ファジィクラスタリングを用いた分野別語義識別方法の提案

山崎　恭史[†]　　木村　昌臣[‡]
芝浦工業大学院工学研究科[†]　　芝浦工業大学工学部情報工学科[‡]

1. はじめに

近年，インターネットの普及やマルチメディアの発達により，我々が扱う情報量は膨大になっており，情報を効率的に管理・活用する技術としてトピックマップが注目されている[1]．トピックマップは，グラフ構造によって情報を管理する技術で，主に主題を表現するトピックと主題間の関係を表現する関連から構成される．トピックマップには，構成要素に対して有効範囲を指定するスコープという機能が存在し，年代や地域などある種の分野で関わりのあるトピック群に設定され，利用者は必要とする部分のトピックマップを取得することが出来る[2]．しかし，現在スコープの設定は手作業で行うしかないため，非常に手間がかかるという問題がある．そこで本研究では，トピックをノード，1つ以上の関連を持つトピック間に1本のエッジを張り，トピックマップをネットワーク（トピックマップネットワーク）として扱い，スコープが設定される関わりのあるトピック同士はネットワークでは繋がりが強いと考え，ネットワークのノードを関わりのある部分集合に分類するクラスタリングを適用することで，仮想的なスコープを自動的に取得することを目指す．また，スコープはひとつのトピックに複数設定されることがあるため，本研究では，1つのデータが1つのクラスタのみに所属するハードクラスタリングではなく，複数のクラスタに所属できるファジィクラスタリングを用いることにする．また，応用として複数の分野に関する記述が存在する論文を対象としてトピックマップネットワークを構築し，提案手法を適用することで評価・検証を行う．

2. 既存のクラスタリング手法

ネットワークをクラスタリングする既存の手法として，Reichard らが提案する手法がある[3]．これは，エッジで接続しているノードは同じクラスタへ，接続していないノードは異なるクラスタへ分類する手法である．しかし，この手法はハードクラスタリングであり，ファジィクラスタリングではない．一方で，ファジィクラスタリングの標準的な手法に，Fuzzy C-Means 法がある．これは，k-Means 法をファジィ分割出来るように拡張した手法で，各ノードはクラスタに対して所属する度合い（帰属度）を保持する．帰属度は0から1の実数をとり，クラスタの中心との距離が近ければ帰属度は高くなり，遠ければ低くなる．しかし，Fuzzy C-Means 法は，座標データを入力とするため，そのままではネットワークに適用出来ない．そこで，本研究では，Fuzzy C-Means 法をネットワークに適用する方法を提案する．

3. 提案手法

本研究で提案するネットワークをファジィクラスタリングする手法の基本手順は以下の通りである．

3.1. ネットワークからの距離行列生成

まず，座標に変換する上で必要となる距離行列を用意するために，ネットワークから距離行列を生成する．本研究では，ネットワークをエッジの重みを1とした無向グラフとして扱い，ダイクストラ法によって距離を計算し，距離行列を生成する．

3.2. 距離行列の座標変換

距離行列を座標に変換するのに，多次元尺度構成法を用いる．多次元尺度構成法は，データ間の類似度をもとにして，2次元あるいは3次元空間に，類似度が高いデータ同士は近くに配置し，類似度が低いデータ同士を遠くに配置する方法である．本研究では，データ間の類似度に3.1節で計算した距離を用い，距離行列から座標データへの変換を行う．

3.3. クラスタ数の算出

Fuzzy C-Means 法では，予めクラスタ数を指定する必要がある．本研究ではクラスタ数を決めるのに，対象データをもとにベイズ情報基準量を用いてクラスタ数を算出する方法を用いる[4]．

3.4. 座標データへのファジィクラスタリング

3.2節の変換方法によって求まる座標データを対象データとし，3.3節で求めたクラスタ数を用いてFuzzy C-Means 法を実行する．

4. 江戸川乱歩トピックマップネットワークへの適用

4.1. 対象データ及びネットワーク

江戸川乱歩のトピックマップ（図1）を対象データとする．トピックをノード，関連をエッジとする．ノードがクラスタに所属しているとする帰属度の閾値は，帰属度の分布から0.1とした．

4.2. 結果・評価

結果は表1のようになった．クラスタ1には作品関係，クラスタ2には少年探偵団関係，クラスタ3には二銭銅貨関係のトピックがそれぞれ分類されている．全てのクラスタに関係する江戸川乱歩が全てのクラスタに所属するファジィノードになっており，作品と少年探偵団に関係のある明智小五郎はクラスタ1,2に所属するファジィノードになっており，提

A Word Sense Disambiguation by utilizing Fuzzy Clustering
[†]Takashi Yamazaki, [‡]Masaomi Kimura
[†]Graduate School of Shibaura Institute of Technology
[‡]Shibaura Institute of Technology

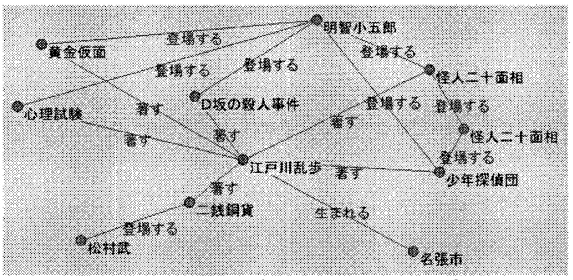

図 1：江戸川乱歩のトピックマップ

表 1：クラスタリング結果

ノードの種類	クラスタ番号	クラスタに含まれるノード
一つのクラスタに属するノード	クラスタ1	黄金仮面, 心理試験, D坂の殺人事件
	クラスタ2	怪人二十面相(小説), 少年探偵団
	クラスタ3	松村武, 二銭銅貨
複数のクラスタに属するファジィノード	クラスタ1, 2	明智小五郎, 怪人二十面相(人)
	クラスタ1, 2, 3	江戸川乱歩, 名張市

案手法により妥当な結果が得られた.

5. 論文トピックマップネットワークへの適用

江戸川乱歩のトピックマップは，小規模であり，より大きな規模のトピックマップで実験する必要がある．しかし，現在公開されているトピックマップの種類が少なく，適したサイズのトピックマップが存在しない．そこで，複数の分野に関する記述が存在する論文を用いて，トピックマップネットワークを作成し，そのネットワークに提案手法を適用する．

5.1. 対象データ及びネットワーク

横幹連合第2回横幹連合総合シンポジウムの論文37編を対象とする．構築するネットワークをトピックマップのモデルとするために，ノードをトピック，エッジを関連と対応付くように設定する．ノードは，TF・IDF値を利用し，対象の中で特徴的な名詞とする．エッジは，同じ文中に存在する名詞間には関係があるとして考え，同じ文中に存在するという共起関係とする．また，帰属度の閾値は帰属度の分布から0.1とした．

5.2. 結果・評価

評価を行うに当り，予めノードとクラスタには正解となる分野 C_i を設定する．まず論文に対して発表セッションに基づき分野を設定し，ノードの分野は，分野 C_i を持つ論文の中でそのノードが2編以上に出現している場合に分野 C_i をノードの分野として設定する．また，分野を1つも持たないノードは，特徴的な語ではないとしてネットワークから削除する．クラスタの分野は，クラスタ内のノードが持つ分野の内，最も出現回数が多い分野とした．

ファジィノードに対して以下の方法を用いて評価を行う．異なる分野を保持する複数のクラスタに所属しており，クラスタの分野 C_i を複数持つノードをファジィノードとし，予め設定された分野を複数持つノードに割り当てられた分野との一致している割合をみることで，ファジィノードの評価を行った．

総ノード数282個のうち，ファジィノードが105個，複数の分野を持つノードが174個，その中で一致しているノードは72個で，精度は68%だった．一致しているノードが持つ，ファジィノードの分野と複数の分野を持つノードの分野が一致している割合は，完全一致は9.7%と低くなったが，部分一致まで含めると約90%となった．また，不一致も9.1%と低かった．

スコープの候補として各クラスタから，帰属度とクラスタ内の次数の積からクラスタの特徴語を抽出したところ，「ナノ」，「医薬品」，「回路」など具体的に分野となりそうな語が抽出された一方で，「列」，「判断」，「関係」など分野としては抽象的な語も抽出された．

5.3. 考察

ファジィノードと複数の分野を持つノードで一致しないノードは，異なる分野を持つ複数のクラスタに所属しているが，ノードの分野を1つしか持たないノードと，ノードの分野を複数持っているが，同じ分野を持つクラスタにしか所属しないノードに分けられる．前者のノードは，ノードの分野を持つクラスタにのみ所属することが望ましいが，座標上，複数の分野に所属する場所に配置されてしまい，異なる分野を持つ複数のクラスタに所属してしまったと考えられる．一方，後者のノードは，ノードの分野を持つ複数のクラスタに所属することが望ましいが，座標上，同じ分野を持つノード群の中心近くに配置されてしまい，同じ分野を持つクラスタだけに所属してしまったと考えられる．また，特徴語の抽出では，スコープとなりうる語が抽出される一方で，無関係な語も抽出しており，次数と帰属度の大小以外も考慮に入れて特徴的な語を求める必要があると考えられる．

6. まとめと今後の展望

本研究では，ネットワークをファジィクラスタリングし，そのクラスタから特徴的な語を取り出す手法の提案を行い，トピックマップネットワークに適用して評価を行った．今後は，実際のトピックマップへ提案手法を適用し，クラスタから抽出した特徴語をスコープとして自動的にトピックマップに付加し，フィルタリングを実現するシステムの構築が今後の展望となる．

参考文献

[1] 内藤求：トピックマップ入門，東京電機大学出版局，(2006).

[2] Lars Marius Garshol : A Theory of Scope, TMRA2007, (2007).

[3] Joerg Reichard, Stefan Bornhold : Statistical mechanics of community detection, Physical Review E, vol.74, 016110, pp1-14(2006).

[4] Julio Ponce, Adem Karahoca : DATA MINING AND KNOWLEDGE DISCOVERY IN REAL LIFE APPLICATIONS, In-Teh, (2009).