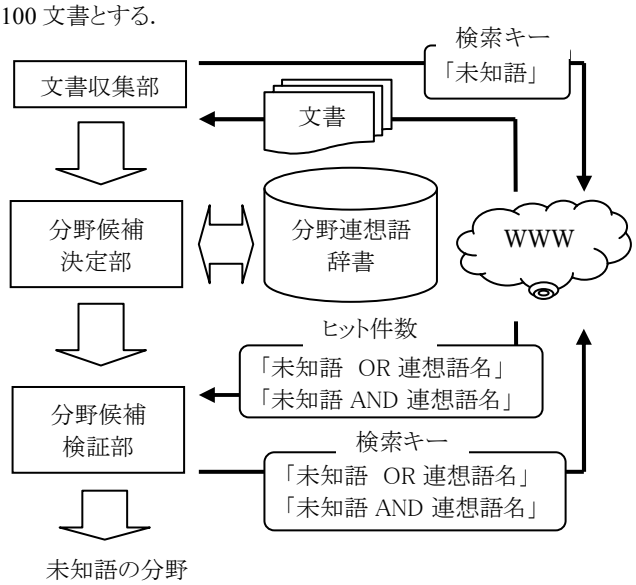

図 1：提案手法の流れ

レッドソックスのベケットはこの大舞台で完封した … 2 位を争う**レッドソックス**に対してヤンキースの中継ぎ 防御率はリーグでも … （網かけ部:<野球>）

図 2:未知語「レッドソックス」,<野球>に関する文書

3.2 分野候補決定部

分野候補決定部では，まず収集された文書 1 件ずつ分野連想語辞書を用いて分野連想語を抽出する．分野連想語辞書とは，分野数 1000，分野連想語 70 万語から成り立っている．次に，抽出された分野連想語の頻度を分野毎に集計し，最も分野連想語の頻度が多い分野を文書の分野として判定する．最後に，収集した全ての文書の分野頻度を集計することにより分野候補を決定する．経験的に上位 5 位までを分野候補とする．

3.3 分野候補検証部

分野候補決定部で獲得した分野候補に対し，共起を利用して未知語の分野として相応しいかどうかを検証する．ここで用いる共起は Jaccard 係数を用いる．また，分野候補検証部をおこなうには，以下の条件に合致しない場合とする．

条件:1 位の分野が 2 位の分野の 2 倍以上の頻度かつ 1 位の分野に含まれる分野連想語 3 種類以上ある場合，分野候補検証は行わない．この時，1 位の分野を未知語の分野とする．

一般的に，文書中には話題に関する分野連想語が 1 語しか含まれていないことは珍しく，複数の分野連想語が含まれている可能性が高い．例えば，図 2 のように＜野球＞に関する文書であれば，「完封」の他に「ヤンキース」，「中継ぎ」などの複数の分野連想語が含まれている可能性が高い．また，分野候補決定部で求めた分野の 1 位と 2 位の頻度を用いる点は，経験的によるものである．

以降，分野候補検証部の詳細について述べる．

ある未知語 U における分野候補を f_i としたとき，分野候補の集合 F を以下のように表す．

$$F = \{f_i \mid 1 \le i \le 5\}$$

各分野 f_i に含まれる連想語を w_k としたとき，連想語集合 W を以下のように表す．

$$W = \{w_k \mid 1 \le k \le 3\}$$

この時，未知語 U と分野 f_i との関連度は以下の式で表わす．

$$\mathrm{Re}\,l(U, f_i) = \frac{\sum_{k<3} Jaccard\ (U, w_k)}{N}$$

上記の式は，未知語 U と分野 f_i における関連度 Rel は，未知語 U と分野 f_i に含まれる連想語 w_k との Jaccard 係数の平均値により求めていることを表す．N は，分野 f_i 内に含まれている連想語数のことである．今回は時間を考慮した上で N=3 とする．また，Jaccard 係数の詳細については，以下の式で表す．

$$Jaccard(U, w_k) = \frac{H(U \cap w_k)}{H(U \cup w_k)}$$

ここで，$H(U \cap w_k)$ は，未知語 U と連想語 w_k の AND 検索，$H(U \cup w_k)$ は，未知語 U と連想語 w_k の OR 検索の Hit 件数である．上記の Jaccard 係数は，$H(U \cap w_k)$ の値が大きいほど，未知語 U と連想語 w_k の関連は大きくなることを示す尺度であり，これは，分野 f_i 内に含まれる連想語 w_k を用いていることから，分野 f_i との関連が大きいともいえる．つまり，Jaccard 係数の平均値を用いて計算した関連度 $\mathrm{Re}l(U, f_i)$ の値が大きければ，分野 f_i は未知語 U と関連のある分野であるといえる．

表 1:実験結果

	正解数	不正解数	正解率(%)
分野検証前	122	40	75.3
分野検証後	135	27	83.5

従って，関連度 $\mathrm{Re}l(U, f_i)$ を降順に並べ替えて，1 位の分野 f_i を未知語 U の分野とする．

4. 実験

提案手法の有効性を示すために実験を行った．

4.1 実験設定

実験データとして，未知語を 162 個用いる．評価方法としては，提案手法により抽出された分野が未知語に対して適切であれば，正解とし，適切でなければ不正解とする．判定は人手により行う．

4.2 実験結果

表 1 に分野候補検証前と分野検証後の結果を示す．実験結果から，分野候補検証後の方が分野候補検証前よりも正解率が向上していることが確認できる．また，分野候補検証後では，正解率が 80%以上と提案手法の有効性を示すことができた．

4.3 考察

分野候補検証部で関連度計算をおこなうことで不正解となる場合があった．例えば，未知語「寛永寺」の場合，分野候補検証前では，1 位が＜神社・仏閣＞，2 位が<交通>となっている．分野候補検証後では，1 位が<交通>，2 位が<神社・仏閣>となっている．これは，「寛永寺」までの行き方についての<交通>に関する文書が Web 上には多かったためと考えられる．ただ，分野候補検証前では，<神社・仏閣>に対する分野連想語の種類・頻度が<交通>に対する分野連想語よりも多かった．ただし，分野候補検証の条件に，今回は 3 種類としていたが，状況に応じて変更しなければならない．分野連想語の種類数・頻度を相対的に考慮する条件を考案する必要がある．

5. まとめと今後の課題

本稿では，文書分類の精度向上を最終目標とし，そのために分野が付いていない未知語の分野を推定する手法を提案し，提案手法の有効性を示すために実験を行った．

今後の課題としては，未知語の分野推定精度を向上するために，4.3 節の考察で述べた分野候補検証での条件に分野連想語の種類と頻度を相対的に考慮した条件を考案する必要がある．そして，未知語に対して分野を推定し，動的に分野連想語辞書を更新することで，文書分類の精度向上をおこなうことも今後の課題である．

参考文献

[1] 辻孝子, 泓田正雄, 森田和宏, 青江順一"複合語の分野連想語の効率的決定法", 自然言語処理, Vol.7, No. 2, pp.3-26 (2000).

[2] 佃 陽平, 森田 和宏, 泓田 正雄, 青江 順一 : Web 検索エンジンを用いた分野 連想語の自動抽出に関する研究, 言語処理学会第 12 回 年次大会, pp.648-651, (2006)

[3] Chikara Hashimoto,Sadao Kurohashi : Construction of Domain Dictionary for Fundamental Vocabulary,2007

[4] Yahoo!デベロッパーネットワーク, http://developer.yahoo.co.jp/