ICT 用および医学用大規模専門辞書の Web 集合知を用いた分類法

A Classification Method for Large Scaled Dictionaries based on Collective Intelligence of Web Pages

千種 康民[*1], 郭 炳君[*1], 服部 泰造[*2]
Yasutami CHIGUSA[*1], Heikun KAKU[*1],Taizoh Hattori[*2]
[*1] 東京工科大学大学院バイオ情報メディア研究科
[*1]Graduate School of Baio, Informatics and Media, Tokyo University of Technology
[*2] 東京国際大学
[*2]Tokyou International University
Email: chigusa@media.teu.ac.jp

あらまし：本稿では、著者らの ICT 用の大規模専門辞書の自動分類法の研究を発展させ、医学用の大規模専門辞書への応用を適用し、それぞれ一定の効果を得た。本研究の特徴は用語の相関に基づく分類法であり、その用語の相関は Web 集合知を用いることにより、専門家による分類を必要とせず、一定以上の正答率を得ることを実現した。

キーワード：Web 集合知、大規模辞書の自動分類、Jaccard 係数、Simpson 係数

1. はじめに

e-learning の導入の最大の目的は自動化と高い効果を得ることの両立である。その実現のためには良質の用語辞書の構築が必要不可欠である。

しかし、近年の e-learning の急速な普及を考える時、旧来からの学問分野においては一定の体制が整っているが、良質な専門辞書を構築する様々な分野における体制が、必要とされる適用範囲の拡大に追い付いていないのが現状である。また、専門辞書を構築するための体制は、人材面における不足だけでなく、コスト面においても不十分であるのも現状である。

そこで本研究では、専門辞書の自動分類に Web 集合知を活用することに注目し、用語間の共起頻度が高い場合ほど、それらの用語は同じカテゴリーに分類されるという仮説を立て、そのルールに則り、専門辞書の自動分類に適用し、その効果を報告する。

また、間違った分類された辞書をメンテナンスすることは非常に困難になるため、用語の自動分類可能性という尺度を導入し、その尺度に基づき、自動分類する用語と、自動分類が困難である用語に自動分別し、自動分類される専門辞書の品質を維持することを実現した。

以上の手法を用いて、具体的には ICT 用専門辞書と医学用専門辞書の２つの異なる学問領域の辞書の分類を実施し、その効果を調査し、一定の成果を得た。

2. 用語の自動分類可能性

分類の対象となる専門用語を自分の検索件数によって 2 種類に分けて分類する。

非分類用語　１単語を指定した際の自身の検索件数がある値α未満の場合、他の用語との共起件数が極めて少なくなり、自動分類の精度が保証されにくくなるため、仮の専門用語として分類し、自動分類されない用語として処理する。 後日、専門家により手動で分類されることを想定している。

可分類用語　自身検索件数がα以上の場合、正常に分類可能な専門用語として処理する。辞書内の専門用語中のすべての既知専門用語との共起を調べ、分類処理を実行する。

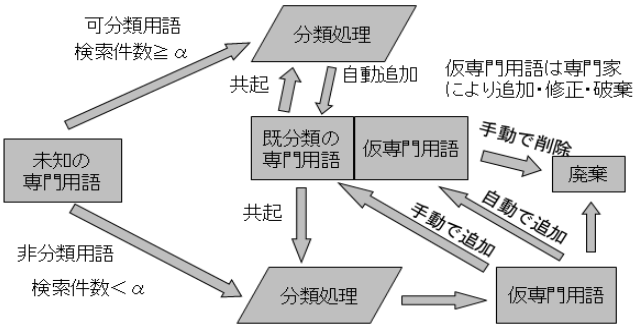

図１．分類処理と非分類用語・可分類用語

3. 分類処理と使用アルゴリズム

分類処理には、Simpson 係数を用いる手法と、横軸係数（本手法）とを検討する。今、２つの単語 ai と bj があるとき、C(x,y)は x と y の共起件数であるとし、シンプソン係数 $Ps(a_i, b_j)$は（１）式で表現される。同様に横軸係数 $Pr(a_i, b_j)$は（２）式で表現される。

$$Ps(a_i, b_j)= C(a_i, b_j)/\min(C(a_i, a_i) , C(b_j, b_j)) \quad (1)$$

$$Pr(a_i, b_j)= C(a_i, b_j)/ C(a_i, a_i) \quad (2)$$

4. 計算機シミュレーション

4.1 ICT 系専門用語に対する予備実験

「IT パスポート試験」に刑されている専門用語を対象にし、第２章～第４章、第６章の４つの章から

分類済みの５語ずつピックアップし、前述の２つの手法で分類した。専門用語は既に分類済みであるが各専門用語が正しく分類されるかどうかを確認した。その結果 Simpson 係数法では正解 17、不正解 3 であるのに対して、横軸係数法では正解 14、不正解 6 であった。Simpson 係数法では検索件数の少ない用語に対しては比較的正確に分類できるが、多い用語の影響を受けやすい。一方、横軸係数法では検索件数の多い用語に対しては比較的正確に分類できるが、少ない用語に対しては不正解になりやすい傾向がある。

4.2 ICT 系専門用語に対する本実験と考察

予備実験を踏まえ、本実験では４つのカテゴリーに対して、５０単語中からランダムに各４０単語・計１６０単語が既分類、各１０単語・計４０単語を未知語として分類した。この処理を 50 回繰り返し集計した。ここではα＝５００とした。

Simpson 係数法では、共起件数の多い専門用語が分類に悪影響を与えていることが分かり、本提案手法では、共起件数が少ない物についてのご分類の影響はあまり見られず、共起件数が多いものについて改善効果が見られた。結果として提案手法（横軸係数法）によれば 5 語以上の間違いがなく、全体的に Simpson 係数法より優れた結果になった。

	0-1語間違い	2-3語間違い	3-5語間違い	5語以上間違い
simpson係数	12%	72%	12%	4%
提案手法	12%	76%	12%	0%

表１．ICT 系専門用語に対する Simpson 係数法と提案手法（横軸係数法）

4.3 医学系専門用語辞典に対する本実験と考察

「大安心　健康の医学大事典」を対象とし、各章から 3 単語を選び分類し、その結果から、比較的正確に分類できそうな 4 章､循環器病気､消化器病気、運動器の病気、皮膚の病気、を対象カテゴリーとして分類の本実験を実施した。

本実験では４つのカテゴリーに対して、５０単語中からランダムに各４０単語・計１６０単語が既分類、各１０単語・計４０単語を未知語として分類した。この処理を 50 回繰り返し集計した。ここではα＝５００とし、ここでは提案手法のみを調べた。

	80%以上正確	70%以上正確	70%以下正確
循環器病気	10	19	11
消化器病気	9	31	0
運動器の病気	37	3	0
皮膚の病気	38	2	0
合計	25	14	1

表２．医学系専門用語の提案手法による分類性能

併発病が web 検索結果からよく共に出現するため、分類結果を影響する。分類結果としては、ICT 系の実験より効果が若干良くないため、併発病への対応など医学事典の自動分類には他の条件を追加する必要があると思われる。

5.　総合評価とまとめ

本稿では、自動分類機能の一手法を提案し、未知専門用語が少ない場合、提案手法の分類精度は Simpson 法より高く、ICT 系と医学辞書の分類実験をし、一定の効果を示すことができた。

今後は、提案手法で使っている検索件数αの値の検討、他の専門用語の分類実験、カテゴリー数を増やした場合の実験、他の研究事例との比較評価、を実施していく予定である。

	IT系		医学系	
	平均値	標準偏差	平均値	標準偏差
カテゴリA	88.67%	4.53	76.28%	8.10
カテゴリB	82.08%	6.49	80.95%	5.00
カテゴリC	86.43%	5.36	86.15%	10.31
カテゴリD	84.76%	5.87	89.23%	7.58
合計	85.17%	4.26	82.61%	4.51

表３．専門辞書の違いにより正答率の差異

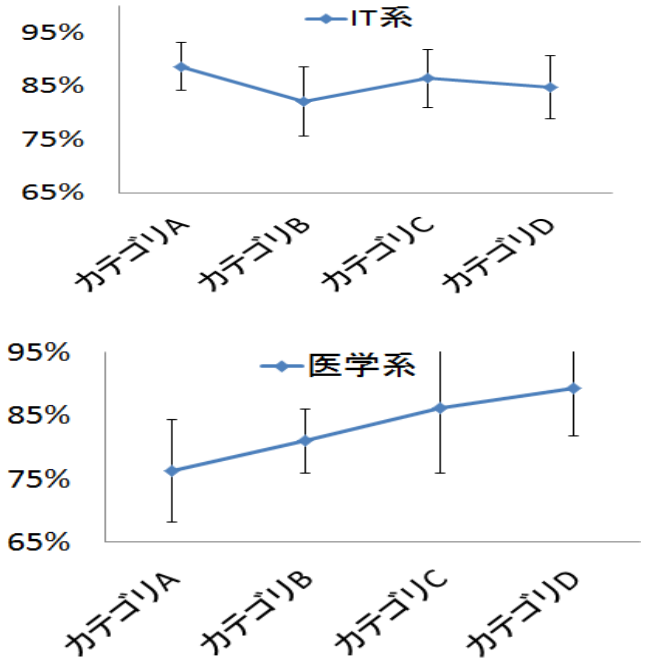

図２．自動分類の正解率と標準偏差

参考文献

[1]蘇　寧，張 暁霞，佘 錦華，服部泰造, 山崎祥行，“日中混在 ICT 問題自動作成システムの開発”, 電子情報通信学会技術研究報告，Vol. 109,No.11，pp 119-124(2010.02).

[2]李 依霖，張 暁霞，佘 錦華，陳 淑梅，千種 康民，亀田 弘之, 大野 澄雄，“個人適応技術中国語 e-ラーニングシステムの構築”，日本 e-Learning 学会，vol.11, pp.4-11 (2011.07).