

文書分類体系間の対応関係の自動抽出 Automatic Extraction of Correspondences between Document Taxonomies

金田 有二[†]
Yuji Kaneda

斉藤 和巳[†]
Kazumi Saito

上田 修功[†]
Naonori Ueda

1. はじめに

近年、インターネットの普及に伴い、膨大な情報が自己組織的に分散されて蓄積されつつある。Webディレクトリに代表されるように、情報は階層的に管理されることが多く、これらは“階層知識”とも呼ばれている。知識体系の階層化は、設計者の意図に依存するので、同一分野、同一ドメインにおいても、多様な知識体系が存在し得る。それ故、それらの階層知識間の“相互翻訳”が実現できれば、各々の特徴、メリットを活かしながら、分散された知識を有効利用することが可能となる。本稿では、二つの階層分類トピック体系間のトピック対応関係を自動抽出する手法について検討する。

2. 階層知識間の対応関係の自動抽出

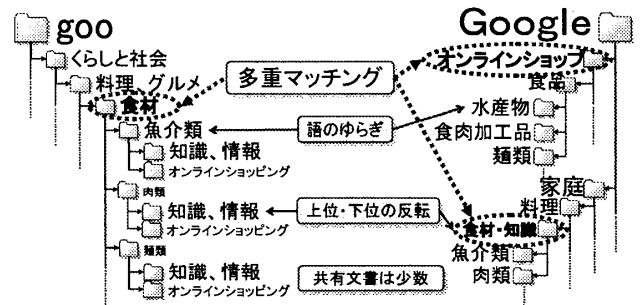
2.1 抽出問題の特徴と従来法の問題点

図1に、“グルメ”における、2つのWebディレクトリ goo と Google のトピック体系の例を示す。同じドメインでも分類体系が異なっていることが確認できる。この例を用いて、二つの階層分類トピック体系の対応問題の特徴と従来法の問題点を以下に整理する。

- 文書は多重のトピックを持ち得るため、“多重対応”を考慮する必要がある。しかし、従来手法 [1, 2] は“一対一対応関係”のみを扱っている。
- 「魚介類」と「水産物」のように、語のゆらぎが存在するため、トピック名のマッチングでは限界がある。
- 二つの体系間で、上位・下位の関係が反転する場合がある。上位・下位構造を用いた木構造マッチング手法 [2] では、正しい対応関係が得られない。
- 二つの体系間で共有する文書は必ずしも多くない。共有文書数からトピックの対応を求める手法 [2] では限界がある。

2.2 提案法のアプローチ

- 単語頻度分布に基づく確率モデルアプローチをとる。二つの体系の各々において、各トピックをそのトピックに属す文書の単語頻度分布で特徴づけ、その分布間の“尤度基準”に基づいてトピックの対応関係を抽出する。但し、対応関係は1対1対応を含む“多重対応”とする。また、MDL基準 [4] を用いて、できるだけ簡潔な多重対応を求める。
- 対応関係を抽出する際、トピックの階層構造を直接用いない。単語頻度分布を算出する際、そのトピックに直接属す文書のみならず、全ての子孫のトピックに属す文書も用いることにより、トピックの階層構造を暗に用いる。



問題: 各体系の文書群を用いて、分類トピック間の多重対応を求める

図1: 文書分類体系間の対応関係の抽出問題

以上から明らかなように、本アプローチでは、対応関係にあるトピック間で名前が異なったり、共有文書数が全くなくても問題ない。また、木構造を直接用いていない為、(c)の問題がなく、また DAG (非巡回有向グラフ) にも適用できる。さらに、計算量もトピックの線形オーダ (構造マッチングでは指数オーダ) で、極めて効率的ゆえ、大規模問題にも適用可能である。

3. 提案法の詳細

2つの分類体系 S^1, S^2 のトピック集合を、それぞれ $\{c_k^1\}_{k=1}^{K^1}, \{c_l^2\}_{l=1}^{K^2}$ とし、2つの分類体系間の対応関係を抽出する。具体的には、 S^1 のトピック c^1 を、 S^2 のトピックに多重を許容して対応付ける。なお、トピックは有向リンクで結ばれ、全体は DAG であるとする。提案法は、以下の処理より成る。

3.1 トピックの NB パラメータの算出

まず、文書 d を単語頻度ベクトル $\mathbf{x}(d) = (x(d, 1), \dots, x(d, V))$ で表現する。すなわち、BOW (Bag-of-Words) 表現を採用する。ここで、 V は全文書での異なる単語数 (語彙規模) を、 $x(d, i)$ は文書 d にて i 番目の単語が現れた頻度を表す。そして、 S^2 のトピック c^2 の NB パラメータ $\theta(c^2) = (\theta(c^2, 1), \dots, \theta(c^2, V))$ を、 $\theta(c^2, i) = \sum_{d \in D(c^2)} x(d, i) / \sum_j \sum_{d \in D(c^2)} x(d, j)$ として算出する。 $D(c^2)$ は、トピック c^2 、および、トピック c^2 の子孫のトピックに属する文集の書合とする。

3.2 リーフトピックでの対応関係の抽出

各 $c_k^1 \in S^1$ に対し、単語頻度ベクトル $\mathbf{y}(c_k^1) = (y(c_k^1, 1), \dots, y(c_k^1, V))$ を、 $y(c_k^1, i) = \sum_{d \in D(c_k^1)} x(d, i)$ として求める。多重トピックテキストの確率モデル [3] の基本アイデアを用い、 S^2 のリーフトピックにおける第 i 単語の生起確率を、リーフトピックでの生起確率の線型混合 $w_0/V + \sum_{\{l|c_l^2 \in G_0\}} w_l \theta(c_l^2, i)$ としてモデル化する。但し、 $w_0 + \sum_{\{l|c_l^2 \in G_0\}} w_l = 1$, G_0 はリーフトピッ

[†]日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

表 1: 評価データの基本統計量と共通ページ数

	トピック数	ページ数	共通ページ数		
			goo	Google	Excite
goo	282	1,942	-	250	304
Google	196	1,149	250	-	160
Excite	245	1,411	304	160	-

ク集合とする。そして、 c_k^1 に属す文書集合 $D(c_k^1)$ での単語群が、このモデルで生成されるとすると、対数尤度は次式と書ける。

$$L_k(\mathbf{w}; G_0) = \sum_{i=1}^V y(c_k^1, i) \log \left(\frac{w_0}{V} + \sum_{\{l|c_l^2 \in G_0\}} w_l \theta(c_l^2, i) \right). \quad (1)$$

従って、対応関係を表す未知パラメータ $\mathbf{w} = \{w_l\}_{l=0}^{K^2}$ の最尤推定値は、 $w_0 + \sum_{\{l|c_l^2 \in G_0\}} w_l = 1$ の制約条件の下で、 $L_k(\mathbf{w}; G_0)$ を最大化する $\hat{\mathbf{w}}$ として求まる。この最大化問題は、 $\mathbf{w}$ に関して上に凸で、解の大域的最適性が理論保証される。

3.3 MDL 基準によるモデル探索

対応先トピックの集合を $G \subset S^2$ とし、簡潔な表現を得るため、次の目的関数を最大にする G を求める。

$$\text{MDL}_k(G) = -L_k(\hat{\mathbf{w}}; G) + \frac{1}{2} |G| \log \sum_{i=1}^V y(c_k^1, i). \quad (2)$$

ここで、 $\hat{\mathbf{w}}$ は式 (1) の最尤推定量を表す。 G を探索する際には、 G_0 を初期値として、上位トピックへのマージを繰り返す。そして、得られた解 $\hat{G}$ を用いて、 $L(\mathbf{w}; \hat{G})$ の最尤推定量 $\hat{w}$ を、抽出された対応関係とする。以上を全ての $k = 1, \dots, K^1$ について行う。

4. 評価実験

提案法を評価するため3つのポータルサイト“goo”、“Google”“Excite”に着目し、グルメをトピックとするそれぞれの分類体系とWebページ群を収集した。表1に、評価データの基本統計量を示す。

4.1 比較方法

提案法の性能を評価するため、本研究の意図を告げない2名の被験者に、ある分類体系にのみ分類されているWebページ群を、別の分類体系によってラベリングさせた。このとき、6通り全ての組み合わせについて、ラベリングを行った。これらのラベリングの際には、一つのページに対して複数トピックの付与を許容し、適当なトピックの付与が困難なページに対して“その他”の付与を許容した。ラベル付与後、分類元の各トピック c^1 に対し、 c^1 を継承する文書群 $D(c^1)$ の文書 d の0-1ラベルベクトルの平均として、対応先でのトピック度ベクトル $\mathbf{h} = (h_0, h_1, \dots, h_{K^2})$ を求めた。ただし、複数ラベルを持つ文書では、和が1になるように各文書毎の正規化を予め行ない、“その他”が付与されたページのラベルは h_0 に対応させた。

4.2 定性評価

図1の例と同様に、gooの“食材”を、Googleの分類体系に対応させた場合の、被験者の $\mathbf{h}$ と、提案法の

表 2: 提案法と比較法の正当率 (%)

S^1 S^2	goo	goo	Google	Google	Excite	Excite
	Google	Excite	goo	Excite	goo	Google
提案法	26	31	36	30	35	24
multi	16	12	6	9	9	18

$\mathbf{w}$ を比較したところ、直感的に良好な結果が得られた。例えば、被験者では、“オンラインショップ”に属するトピックの h_l の和が0.58，“食材・知識”に属するトピックの h_l の和は0.12であった。一方、提案法では、“オンラインショップ”に属するトピックの w_l の和は0.50，“食材・知識”に属する w_l の和は0.04であった。

4.3 定量評価

被験者のラベリングから求まる $\mathbf{h}$ と、提案法で求まる $\mathbf{w}$ の一致正当率 E を以下で定義する。

$$E = \sum_{k=1}^{K^1} \frac{E(c_k^1)}{K^1}, \quad E(c_k^1) = \sum_{l=0}^{K^2} \min(h_l, w_l). \quad (3)$$

この評価尺度は、トピックの階層構造を考慮しない厳しい評価量である。例えば、“水産物”に対応付けるべきトピックを、その上位トピックである“食品”に対応づけても全く評価しない。

比較のため、ナイーブな手法として、多項分布に基づく比較法 (multi 法) を独自に構築した。multi 法ではトピック度 $\mathbf{w}$ を $w_l \propto \exp\{(\beta / \sum_{i=1}^V y(c^1, i)) \sum_{i=1}^V y(c^1, i) \log \theta(c_l^2, i)\}$ により求める。 β はトピック度ベクトルの滑らかさを制御するパラメータである。また、 $\theta(c_0^2) = (1/V, \dots, 1/V)$ とした。

4.4 定量評価の結果

6通りの $\{S^1, S^2\}$ の組み合わせについて、提案法と multi 法を適用し、被験者の対応結果との一致度を比較した。表2に評価結果を示す。なお、multi 法については、 $\beta = 0, 1, 2, \dots, 50$ と、 β を動かしたときの最大の正当率を示す。提案法の正当率は24%から36%であった。トピック数 (282, 196, 245) に比べれば、それなりの性能が実現できたと考える。また、いずれの組み合わせにおいても、提案法は、最適な β を用いた multi 法よりも高い正当率を示した。

5. まとめ

本稿では、テキスト数理モデルを土台とし、文書分類体系間のマッチングを行なうための問題設定と初期アルゴリズムについて述べた。実データを用いた評価実験では、提案手法の有望性を示す結果が得られた。

参考文献

- [1] A. Doan, et al., “Learning to map between ontologies on the semantic web,” Proc. WWW’02, 2002.
- [2] R. Ichise et al., “Rule induction for concept hierarchy alignment,” In Proc. IJCAI’01, 2001.
- [3] 上田修功, 齊藤和巳, “多重トピックテキストの確率モデルパラメトリック混合モデル,” 電子情報通信学会論文誌, vol. J87-D-II, no. 3, pp. 872–883, 2004.
- [4] J. Rissanen, “Stochastic Complexity in Statistical Inquiry,” World Scientific, 1989.