

確率モデルを用いた文書分類体系間の構造マッチング

斉藤 和巳 上田 修功 金田 有二

日本電信電話(株) NTT コミュニケーション科学基礎研究所

概要: 確率モデルを土台として、階層構造を持つ異なる文書体系のマッチングを求める新手法を提案する。提案法の特徴は、多重トピック文書を扱うパラメータ混合モデル PMM に基づき、最小記述長基準 MDL を評価尺度として、モデル探索によりマッチングを求めることにある。実データを用いた評価実験では、提案法の有望性を示す初期結果が得られたことを示す。

Matching between Document Taxonomies using Probabilistic Models

Kazumi SAITO Naonori UEDA Yuji KANEDA

NTT Communication Science Laboratories, NTT Corporation

abstract: We propose a new method for obtaining mappings between document taxonomies by using probabilistic models. The main characteristics of the proposed method is that based on parametric mixture models for multi-topic documents, the matching problem is formalized as a model search problem by using a minimum description length criterion. In our experiments using real Web data, we show that our approach can produce promising initial results.

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。オンライン辞書や 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。

階層知識の体系は、異なる分野間では当然異なるが、関連する知識も多く存在する。例えば、計算機科学、脳科学、認知科学という分野において、“学習”というテーマでは相互に関連する研究が数多く存在する。この時、異分野間で関連テーマに関する整合がとれていれば、異分野で行われている研究の理解と成果の有効利用が可能となり、研究の活性化を図ることができる。

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。例えば、“グルメ”において、二つの Web ディレクトリ、goo と Google では、トピック階層の上位/下位の反転やトピック名の揺らぎ等によって異なる体系を持つ。セマンティック Web で用いられるオントロジーもその典型例であ

る [6][7][8]。

このように、多様な階層知識は各々独自の特徴やメリットをもつ。それ故、それらの知識間での階層知識の“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。しかしながら、人手で上記対応付けを行うのは、コストと労力の面で困難であり、何らかの自動化手法が必要である。

こうした背景の下、階層知識間の“一対一対応関係”の自動抽出手法が数多く提案されている [1]-[9]。ところが、一般に文書は多重のトピックを持ち得るため、“一対一対応関係”では不十分で、“多重対応”を考慮する必要がある。実際、Web ディレクトリの場合、同一ページがハイパーリンクにより複数のトピックにリンクされている。この場合、あるトピック階層のトピックが別のトピック階層の複数のトピックに対応すると考えるのが自然である。従来法はこの多重対応の視点が欠けている。

筆者らは先に、多重トピックをもつテキストモデル(パラメトリック混合モデル: PMM)を考案し、多重テキスト分類 [10][11] およびテキスト検索 [12] での有効性を示した。本稿では、PMM を階層情報

源の多重トピック対応に発展すべく、インターネットディレクトリのように、トピック階層と各トピックに属す文書群からなる構造化された階層情報源を対象に、2つのトピック階層の多重対応関係を、両者に帰属する文書群に現れる語彙情報のみを用いて効率良く求める手法を提案する。

以下の本文では、まず、問題設定とその解法を詳述し、次いで、goo と excite の二つの Web ディレクトリ間のマッチングへの適用例と手法の有効性の評価を行う。

2 問題設定と解法

文書分類体系の多くは階層的構造で、上位に汎用概念の分類トピックがあり、それらを再帰的に詳細化することで構成される。ただし、既に述べているように、同じドメインでも、分類目的や作成者の意図に応じて多様な分類体系が存在する。図 1 に、2つのグルメ関連 Web ディレクトリの例を示す。まず、goo の“ノンアルコールドリンク”と google の“嗜好品”のように、トピックとして比較的近くても名称は異なることがある。また、共通に使われる“パン”のように、同じ名称でも置かれる階層レベルが異なる。さらに、“パン”と“レシピ”のように、階層での上下関係が逆転する場合もある。

このような多様性があるため、2つの文書分類体系を“多重対応”させるには、トピック名称や階層構造のみに着目するだけでは限界があり、本稿では実際に分類されている文書群を使うアプローチを採用する。ただし、全ての文書が2つの体系でともに分類されている状況は前提としない。以下では、トピックに基づく文書確率モデルを土台に、文書分類体系の構造マッチングを行なう問題設定と解法について述べる。

2.1 単語頻度によるトピック表現

分類体系 S のトピック集合を $\{c_k : k = 1, \dots, K\}$ とし、分類されている文書集合を $\{d_n : n = 1, \dots, N\}$ とする。これらの各要素は有効リンクで結ばれ、全体は DAG (Directed Acyclic Graph) 構造であるとする。図 1 に示したように、Web ディレクトリに代表される文書分類体系で標準的な構造である。た

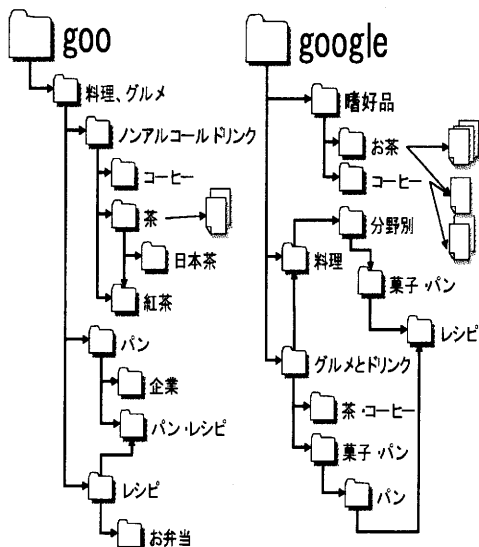


図 1: 文書分類体系 (Web ディレクトリ) の例

だし、文書 d から他のノードへのリンクは考慮しないが、任意の複数トピックからの文書 d へのリンク (文書の多重トピック) は許容する。

文書 d を単語頻度ベクトル $\mathbf{x} = (x_1, \dots, x_V)$ で表現する。ここで、 V は全文書での異なる単語数 (語彙規模) を、 x_i は文書 d にて i 番目の単語が現れた頻度を表す。すなわち、文書中出现する単語順序を無視し、その出現頻度のみに着目してテキストを記述する BOW (bag-of-words) 表現 (e.g., [14]) を採用する。

分類体系 S にて、上位ノードのトピックは、有効リンクに基づき下位ノードに多重継承されるとする。すなわち、あるノードから有効リンクを手繰って到達できる文書群は、そのノードのトピックを持つと考える。いま、ノード c のトピックを継承する文書集合を $D(c)$ とする。 $D(c)$ に含まれる文書の単語頻度ベクトルを加算すれば、文書の場合と同様に、トピック c を単語頻度ベクトル $\mathbf{y} = (y_1, \dots, y_V)$ で表現できる。

ここで、本稿で対象とするような BOW に基づくテキスト分類などで、シンプルでロバストな手法の一つとして広く採用されている NB (naive Bayes) モデル [13] について述べる。NB 文書生成モデルでは、トピック c を持つ文書 $\mathbf{x}$ は、語彙に関する多項 (multinomial) 分布により生成されると仮定す

る。すなわち、第 i 単語の生起確率を $\theta_{c,i}$ とすれば、

$$P(\mathbf{x} | c) \propto \prod_{i=1}^V \theta_{c,i}^{x_i}, \quad \theta_{c,i} \geq 0, \quad \sum_{i=1}^V \theta_{c,i} = 1, \quad (1)$$

に基づいてトピック c の文書 $\mathbf{x}$ が生成されるとする。したがって、単語頻度ベクトル $\mathbf{y}$ から、トピック c の NB パラメータを θ を求めることができる。

2.2 確率モデルを用いたマッチング法

分類体系 S' の S へのマッチングの基本として、 S' の一つのトピック c' を、 S のトピック集合 $\{c_k : k = 1, \dots, K\}$ へ確率的に対応付ける問題を考える。既に述べたように、トピック c' は S の複数トピックに対応するケースも起こる。本稿では、このような多重トピックを扱う文書確率モデルとして、パラメータ混合モデル (PMM: Parametric Mixture Model) [10] の基本アイデアを用いる。PMM では、多重トピックを持つ文書には、各トピックに特徴的単語が混合して出現すると考え、これらトピックの NB パラメータを線形結合させて多項分布を構成する。

いま、トピック c' の単語頻度ベクトルを $\mathbf{y}$ で、 c_k の NB パラメータを θ_k で表す。PMM の考え方に基づけば、トピック度ベクトル $\mathbf{w} = (w_0, \dots, w_K)$ に対して、以下の最尤推定問題を解けばトピック c' の S への対応関係が得られる。

$$L_0(\mathbf{w}) = \sum_{i=1}^V y_i \log \left(\frac{w_0}{V} + \sum_{k=1}^K w_k \theta_{k,i} \right). \quad (2)$$

ただし、 $w_0 + \dots + w_K = 1$ の制約を持つ。しかるに、この基本アイデアを実現するには、以下の 2 つの問題を解決する必要がある。

第一に、式 (2) の解は一般に唯一に定まらない。 $(V-1)$ -次元単体上で、 $\{\theta_k : k = 1, \dots, K\}$ が互いに線形独立ならば、 $L_0(\mathbf{w})$ は凸領域で定義された凸関数なので、唯一の最適解を持つことが保証されるが、上記対応付け問題では、一般に線形独立とはならない。例えば、トピック c が下位ノード c_1 と c_2 を持つならば、これらトピックの単語頻度ベクトル間に $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$ の関係が成り立つことが容易に起こり得る。よって、これらを正規化して得られる NB パラメータは線形従属となる。

第二に、式 (2) を解いても陽に簡潔な対応関係は得られない。つまり、対応関係を直観的に理解するのが困難になる。学習問題の観点からは、単語頻度

ベクトル $\mathbf{y}$ に対して、全ての NB パラメータの混合モデルを用いては一般に過剰適合するので、適切なモデル選択を行なう必要がある。

以下では、このような問題を克服する提案法について述べる。まず、NB パラメータの線形従属性を克服するため、トピックの部分集合 $G \subset \{1, \dots, K\}$ を導入して、以下の最尤推定問題を考える。

$$L(\mathbf{w}) = \sum_{i=1}^V y_i \log \left(\frac{w_0}{V} + \sum_{k \in G} w_k \theta_{k,i} \right). \quad (3)$$

ただし、 $w_0 + \sum_{k \in G} w_k = 1$ の制約を持つ。そして、最小記述長 (MDL: Minimum Description Length) 基準 [15] を導入して、以下の目的関数に基づくモデル探索問題を考える。

$$\text{MDL}(G) = -L(\hat{\mathbf{w}}) + \frac{1}{2} |G| \log \sum_{i=1}^V y_i. \quad (4)$$

ここで、 $\hat{\mathbf{w}}$ は式 (3) の最尤推定量を表す。 $|G|$ は集合 G の要素数で、モデルの自由パラメータ数に他ならない。 $\sum_{i=1}^V y_i$ は出現した総単語数で、確率モデルの試行回数 (MDL 適用におけるサンプル数) に対応する。

式 (4) に基づく探索アルゴリズムは以下である。ただし、トピック番号 k のノードが直接リンクしているノード群のトピック番号集合を $A(k)$ で、出力用の結果集合を G^* で、探索制御用の集合を NG で、そして、ループ制御変数を t で表す。

- S1:** 文書に直接リンクしているノードのみ選択して $G(0)$ を構成し、 $G^* = G(0)$, $NG = \emptyset$, $t = 1$ に初期化する。
- S2:** 条件 $k \notin NG \wedge A(k) \subseteq G(t-1)$ を満たすトピック番号 k を任意に選び、 $G(t) = (G(t-1) - A(k)) \cup \{k\}$ とする。もしこの条件を満たす k がなければ $NG = \emptyset$ として **S4:** に進む。
- S3:** もし $\text{MDL}(G(t)) < \text{MDL}(G(t-1))$ なら $t = t+1$ とし、さもなければ $NG = NG + \{k\}$ として **S2:** に戻る。
- S4:** 条件 $k \in (G(t-1) - NG)$ を満たすトピック番号 k を任意に選び、 $G(t) = G(t-1) - \{k\}$ とする。もし条件を満たす k が存在しなければ G^* を出力して終了する。
- S5:** もし $\text{MDL}(G(t)) < \text{MDL}(G(t-1))$ なら $t = t+1$ とし、さもなければ $NG = NG + \{k\}$ として **S4:** に戻る。

表 1: 評価データの基本統計量

サイト名	トピック数	リンク数	ページ数
goo	282	2,521	1,942
excite	245	1,846	1,411

上記アルゴリズムでは、集合 G で規定する NB パラメータ群の線形独立性を保持しつつ、S2:と S3:で下位ノード群を上位ノードに汎化できるか、S4:と S5:で不要ノードを枝刈りできるかを調べながら、簡潔かつ適切なモデルの探索を実行している。最終的なトピック度ベクトル w_k は、 $k \in G^*$ ならば式 (3) で求まる推定量、さもなければ 0 とする。明らかに、 $w_0 + \dots + w_K = 1$ の制約は保持される。

3 評価実験

提案法を評価するため2つのポータルサイト “goo” と “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 ページ群の収集を行なった。これらデータ収集には Gnu Wget と呼ばれるインターネットロボットを用い、それぞれのサイトから直接リンクされるサイト外の Web ページの実際の文章と、分類体系を規定するリンク情報を得た。

本実験のため収集した Web ページの単語には、動詞の活用などを基本形に統一する語末処理を施し、冠詞などに代表される 205 個の不要語 (stop words) を削除し、出現単語数が 50 未満のページを除外し、そして頻度 10 未満の単語を無視した。その結果、単語の総数は 10,659 となった。この処理は2つのサイトで同一とした。表 1 に、評価データの基本統計量を示す。ただし、リンク数にはトピックから文書へのリンクも含まれる。なお、両サイトに共通して含まれるページ数は 304 であった。

3.1 評価尺度と比較法

提案法の性能を定量的に評価するため、goo にのみ登録されている Web ページ群に対して、excite の分類体系ならばどのトピックを付与するかのラベリングを、本研究の意図を告げない2名の被験者により行なった。ただし、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 Web ページに対して “その他” の付与を許容した。ラベル付与完了後、goo の各トピック c に対し、このトピックを継承する文書群 $D(c)$ の文書 d の 0-1 ラベルベクトルの平均として、対応付け先 excite でのトピック度ベクトル $\mathbf{h} = (h_0, h_1, \dots, h_K)$ 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。

被験者のラベリングから求まる $\mathbf{h}$ と提案法で求まる $\mathbf{w}$ の一致正当率を以下のように求めた。

$$E = \sum_{m=1}^{K'} \frac{E(c_m)}{K'}, \quad E(c_m) = \sum_{k=0}^K \min(h_k, w_k). \quad (5)$$

評価尺度 $E(c)$ は、トピック度ベクトル $\mathbf{h}$ と $\mathbf{w}$ 分布の共通部であり、明らかに、両者が完全に一致する最良ケースで 1 に、全く一致しなければ 0 になる。

BOW に基づくテキスト処理で幅広く採用されるコサイン類似度、および多項分布に基づく比較法を構築し、提案法の有効性を評価した。コサイン類似度が土台の cos 法では、トピック度ベクトルを以下で求めた。

$$u_k \propto \exp \left(\frac{\beta}{\|\mathbf{y}\| \|\boldsymbol{\theta}\|} \sum_{i=1}^V y_i \theta_k \right). \quad (6)$$

一方、多項分布が土台の multi 法では以下で求めた。

$$v_k \propto \exp \left(\frac{\beta}{\sum_{i=1}^V y_i} \sum_{i=1}^V y_i \log \theta_k \right). \quad (7)$$

ここで、 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、実験では β の値を適当に変化させて評価を行なった。なお、multi 法にて $\sum_{i=1}^V y_i$ の除算を行なっているのは、両手法で有効な β のレンジを近い範囲にするためである。ただし、multi 法の NB パラメータにはラプラススムージングを施している。また両手法とも、 $\theta_0 = (1/V, \dots, 1/V)$ として、それぞれの u_0 と v_0 を求めて評価した。

3.2 評価結果

図 2 に評価結果を示す。提案法の正当率は 33% 程度であった。トピック数は 245 もあるので、それなりの性能が実現できたと考える。また、どの β を用いた比較法よりも、提案法は高い正当率を示した。なお、multi 法では、NB パラメータを求めるスムー

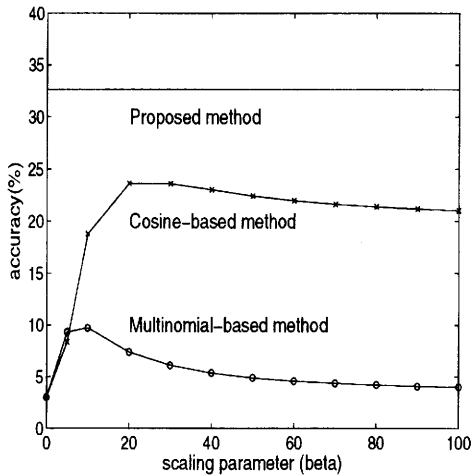


図 2: 全トピックを対象とした評価結果

ジングを改良すれば、ある程度の性能改善できると考える。

提案法ではモデル選択によりトピック群を限定するのに対し、比較法では全トピックを対象とするので、単純な比較は公平でない面もある。よって、被験者のラベリングから求まるトピック度ベクトルも含め、全中間ノードにおいて、トピック度を下位ノード群にリンク数で割って足し込む操作を再帰的に繰り返し、リーフノードのみから構成するトピック度ベクトルによる評価を行なった。

図 3 に評価結果を示す。提案法を含む全手法で正当率が向上した。トピック数が少なくなり、問題の難しさが若干緩和されたためと考える。また、図 2 と比べれば、提案法と比較法の差は若干縮まったものの、どの β を用いた比較法よりも、提案法は高い正当率を示した。実データを用いたこれらの実証実験より、提案法で採用するパラメータ混合モデルの考え方、およびモデル選択アプローチの有望性が示唆されたと考える。

4 議論と今後の課題

本報告では、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初期結

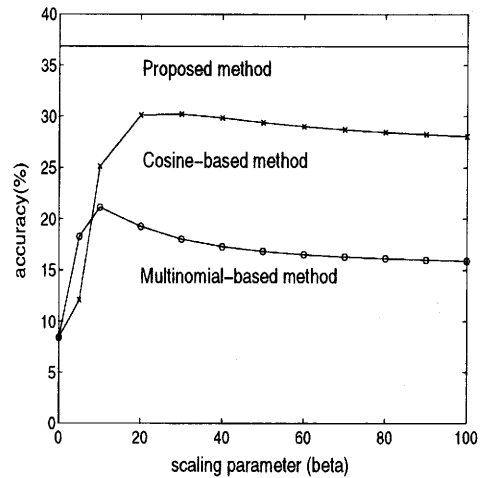


図 3: リーフトピックのみでの評価結果

果が得られたものの、これを有用なマッチング手法として発展させるには多くの課題も残されている。

本検討の次のステップとしては、モデル選択における評価尺度と探索法の高度化が挙げられる。評価尺度として、本提案法では単純な MDL を採用したが、文書分類体系間の対応付けに適したモデルの事前分布を探索し、ベイズ学習のフレームワークでの定式化が考えられる。一方、現時点の探索は欲張り (greedy) 法に基づくが、最良優先探索や simulated annealing などの適用も可能である。

本報告では、単一のトピックを別の分類体系へ対応付けすることに主眼を置いた。これに対し、それぞれの文書分類体系から、マッチング精度の高い 2 つのトピック群を自動抽出し、このようなトピックのペア集合として、簡潔な対応関係の記述が構築できれば、2 つの文書分類体系の違いは容易に理解できる。本提案法は、これを実現させるための重要な核技術であると考ええる。

一方で、本提案法の有効性を評価するには、幅広い実データへの適用が必要である。Web ページはもとより、電子メール、社内文書、学術論文などの文書分類体系間の構造マッチングでの検証も重要と考える。さらに今後は、文書だけでなく、頻度情報で記述されるサンプルを分類する体系間のマッチング手法としても発展させたい。

謝辞

Web データの収集, および実験評価システムの作成は, NTT コムウェア西日本株式会社の飯野斉, 坂口章の両氏により行なわれた. また, Web ページに対するラベリングは, 中俣麻佐美, 島村久美子の両氏により行なわれた.

参考文献

- [1] T. Milo and S. Zohar, “Using schema matching to simplify heterogeneous data translation,” Proc. VLDB’98, 1998.
- [2] N. Noy and M. Musen, “Algorithm and tool for automated ontology merging and alignment,” Proc. AAAI’00, pp. 450–455, 2000.
- [3] D. McGuinness, R. Fikes, J. Rice, and S. Wilder, “The chimaera ontology environment,” Proc. AAAI’00, 2000.
- [4] H. Chalupsky, “Ontomorph: a translation system for symbolic knowledge,” Proc. Principles of Knowledge Representation and Reasoning, 2000.
- [5] E. Rahm and P. Bernstein, “On matching schemas automatically,” VLDB journal, vol. 10, no. 4, 2001.
- [6] A. Doan, P. Domingos, and A. Halevy, “Reconciling schemas of disparate data sources: a machine learning approach,” Proc. ACM SIGMOD’01, 2001.
- [7] A. Doan, J. Madhavan, P. Domingos, and A. Halevy, “Learning to map between ontologies on the semantic web,” Proc. WWW’02, 2002.
- [8] R. Ichise, H. Takeda, and S. Honiden, “Rule induction for concept hierarchy alignment,” In Proc. IJCAI’01, 2001.
- [9] S. Sarawagi, S. Chakrabarti, and S. Godbole, “Cross-training: learning probabilistic mappings between topics,” Proc. ACM SIGKDD’03, 2003.
- [10] N. Ueda and K. Saito, “Single-shot detection of multiple topics using parametric mixture models,” Proc. ACM SIGKDD’02, pp. 626–631, 2002.
- [11] 上田修功, 斉藤和巳, “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.
- [12] 上田修功, 斉藤和巳, “類似テキスト検索のための多重トピックテキストモデル,” 情報処理学会論文誌, Vol. 44, No. SIG14(TOM9), pp.1–8, 2003.
- [13] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification (2nd Ed.),” John Wiley & Sons, 2000.
- [14] C. D. Manning, and H. Schütze, “Foundations of statistical natural language processing,” MIT press, 1999.
- [15] J. Rissanen, “Stochastic Complexity in Statistical Inquiry,” World Scientific, 1989.