2C1-NFC-2-2

ナイーブベイズによる文書分類のためのWikipediaカテゴリグラフ解析

Wikipedia Category Graph Analysis for Document Classification Using Naive Bayes

白川 真澄[*1] 中山 浩太郎[*2] 原 隆浩[*1] 西尾 章治郎[*1]
Masumi Shirakawa Kotaro Nakayama Takahiro Hara Shojiro Nishio

[*1]大阪大学 大学院情報科学研究科
Grad. Sch. of Information Science and Technology, Osaka Univ.

[*2]東京大学 知の構造化センター
The Center for Knowledge Structuring, The Univ. of Tokyo

This paper describes a method to estimate $P(c|t)$, the conditional probability of a class c given a term t, by analyzing the category structure of Wikipedia. Our method regards the category structure as a graph and formulates the way to estimate the random walk probability without supervised data. Estimated probability then can be used for document categorization by Naive Bayes. From the results of Web snippet categorization, we confirmed that our method had potential to be competitive to the supervised Naive Bayes method.

1. はじめに

文書分類とは，あらかじめ指定したカテゴリ（経済，スポーツ，エンターテイメントなど）に対し，入力となる文書がどのカテゴリに属するかを決定するタスクであり，従来は教師データを作成し，ナイーブベイズ (NB) やサポートベクタマシン (SVM) などの機械学習手法を用いることで達成されてきた．しかし，様々なカテゴリや粒度で Web 検索結果のスニペットや Twitter の投稿メッセージなどの短いテキストを分類したいという要求が高まっており，教師データを用いて高精度を達成するよりも，教師データなしでそれなりの精度を達成可能なテキスト分類手法が求められている．

本研究では大規模オンライン百科事典である Wikipedia に注目する．Wikipedia は固有名詞や専門用語，新語などを多数定義しており，それらの語句はカテゴリ構造によって整理・分類されているため，テキストを様々なカテゴリに分類するための知識リソースとして優れている．これまでに Wikipedia を用いたテキスト分類に関する研究 [Schönhofen 06, Phan 08] が行われているが，これらの研究の多くは Wikipedia のカテゴリを教師あり学習の素性として用いている．

本研究では，Wikipedia のカテゴリ構造をグラフとみなし，確率的な手法を用いて解析することにより，ナイーブベイズの教師データを自動で生成するための手法を提案する．すなわち，語句 t をカテゴリ c に確率的に分類し，条件付き確率 $P(c|t)$ を算出する．Wikipedia のカテゴリ構造では，ある記事から親カテゴリをランダムに辿ったとき，そのパス上でより確実に出現するカテゴリに対して，より強く所属していると考えられる．そこで，親カテゴリを辿る際に確率的にスコアを割り当て，より大きいスコアを持つカテゴリに強く所属するとみなす．これは，隣接ノードのいずれかに等確率で遷移するランダムウォークを用いて表現できる．提案手法では，あらかじめ複数のカテゴリ（基底カテゴリ）を指定し，ある語句 t から親カテゴリを辿ったとき基底カテゴリ c に到達する確率 $P(c|t)$ を，ランダムウォークにより算出する．従来の研究では Wikipedia を用いた教師なしのテキスト分類が課題であったが，本研究では語句のカテゴリへの所属を確率 $P(c|t)$ として表現することにより，ナイーブベイズの教師データとして利用できる．

連絡先: 白川 真澄，大阪大学 大学院情報科学研究科，〒 565-0871 大阪府吹田市山田丘 1-5，06-6879-4513，06-6879-4514，shirakawa.masumi@ist.osaka-u.ac.jp

2. ナイーブベイズによる文書分類

現時点において最も実用的な文書分類アルゴリズムの一つとして，ナイーブベイズ [Domingos 97] が挙げられる．ナイーブベイズでは，テキスト中に含まれる語句が互いに独立に発生したものであるというナイーブ（単純）な仮定を置き，それらの語句を含むテキストのカテゴリへの所属確率を，ベイズの定理に基づき算出する．ナイーブベイズはシンプルでありながら学習や推論において高速に動作し，精度も高いため，実用的な文書分類手法として一般に認識されている．

語義の曖昧性解消タスク [Shirakawa 11] で用いられている拡張ナイーブベイズは，入力系列が確率的に予測可能な場合に適用できる手法であり，自然文の入力に対して有効である [白川 12]．通常のナイーブベイズと比べて，入力テキストに含まれる特徴語の影響をより大きく反映できるため，本研究では拡張ナイーブベイズを用いてテキストの分類を行うことを考える[*1]．すなわち，入力テキストが与えられたとき，そこからキーフレーズ集合 T を確率的に予測し，あらかじめ設定したカテゴリ（基底カテゴリ）の一つ c への所属確率 $P(c|T)$ を算出する．基底カテゴリ c ごとに $P(c|T)$ を比較することにより，入力テキストが最も所属しそうなカテゴリ c を決定する．具体的には，以下の式を用いる．

$$P(c|T) \propto \frac{\prod_{k=1}^{K}\Big(P(t_k \in T)P(c|t_k) + \big(1 - P(t_k \in T)\big)P(c)\Big)}{P(c)^{K-1}} \tag{1}$$

K は入力テキストに含まれるキーフレーズ候補の数，$P(t_k \in T)$ は語句 t_k がキーフレーズ集合 T に含まれる確率，$P(c|t_k)$ は語句 t_k が与えられたときにそれが基底カテゴリ c に属する確率，$P(c)$ は基底カテゴリ c の事前確率である．また，E を Wikipedia で定義されているエンティティ（記事）集合とすると，$P(c|t_k) = \sum_{e_i \in E} P(c|e_i)P(e_i|t_k)$ である．$P(e_i|t_k)$ および $P(t_k \in T)$ については文献 [白川 12] と同様，Wikipedia の情報を用いてそれぞれ以下の式により算出する．

$$P(t_k \in T) \approx \frac{CountDocuments(t_k \in Key)}{CountDocuments(t_k)} \tag{2}$$

[*1] なお，通常のナイーブベイズを用いても問題なく動作する．

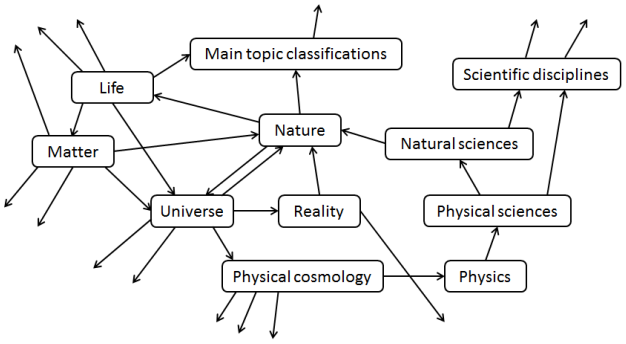

図 1: Wikipedia のカテゴリ構造の例

$$P(e|t_k) \approx \frac{CountAnchortexts(t_k, e)}{\sum_{e_i \in E} CountAnchortexts(t_k, e_i)} \quad (3)$$

$CountDocuments(t_k)$ は語句 t_k が出現する記事数，$CountDocuments(t_k \in Key)$ は語句 t_k がアンカーテキストとして出現する記事数，$CountAnchortexts(t_k, e)$ は語句 t がアンカーテキストとしてエンティティ e の記事にリンクされている回数である．なお，式 (2) は Mihalcea らの研究 [Mihalcea 07] の Keyphraseness，式 (3) は Milne らの研究 [Milne 08] の Commonness である．

ここで，$P(c|e_i)$ は Wikipedia のカテゴリ構造を解析することで算出する（第 3. 章）．また，$P(c)$ は基底カテゴリ c の一般度を表すものであることから，どの程度所属されやすいか，を算出することでおよその値が得られる．具体的には，以下の式により算出する．

$$P(c) \approx \sum_{e_i \in E} P(c|e_i) \quad (4)$$

これらの情報と式 (1) を用いることで，指定した基底カテゴリに対するテキストの分類が可能となる．次章では，Wikipedia のカテゴリグラフを解析し，$P(c|e_i)$ を算出するための手法について説明する．

3. Wikipedia からの教師データ生成手法

本研究では，Wikipedia のカテゴリ構造をグラフ理論に基づいて解析することにより，語句を確率的にカテゴリに分類し，文書分類の教師データとして利用する．以下ではまず，Wikipedia のカテゴリ構造について説明し，その後，提案手法のアプローチついて詳述する．

3.1 Wikipedia のカテゴリ構造

Wikipedia では，基本的に各記事（エンティティ）に対して一つ以上のカテゴリ（親カテゴリ）が割り当てられている．また，カテゴリにも同様に親カテゴリが割り当てられており，カテゴリ構造を成している．この親カテゴリは，該当の記事あるいはカテゴリが所属すると思われるカテゴリであり，上位下位関係や全体部分関係を表すこともあれば，トピックや関連を表す場合もある．そのため，ある記事から親カテゴリを辿っていくと，ほとんど関係のないカテゴリに到達することが頻繁に起こりうる．例えば，動物の “Lion” についての記事から親カテゴリを辿っていくと，“Lions,” “Panthera,” “Pantherinae,” “Felids,” “Cats,” “Domesticated animals,” “Agriculture” とあまり関係のないカテゴリに到達できる．さらに親カテゴリを辿れば “Humans,” “Economics,” “Education” などのカテゴリにも到達可能である．このようなカテゴリに対する緩い制約により，Wikipedia のカテゴリ構造は図 1 のような複数の親やループを許容するネットワーク構造となっている．なお，図 1 は全てカテゴリであり，Wikipedia の各記事はこのようなカテゴリ構造において一つ以上のカテゴリに属している．このようなカテゴリ構造のため，あるエンティティがどのカテゴリに属しているかという情報を，単純に親カテゴリや子カテゴリを辿るだけでは抽出できない．このことは，Wikipedia のカテゴリ構造を直接テキスト分類に利用することが難しい要因になっていると考えられる．

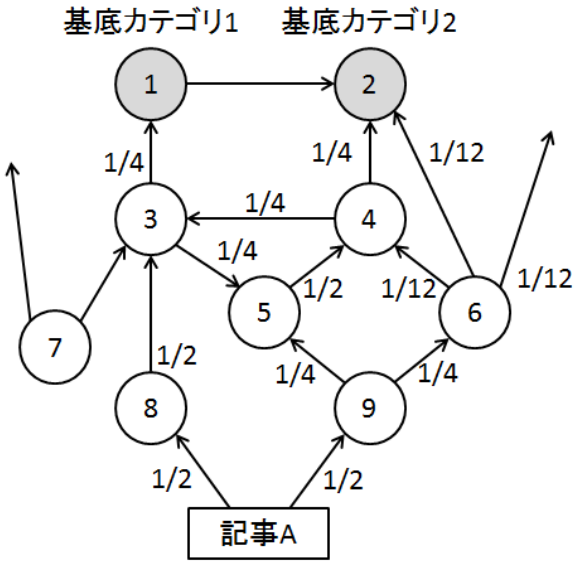

図 2: ランダムウォークによる記事 A の基底カテゴリ 1, 2 への所属確率計算の例

3.2 ランダムウォークによる語句の確率的分類

前節で述べたように，Wikipedia のカテゴリ構造はネットワーク構造であるため，ある記事に対し，指定したカテゴリ（基底カテゴリ）に属するか否かを判断することが困難である．そこで本研究では，ナイーブベイズを用いたテキスト分類において，式 (1) に示すように，語句 t がカテゴリ c に属する確率 $P(c|t)$ が計算できればよいということに着目する．すなわち，ある記事について，カテゴリに属するか否かではなく，どの程度の確率で属するかという数値として表現し，その情報をナイーブベイズによるテキスト分類に利用する．

Wikipedia のカテゴリ構造では，ある記事から親カテゴリを辿るとき，そのパス上で出現しやすいカテゴリに対してより強く所属していると考えられる．この考え方に基づき，親カテゴリを辿るときに確率的にスコアを割り当て，より大きいスコアを持つカテゴリに強く所属するとみなす．これは，隣接ノードのいずれかに等確率で遷移するモデルであるランダムウォークを用いて表現できる．提案手法では，カテゴリをノード，親カテゴリへのリンクを有向リンクとしたグラフに対して，ランダムウォークを適用する．十分な時間が経過した後のランダムウォークによるスコアは，あるノードから出発したときに，そのカテゴリに到達する確率を表す．具体的には，記事 e_i から親カテゴリを辿ったときにカテゴリ c に到達する確率を所属確率 $P(c|e_i)$ として用いる．

提案手法のアプローチについて，例を図 2 に示す．図 2 では，記事 A から確率優先探索（確率が同じ場合はノード番号順）により各カテゴリへの所属確率を算出している．まず，記

表 1: Web データセット

基底カテゴリ	トレーニングセット		テストセット	
	#Phrs.	#Snip.	#Phrs.	#Snip.
Business	60	1,200	10	300
Computers	60	1,200	10	300
Culture-Arts-Ent.	94	1,880	11	330
Education-Science	118	2,360	10	300
Engineering	11	220	5	150
Health	44	880	10	300
Politics-Society	60	1,200	10	300
Sports	56	1,120	10	300
合計		10,060		2,280

事 A は親カテゴリを二つ持っているため，それぞれのカテゴリ（カテゴリ 8, 9）への所属確率をそれぞれ $\frac{1}{2}$ とする．カテゴリ 8 は親カテゴリを一つしか持たないため，カテゴリ 3 への所属確率をそのまま $\frac{1}{2}$，また，カテゴリ 9 は親カテゴリを二つ持っているため，カテゴリ 5, 6 への所属確率をそれぞれ $\frac{1}{4}$ とする．このような処理を繰り返すことにより，基底カテゴリへの所属確率を算出する．結果的に，記事 A は基底カテゴリ 1 に $\frac{1}{4}$，基底カテゴリ 2 に $\frac{1}{3}$ の確率で所属していることが分かる．なお，ここでは基底カテゴリに到達するか，ループの発生を検知した場合，親カテゴリの探索を中止しているが，ループに対して同様の処理を繰り返すことで，より厳密な所属確率を算出できる．評価実験では，所属確率の変動が十分小さくなるまで計算したときの値を使用している．

4. 評価

4.1 評価環境

提案手法の有効性を客観的に評価するため，テキスト分類において評価を行った．データセットとして，Phan らの研究 [Phan 08] で用いられている Web 検索結果のスニペット（Web データセット）を用いた．各データセットの統計データを表 1 に示す（#Snip. はスニペットの数，#Phrs はスニペットを取得するための検索クエリの数）．Web データセットは，各カテゴリに対して排他的になるよう選択された検索クエリによってそれぞれ 20 件または 30 件の検索結果のスニペットを取得したものである．基底カテゴリは Wikipedia の中から 13 カテゴリ “Business,” “Economics,” “Computing,” “Culture,” “Arts,” “Entertainment,” “Education,” “Science,” “Engineering,” “Health,” “Politics,” “Society,” “Sports” を選択した*2．

比較手法として，語句の分類を親カテゴリを辿る際のホップ数に応じて決定する手法，WordNet を用いた手法，教師ありナイーブベイズ（NB）手法とした．ホップ数ベースの手法では，N ホップ ($N = 1, \cdots, 6$) までの祖先カテゴリに所属するとみなし，語句の重要度を表す Keyphraseness[Mihalcea 07] の重み付き和としてテキストの分類を行った．また，一般語を定義した辞書である WordNet を用いた手法では，祖先カテゴリに全て所属するとみなし，最も出現回数の多いカテゴリに分類した．これは，WordNet では DAG 構造により正確に上位下位関係が定義されており，単純に親カテゴリを辿る手法がうまく機能するためである．教師ありナイーブベイズでは，トレーニングセットを教師データとして利用し，使用するスニペット数を変化させた．これらの手法では，テキスト入力に対して基底カテゴリの順位付きリストを出力として返すため，評価指標として最上位の適合率に加え，正解のカテゴリの順位の逆数の平均（MRR）を用いた．MRR は，順位付けのタスクの評価指標としてよく用いられ，正解のカテゴリが上位であればあるほど高いスコアが与えられる．

*2 Wikipedia のカテゴリにおいて “Business” は主に「企業」という意味で用いられているため，「経済」の意味を包含する目的で “Economics” も “Business” の基底カテゴリとして選択した．

4.2 評価結果

評価結果を表 2 に示す．表ではそれぞれの基底カテゴリごとの評価指標と全体の評価指標を計算している．親カテゴリを辿る際のホップ数で所属を決定する方法と比較して，所属を確率として表す提案手法のほうが全体的に安定して高い精度で分類できている．ホップ数ベースの手法では，ある一つのホップ数では全ての基底カテゴリに対して高い精度を達成するのが難しいことが分かる．また，ホップ数が大きくなると，ほとんどの入力に対して少数の支配的なカテゴリ（“Culture” や “Society”）のスコアが高くなることが問題となっている．一方，提案手法では，確率的に語句を分類することにより，ナイーブベイズの教師データ $P(c|t)$ としてシームレスに組み込むことが可能となり，精度向上や精度安定につながっていると考えられる．

WordNet を用いた手法についてみると，WordNet はスニペットの分類に対してあまり効果的でないことが分かる．これは，WordNet では固有名詞，専門用語，新語をあまり定義していないことや，親カテゴリが基本的に上位下位関係を表すものであることに由来する．実際，多くのスニペットに対して，トピックの分類に利用できる語句が WordNet に全く存在していなかった．WordNet は，語句間の上位下位関係により，推論を用いた様々なアプリケーションに適用できるが，実データ（特にテキストが短い場合）に対してトピックによる分類を行うには情報量が少ないと考えられる．

提案手法と教師ありのナイーブベイズによるテキスト分類手法を比較すると，教師データを 1,000 件程度用いた場合と同等の適合率および MRR となっている．提案手法では教師データを用いていないことから，Wikipedia の記事の分類を確率的に行う手法が，テキスト分類に対する正解データとして有効であるといえる．この結果から，教師データが十分に用意できない場合，あるいはそこまで高い精度が要求されない場合においては，提案手法を用いたテキスト分類が効果的であることが分かる．例えば，Web 検索のスニペットをいくつかのカテゴリに分類することで，検索結果を見やすく表示するようなアプリケーションに対して有効であると考えられる．

5. おわりに

本研究では，Wikipedia のカテゴリ構造をグラフとみなして解析し，確率的に分類した語句を用いてテキストを分類する手法を提案した．具体的には，Wikipedia のある記事から親カテゴリを辿るときにランダムウォークを適用することにより，語句 t のカテゴリ c への所属確率 $P(c|t)$ を算出し，得られた確率 $P(c|t)$ を用いてナイーブベイズによる文書分類を行った．評価実験により，教師データを必要としない提案手法が，教師ありのナイーブベイズと比較して十分高い精度を達成できることを確認した．今後の課題として，分類したいカテゴリと Wikipedia から選択するカテゴリについて意味の相違が発生しないよう，ユーザが正しく基底カテゴリを選択できるような仕組みを検討する予定である．

表 2: 適合率（上）と MRR（下）

基底カテゴリ	Bus.	Comp.	Cult.	Sci.	Eng.	Heal.	Pol.	Spo.	All
教師ありナイーブベイズ									
全部 (10,060)	0.787	0.837	0.879	0.853	0.773	0.830	0.700	0.883	0.821
1/2 (5,030)	0.727	0.760	0.836	0.760	0.640	0.780	0.660	0.857	0.761
1/5 (2,012)	0.720	0.777	0.785	0.760	0.700	0.780	0.563	0.817	0.741
1/10 (1,006)	0.623	0.653	0.752	0.743	0.620	0.697	0.520	0.793	0.680
1/20 (503)	0.600	0.657	0.621	0.740	0.593	0.627	0.330	0.730	0.614
1/50 (201)	0.537	0.427	0.482	0.583	0.027	0.563	0.327	0.623	0.474
1/100 (100)	0.527	0.370	0.376	0.663	0.033	0.580	0.160	0.447	0.418
WordNet	0.417	0.240	0.200	0.217	0.033	0.100	0.027	0.553	0.236
Wikipedia									
1 ホップ	0.363	0.273	0.358	0.183	0.080	0.193	0.283	0.177	0.263
2 ホップ	0.627	0.457	0.545	0.350	0.047	0.197	0.507	0.580	0.412
3 ホップ	0.703	0.723	0.685	0.520	0.120	0.500	0.610	0.667	0.594
4 ホップ	0.563	0.727	0.791	0.437	0.047	0.267	0.797	0.613	0.522
5 ホップ	0.303	0.610	0.879	0.410	0.013	0.097	0.873	0.337	0.436
6 ホップ	0.140	0.427	0.842	0.397	0.000	0.013	0.950	0.173	0.349
提案手法（確率的な語句の分類）	0.737	0.837	0.658	0.630	0.547	0.713	0.513	0.797	**0.687**

基底カテゴリ	Bus.	Comp.	Cult.	Sci.	Eng.	Heal.	Pol.	Spo.	All
教師ありナイーブベイズ									
全部 (10,060)	0.867	0.909	0.926	0.915	0.861	0.895	0.825	0.925	0.893
1/2 (5,030)	0.827	0.858	0.895	0.849	0.788	0.863	0.792	0.906	0.852
1/5 (2,012)	0.825	0.868	0.858	0.853	0.815	0.849	0.718	0.874	0.834
1/10 (1,006)	0.743	0.787	0.839	0.837	0.747	0.790	0.675	0.857	0.788
1/20 (503)	0.726	0.786	0.746	0.841	0.727	0.739	0.519	0.810	0.738
1/50 (201)	0.673	0.590	0.657	0.734	0.295	0.694	0.537	0.745	0.637
1/100 (100)	0.662	0.551	0.578	0.791	0.280	0.699	0.365	0.619	0.587
WordNet	0.452	0.263	0.221	0.268	0.037	0.108	0.047	0.608	0.264
Wikipedia									
1 ホップ	0.418	0.300	0.391	0.201	0.090	0.207	0.301	0.180	0.274
2 ホップ	0.715	0.541	0.577	0.438	0.081	0.288	0.585	0.629	0.509
3 ホップ	0.812	0.817	0.777	0.660	0.257	0.637	0.760	0.729	0.710
4 ホップ	0.729	0.813	0.889	0.644	0.237	0.492	0.888	0.743	0.711
5 ホップ	0.573	0.731	0.936	0.638	0.221	0.398	0.935	0.572	0.656
6 ホップ	0.473	0.586	0.917	0.634	0.211	0.320	0.974	0.429	0.596
提案手法（確率的な語句の分類）	0.827	0.889	0.785	0.782	0.698	0.779	0.722	0.862	**0.799**

参考文献

[Domingos 97] Domingos, P. and Pazzani, M.: On the Optimality of the Simple Bayesian Classifier under Zero-One Loss, *Machine Learning*, Vol. 29, No. 2-3, pp. 103–130 (1997)

[Mihalcea 07] Mihalcea, R. and Csomai, A.: Wikify! Linking Documents to Encyclopedic Knowledge, in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pp. 233–241 (2007)

[Milne 08] Milne, D. and Witten, I. H.: Learning to Link with Wikipedia, in *Proceedings of ACM Conference on Information and Knowledge Management (CIKM)*, pp. 509–518 (2008)

[Phan 08] Phan, X.-H., Nguyen, L.-M., and Horiguchi, S.: Learning to Classify Short and Sparse Text & Web with Hidden Topics from Large-scale Data Collections, in *Proceedings of International World Wide Web Conference (WWW)*, pp. 91–100 (2008)

[Schönhofen 06] Schönhofen, P.: Identifying Document Topics Using the Wikipedia Category Network, in *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pp. 456–462 (2006)

[Shirakawa 11] Shirakawa, M., Wang, H., Song, Y., Wang, Z., Nakayama, K., Hara, T., and Nishio, S.: Entity Disambiguation based on a Probabilistic Taxonomy, Technical Report MSR-TR-2011-125, Microsoft Research (2011)

[白川 12] 白川 真澄, 中山 浩太郎, 原 隆浩, 西尾 章治郎 : Wikipedia とナイーブベイズを用いた自然文に対する関連語句取得手法, 第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012) (2012)