3A2-3

分散表現を利用したタグ集合の階層化

Extraction of Hierarchical Structures on Tag Sets using Distributed Representation

鈴木宏明　　尾崎知伸
Hiroaki Suzuki　Tomonobu Ozaki

日本大学文理学部
College of Humanities and Sciences Nihon University

In recent years, with the development of SNS, user generated contents posted on the Web are increasing rapidly. While tags are attached to the posts to represent contents breifly, no detailed relationships among posts are provided by tags in general. In this paper, we propose methods to obtain hierarchical structures among sets of tags using distributed representation in order to discover the relationships among posts. The experimental results using Nicovideo datasets show the effectiveness of the proposed methods.

1. はじめに

近年，SNS の台頭などにより Web 上へ投稿されるコンテンツが増加している．これらのコンテンツには，その内容を表すタグが付与されることも多いが，一般的にコンテンツ間の詳細な関係性が与えられることは少ない．本研究では，タグが付与されたコンテンツを構造化することを目的に，タグ集合間の概念階層を獲得することを目的とする．具体的には，分散表現を利用した語の is-a 関係モデル化手法 [市瀬 15] を拡張し，与えられたタグ集合からその上位概念に相当するタグ集合を獲得する手法を提案する．

2. 既存研究

これまでに，タグの階層化に関していくつかの研究が報告されている．

文献 [村上 10] では，タグを対象とした階層構造構築手法として，出現頻度と共起確率を用いて単語間の上位下位関係を特定する ISR(Intersection Ratio) 法が提案されている．一般的なソーシャルブックマークサービスを対象に良好な結果が得られているが，付与できるタグ数等に制限のあるニコニコ動画に対しては，同義語タグの共起が淘汰される傾向があることから，必ずしも ISR 法が適しているとは言えず，別の観点からの手法が必要であると述べている．

文献 [Heymann 06] では，木構造としての構造化手法が提案されている．しかし抽出する構造を木構造に限定することは，計算等の簡略化が達成される一方で，各タグに対する上位概念が一つに限定されてしまうため不自然な個所が発生することになる．この問題を解決するため，文献 [高橋 13] では，閾値以上の関連性を持つタグすべてを上位概念として配置することを提案している．

文献 [市瀬 15] では，語の分散表現と既知の is-a 関係を用い，下位概念を入力として上位概念を予測する手法を提案している．is-a 関係が成り立つ 2 つの語の間の関係を，各語のベクトルの差でモデル化するとともに，複数の差ベクトルの平均を用いることで，ある単語に対してその上位概念に相当するベクトルを算出する．実験では，算出されたベクトルの最近傍の単語としては適切に is-a 関係を取り出すことが困難であったとしながらも，近傍上位 100 単語までを考慮すると，25%弱の単語に対して適切に上位語が獲得できたとしている．

連絡先: 尾崎 知伸，日本大学 文理学部 情報科学科，〒 156-8550
東京都世田谷区桜上水 3-25-40，tozaki@chs.nihon-u.ac.jp

3. 提案手法

3.1 手法の概要

本研究では，文献 [市瀬 15] で提案された手法を基に，複数の上位語ベクトル算出手法を提案するとともに，対象を語から語の集合へと拡張することで，タグ集合の階層化を試みる．具体的には，word2vec[Mikolov13] や Doc2vec[Le14] などの分散表現技術をコーパスに適用し，タグやタグ集合のベクトル表現を獲得するとともに，タグ集合に含まれる語の is-a 関係から，上位下位関係に相当するタグ集合の対を抽出する．これらの対からそれぞれ差ベクトル（上位ベクトル−下位ベクトル）を計算し，下位概念に対応するベクトルを入力とし上位概念に相当するベクトルを出力するモデルを構築する．

3.2 タグ集合に対する上位下位関係の抽出

提案手法では，一定数のタグ集合（語の集合）に対し，上位下位関係を与える必要がある．人手によるデータの作成は時間的な制約から困難であるため，今回は，タグ集合を構成するタグの is-a 関係を用いて作成することとした．具体的には，タグ w_1 と w_2 の間に is-a 関係が成り立つとき，2 つのタグ集合 S_1 と S_2 の間に $w_1 \in S_1 \wedge w_2 \notin S_1$ かつ $w_1 \notin S_2 \wedge w_2 \in S_2$ が成り立つ，すなわち S_1 は w_1 を含むが w_2 を含まず，逆に S_2 は w_1 を含まず w_2 を含むとき，S_1 と S_2 の間に is-a 関係が成り立つと設定した（図 1 参照）．

3.3 差ベクトルを利用した上位概念ベクトルの獲得

本研究では，差ベクトルを利用して上位概念を獲得する手法を 3 種提案する．

近傍平均法

一つ目の提案手法は，タグ集合ベクトルに対し，その K 近傍タグ集合の差ベクトルの平均値を関係ベクトル R として求め，それを対象となるタグ集合ベクトルに加えることで，上位概念に相当するベクトルを獲得するものである．ここで k はパラメタである．手法の概要を図 2 に示す．この手法は，分散表現では意味的に近い語は近いベクトルを持つことから，関係ベクトルも似たようなものになるのではないかという考えによるものである．

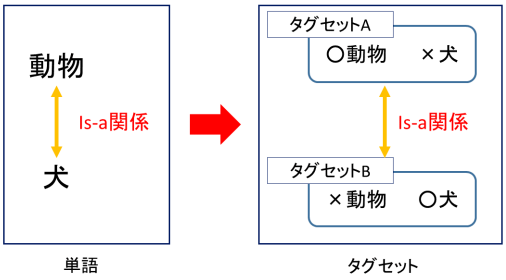

図 1: タグ集合における上位下位関係の例

表 1: 動画メタデータ

動画総数	8,305,696
タグ「アニメ」を持つ動画	322,090
タグ種類数	20,004
日本語 WordNet 登録タグ数	1,261(6.3%)
日本語 WordNet 未登録タグ数	18,743(93.7%)

セントロイド法

二つ目の提案手法は，クラスタリング手法を適用することで各タグ集合対の差ベクトルからセントロイドを抽出し，それらを用いて関係ベクトル R を獲得するものである．タグ集合対の差ベクトルには様々な向きがあり，例えば正反対の向きのベクトルの平均を取った場合は 0 ベクトルとなってしまう．この問題を回避し，似たような向きの差ベクトルをまとめることで，予測精度の向上が期待できる．手法の概要を図 3 に示す．具体的な上位集合の予測には，近傍集合が最も多く所属するクラスタのセントロイドを使用する．またクラスタリングには X-means 法を採用している．

関数学習法

三つ目の提案手法は，下位タグ集合ベクトルを入力，上位タグ集合ベクトルを出力とし，フィードフォワードニューラルネットワークを用いて関係ベクトル R を学習する方法である．手法の概要を図 4 に示す．非線形な学習を行うことで，予測精度向上を目指す．

4. 評価実験

4.1 実験用データセット

実験には，NII とドワンゴが提供するニコニコデータセット [*1] の動画メタデータと大百科記事本文データ，NICT の提供する日本語 WordNet [*2] を用いた．約 830 万件の動画データから，タグ「アニメ」を含め 2 つ以上のタグを持つ動画約 32 万件を対象とした．また，タグの is-a 関係特定には日本語 WordNet の Hype と Hypo を使用し，タグセットに含まれる単語を対象に 96,812 組を得た．加えて，タグ間の上位下位関係を利用し，1,005,395 対のタグ集合対を獲得した．表 1 に対象データに関する情報を示す．

4.2 タグに関する上位概念獲得実験

提案手法の有効性を評価するために，is-a 関係によるタグ対とタグのベクトルを用いて既存手法との比較実験を行った．

*1 http://www.nii.ac.jp/dsc/idr/nico/nico.html
*2 http://compling.hss.ntu.edu.sg/wnja/

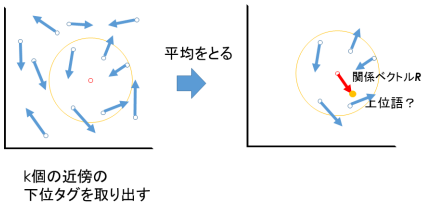

図 2: 近傍平均法の概要：$k = 5$ の例

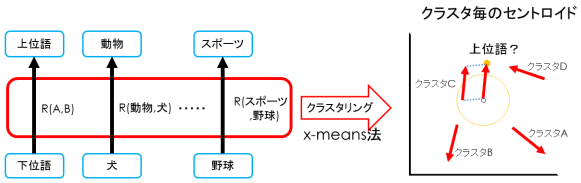

図 3: 近傍セントロイド法の概要

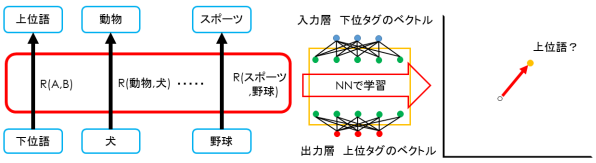

図 4: 関数学習法の概要

[市瀬 15] による手法をベースラインとし，三つの提案手法で上位概念に相当するベクトルを求め，その近傍を上位タグとして予測する．なお，大百科記事本文データに word2vec（次元数 300）を適用することでタグのベクトル化を行った．実験結果を表 2 に示す．

ベースライン，近傍平均法，セントロイド法では，データを半分に分け 10-fold 交差検証法を用いて正しい上位タグが抽出できるかを調査した．なお，上位タグに相当するベクトルの 100 近傍内に正しい上位タグが含まれている場合を正解としている．関数学習法では，データ全体を訓練例集合 7，テスト例集合 3 の割合で分けた上で，1 回の学習ごとに訓練例の 10%を検証データとして学習を行い，テスト例集合を用いて学習したモデルの評価を行った．

実験の結果を表 2 に示す．表中において，平均順位とは，上位タグに相当するベクトルの近傍幾つ目に正しい上位タグが出

表 2: タグに対する実験結果

手法	データ数	近傍	クラスタ数	正解数	正解率	平均順位
ベースライン	4980			197.47	4.08	36.89
近傍平均法	4980	1		206.89	4.23	18.95
近傍平均法	4980	5		300.83	6.16	24.75
近傍平均法	4980	10		263.17	5.38	29.47
セントロイド法	4980	1	17.4	534.6	10.99	31.4
関数学習法	29044			102	0.35	

表 3: タグ集合に対する実験結果

手法	データ数	近傍	クラスタ数	正解数	正解率	平均順位
近傍平均法	10053	1		139.8	1.36	48.01
近傍平均法	10053	5		132.1	1.27	48.49
近傍平均法	10053	10		121.2	1.16	48.23
セントロイド法	10053	1	20	102.3	0.96	50.86
関数学習法	60324			106	0.18	

現したかを示す．結果より，セントロイド法では正解率はベースライン手法の 2 倍以上となり，平均順位も向上した．また近傍平均法は，正解率の向上は僅かであったが平均順位が大きく上昇する結果となった．一方で，関数学習法は，正解率が 1%未満と十分な精度を得ることができなかった．

実験結果全体として，ベースラインより精度が向上したものの必ずしも十分ではなく，データ構築そのものを見直す必要があることが示唆された．今回の実験では日本語 WordNet を利用して is-a 関係の抽出を行ったが，動画タグに対しての抽出精度は必ずしも高くなく，日本語 WordNet 以外のオントロジーも併用する必要があると考えられる．

4.3 タグ集合に関する上位概念獲得実験

タグを対象とした実験では，セントロイド法と近傍平均法の有効性が確認できた．次に，タグ集合を対象に提案手法の性能を評価する．一つのタグ集合を一文書とみなし，次元数を 300 とする Doc2Vec[Le14] により，タグ集合をベクトル化した．なお実験の設定は，データを 10 分の 1 ずつに分けて用いた以外はタグを対象とした実験と同様である．

実験の結果を表 3 に示す．近傍平均法の正解率が僅かに高い結果となったが，タグを対象とした実験とは異なり，いずれの手法も十分な精度を得ることができなかった．タグを対象とした実験同様，この原因の一つはタグ集合間の上位下位の設定方法そのもののにあると考えられ，今後の改善が必要とされる．また，関数学習法は特に精度が低い結果となったが，ニューラルネットワーク構造やその学習方法の改善が必要である．

5. まとめと今後の課題

本論文では，動画に付与されたタグ集合を対象に，与えられたタグ集合の上位概念に相当するタグ集合を抽出する手法を提案した．実験では，必ずしも十分な精度が得られておらず，データ構築の段階からの見直しを行うとともに，手法の改良が必要であると考えている．

参考文献

[村上 10] 村上直至，伊藤栄典:動画投稿サイトで付与された動画タグの階層化，研究報告数理モデル化と問題解決（MPS），2010-MPS-81(17)，pp.1-6 (2010).

[高橋 13] 高橋文彦，山本雅人，古川正志：ニコニコ動画における共起関係を用いたタグの階層化，研究報告知能システム（ICS），2013-ICS-170(3)，pp.1-6 (2013).

[Heymann 06] Paul Heymann and Hector Garacia-Molina : Collaborative creation of communalhierarchical taxsonomies in social tagging systems, Techniqual Report 2006-10, Stanford University (2006).

[市瀬 15] 市瀬龍太郎，荒川直哉：分散表象とオントロジーの関係，第 29 回 人工知能学会全国大会，2I4-OS-17a-5 (2015).

[Mikolov13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado and Jeffrey Dean: Distributed Representations of Words and Phrases and their Compositionality, *Advances in Neural Information Processing Systems* (2013)

[Le14] Quoc V. Le and Tomas Mikolov : Distributed Representations of Sentences and Documents, *CoRR*, (2014)