Wikipediaのカテゴリ構造解析とクラスタリングによる概念ベクトルの生成

Concept Vectorization using Wikipedia Category Structure Analysis and Clustering

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

[*1]大阪大学 Osaka University　[*2]東京大学 Tokyo University

Existing works on automatic dictionary building have accuracy problems due to the technical limitation of statistical NLP (Natural Language Processing) and noise data on the WWW. To solve these problems, we have proposed concept vectorization methods by analyzing the category network structured in Wikipedia. In this work, to achieve higher accuracy and versatility, we propose a clustering-based method for selecting bases in the concept vector.

1. はじめに

国語学や言語学，分類学等の研究領域において，概念（語彙）を一つの大系として整理することは有益であり，多くの国語辞典や対訳辞書，百科事典などの辞書が人の手によって製作されてきた．分類辞書（タクソノミ）は，概念がどのようなカテゴリに所属しているかという情報を，木構造や DAG(Directed Acyclic Graph) 構造などで表現した辞書である．高い網羅性と精度を併せ持つ分類辞書の構築は，現在の Web が概念の意味を処理できる次世代 Web へと進歩するための重要な基盤技術として必要とされている．しかし，分類辞書の自動構築には，自然言語処理の技術的限界やノイズデータに起因する精度低下の問題がある．これらの問題を解決するため，筆者らは大規模な Web 百科事典である Wikipedia の，知識コーパスとしての有用性に注目し，Wikipedia のカテゴリ構造解析による概念のベクトル化手法 [Shirakawa 09] を提案してきた．その結果，概念とカテゴリの多対多の所属関係を概念ベクトルによって精度よく表現できることを確認したが，一方で，概念ベクトルを実際のアプリケーションに適用する場合に，概念ベクトルの基底をどのように選択するかという問題に直面した．概念ベクトルを適用するアプリケーションによって，適切な基底やその数が異なってくる．そこで本研究では，クラスタリング技術を応用し，概念ベクトルの基底を選択する手法を提案する．

2. 概念ベクトルの生成手法

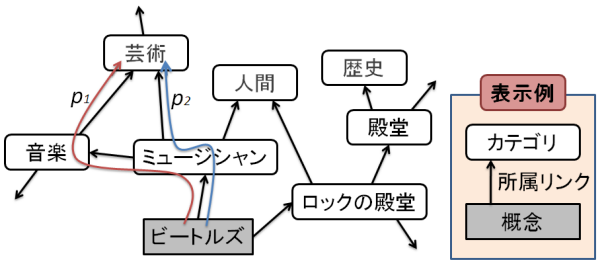
筆者らはこれまで，Wikipedia のカテゴリ構造から概念とカテゴリの所属関係を解析し，概念をベクトル化する手法 [Shirakawa 09] を提案してきた．ここでは，概念ベクトルの定義と，基本的な概念ベクトルの生成手法である BVG 法について説明する．

2.1 概念ベクトル

概念ベクトルとは，概念がどのようなカテゴリに所属しているかという情報を，所属しているか否かではなく，どの程度所属しているかという度合（所属の強さ）で表現し，ベクトル化したものである．例えば図 1 に示すように，概念「ジャズ」はカテゴリ「芸術」に高い値を持つ概念ベクトル，概念「ビートルズ」はカテゴリ「芸術」とカテゴリ「人間」に高い値を持つ概念ベクトルとして表現できる．このように概念をベクトル化することで，概念とカテゴリの多対多の所属関係を，所属の強さを持った形で表現できる．

	芸術	歴史	地理	思想	人間	自然	·····
ジャズ	0.375	0.125	0	0	0	0	·····
ビートルズ	0.375	0.125	0	0	0.5	0	·····

図 1: 「ジャズ」と「ビートルズ」の概念ベクトル



$$I(\text{ビートルズ}, \text{芸術}) = \frac{1}{d(t_1)} + \frac{1}{d(t_2)} = \frac{1}{2^3} + \frac{1}{2^2} = 0.375$$

図 2: Basic Vector Generation（BVG）法の適用例

2.2 Basic Vector Generation（BVG）法

BVG 法は，Wikipedia のカテゴリ構造に適用できる基本的な概念のベクトル化手法である．Wikipedia は，概念が一つまたは複数のカテゴリに所属し，また，カテゴリ同士も所属関係を持った形でリンクしているネットワーク構造を構成している．これらの所属関係は所属リンクによって表現されており，Wikipedia のカテゴリ構造は，概念をノード集合 W，カテゴリをノード集合 V，所属リンクをエッジ集合 E とする有向グラフ $G = \{W, V, E\}$ で表現できる（図 2）．このとき，概念 w_i のカテゴリ v_j への所属の強さを計測する問題を考えた場合，所属の強さは，w_i から v_j へのパスの多さと，w_i から v_j への各パスの短さに影響を受けると考えられる．つまり，w_i から v_j へのパスが多ければ多いほど，またそのパスの長さが短ければ短いほど，w_i は v_j に強く所属する．ここで，パスとは所属リンクを伝って w_i から v_j へと移動可能な経路を示す．そこで，w_i から v_j への全パス $P = \{p_1, p_2, ..., p_n\}$ が与えら

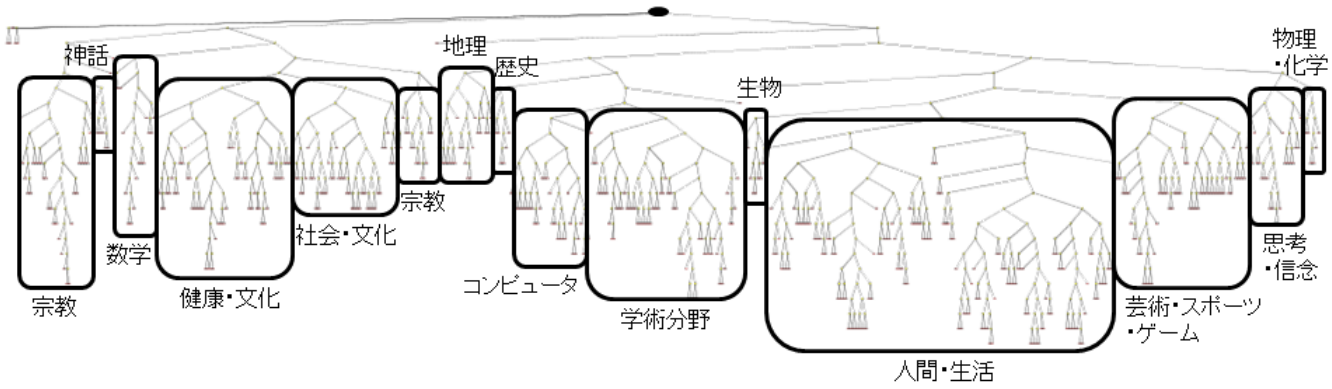


図 3: 階層クラスタリングを用いた概念ベクトルの基底の選択（類似度の閾値 0.3）

れたとき，w_i の v_j への所属の強さ $I(w_i, v_j)$ を以下の式により表現する．

$$I(w_i, v_j) = \sum_{l=1}^{n} \frac{1}{d(t_l)} \quad (1)$$

d はパス p_l のホップ数 t_l に応じて増加する関数であり，指数関数などの単調増加関数が利用できる．

図 2 では，概念「ビートルズ」のカテゴリ「芸術」への所属の強さを計測している（d は指数関数 2^{t_l}）．t_1 が 3，t_2 が 2 であるため，所属の強さは 0.375 となる．

3. 概念ベクトルの基底の選択手法

概念のベクトル化手法は，前もって何らかの方法で基底を選択する必要がある．本研究では，概念ベクトルの基底を選択する手法として，クラスタリング技術を応用した手法を提案する．

提案手法ではあらかじめデータセットとして，適当な数のカテゴリ群を用いる．また，それらのカテゴリの，Wikipedia のカテゴリ構造における数ホップ以内の祖先カテゴリの情報も使用する．データセットの各カテゴリについて，祖先カテゴリを基底として祖先カテゴリの出現回数をベクトル化し，このベクトルの類似度を基に階層クラスタリングを行う．クラスタリングにより，類似度の高いカテゴリ（または併合されたクラスタ）同士から順に併合していき，徐々に大きなクラスタを形成していく．その結果，全カテゴリが併合されたクラスタをルートノード，部分的に併合されたクラスタを類似度付きの中間ノード，各カテゴリを葉ノードとするツリーが形成される．ここで，類似度に関する閾値を設定することで，概念ベクトルの基底となるクラスタ（歴史クラスタ，宗教クラスタなど）を決定していく．

4. 出力例と考察

データセットとして，Wikipedia の Categorical index [*1] に記載されているカテゴリを使用し，Wikipedia のカテゴリ構造における 3 ホップ以内の祖先カテゴリをベクトル化して HCM (Hard c-Means) クラスタリングを適用した．類似度の閾値を 0.3 に設定してクラスタを形成した結果を図 3 に示す．図 3 では，人間・生活のクラスタ内に多くのカテゴリがまとめられている一方，歴史のクラスタ内には少数のカテゴリしか存在していない．これはデータセットに，人間・生活に所属するカテゴリが多く存在していたことや，歴史に所属するカテゴリが少なかったことが原因である．つまり，提案手法では，データセットのカテゴリに偏りがある場合でも，類似度の大きさによってクラスタの専門性の度合いが統一できていることを示している．

また，データセットの偏りがアプリケーションにとって重要な意味を持つ場合，より多くのカテゴリを擁するクラスタがこのアプリケーションにとって重要な属性であることを示しており，概念ベクトルの基底を選択する際の基準となる．例えば，ニュース記事の分類を想定すると，ニュース記事を分類するのに適した，経済やエンターテインメントに関するカテゴリがデータセットに多く含まれ，クラスタリングの結果，それらのクラスタが大きく形成される．大きく形成されたクラスタを概念ベクトルの基底に設定することで，結果的にニュース記事の分類に適した概念ベクトルが生成できるものと考えられる．

5. まとめと今後の課題

本稿では，Wikipedia のカテゴリ構造解析によって生成された概念ベクトルをアプリケーションに適用する場合において，概念ベクトルの基底を，クラスタリング技術を用いて選択する手法を提案した．出力例から，数ホップ以内の祖先カテゴリとその出現数をベクトル化してクラスタリングを適用する手法が，概念ベクトルの基底の選択に有効であることを確認した．今後の課題として，具体的な評価指標や，実アプリケーションに適用することによって，提案手法の有効性を客観的に評価することが挙げられる．

謝辞 本研究の一部は，科学研究費補助金基盤研究 C(20500093)，およびマイクロソフト産学連携研究機構 CORE 連携研究プロジェクトの助成によるものである．ここに記して謝意を表す．

参考文献

[Shirakawa 09] Shirakawa, M., Nakayama, K., Hara, T., and Nishio, S.: Concept Vector Extraction from Wikipedia Category Network, in *Proc. of International Conference on Ubiquitous Information Management and Communication (ICUIMC)*, pp. 71–79 (2009)

∗1 http://en.wikipedia.org/wiki/Portal:Contents/Categorical_index